Data Science Quarterly: Manuscript for Cases Section
Volume 1, Issue 1, January 2021
ISSN 2693-8871 (print) ISSN 2693-8863 (online)

Model of an Epidemic

Thomas W. Miller

Revised 25 August 2020

Abstract This case shows how to build a simple model of an epidemic, identifying four disease classes and conditional probabilities of going from one disease class to another. The model, a finite Markov chain, may be extended to accommodate differences in rates of contagion and fatality across age groups, as well as alternative social networks. Students are encouraged to develop a more accurate and more flexible model using discrete event simulation.

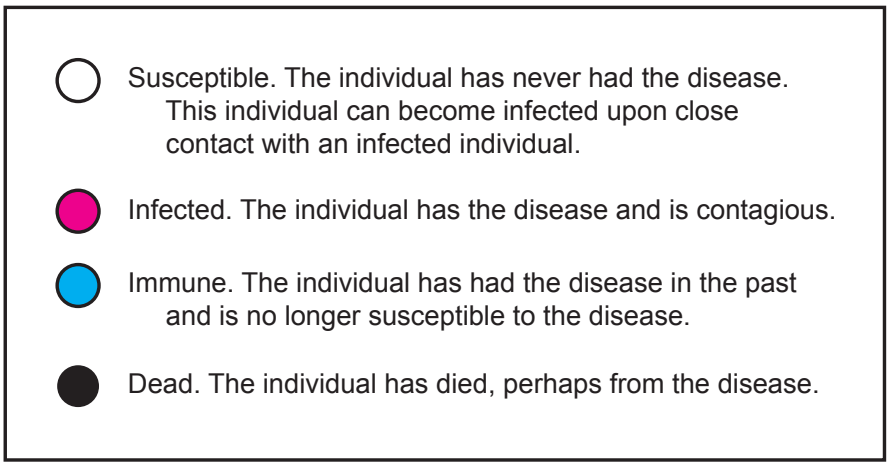
Keywords COVID-19 · Disease states · Epidemiology · Finite Markov chain · Stochastic process

1 What factors affect the spread of disease?

What is needed to develop a model of an epidemic or pandemic? We would need information about the nature of the disease, probabilities, and conditional probabilities relating to contagion and fatality rates. We would be most uncertain about estimated contagion and fatality rates when a disease first appears. As the number of tested individuals increases with time and as we learn about the conditions of contagion and fatality, our uncertainty about probabilities and conditional probabilties will diminish, allowing us to obtain better estimates of disease parameters.

Stochastic models of epidemics are well understood, as evidenced by a large number of works in the area (Andersson and Britton, 2000; Diekmann et al., 2012; Martcheva, 2015; Kiss et al., 2017), including work on the modeling the sequence of COVID-19 symptoms (Larson et al., 2020). Heesterbeek and Dietz (1996) review the basic reproduction number R_0, which reflects the expected number of secondary cases of an infectious disease that one infected case generates in a completely susceptible population. Finite Markov chain modeling, as employed in this case study, is reviewed in Kemeny and Snell (1960), Allen (2010), Murphy (2012), and Ross (2019). Spedicato (2017) discusses the R markovchain package, which can be used to estimate disease class distributions across time. Estimates of COVID-19 disease parameters as of April 29, 2020 are provided by the Centers for Disease Control and Prevention (CDC) (2020).

Fig. 1 Disease Classes for an Epidemic



To build a probability model for an epidemic, start by identifying disease classes associated with each individual. Suppose we define four mutually exclusive disease classes, as shown in figure 1. We assume a completely susceptible population—all individuals are initially susceptible to the disease. We also assume that all infected individuals are immediately contagious—there is no period of incubation during which an individual is infected but not contagious.

Consider a group of 100 individuals in a small rural community. At any point in time, there will be specific numbers of individuals in each of the four classes. Let us use t to represent the time period, so p_t is the probability at time t. Prior to the onset of disease (at time $t = 0$), no one has been infected, yielding relative frequencies or probabilities across the four disease classes as follows:

$$p_0(\text{susceptible}) = 100/100 = 1.00$$
$$p_0(\text{infected}) = 0/100 = 0.00$$
$$p_0(\text{immune}) = 0/100 = 0.00$$
$$p_0(\text{dead}) = 0/100 = 0.00$$

There would be no epidemic if the initial state of the community had no one infected. Let us assume that one person has become infected with a virus, perhaps through travel to another community. Upon that person's return, we have an initial state ($t = 0$) with the following relative frequencies or probabilities:

$$p_0(\text{susceptible}) = 99/100 = 0.99$$
$$p_0(\text{infected}) = 1/100 = 0.01$$
$$p_0(\text{immune}) = 0/100 = 0.00$$

$$p_0(\text{dead}) = 0/100 = 0.00$$

And so the epidemic begins with a disease class distribution [0.99, 0.01, 0.00, 0.00], which is a discrete probability distribution for the initial state of the disease process.

