Chapter 5: Modeling infectious disease epidemics

In previous Chapters, we constructed Markov models that assumed a person's risk of a disease was a constant value (i.e., a fixed probability per unit time). The assumption that risk is constant, rather than changing with the environment a person is living in, ignores a critical aspect of modeling *infectious* diseases. For infectious diseases, the risk of getting the disease is related to how many people are infectious at a given time; the more infectious people in the area, the higher the risk of infection for a susceptible person. In a typical Markov model, we can't account for this basic feature of infectious diseases, because the risk of moving from one state (healthy) to another state (diseased) is assumed to be constant and independent of how many infectious people might be around. In this Chapter, we introduce a simulation modeling framework that has been used for nearly a century to overcome this problematic assumption of Markov models, and simulate infectious disease epidemics and the interventions we can introduce to address them.

Understanding rates

To model infectious disease epidemics, we first need to have a good understanding of epidemiological *rates*. By definition, a rate refers to a number of people or objects per unit time—such as ten people per hour eating in a café, or twenty cars per minute traveling through an intersection.

To understand how rates are useful for constructing public health models, let's begin with a silly example. Suppose you are a businessperson who manages a warehouse

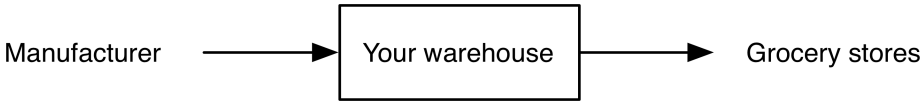
business. Your warehouse business serves to store and supply "Peeps"—those brightly-colored marshmallow candies that appear in trick-or-treat containers every Halloween (Figure 5.1).

First of all, shame on you for giving little children excess sugar!

But that problem aside, suppose you want to be a savvy businessperson and optimize your warehouse business. Your business is based on the premise that you need to buy a certain number of Peeps from the manufacturer, then sell a certain number of Peeps to grocery stores who buy the products from you, hopefully providing you with a profit.

As a savvy businessperson, you know that you shouldn't order too many Peeps from the manufacturer and have them sitting on the shelf for weeks or months because you'll be wasting your money shelving a product rather than selling it. On the other

hand, you also know that you shouldn't order too few Peeps from the manufacturer, because the grocery store will want more than you have available, and you will lose the opportunity to make money if you can't meet the high demand. Your business model can be captured in a simple diagram, shown in FIGURE 5.2.



You need to determine the optimal number of Peeps to order from the manufacturer to satisfy your marshmallow-craving customer base. To determine the optimal number to order, you need to estimate the demand that people are placing on your warehouse. Suppose you do a simple survey at your warehouse, in which you ask your manager: "How long does the average peep stay on the shelf in my warehouse before it gets ordered and shipped off to the grocery store?"

The manufacturer can tag some Peeps with a label as soon as they come from the manufacturer, and keep track of how many days it takes before the labeled Peeps leave your warehouse shelf and get shipped to grocery stores. Suppose the average Peep, according to this experiment, stays on the shelf three days. In that case, what is the average *per Peep rate* of leaving the warehouse?

We can say that 1 Peep lasts on average 3 days, hence the average per Peep rate of leaving the warehouse is 1 Peep/3 days = 1/3 of a Peep per day. Here, we can see that for rates of transfer, we have a value of time that is in the denominator of our

equation, which is why epidemiological rates are often confusing—they typically have units such as days^{-1} or 1/days, which simply means that they are reflecting some rate of movement of people (or Peeps) per unit time (e.g., from a healthy to a diseased state).

Suppose that instead of having each Peep last for 3 days, you have a sudden surge in business because it's the week before Halloween. Peeps are flying off of your shelves at a rate of 5 Peeps per day on average. What's the average *survival time* of one Peep on your warehouse shelf?

We can say that the survival time is 1 day for each set of 5 Peeps, or 1/5 of a day of survival time on the shelf of the warehouse.

To generalize, we see that the duration of time that a Peep survives on the warehouse shelf is the inverse of the duration of time it spent on the shelf. Equation 5.1 provides the general nomenclature for this relationship.

[Equation 5.1] $$E(T) = \frac{1}{\mu}$$

In this equation *E(T)* refers to the expected (or average) duration of the time T that a Peep survives on the shelf in your warehouse, and is the average per peep rate of leaving the warehouse. Hence, the duration of time spent in a given state is the inverse of the rate of leaving that state. For example, the duration of life (a.k.a., life expectancy) is the inverse of the rate of death (the mortality rate). As another example, suppose the average time that a peep stays on your warehouse shelf is 0.2 days. From Equation 5.1, *E(T)* = 0.2 days. How many Peeps should you order from the factory per day to perfectly meet the demand? Using the formula to calculate the rate of peeps

leaving the warehouse, you can estimate that you need to deliver Peeps at a rate of 1/0.2, or 5 Peeps/day. Hence, you should order 5 Peeps per day, on average, to keep up with the demand.

Of course, all of these expressions are just averages. Suppose we want to come up with a way of determining what the probability is that a Peep "survives" on the shelf for 1 day, or 2 days, or 15 days. Then we need to find a function to describe the probability that a Peep stays on your warehouse shelf for at least t units of time. We can write that as $\Pr\{T>t\}$, or the probability that a Peep's survival time T is greater than some value t. This is a type of *survival function*, designated $S(t)$.

