Chapter Two

Introduction to Simple Epidemic Models

The process of modeling in epidemiology has, at its heart, the same underlying philosophy and goals as ecological modeling. Both endeavors share the ultimate aim of attempting to understand the prevalence and distribution of a species, together with the factors that determine incidence, spread, and persistence (Anderson and May 1979; May and Anderson 1979; Earn et al. 1998; Shea 1998; Bascompte and Rodriguez-Trelles 1998). Whereas in ecology the precise abundance of a species is often of great interest, establishing or predicting the exact number of, for example, virus particles in a population (or even within an individual) is both daunting and infeasible.[1] Instead, modelers concentrate on the simpler task of categorising individuals in the "host" population according to their infection status. As such, these epidemiological models can be compared to the metapopulation models used in ecology (Levins 1969; Hanski and Gilpin 1997), where each individual host is considered as a patch of resource for the pathogen, with transmission and recovery analogous to dispersal and extinction (Nee 1994; Rohani et al. 2002).

There is a long and distinguished history of mathematical modeling in epidemiology, going back to the eighteenth century (Bernoulli 1760). However, it was not until the early 1900s that the increasingly popular dynamical systems approaches were applied to epidemiology. Since then, theoretical epidemiology has witnessed numerous significant conceptual and technical developments. Although these historical advances are both interesting and important, we will side-step a detailed account of these progressions and instead refer interested readers to the lucid texts by Bailey (1975), Anderson and May (1991), Grenfell and Dobson (1995), Daley and Gani (1999), and Hethcote (2000).

In this chapter, we start with the simplest epidemiological models and consider both infections that are strongly immunizing as well as those that do not give rise to immunity. In either case, the underlying philosophy is to assume individuals are either susceptible to infection, currently infectious, or recovered (previously infected and consequently immune). Although the progress between these classes could be presented as a verbal argument, to make quantitative predictions we must translate them into formal mathematical terms. This chapter presents the mathematical equations describing these models, together with the kinds of model analyses that have proved useful to epidemiologists. These approaches encompass both deterministic and probabilistic frameworks. The preliminary models will, of necessity, be somewhat primitive and ignore a number of well-known and important heterogeneities, such as differential susceptibility to infection, contact networks, variation in the immunological response, and transmissibility. Many of these complexities are addressed in subsequent chapters.

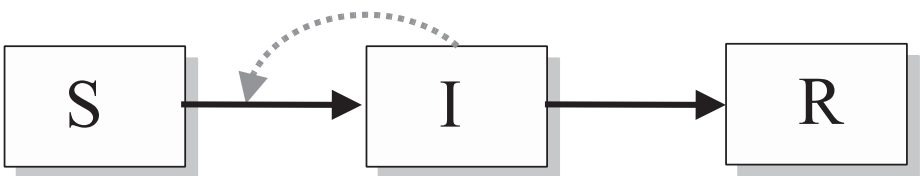
[1] A clear exception to this is the study of macroparasitic infections (such as helminthic worms) where the worm burden is of great interest because it can substantially affect both host and parasite demography. In this book, we do not deal with such systems and refer interested readers to Anderson and May's (1991) thorough treatment of the subject.

2.1. FORMULATING THE DETERMINISTIC *SIR* MODEL

In order to develop a model, we first need to discuss terminology. Infectious diseases are typically categorized as either acute or chronic. The term acute refers to "fast" infections, where relatively rapid immune response removes pathogens after a short period of time (days or weeks). Examples of acute infections include influenza, distemper, rabies, chickenpox, and rubella. Chronic infections, on the other hand, last for much longer periods (months or years) and examples include herpes and chlamydia. We start the development of models by focusing on acute infections, assuming the pathogen causes illness for a period of time followed by (typically lifelong) immunity. This scenario is mathematically best described by the so-called *S*-*I*-*R* models (Dietz 1967). This formalism, which was initially studied in depth by Kermack and McKendrick (1927), categorizes hosts within a population as **S**usceptible (if previously unexposed to the pathogen), **I**nfected (if currently colonized by the pathogen), and **R**ecovered (if they have successfully cleared the infection).