Suppose one time period is the period of infection and the period of being contagious. The rate of contagion is a key disease parameter. Suppose each person infected with the virus infects three other people, and suppose that, after being infected in time period $(t-1)$, a person becomes immune in time period t. Then we would have these relative frequencies or probabilities at time period $t = 1$, one period after time $t = 0$:

$$p_1(\text{susceptible}) = 96/100 = 0.96$$
$$p_1(\text{infected}) = 3/100 = 0.03$$
$$p_1(\text{immune}) = 1/100 = 0.01$$
$$p_1(\text{dead}) = 0/100 = 0.00$$

Continuing to the next time period $(t = 2)$, if each infected individual infects three other individuals while simultaneously building up immunity, and assuming none of the three infected individuals dies from the disease, we would have these relative frequencies or probabilities:

$$p_2(\text{susceptible}) = 87/100 = 0.87$$
$$p_2(\text{infected}) = 9/100 = 0.09$$
$$p_2(\text{immune}) = 4/100 = 0.04$$
$$p_2(\text{dead}) = 0/100 = 0.00$$

A complete model of an epidemic or pandemic would include information about of the fatality rate. If we estimate that two out of every one hundred infected individuals dies from the disease, then we would obtain the matrix of transition probabilities shown in table 1. Transition probabilities show the probability of going from one disease class to another across a single time period.

Table 1 shows no one going directly from susceptible to immune, which suggests that there is no vaccine to prevent infection. The table also shows no one going directly from susceptible to dead. In fact, the transition probability of going from susceptible to dead would be a small positive value because people die from many causes not related to the disease.

Table 1 provides a convenient summary of sixteen transition probabilities, which are conditional probabilities describing the disease process. From the first row of the table, we see that three of every one hundred susceptible individuals become infected with the disease. Thus the contagion rate is 0.03. From the second row of the table, we see that for every one hundred people infected with the disease, two will die—the fatality rate is 0.02. A transition probability is a conditional probability: the probability of an event or state at time t, given a previous event or state at time $(t-1)$. The fatality rate is a

Table 1 Transition Probabilities for a Four-Class Disease Model

Disease Class at time (t-1) \ Disease Class at time t	Susceptible	Infected	Immune	Dead
Susceptible	0.97	0.03	0.00	0.00
Infected	0.00	0.00	0.98	0.02
Immune	0.00	0.00	1.00	0.00
Dead	0.00	0.00	0.00	1.00

person's conditional probability of being dead at time t, given that he or she was infected at time $(t-1)$, as shown in this expression:

$$\text{fatality rate} = p(\text{dead at time } t \mid \text{infected at time } t-1) = 0.02$$

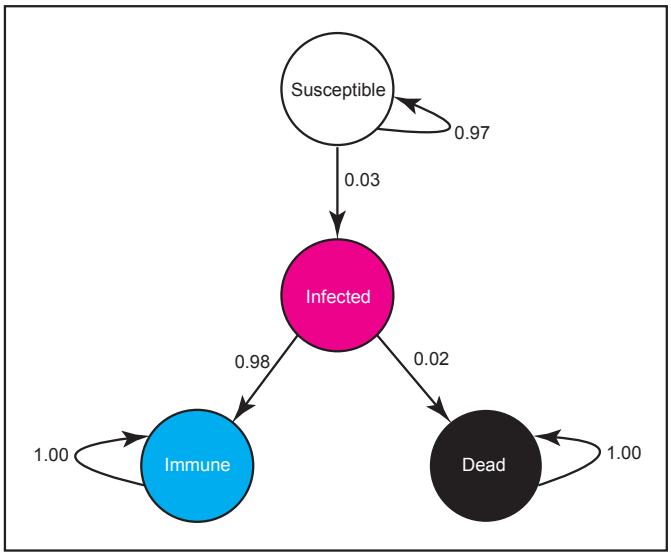
By comparison with the disease represented in table 1, the fatality rate of the common flu is around 0.001—one out of one thousand people infected with the common flu dies. For the coronavirus pandemic (COVID-19 caused by the SARS-CoV-2 virus) in the United States as of April 29, 2020, the estimated fatality rate (ratio) for symptomatic individuals was 0.004, with 0.65 of infections being symptomatic. Accordingly, the estimated fatality rate for coronavirus would be $0.004 \times 0.65 = 0.0026$.