How can we derive such a function? We can reason that the average shelf life in the warehouse, $E(T)$, is a sum of 1 day*Pr{Peep on shelf 1st day}+1day*Pr{Peep still on shelf 2nd day}+1day*Pr{peep still on shelf 3rd day}+...all the way to infinity days (by which time, presumably, the probability of surviving on the shelf has reached 0). In other words, our average shelf life $E(T)$ is the sum of $\Pr\{T>t\}$ across all non-negative values of t. We want to find some survival function $S(t)$ for which this is true. Writing down this expression, we have Equation 5.2.

[Equation 5.2] $$E(T) = \int_0^\infty \Pr\{T > t\}\,dt = \int_0^\infty S(t)dt = \frac{1}{\mu}$$

What function $S(t)$ could satisfy this equation, such that its sum from 0 to infinity is 1/ ? If you remember calculus, you'll recall that the exponential function will do the job beautifully, as shown in Equation 5.3.

[Equation 5.3] $$\int_0^{\infty} e^{-\mu t} dt = -\frac{e^{-\mu \times \infty} - e^{-\mu \times 0}}{\mu} = -\frac{0-1}{\mu} = \frac{1}{\mu}$$

So now we have a function to describe the probability that a peep survives to time t, summarized in Equation 5.4.

[Equation 5.4] $$S(t) = e^{-\mu t}$$

To operationalize this equation, we can reason that if the typical duration of shelf-life for a peep is 3 days, then the probability a peep survives 7 days is $e^{-(1/3)*7} = 0.097$, because 1/3 is the rate of leaving the shelf (1 peep/3days) and t is 7 days. Hence, there is a 9.7% chance of a peep surviving a week on the warehouse shelf.

Through our peep example, we have discovered two key learning points about rates of survival and durations of survival: the survival time in a state can be estimated as the inverse of the rate of leaving a state; the probability of surviving in a state until time t can be estimated as an exponential rate of leaving the state.

How does optimizing our warehouse of peeps help us to find solutions to stop infectious disease epidemics? We first have to determine how to adjust our equations to address the case of *fluctuating demand*. In other words, if we will simulate an infectious disease where the number of people infected per unit time will increase when more people are infectious, we no longer have a constant rate but rather have fluctuating rates of infection that could vary widely across time periods. Hence, the survival function—the probability of surviving at least to time t—will not simply be an exponential function of the constant summed over all times up to t (i.e., t), but rather the exponential function of individual rates for each time, (t), summed over all times up to time t:

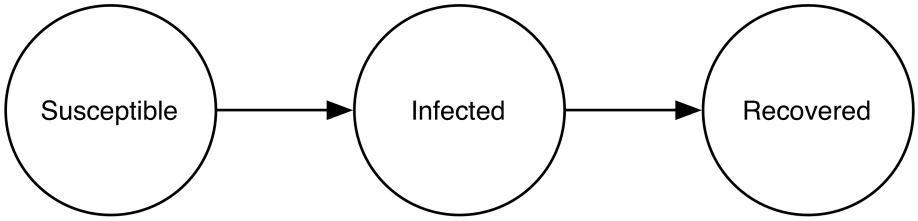
[Equation 5.5] $$S(t) = e^{-\int_0^t \mu(t)dt}$$

The function $\mu(t)$ is commonly referred to as a *hazard function*, meaning the hazardous probability that a peep will "end its life" at the warehouse (by leaving the warehouse, not via peep suicide) at a given time t, given that it survived in the warehouse until time t.

Now that we have a survival function that can accommodate a fluctuating rate across time periods, we can convert our model of peep consumption into a model of infectious disease epidemics.

The Kermack-McKendrik Model

Suppose we want to conceptualize the process of an epidemic using the classic *Kermack-McKendrick model*, which is also known as the *SIR model* of disease. People get shipped around between these states when they are infected with the disease being modeled (moving them from the warehouse of susceptible people, *S*, to the warehouse of infected people, *I*), or recover from infection (moving from the warehouse of infected people, *I*, to the warehouse of recovered people, *R*). The model assumes that the infection is not fatal, that infect*ed* people are infect*ious* people, and that recovered people have lifelong immunity. We'll modify these assumptions later in more complex models, after we learn the basics. (Note that in some infectious disease literature, the states of *S*, *I*, and *R* are referred to as *compartments*, and the SIR model as a *compartmental model*).



How do we use this simple structure to answer complex questions about controlling infectious disease epidemics? One systematic approach is to label and define all of the rates of flow between states of the model, so that we can construct a series of equations to describe the process of infection and recovery, which will help us calculate vital facts about an epidemic, such as how many vaccines to purchase to prevent it from spreading.

Based on Equation 5.5, we can reason that the rate of infection can be expressed as the reciprocal of the expected time until a person moves from the susceptible state to the infected state (let's call that $E(T_S)$). The expected time until a person moves from susceptible to infected is the average length of time between birth (which we assume is when a person enters the susceptible state) and the time a person gets infected. Similarly, the rate of recovery can be expressed as the reciprocal of the expected length of time between the infected state and the recovered state; let's call this expected length of time $E(T_I)$.

By this reasoning, if the average age of infection is 10 years old, then the average incidence rate of disease per person would be 1/10 years^{-1}. The average incidence rate per person per unit time is typically called the *force of infection*, and is typically abbreviated with the Greek letter lambda, .

