Simulating Infectious Diseases

Consider a well-mixed sub-population in a finite area with a fixed size. Suppose that a disease broke out within the sub-population with three non-intersecting states, (S) Susceptible, (I) Infected, and (R) Recovered. The meanings of these three states are as follows:

- (S) Susceptible - an individual that was susceptible to the disease.
- (I) Infected - an individual infected by the disease.
- (R) Recovered - an individual that has recovered from the disease.

We can model the fraction of SIR individuals using a system of differential equations. Note that we denote the number as a uppercase S while the fraction as lowercase s, meaning $N = S + I + R$ and $1 = s + i + r$ for all t. Below, we define the simplest SIR model that simulates the spread of an infectious disease - this model is also known as the Kermack–McKendrick epidemic model proposed by Kermack and McKendrick in 1927:

$$\begin{aligned}
\text{The change in (S)usceptible individuals} &\longrightarrow \frac{ds(t)}{dt} = -\beta si \\
\text{The change in (I)nfected individuals} &\longrightarrow \frac{di(t)}{dt} = (\beta s - \gamma)i \\
\text{The change in (R)ecovered individuals} &\longrightarrow \frac{dr(t)}{dt} = \gamma i
\end{aligned}$$

where the parameters are defined as

- β = contact rate $\times$ transmission probability, and
- $\gamma = 1/$infectious period.

We can then compute the *basic reproductive number* $\mathbf{R_0}$, which is the estimated number of infections caused by a single case in a susceptible group. This is mathematically defined as

$$\mathbf{R_0} = \text{contact rate} \times \text{transmission probability} \times \text{infectious period}.$$

There are three basic assumptions for this model:

1. Total number of sub-population is fixed, meaning no births or deaths, and homogeneous, meaning there are no social structures, spatial and age distributions.
2. Incubation period of the disease is instantaneous, meaning if a person is exposed to the disease, they show symptoms immediately and the agent that causes the disease is present within the individual.
3. Duration of disease infectiousness has the same length of the disease duration.

While the above assumptions may look unrealistic and oversimplifies the complexity of a real disease, we start from these assumptions in order for us to make sense of what the data is doing. Note that the real goal of mathematical/statistical modeling is NOT to exactly mimic our reality but to use it to makes sense of it. We start from this very simple model and build from it to understand what the data is telling us. This is one of the fundamentals of Data Science lifecycle.

In addition, the SIR model is not limited to modeling infectious diseases. The SIR model is a type of model that uses numerous characteristics to forecast how an infectious disease would spread through a society. Each SIR model is tailored to a particular disease (e.g. Malaria, COVID, HIV, etc.). While the model is intended for infectious disease, it may also be used to how information spreads throughout a community if the premise is that information is a "agent" capable of "infecting" an individual and using that information to interact with others (e.g. word usage, sources used, rumors etc.).

In this mini-assignment, we will focus on observing different behaviors of the SIR model using different values of parameters.

```
# load packages
library(tidyverse)
library(deSolve)
library(ggthemes)
library(gridExtra)
```

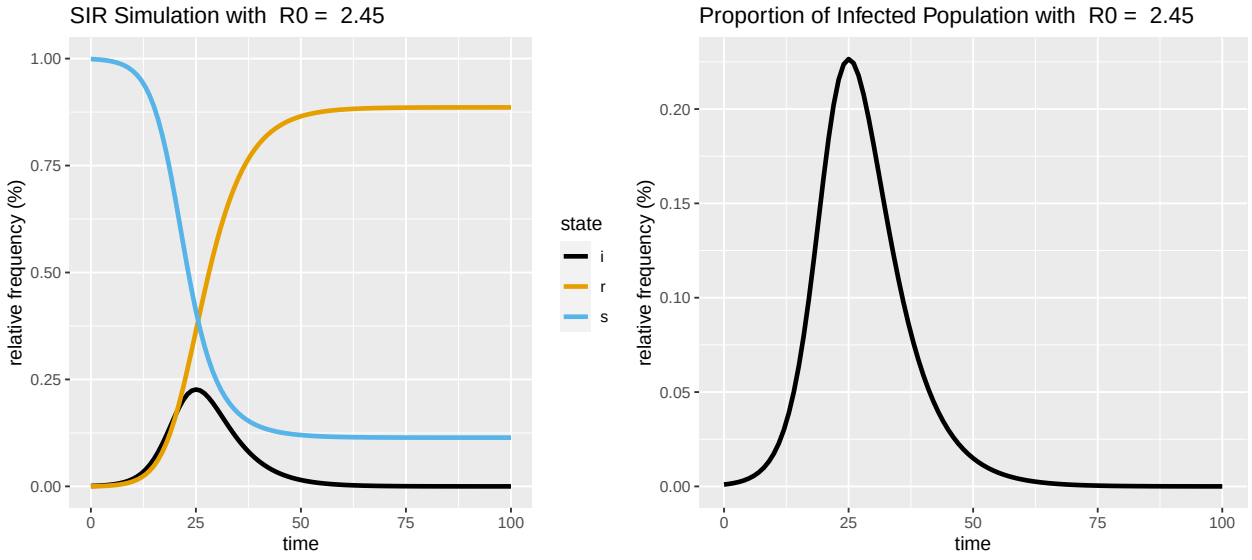
Below is an R code block that defines the SIR model as an R function.

```
# define the SIR model system of differential equations
SIR_mod = function(Time, State, Pars){
  with(as.list(c(State, Pars)),
         {ds = (-beta * s * i)        # change in susceptible
          di = ((beta * s) - gamma)*i # change in infectious
          dr = (gamma * i)            # change in recovered
          results = c(ds, di, dr)
          return(list(results))
       })
}
```