Implicit in the four-disease-class probability model is the assumption that infection/illness lasts for only one time period with people moving from infected to either immune or dead states. Another simplifying (perhaps unrealistic) assumption, as previously noted, is that there is no period of incubation—people move directly from susceptible to infected and contagious.

In figure 2, we represent the four-disease-class probability model as a directed graph. From the third and fourth rows in the matrix of transition probabilities and from the corresponding immune and dead nodes of the directed graph, we can see that there is no transition out of immune and dead disease classes. Notice how links from immune and dead nodes point back to themselves. An immune individual stays immune, and a dead individual stays dead. The probability or stochastic process we have defined here is a finite Markov chain, and the immune and dead disease classes are called absorbing states.

A finite Markov chain obeys the Markov property. To know the state or distribution of disease classes in the current time period, we need only consider the distribution of disease classes in the immediately preceding time period, along with the matrix of transition probabilities. We say that a model obeying the Markov property "has no memory."

Fig. 2 Directed Graph for a Four-Class Disease Model



Modeling is both art and science with probability and statistics at its core. We use our best judgment to make simplifying assumptions so models are not overly complex. We use data to estimate model parameters such as contagion and fatality rates. We revise models as we collect more data and as we learn more about the world.

A finite Markov chain is fully defined by its initial state and matrix of transition probabilities. With such a model in hand, we can predict the percentage of the population in each disease class at any point in time in the future. We can predict the progress of an epidemic or pandemic. While the distribution of disease classes changes from one time period to the next, the matrix of transition probabilities remains fixed or stable.

What can we expect for the final distribution of disease classes? To answer this question, we let the Markov chain model play out or run its course. Table 2 shows how the population moves from its initial state to its final or steady state—a process that takes more than two hundred time periods given the initial state and transition probabilities we have specified.

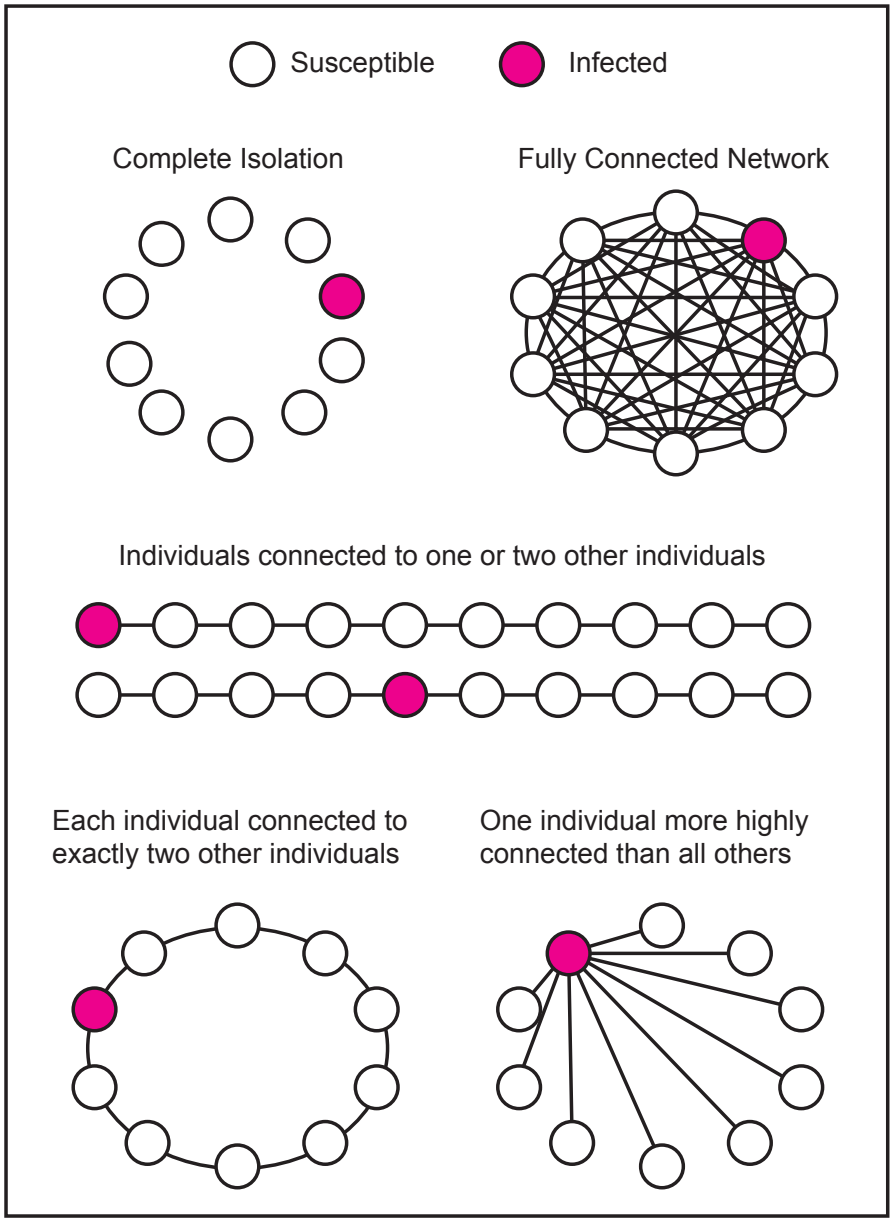
Without a vaccine to protect susceptible individuals, all members of the community ultimately become infected with the disease, after which they become immune or die. Interestingly, the proportion of the population expected to die from the disease is precisely the fatality rate from the matrix of transition probabilities. Note the steady state value for the dead class at the bottom right of table 2. Appendix 1 shows the R program for making the predictions in table 2. This may be a good place to start creating a more flexible modeling framework.