Analogously, if the average duration of disease is 6 months (0.5 years), then the rate of recovery from the disease per infected person per year is $1/(0.5) = 2$ years^{-1}. Let's call this rate v.

There are two more key rates that haven't been drawn into our diagram, but that we'll need to account for in many infectious disease models. One is the birth rate, and the other is the death rate. If the population we are simulating is demographically stable—that is, if it's not rising or declining in population size—and the disease is not lethal, then the birth and death rates will be equal. If the average lifetime in the population (let's call it $E(T_L)$) is 75 years old, then if we assume an exponential distribution of life expectancy as per Equation 5.5, the rate of death in the population per person would 1/75 years^{-1}, which in a demographically stable population (i.e., a population in which the overall number of people is not changing) will be equal to the rate of birth. We'll call this rate .

Now that we have all rates of flow that we need between states, we can write down the equations to describe our model. The equations are easier to derive if we first create a table in which we itemize every rate of flow that enters or leaves each state of our model. We can go through the model state by state to write down the people who enter the state and the people who leave the state, as shown in Table 5.1. For the purposes of the Table and equations, letter N is the total population size (equal to $S + I + R$).

Table 5.1

Equations representing the number of people who enter and leave a given state in the standard Kermack-McKendrick model.

State	People entering state (+)	People leaving state (-)

Susceptible (S)	Births: birth rate per person in population x population size N	Deaths: death rate per person in this warehouse x population of this warehouse
		Infections: infection rate per susceptible x susceptible pop
Infected (I)	Infections: infection rate per susceptible x susceptible pop	Deaths: death rate per person in this warehouse x population of this warehouse
		Recoveries: recovery rate ν per infected x infected pop
Recovered (R)	Recoveries: recovery rate ν per infected x infected pop	Deaths: death rate per person in this warehouse x population of this warehouse

Translating the table into equation format, we can arrive at the differential equations below, where t reflects time:

[Equation 5.6] $\frac{dS(t)}{dt} = \mu N - \lambda S(t) - \mu S(t)$

[Equation 5.7] $\frac{dI(t)}{dt} = \lambda S(t) - vI(t) - \mu I(t)$

[Equation 5.8] $\frac{dR(t)}{dt} = vI(t) - \mu R(t)$

We can quickly check that we have included all of the important rates of disease by adding each of the equations together. Since we have assumed a constant population size, the sum of the right-hand sides of the equations should equal zero, meaning that there is no change in the overall population size. We can sum the equations to find that they sum to the expression x (N-(S+I+R)), but since N = (S+I+R), the overall expression does, in fact, add to zero.

As with Markov models, we can solve our infectious disease model at a steady state to get a sense of the long-term prevalence of being in each state. Under typical steady-state conditions, meaning that the infectious disease infects about the same number of people each year, the prevalence of disease is equal to the incidence of disease multiplied by the duration of disease. To understand this relationship, imagine the analogy of a restaurant: if customers in a restaurant come to eat at a rate of 10 per hour (the incidence of customers) and each customer stays for 2 hours (the duration of being a customer), then at any given time, we could reason that the prevalence of customers would be an old batch of 10 customers from the previous hour (who are now in their 2$^{\text{nd}}$ hour of eating), and a new batch of 10 from the current hour (who are still in their 1$^{\text{st}}$ hour of eating), for a total of 20 customers.

Analogously, we can derive an expression to estimate the steady state population of susceptible, infected, and recovered people in a community. First, we can estimate the prevalence of susceptible people at steady state. This prevalence will be the "incidence" of being susceptible (i.e., births, N) multiplied by the "duration" of being in the susceptible state (i.e., the length of time before either dying at rate or getting infected at rate , which are the only two ways to leave the susceptible state in this model). The expected duration of being in the susceptible state will be E[minimum(T_L or T_S)], since a person will either die first or get infected first, if moving out of the susceptible state. Making use of our survivor function (Equation 5.5), we get:

[Equation 5.9] $$\mathrm{E}(\min[\mathrm{T_L}, \mathrm{T_S}]) = \int_0^\infty e^{-(\mu+\lambda)} dt = \frac{1}{\mu+\lambda}$$

The number of susceptible people will be the incidence of susceptible people N times the duration of susceptibility $1/($ $)$, which produces Equation 5.10:

[Equation 5.10] $$S = \frac{\mu N}{\mu+\lambda}$$

Similarly, we can estimate the prevalence of infected people at steady state, which will be the incidence of infection (times the susceptible population given by equation 5.10), and the duration of infection ($1/(v+$), by the same logic as above). Hence, the number of infected people will be expressed by Equation 5.11:

[Equation 5.11] $$I = \frac{\mu N}{\mu+\lambda} \times \lambda \times \frac{1}{v+\mu} = \frac{N\mu\lambda}{(\lambda+\mu)(v+\mu)}$$

Equations 5.10 and 5.11 provide us with a useful strategy to estimate prevalence rates for stable diseases in a population that do not wildly fluctuate in incidence each year (*endemic*, rather than epidemic, diseases). For example, if we have a population of 100,000 people with an average life expectancy of 75 years, a force of infection of the disease of

interest of 1/10 years^{-1}, and a rate of recovery of 2 years^{-1}, we can calculate the number of susceptible people in the context of this non-fatal, endemic disease at steady state:

[Equation 5.12] $S = \frac{\mu N}{\mu+\lambda} = \frac{(\frac{1}{75})\times 100000}{(\frac{1}{75})+(\frac{1}{10})} = 11{,}765$

Around 11,765 people would be expected to remain susceptible to the disease at steady state. Analogously, the number of infected people would be:

[Equation 5.13] $I = \frac{\mu N}{\mu+\lambda} \times \lambda \times \frac{1}{v+\mu} = \frac{N\mu\lambda}{(\lambda+\mu)(v+\mu)} = \frac{100000\times(\frac{1}{75})\times(\frac{1}{10})}{(\frac{1}{10}+\frac{1}{75})(2+\frac{1}{75})} = 584$

Finally, the number of people who have recovered would be the population size N minus the number of susceptible people, and minus the number of infected people, which is 87,651.