Now that we know how many categories there are and how these categories are defined, the question becomes how individuals move from one to the other. In the simplest case (ignoring population demography—births, deaths, and migration), we have only the transitions $S \to I$ and $I \to R$. The second of these is easier, so we deal with it first. Those infected can move to the recovered class only once they have fought off the infection. For acute infections, it is generally observed that the amount of time spent in the infectious class (the "infectious period") is distributed around some mean value, which can be often estimated accurately from clinical data. From a modeling perspective, this translates into the probability of an individual moving from I to R being dependent on how long they have been in the I class. However, modelers often make the simplifying assumption that the recovery rate γ (which is the inverse of the infectious period) is constant; this leads to far more straightforward equations and exponentially distributed infectious periods. In Section 3.3 we deal with the dynamical consequences of alternative, more realistic formulations of the infectious period.

The progression from S to I clearly involves disease transmission, which is determined by three distinct factors: the prevalence of infecteds, the underlying population contact structure, and the probability of transmission given contact. For a directly transmitted pathogen, there has to be contact between susceptible and infected individuals and the probability of this happening is determined by the respective levels of S and I, as well as the inherent contact structure of the host population. Finally, we need to take into account the likelihood that a contact between a susceptible and an infectious person results in transmission.



These conceptual descriptions of the model can be represented by a flow diagram. The flow diagram for the *SIR* model uses black arrows to show the movement between the S and I classes and the I and R classes. The fact that the level of the infectious disease influences the rate at which a susceptible individual moves into the infected class is shown by the dotted gray arrow. This book uses such flow diagrams to illustrate the essential

epidemiological characteristics. In general, demography will be ignored in these diagrams to reduce the number of arrows and hence improve their clarity.

Flow diagrams provide a useful graphical method of illustrating the main epidemiological assumptions underlying a model.

The previous paragraphs make the derivation of the transmission term seem relatively straightforward. Unfortunately, the precise structure of the transmission term is plagued by controversy and conflicting nomenclature. To explain some of these issues, we start by defining the *force of infection*, λ, which is defined as the *per capita* rate at which susceptible individuals contract the infection. Thus, the rate at which new infecteds are produced is λX, where X is the *number* of individuals in class S. This force of infection is intuitively proportional to the number of infectious individuals. For directly transmitted pathogens, where transmission requires contact between infecteds and susceptibles, two general possibilities exist depending on how we expect the contact structure to change with population size: $\lambda = \beta Y/N$ and $\lambda = \beta Y$ (where Y is the number of infectious individuals, N is the total population size, and β is the product of the contact rates and transmission probability). The first of these formulations will be referred to as frequency dependent (or mass action) transmission and the second as density dependent (or pseudo mass action) transmission. (We note, however, that Hamer (1906) refers to $\lambda = \beta Y$ as mass-action; this duality of nomenclature causes much confusion). A mechanistic derivation of the transmission term is provided in Box 2.1.

It is important to distinguish between these two basic assumptions in terms of the underlying structure of contacts within the population. Frequency-dependent transmission reflects the situation where the number of contacts is independent of the population size. At least as far as directly transmitted diseases are concerned, this agrees with our natural intuition about human populations. We would not expect someone living in, for example, London (population 7 million), or New York (population 8 million), to transmit an infectious disease over 50 times more than someone living in Cambridge, United Kingdom (population 130,000) or Cambridge, Massachusetts (population 100,000). The number of close contacts that are likely to result in disease transmission will be determined by social constraints, resulting in similar patterns of transmission in any large town or city. Indeed, estimates of measles transmission rates in England and Wales demonstrate no relationship with population size (Bjørnstad et al. 2002). In contrast, density-dependent transmission assumes that as the population size (or more accurately, as the density of individuals) increases, so does the contact rate. The rationale is that if more individuals are crowded into a given area (and individuals effectively move at random), then the contact rate will be greatly increased. As a rule of thumb, frequency-dependent (mass action) transmission is considered appropriate for vector-borne pathogens and those with heterogeneous contact structure. Density-dependent (psuedo mass action) transmission, however, is generally considered to be more applicable to plant and animal diseases, although care must be taken in the distinction between number and density of organisms (for further discussion, we refer interested readers to McCallum et al. 2001; Begon et al. 2002).

The distinction between these two transmission mechanisms becomes pronounced when host population size varies, otherwise the $1/N$ term can be absorbed into the parameterization of β in the mass-action term. As a simplification to our notation, it is convenient to let $S(= X/N)$ and $I(= Y/N)$ to be the *proportion* of the population that are susceptible or infectious, respectively. In this new notation our mass-action

(frequency-dependent) assumption becomes βSI, which informs about the rate at which new infectious individuals (as a proportion of the total population size) are infected.