Table 2 Progress of Disease from Initial to Steady State

	Disease Class			
Time period	Susceptible	Infected	Immune	Dead
Initial state t = 0	0.990	0.010	0.000	0.000
t = 1	0.960	0.030	0.010	0.000
t = 2	0.931	0.029	0.039	0.001
t = 4	0.876	0.027	0.095	0.002
t = 8	0.776	0.024	0.196	0.004
t = 16	0.608	0.019	0.366	0.007
t = 32	0.374	0.012	0.603	0.012
t = 64	0.141	0.004	0.838	0.017
t = 128	0.020	0.001	0.960	0.020
Steady state t = 256	0.000	0.000	0.980	0.020

The model we have developed assumes that members of the population are fully interconnected, so that everyone who is susceptible to the disease eventually becomes infected. A more accurate model for an epidemic would consider the extent to which individuals come in contact with one another (the social network or degree of connectedness). Communities vary greatly in the degree to which people are connected to one another. There will be higher rates of contagion with higher connectivity. Alternative network stuctures, as illustrated in figure 3, will affect the transmission of the disease. Disease transmission varies by social network structure, geographic location (rural versus urban), and governmental regulations about social distancing and wearing face masks.

Fig. 3 Disease and Social Network Structure



As stated previously, for the coronavirus pandemic in the United States as of April 29, 2020, the estimated fatality rate (ratio) for symptomatic individuals was 0.004, with 65 percent of infections being symptomatic. So the estimated fatality rate for coronavirus was $0.004 \times 0.65 = 0.0026$. But estimated fatality rates (ratios) for symptomatic individuals vary greatly by age. For symptomatic individuals under 50 years of age, the estimated fatality rate was only 0.0005. For symptomatic individuals 50 to 64 years of age, the estimated rate was 0.002. And for symptomatic individuals 65 years of age and older, the estmated rate was 0.013. With 65 percent of infections being symptomatic, estimated fatality rates would be 0.000325, 0.0013, and 0.00845 for age groups under 50, 50 to 64, and 65 or older, respectively.

Is a simple stochastic model with four disease classes sufficient for predicting the effects of the coronavirus pandemic on the U.S. population? No, because it fails to represent the connectedness and age structure of the population, among other things.

To develop a more accurate, more flexible model of the coronavirus pandemic, we would use the most recent estimates of contagion and fatality rates. We might add a disease class for people exposed to the disease but not yet infectious, allowing for a period of incubation. It may also be wise to distinguish between symptomatic and asymptomatic carriers of the disease. Discrete event simulation provides a flexible modeling framework in which periods of infection, incubation, and contagiousness can vary in duration according to predefined probability distributions, rather than being of fixed duration.