The Law of Mass Action

While it may be of interest that we can calculate the steady state prevalence of non-fatal endemic diseases using only a piece of paper and a pencil, we want to understand how to control diseases that are not at a simple endemic, non-fatal steady state condition.

First, we want to address infectious diseases that are fatal. We can incorporate the fatality from an infectious disease by modifying Equations 5.6 through 5.8 by adding an additional rate of death from the infection itself, to indicate the extra death caused by the disease. This might change the demography of the population, if the additional deaths from the disease are not exactly compensated by new births.

Second, we want to address diseases that are *epidemic* and increasing, not just *endemic* and stable. We can modify our equations to understand how disease may start out as rare but emerge into an epidemic. To do so, we need to account for the fact that,

by definition, infectious diseases often do not have constant rates of infection, but have changing rates of infection based on the number of infectious people in the population. If there are a lot of infectious people in the population, the risk of infection will be high. If a large number of people are cured of their disease and in the recovered state, then the risk of infection will be low.

To address this goal, we can allude to the challenge we faced when we encountered a changing demand for peeps in our warehouse analogy; we derived Equation 5.5 to describe how the rate of demand might change over time. Now we can use that equation to describe how the rate of infection can change over time, to deal with epidemic diseases for which the infection rate is dynamic.

The most common approach to introduce a dynamic rate of infection is called the *Law of Mass Action*. The Law of Mass Action states that in a homogeneously-mixed population (i.e., a population in which people are circulating around and of generally equivalent risk of being infected with the disease), then the incidence rate of the disease will be proportional to the population of infectious people (e.g., the infectious people will be dispersed throughout the community and collide with susceptible people):

[Equation 5.14] $\lambda(t) = \beta I(t)$

For example, suppose that we have a disease for which the force of infection is 1/10 years^{-1} and the number of people infected currently is 584. We would calculate β=(1/10)/584 = 0.000171. It is typically the case that β is a small number, because it expresses the risk of disease upon one contact between an infectious and a susceptible person per unit time, which is hopefully small.

In Equation 5.14, *(t)* is our hazard function for the incidence rate, which changes over time depending on how many infectious *I(t)* people there are at the time. The constant is a transmission rate per person per unit time. It tells us how likely it is that a contact between one susceptible person and one infectious person will result in transmission of the disease in question. As a result, it doesn't change for a specific disease in a specific community, but indicates the "contagiousness" of the specific disease being studied in a particular community.

Using the Law of Mass Action, we can re-write our series of differential equations (equations 5.6 through 5.9) by replacing all of the 's with *I*'s, to keep track of an epidemic disease whose infection rate changes:

[Equation 5.15] $\frac{dS(t)}{dt} = \mu N(t) - \beta I(t)S(t) - \mu S(t)$

[Equation 5.16] $\frac{dI(t)}{dt} = \beta I(t)S(t) - vI(t) - \mu I(t)$

[Equation 5.17] $\frac{dR(t)}{dt} = vI(t) - \mu R(t)$

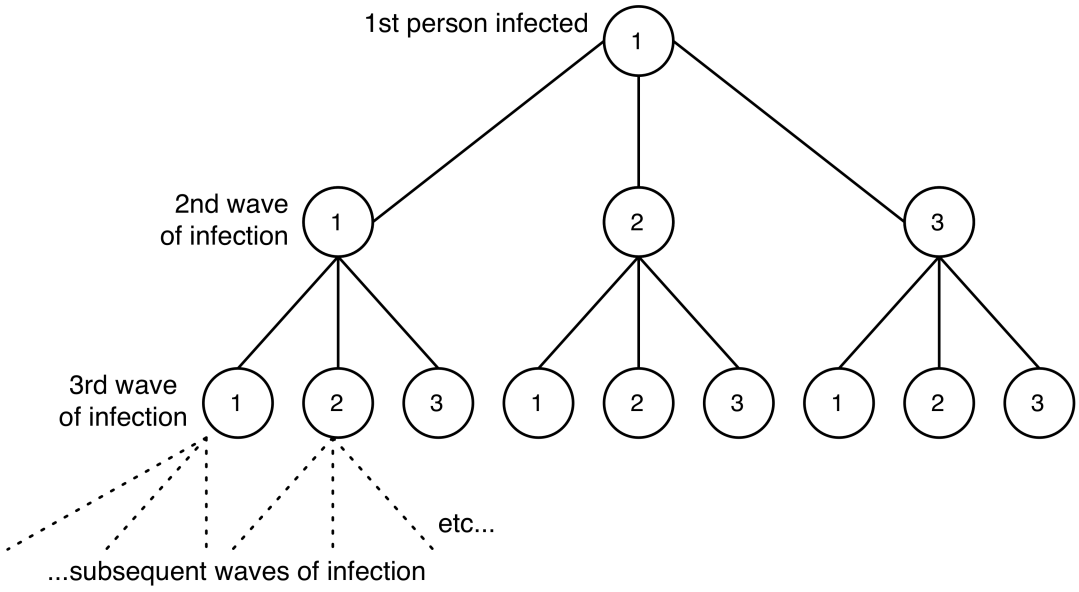
Of course, we can still use these equations for an endemic disease at steady-state, for which *I(t)* will be (by definition) constant, such that the force of infection will be constant at all times *t*.