In some instances, such as when we need to employ integer-valued stochastic models (Chapter 6), variables need to reflect numbers rather than proportions. To distinguish between these different approaches, this book will consistently use X, Y, and Z to represent the *numbers* in each class and S, I, and R to represent *proportions* (see Parameter Glossary).

Box 2.1 The Transmission Term

Here, we derive from first principles the frequency-dependent (mass action) transmission term (also called proportionate mixing; Anderson and May 1992), which is commonly used in epidemic models. It assumes homogenous mixing in the population, which means everyone interacts with equal probability with everyone else; it discards possible heterogeneities arising from age, space, or behavioral aspects (see Chapters 3 and 7).

Consider a susceptible individual with an average κ contacts per unit of time. Of these, a fraction $I = Y/N$ are contacts with infected individuals (where Y is the *number* of infectives and N is the total population size). Thus, during a small time interval (from t to $t + \delta t$), the number of contacts with infecteds is $(\kappa Y/N) \times (\delta t)$. If we define c as the probability of successful disease transmission following a contact, then $1 - c$ is the probability that transmission does not take place. Then, by independence of contacts, the probability (denoted by $1 - \delta q$) that a susceptible individual escapes infection following $(\kappa Y/N \times \delta t)$ contacts is

$$1 - \delta q = (1 - c)^{(\kappa Y/N)\delta t}.$$

Hence, the probability that the individual is infected following any of these contacts is simply δq.

We now define $\beta = -\kappa \log(1 - c)$ and substitute into the expression for $1 - \delta q$, which allows us to rewrite the probability of transmission in a small time interval δt as

$$\delta q = 1 - e^{-\beta Y \delta t/N}.$$

To translate this probability into the *rate* at which transmission occurs, first we expand the exponential term (recalling that $e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} \ldots$), divide both sides by δt, and take the limit of $\delta q/\delta t$ as $\delta t \to 0$. This gives:

$$\frac{dq}{dt} = \beta Y/N,$$

which is the transmission rate per susceptible individual. In fact, this quantity is often represented by λ and referred to as the "force of infection"—it measures the per capita probability of acquiring the infection (Anderson and May 1991). Then, by extention, the total rate of transmission to the entire susceptible population is given by

$$\frac{dX}{dt} = -\lambda X = -\beta XY/N,$$

where X is defined as the *number* of susceptibles in the population. If we rescale the variables (by substituting $S = X/N$ and $I = Y/N$) so that we are dealing with *fractions* (or densities), the above equation becomes

$$\frac{dS}{dt} = -\beta IS.$$

The transmission term is generally described by frequency dependence $\beta XY/N$ (or βSI), or by density dependence βXY.

The differences between frequency- and density-dependent transmission become important if the population size changes or we are trying to parameterize disease models across a range of population sizes.

2.1.1. The SIR Model Without Demography

To introduce the model equations, it is easiest to consider a "closed population" without demographics (no births, deaths, or migration). The scenario we have in mind is a large naive population into which a low level of infectious agent is introduced and where the resulting epidemic occurs sufficiently quickly that demographic processes are not influential. We also assume homogeneous mixing, whereby intricacies affecting the pattern of contacts are discarded, yielding βSI as the transmission term. Given the premise that underlying epidemiological probabilities are constant, we get the following SIR equations:

$$\frac{dS}{dt} = -\beta SI, \tag{2.1}$$

$$\frac{dI}{dt} = \beta SI - \gamma I, \tag{2.2}$$

$$\frac{dR}{dt} = \gamma I. \tag{2.3}$$

This is online program 2.1

The parameter γ is called the removal or recovery rate, though often we are more interested in its reciprocal $(1/\gamma)$, which determines the average infectious period. For most diseases, the infectious period can be estimated relatively precisely from epidemiological data. Note that epidemiologists typically do not write the equation for the R class because we know that $S + I + R = 1$, hence knowing S and I will allow us to calculate R. These equations have the initial conditions $S(0) > 0$, $I(0) > 0$, and $R(0) = 0$. An example of the epidemic progression generated from these equations is presented in Figure 2.1; the conversion of susceptible to infectious to recovered individuals is clear.

Despite its extreme simplicity, this model (equations (2.1) to (2.3)) cannot be solved explicitly. That is, we cannot obtain an exact analytical expression for the dynamics of S and I though time; instead the model has to be solved numerically. Nevertheless, the model has been invaluable for highlighting at least two very important qualitative epidemiological principles.