Next, we use set the initial conditions and the parameters of the model. We then use the `ode` function in the `deSolve` package to find a solution to the SIR model system of differential equations. Below we set the parameters so that $2 < \mathbf{R_0} < 3$.

```
# disease dynamics
contact_rate = 7                # number of contacts per day
transmission_probability = 0.07 # transmission probability
infectious_period = 5           # infectious period

# basic reproductive number
R0 = contact_rate * transmission_probability * infectious_period

# parameters of the model
SIR_parameters <- c(beta = contact_rate*transmission_probability,
                    gamma = 1/infectious_period)

# initial population sizes
S_initial = 999
I_initial = 1
R_initial = 0
N = S_initial + I_initial + R_initial

# convert initial states to relative frequency
initial_state <- c(s = S_initial/N, i = I_initial/N, r = R_initial/N)

# time vector
T_upper = 100
time <- seq(0,T_upper,1)

# solve the system of differential equations
SIR_sim <- as.data.frame(ode(func = SIR_mod, times = time,
                             y = initial_state, parms =SIR_parameters))

# transform the data for plotting
df_all_pivot <- SIR_sim %>%
  pivot_longer(cols = c("s","i","r"),
    names_to = "state", values_to = "count_sim")
```

```r
# plot the simulations
p1 <- ggplot(data = df_all_pivot, aes(x = time, y = count_sim, color = state)) +
  geom_line(size = 1.25) +
  scale_color_colorblind() +
  labs(x = "time",
       y = "relative frequency (%)",
       title = paste("SIR Simulation with ",
                     "R0 = ",as.character(R0)))
p2 <- ggplot(data = df_all_pivot %>% filter(state == "i"), aes(x = time, y = count_sim)) +
  geom_line(size = 1.25) +
  scale_color_colorblind() +
  labs(x = "time",
       y = "relative frequency (%)",
       title = paste("Proportion of Infected Population with ",
                     "R0 = ",as.character(R0)))

grid.arrange(p1,p2,ncol=2)
```



Trying out different $\mathbf{R_0}$. Use the provided code above so that you can plot the simulations with $\mathbf{R_0}$ values:

a. $\mathbf{R_0} < 1$
b. $\mathbf{R_0} = 1$
c. $1 < \mathbf{R_0} < 2$

Plot your simulations under one plot for each state (S,I,R) - there should be three separate plots for each state with three curves (label the legend accordingly) and you may need to adjust the upper time limit - and provide a paragraph describing your observations comparing the results for each $\mathbf{R_0}$ case. In your paragraphs, it is important to answer two key questions:

1. What are your thoughts on the comparatively low level of infection at the peak of the epidemic compared to the recovered population?
2. Can you see how a low peak level of infection can cause more than half of the population to become ill? Explain.

```
### [BEGIN] - PERFORM SIR SIMULATIONS CODES HERE

### [END] - PERFORM SIR SIMULATIONS CODES HERE
```

```
### [BEGIN] - CREATE GGPLOT PIPELINE HERE

### [END] - CREATE GGPLOT PIPELINE HERE
```