As a side note, some authors will make optional modifications to Equation 5.15 through 5.17. For example, some will replace with two variables multiplied together: the probability of contact between infectious and susceptible people, and the probability that this contact will result in disease transmission. This is useful, for example, when modeling some sexually-transmitted diseases, as we will see later in the book. Alternatively, it is possible to change to equal *I(t)/N*, so that is multiplied by the fraction of the

population that is infectious. The expression scales up the value of based on the population size.

The reproductive number

The *basic reproductive number*, or R_0 (commonly pronounced "are-naught", due to its British origins) is the mean number of secondary infections a single infected person will cause in a population with no immunity to the disease, in the absence of interventions to control the infection. If we imagine an island of people who have never been affected by the disease, then R_0 is the number of people who would be infected by one infected person who lands on the island (a fully susceptible population), before the infected person recovers or dies (is no longer infectious). Figure 5.4 illustrates a disease with an R_0 of 3.



It is important not to confuse R_0 with the *effective reproductive number* (R or R_e, depending on the author), which is the number of secondary cases generated by an infectious case once the epidemic is underway (e.g., once the population is no longer fully susceptible, but there are some immune persons, or after some interventions have been introduced such as medical treatment). Of course, we're interested in driving down the effective reproductive number by creating public health programs, but the basic reproductive number is a fundamental property of a disease spreading in a given population.

We can reason that if the R_0 of a disease is greater than 1, then one infected person will produce more than one infected person in the next generation of infections on average, hence the disease will spread. Conversely, if R_0 is less than 1, then one infected person will produce less than one infected person in the next generation of infections on average, hence the epidemic will eventually burn out.

For epidemic diseases that initially expand, there is some point at which the number of susceptible people will be so few that infected people will not be able to transfer the disease onto as many susceptible people (most people who are in contact with the infected person are either recovered or infected). Hence, the disease transmission process will reach a point when the effective R will be 1, meaning there is some population size S such that:

[Equation 5.18] $R_0 \times \frac{S}{N} = 1$

Hence:

[Equation 5.19] $R_0 = \frac{N}{S}$

The point at which the effective R will be equal to 1 is effectively the point we solved for earlier in the endemic steady state disease situation, in which the prevalence of being in a given state is equal to the incidence of moving into the state times the duration of being in the state. The "incidence" of being in the population is the birth rate times the population that can give birth N, and the "duration" of being in the population is the life expectancy $E(T_L)$. Similarly, the "incidence" of being susceptible is the birth rate times the population that can give birth N, and the "duration" of being susceptible is E[minimum(T_L or T_S)], since a person will either die first or get infected first, if moving out of the susceptible state. Therefore, Equation 5.19 becomes Equation 5.20:

[Equation 5.20] $$R_0 = \frac{N}{S} = \frac{\mu NE(T_L)}{\mu NE[\min(T_L,T_S)]} = \frac{E(T_L)}{E[\min(T_L,T_S)]} \approx \frac{E(T_L)}{E(T_S)} = \frac{\frac{1}{\mu}}{\frac{1}{\mu+\lambda}} = 1 + \frac{\lambda}{\mu}$$

The approximation in the midst of this derivation is because T_S is far smaller than T_L in the vast, vast majority of cases (since life expectancy would be longer than the time to get disease).

We can simplify our expression for R_0 further by remembering that I, and using Equation 5.13:

[Equation 5.21] $$\lambda(t) = \beta I(t) = \frac{\beta N \mu \lambda}{(\lambda+\mu)(v+\mu)}$$

After some simplification and rearrangement, we can substitute Equation 5.21 into Equation 5.20 and end up with this very well-known formula for R_0:

[Equation 5.22] $$R_0 = 1 + \frac{\lambda}{\mu} = \frac{\beta N}{v+\mu} \approx \frac{\beta N}{v}$$

The last approximation is because v is much larger than usually.

What we've accomplished here is to use the endemic steady-state expressions to derive a generalizable expression for R_0. This can help us to understand the dangers of a disease and the conditions under which we might be able to control the disease. For example, suppose we have a population of 100,000 people and a disease for which β=0.000171 and ν=2 years^{-1}, then using Equation 5.22, $R_0 = (0.000171*100000)/2 = 8.6$. Each infected person will on average infect 8.6 others in a fully susceptible population.