2.1.1.1. The Threshold Phenomenon

First, let us consider the initial stages after $I(0)$ infectives are introduced into a population consisting of $S(0)$ susceptibles. What factors will determine whether an epidemic will occur or if the infection will fail to invade? To answer this, we start by rewriting equation (2.2) in the form

$$\frac{dI}{dt} = I(\beta S - \gamma). \tag{2.4}$$

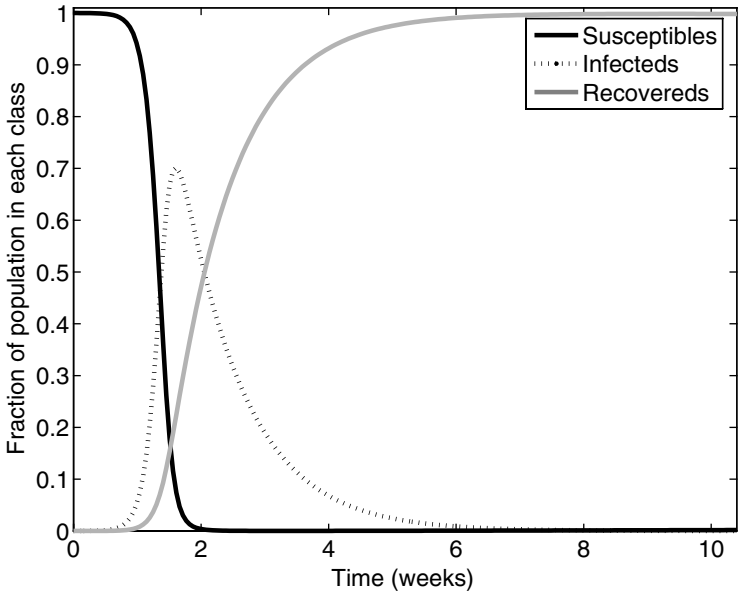


Figure 2.1. The time-evolution of model variables, with an initially entirely susceptible population and a single infectious individual. The figure is plotted assuming $\beta = 520$ per year (or 1.428 per day) and $1/\gamma = 7$ days, giving $R_0 = 10$. (See the following pages for a definition of the crucial parameter R_0)

If the initial fraction of susceptibles ($S(0)$) is less than γ/β, then $\frac{dI}{dt} < 0$ and the infection "dies out." This is a famous result due to Kermack and McKendrick (1927) and is referred to as the "threshold phenomenon" because initially the proportion of susceptibles in the population must exceed this critical threshold for an infection to invade (see also Ransom 1880, 1881; Hamer 1897). Alternatively, we can interpret this result as requiring γ/β, the relative removal rate, to be small enough to permit the disease to spread. The inverse of the relative removal rate is called the **basic reproductive ratio** (universally represented by the symbol R_0) and is one of the most important quantities in epidemiology. It is defined as:

> *the average number of secondary cases arising from an average primary case in an entirely susceptible population*

and essentially measures the maximum reproductive potential for an infectious disease (Diekman and Heesterbeek 2000). We can use R_0 to re-express the threshold phenomenon; assuming everyone in the population is initially susceptible ($S(0) = 1$), a pathogen can invade only if $R_0 > 1$. This makes very good sense because any infection that, on average, cannot successfully transmit to more than one new host is not going to spread (Lloyd-Smith et al. 2005). Some example diseases with their estimated R_0s are presented in Table 2.1; due to differences in demographic rates, rural-urban gradients, and contact structure, different human populations may be associated with different values of R_0 for the same disease (Anderson and May 1982). The value of R_0 depends on both the disease and the host population. Mathematically, we can calculate R_0 as the rate at which new cases are produced by an infectious individual (when the entire population is susceptible)

TABLE 2.1.
Some Estimated Basic Reproductive Ratios.