2 About the Author

Thomas W. Miller is faculty director of the data science program at Northwestern University, author of six books about data science, and owner of Research Publishers LLC.

References

Allen, Linda J. S. 2010. *An Introduction to Stochastic Processes with Applications to Biology*, 2nd edn. Boca Raton, Fla.: CRC Press.

Andersson, Håkan, and Tom Britton. 2000. *Stochastic Epidemic Models and Their Statistical Analysis*. New York: Springer.

Centers for Disease Control and Prevention (CDC). 2020. COVID-19 Pandemic Planning Scenarios. Retrieved from the World Wide Web on May 23, 2020, at https://www.cdc.gov/coronavirus/2019-ncov/hcp/planning-scenarios.html.

Diekmann, Odo, Hans Heesterbeek, and Tom Britton. 2012. *Mathematical Tools for Understanding Infectious Disease Dynamics*. Princeton, N.J.: Princeton University Press.

Heesterbeek, J. A. P., and K. Dietz. 1996. "Emerging Trends in Geospatial Artificial Intelligence (geoAI): Potential Applications for Environmental Epidemiology". *Statistica Neerlandica* 50 (1): 89–110. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9574.1996.tb01482.x.

Kemeny, John G., and J. Laurie Snell. 1960. *Finite Markov Chains*. New York: D. Van Nostrand.

Kiss, István Z., Joel C. Miller, and Péter L. Simon. 2017. *Mathematics of Epidemics on Networks: From Exact to Approximate Models*. New York: Springer.

Larsen, Joseph R., Margaret R. Martin, John D. Martin, Peter Kuhn, and James B. Hicks. 2020. "Modeling the Onset of Symptoms of COVID-19". *Frontiers in Public Health* 8. Retrieved from the World Wide Web, August 19, 2020 at https://www.frontiersin.org/articles/10.3389/fpubh.2020.00473.

Martcheva, Maia. 2015. *An Introduction to Mathematical Epidemiology*. New York: Springer.

Murphy, Kevin P. 2012. *Machine Learning: A Probabilistic Perspective*. Cambridge, Mass.: MIT Press.

Ross, Sheldon M. 2019. *Introduction to Probability Models*, twelfth edn. San Diego: Academic Press.

Spedicato, Giorgio Alfredo. 2017. Discrete Time Markov Chains with R. *The R Journal*. R package version 0.6.9.7. https://journal.r-project.org/archive/2017/RJ-2017-036/index.html.

Appendix 1. Model of an Epidemic (R Program)

```
# Finite Markov Chain Model of an Epidemic with Four Disease Classes
# Revised May 19, 2020 by Thomas W. Miller

# install markovchain package prior to running
library(markovchain)

stateNames = c("susceptible", "infected", "immune", "dead")

# set disease parmeters
# contagion depends on susceptibility and network structure
# must be a value in the open interval (0,1)
contagionRate = 0.03

# fatality depends on biology of disease and medical treatments
# must be a value in the open interval (0,1)
fatalityRate = 0.02

# define matrix of transition probabilities
# all elements must be in the closed interval [0,1]
# sum of every row must be 1
transitionMatrix = matrix(c((1-contagionRate),contagionRate, 0, 0,
                            0,    0, (1-fatalityRate),fatalityRate,
                            0,    0,     1,    0,
                            0,    0,     0,    1),
    byrow = TRUE, nrow = 4,
    dimnames = list(stateNames, stateNames))

# define model with two parameters
epidemic = new("markovchain", transitionMatrix = transitionMatrix)
print(epidemic)

# define initial state distribution
# susceptible and infected in the open interval (0,1)
# immune and dead values at zero
# sum of the four state probabilities must be 1
initialState = c(0.99,0.01,0.00,0.00)
# show initial state
cat('time: 0  state:', round(initialState, 3), '\n')

# report state distribution across the first ten time periods
for (i in 0:10){
    state = initialState * (epidemic ^ i)
    cat('time:', i, ' state:',round(state, 3), '\n')
 }

# report state distribution across a wide range of time periods
for (i in 0:10){
    state = initialState * (epidemic ^ (2^i))
    cat('time:',(2^i), ' state:',round(state, 3), '\n')
 }
```