Using the Kermack-McKendrick model to plan a vaccination campaign

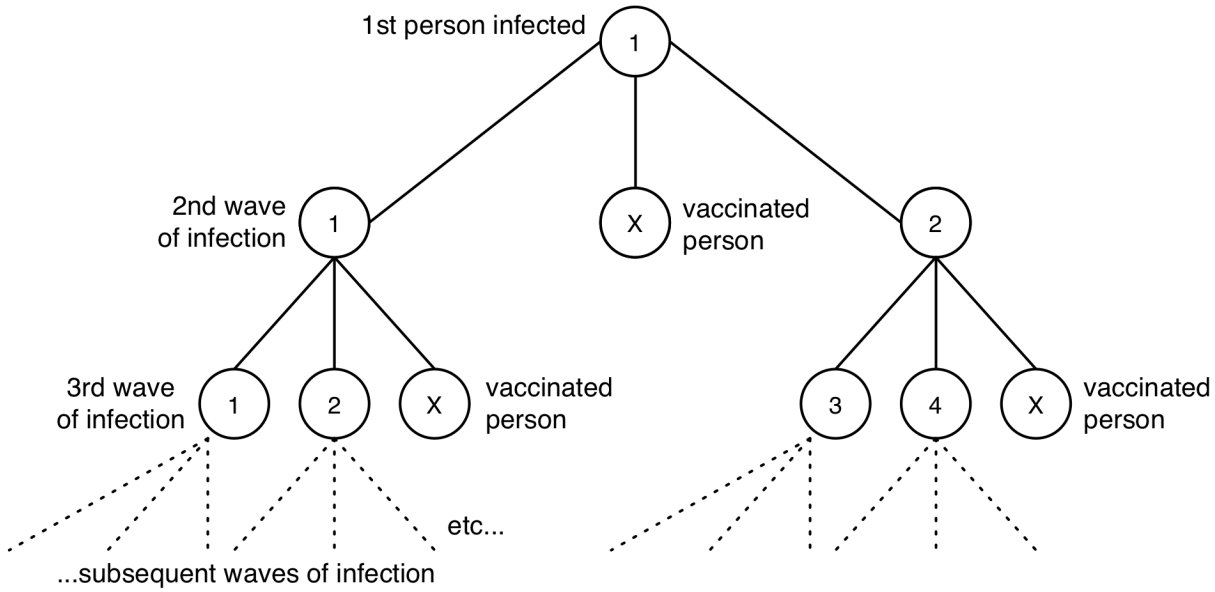
Following a recent terrible earthquake, the country of Haiti experienced an outbreak of cholera after sewer systems were damaged and drained into water sources. Cholera causes potentially-fatal diarrhea, and is usually prevented by maintaining clean water supplies. To prevent the further spread of the disease, United Nations officials not only supplied clean water, but also purchased supplies of a cholera vaccine.

Suppose we are charged with providing cholera vaccines for a village of 1,000 people in Haiti. Based on data from prior outbreaks, we know that each person in Haiti who becomes infected with cholera typically infects three other people with the disease before recovering or dying from the illness ($R_0 = 3$), causing the outbreak to continue spreading. What is the minimum number of vaccines that should be purchased to stop the disease from continuing to spread in the village? Note that because of the biological concept of *herd immunity*, which refers to the ability of a group of unvaccinated people to be protected from infection because they are surrounded by vaccinated people, we

don't need to buy 1,000 vaccines. The vaccinated people act like a virtual wall surrounding each unvaccinated person, preventing a chain of infection from connected unvaccinated people to each other and perpetuating an epidemic. The extra money saved from vaccine purchases can be directed to other important public health activities such as maintaining clean water supplies, so we wish to calculate the minimum number of vaccines necessary to stop the outbreak to optimize use of our precious resources.

Figure 5.4 illustrates how we might visualize the chain of infection from a single person to the people who they subsequently infect.

Our goal in this problem is to convert a chain of transmission from one that enables perpetual transmission, such as shown in Figure 5.4, to a scenario in which sufficient people are vaccinated that the chain of transmission cannot continue. Figure 5.5 illustrates how a chain of transmission can be interrupted when people are vaccinated; but as shown in the Figure, if only a few people are vaccinated, the chain of transmission can continue; our task is to determine what fraction of people must be vaccinated for the chain to eventually die out.



To solve this problem, we can derive an equation describing how vaccination affects the chain of transmission. If no people are vaccinated (Figure 5.4), then with every next wave of infection, the number of people who are newly infected with cholera are the number who were previously infected, multiplied by three (the value of R_0). If one person is currently infected, then in the next wave of infection, three people will be infected, and in the next wave of infection, nine people will be infected, and so on.

However, if people are vaccinated, then with every next wave of infection, the number of people who are newly infected with cholera are the number who were previously infected, multiplied by three, and then multiplied by the fraction of people who are unvaccinated (the fraction who remain susceptible to the disease). If one person is currently infected, and one third of people are vaccinated against the disease, then in

the next wave of infection, only two people will be infected instead of three, and in the next wave of infection, only four people will be infected instead of nine (Figure 5.5), and so on. That is, if f is the fraction of people who have been vaccinated in the village, then:

[Equation 5.23] $$new\ infections = (old\ infections) \times\ R_0\ \times\ (1 - f)$$

To stop an epidemic, we need the number of new people infected in the next wave of disease to be less than the number of people previously infected, so that each next wave of disease has less and less people infected, and the epidemic "dies out". Put another way:

[Equation 5.24] $$\frac{new\ infections}{old\ infections} < 1$$

The number of new infections divided by the number of old infections must be less than 1, meaning that fewer infections must occur with each subsequent wave of disease for the chain of transmission to eventually die out. We know from Equation 5.23 that the number of new infections divided by the number of old infections is $R_0\ \times\ (1 - f)$, hence by combining Equations 5.23 and 5.24, we get:

[Equation 5.25] $$R_0\ \times\ (1 - f) < 1$$

Solving for the value of f, we find that:

[Equation 5.26] $$f > 1 - 1/R_0$$

to achieve herd immunity. In this case, we want $f > 0.666$, or 66.6% of the population, to achieve herd immunity. In a village of 1,000 people, this means we must vaccinate at least 666 people to stop the cholera epidemic.