Infectious Disease	*Host*	*Estimated* R_0	*Reference*
FIV	Domestic Cats	1.1–1.5	Smith (2001)
Rabies	Dogs (Kenya)	2.44	Kitala et al. (2002)
Phocine Distemper	Seals	2–3	Swinton et al. (1998)
Tuberculosis	Cattle	2.6	Goodchild and Clifton-Hadley (2001)
Influenza	Humans	3–4	Murray (1989)
Foot-and-Mouth Disease	Livestock farms (UK)	3.5–4.5	Ferguson et al. (2001b)
Smallpox	Humans	3.5–6	Gani and Leach (2001)
Rubella	Humans (UK)	6–7	Anderson and May (1991)
Chickenpox	Humans (UK)	10–12	Anderson and May (1991)
Measles	Humans (UK)	16–18	Anderson and May (1982)
Whooping Cough	Humans (UK)	16–18	Anderson and May (1982)

multiplied by the average infectious period:

For an infectious disease with an average infectious period given by $1/\gamma$ and a transmission rate β, its basic reproductive ratio R_0 is determined by β/γ.

In a closed population, an infectious disease with a specified R_0 can invade only if there is a threshold fraction of susceptibles greater than $1/R_0$.

Vaccination can be used to reduce the proportion of susceptibles below $1/R_0$ and hence eradicate the disease; full details are given in Chapter 8.

2.1.1.2. Epidemic Burnout

The above observations are informative about the initial stages, after an infectious agent has been introduced. Another important lesson to be learned from this simple SIR model concerns the long-term (or "asymptotic") state. Let us first divide equation (2.1) by equation (2.3):

$$\begin{aligned} \frac{dS}{dR} &= -\frac{\beta S}{\gamma}, \\ &= -R_0 S. \end{aligned} \tag{2.5}$$

Upon integrating with respect to R, we obtain

$$S(t) = S(0)e^{-R(t)R_0}, \tag{2.6}$$

assuming $R(0) = 0$. So, as the epidemic develops, the number of susceptibles declines and therefore, with a delay to take the infectious period into account, the number of recovereds increases. We note that S always remains above zero because e^{-RR_0} is always positive; in fact given that $R \leq 1$, S must remain above e^{-R_0}. Therefore, there will always be some susceptibles in the population who escape infection. This leads to the another important

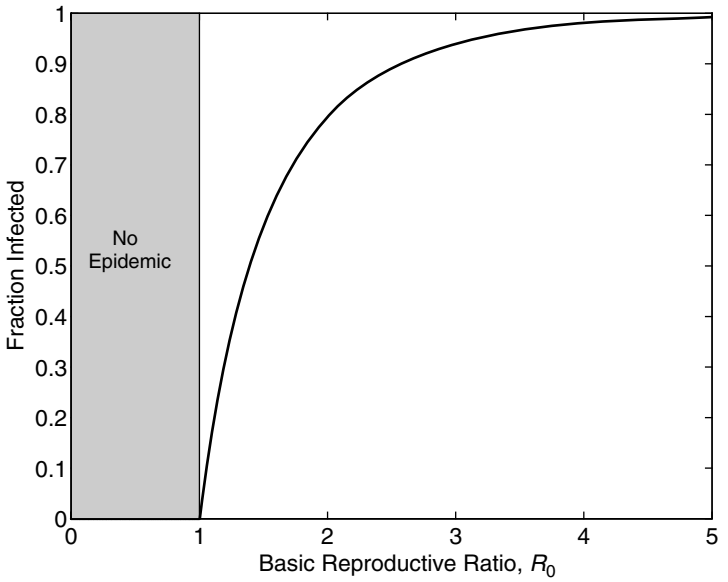


Figure 2.2. The total fraction of the population infected as a function of disease R_0. The curve is obtained by solving equation (2.7) using the Newton-Raphson method, and assuming that initially the entire population is susceptible, $S(0) = 1$, which generates the largest epidemic size.

and rather counter-intuitive conclusion that emerges from this simple model:

The chain of transmission eventually breaks due to the decline in infectives, *not* due to a complete lack of susceptibles.

This approach to model analysis can also shed some light on the fraction of the population who eventually contract an infection (Kermack and McKendrick 1927, Waltmann 1974). As shown in the steps leading to equation (2.6), it is possible to remove the variable I from the system by the division of equation (2.1) by equation (2.3), which (after integrating) gave an expression for S in terms of R. Bearing in mind that by definition $S + I + R = 1$, and that the epidemic ends when $I = 0$, we can rewrite the long-term behavior of equation (2.6):

$$S(\infty) = 1 - R(\infty) \;=\; S(0)e^{-R(\infty)R_0}$$

$$\Rightarrow\; 1 \;-\; R(\infty) \;-\; S(0)e^{-R(\infty)R_0} \;=\; 0. \tag{2.7}$$

where $R(\infty)$ is the final proportion of recovered individuals, which is equal to the total proportion of the population that gets infected.