<2>Implementing the Kermack-McKendrick model in R

An example of *R* code to program the model is provided on the textbook website. To program the Kermack-McKendrick model in *R*, we specify four key elements of the model: the parameters that will go into our equations, how long we want to run the simulation, the initial conditions, and the equations themselves.

First, let's input some parameters for a non-fatal SIR disease epidemic among a population of 100,000 people, with a life expectancy of 75 years, $\beta = 0.000171$ years^{-1}, and $v = 2$:

```
N = 100000
mu = 1/75
beta = 0.000171
v = 2
```

Second, we specify how long we want to run the model, and with what time steps. Let's say we want to simulate 5 years of time, with small time increments of 0.01 years between simulated periods. The time step reflects a small period of the simulation over which to integrate the differential equations; it must be small relative to the simulated period to ensure accuracy of approximating differential equations with difference equations (remember Riemann sums from calculus, which approximate the sum of the area under a curve by drawing thin rectangles under the curve; here, the curve is the

number of people over time in each state *S, I* and *R*, and the rectangles are the small areas under those curves defined by each time step):

```
time = 5
dt = 0.01
```

Third, we input our initial conditions, meaning the number of people who are in each of the states *S, I* and *R* at the start of the epidemic. For illustrative purposes, let's have 1 infected person and everyone else be susceptible. We'll also create some vectors named after each state (*Svec, Ivec* and *Rvec*), which we'll later expand for future time points, but which for now will only contain the initial conditions:

```
S = 99999
I = 1
R = 0

Svec = S
Ivec = I
Rvec = R
```

Finally, we write our equations and insert them into two for loops to have *R* update over each period of time (first for loop) and across all small time steps (second for loop). In each loop, we attach the current value of the state to the list of previous values, expanding the vector for each state (*Svec, Ivec* and *Rvec*) with each time step across all simulated time points. Note that there are more efficient ways to code this using the library 'deSolve', but this way is more intuitive for learning purposes:

```
for (i in 1:time){
  for (i in 1:(1/dt)){
    S = S + mu*N*dt - beta*S*I*dt - mu*S*dt
    I = I + beta*S*I*dt - v*I*dt - mu*I*dt
    R = R + v*I*dt - mu*R*dt
    Svec = c(Svec, S)
    Ivec = c(Ivec, I)
    Rvec = c(Rvec, R)
  }
}
```

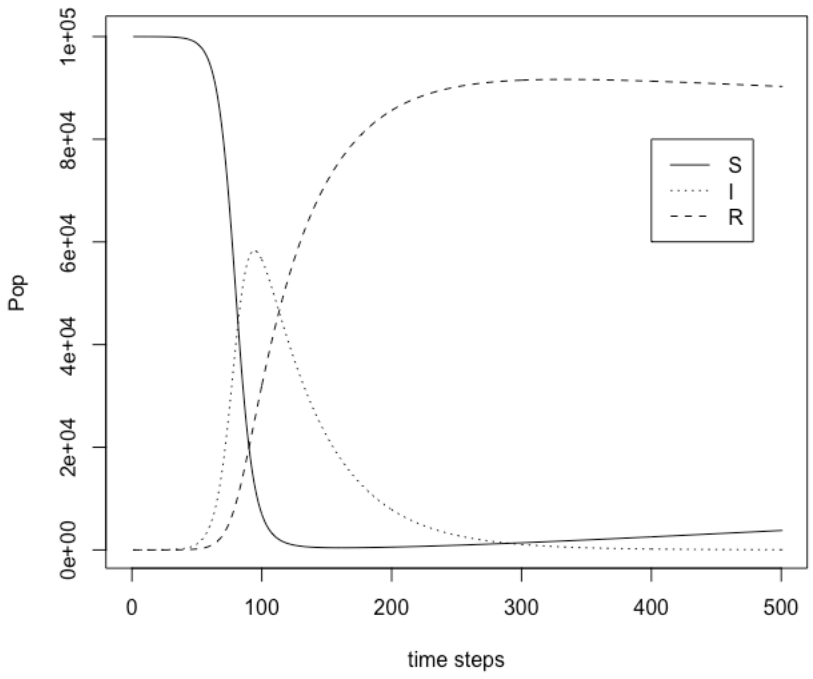
Now we can plot the result; here, we use the Quick-R website (http://www.statmethods.net/graphs/line.html) to teach ourselves of how to draw a smooth line (lty=l), or different types of dashed lines (lty=2 or lty=3), and label our axes (xlab="time steps", ylab="Pop"). We choose one vector to start the plot with the plot command, then add the other two lines using the lines command:

```
plot(Svec,lty=1,type="l",xlab="time steps",ylab="Pop",xlim=c(0,500))
lines(Ivec,lty=3)
lines(Rvec,lty=2)
```

The Quick-R site also reminds us how to add a legend by specifying the x-axis location for the left corner of the legend box, the y-axis location for the top of the legend box, the labels for the legend, and the type of lines we want in the legend:

```
legend(400,80000,c("S","I","R"),lty=c(1,3,2))
```

The plot commands produce Figure 5.6:



We can use the power of R to extend the model to do a large-scale uncertainty analysis. Suppose that we don't know exactly what the value of the input parameters are. We might know, for example, that = 0.000171 on average, but has a standard deviation of 0.00001, and $v = 2$ on average, but has a standard deviation of 0.1. How much would this affect the number of people infected in the first year of the epidemic?

We can first create a vector to keep track of how many total people get infected over the simulation, by adding a vector Totinf = 0 before the "for loop" and adding within the for loop the code:

```
Totvec = c(Totvec, beta*S*I*dt)
```

This code keeps track of how many people are newly infected in each time step. Hence, the sum from time 0 to time 1/dt will be the total number of people infected over the course of the epidemic:

```
> sum(Totvec)
[1] 101842.4
```

Next, we can make our model *stochastic* instead of *deterministic*. That is, rather that having the parameters equal pre-determined set values, we can have the model incorporate parameters that are stochastic, or chosen from random distributions we specify. In this case, our initial parameters will now be:

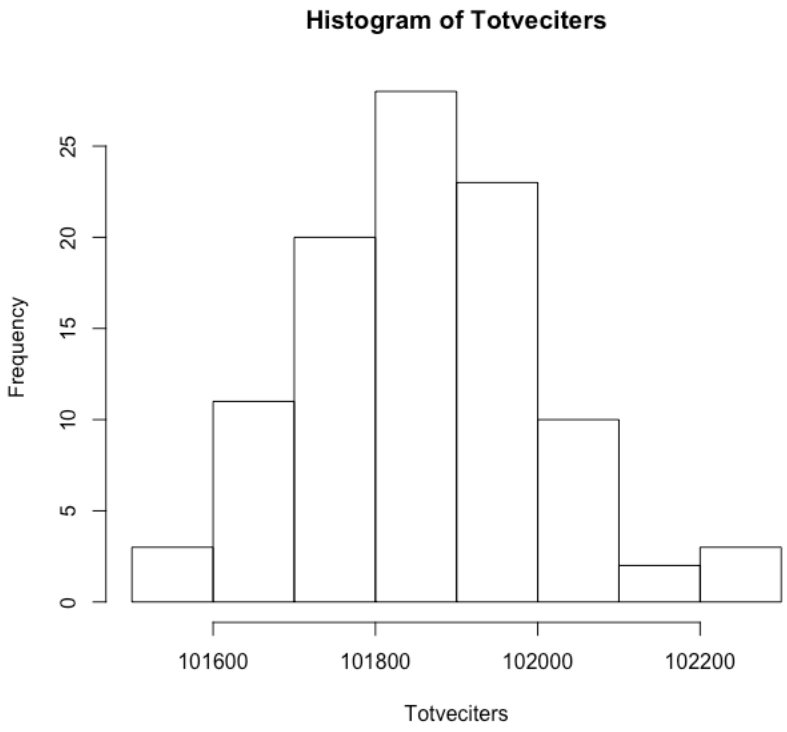
```
N = 100000
mu = 1/75
beta = rnorm(1,mean=0.000171,sd=0.00003)
v = rnorm(1,mean=2,sd=0.2)
```

The rnorm commands indicate to R that the program should sample once from a normal distribution with means and standard deviations as specified. Now, we want to

know how much the final size of the epidemic (the total number of people infected over the course of the epidemic) might change due to uncertainty in the parameter values. We can wrap the entire code in a larger for loop and run it 100 times, storing the value of sum(Totvec) in a new vector called Totveciters, which will include the value of Totvec for each of the 100 iterations:

```
Totveciters=c()
for (iters in 1:100){

N = 100000
mu = 1/75
beta = rnorm(1,mean=0.000171,sd=0.00001)
v = rnorm(1,mean=2,sd=0.1)

time = 5
dt = 0.01

S = 99999
I = 1
R = 0

Svec = S
Ivec = I
Rvec = R
Totvec = 0

for (i in 1:time){
  for (i in 1:(1/dt)){
    S = S + mu*N*dt - beta*S*I*dt - mu*S*dt
    I = I + beta*S*I*dt - v*I*dt - mu*I*dt
    R = R + v*I*dt - mu*R*dt
    Svec = c(Svec, S)
    Ivec = c(Ivec, I)
    Rvec = c(Rvec, R)
    Totvec = c(Totvec, beta*S*I*dt)
  }
}

plot(Svec,lty=1,type="l",xlab="time steps",ylab="Pop",xlim=c(0,500))
lines(Ivec,lty=3)
lines(Rvec,lty=2)
legend(400,80000,c("S","I","R"),lty=c(1,3,2))

Totveciters=c(Totveciters,sum(Totvec))
}
```

We can finally plot the histogram that tells us how much the size of our epidemic in the first year varied due to the uncertainty in our input parameters, by typing in hist(Totveciters), which gives us Figure 5.7:



We see in Figure 5.7 that while the distribution of the final size of the epidemic centers around 101,900 people, there can be as few as 101,500 or as many as 102,300 people infected simply due to uncertainties in the two main parameters describing the infectious disease.