This equation is transcendental and hence an exact solution is not possible. However, by noting that when $R(\infty) = 0$, equation (2.7) is positive, whereas if $R(\infty) = 1$ then the equation is negative, we know that at some point in between the value must be zero and a solution exists. Using standard methods, such as the Newton-Raphson (Press et al. 1988), or even by trial and error, it is possible to obtain an approximate numerical solution for equation (2.7); this is shown for the standard assumption of $S(0) = 1$ in Figure 2.2. This figure reinforces the message that if $R_0 < 1$, then no epidemic occurs. It also demonstrates the principle that whenever an infectious disease has a sufficiently large basic reproductive

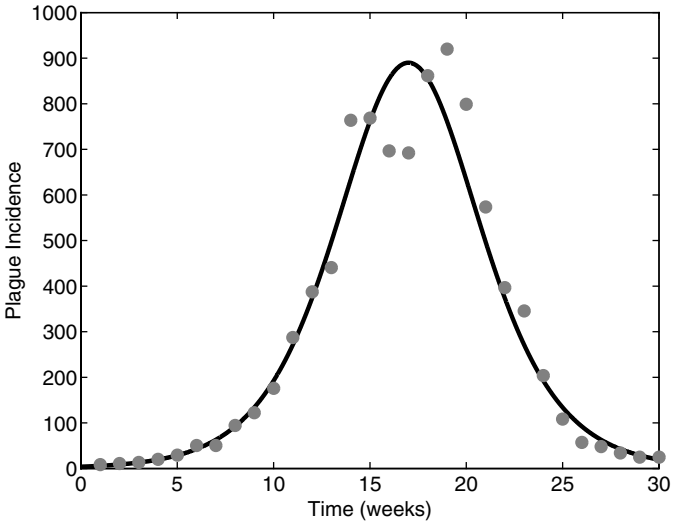


Figure 2.3. The epidemic curve. The filled circles represent weekly deaths from plague in Bombay from December 17, 1905 to July 21, 1906. The solid line is Kermack and McKendrick's approximate solution given by $dR/dt = 890\,\mathrm{sech}^2(0.2t - 3.4)$.

ratio (R_0 is larger than approximately 5), essentially everyone (>99%) in a well-mixed population is likely to contract it.

Note that the expression derived in equation (2.7) is not specifically dependent on the structure of the SIR model. It can be alternatively derived from a probabilistic argument as follows: If a single individual has been infected, then assuming the rest of the population are susceptible, they will on average infect R_0 others. Therefore, the probability of a randomly selected individual escaping infection and remaining susceptible is $\exp(-R_0/N)$. Now, if Z individuals have been infected, then the probability of an individual escaping infection is $\exp(-ZR_0/N)$. If at the end of the epidemic a proportion $R(\infty) = Z/N$ have been infected, then the probability of remaining susceptible is clearly $S(\infty) = \exp(-R(\infty)R_0)$, which again must be equal to $1 - R(\infty)$. Therefore, we once again find that: $1 - R(\infty) = \exp(-R(\infty)R_0)$, which is independent of the exact structure of the model.

As we mentioned above, an exact solution of the SIR model (equations (2.1) to (2.3)) is not feasible due to the nonlinear transmission term, βSI. It is possible, however, to obtain an approximate solution for the "epidemic curve," which is defined as the number of new cases per time interval (Waltmann 1974; Hethcote 2000). A classic example of the epidemic curve is provided in Figure 2.3 which shows the number of deaths per week from the plague in Bombay during 1905–1906. Assuming new cases are identified once an individual exhibits the characteristic symptoms of the infection, we can get a handle on the epidemic curve by exploring the equation involving dR/dt (details provided in Box 2.2). This gives the following inelegant but useful approximation:

$$\frac{dR}{dt} = \frac{\gamma\alpha^2}{2S(0)R_0^2}\,\mathrm{sech}^2\left(\frac{1}{2}\alpha\gamma t - \phi\right). \tag{2.8}$$

The quantities α and ϕ depend in a complex way on the parameters and initial conditions, and are defined in Box 2.2. For any specific epidemic, the parameters in equation (2.8) can

Box 2.2 The Epidemic Curve

To obtain an expression for the epidemic curve, we start by considering equations (2.1)–(2.3). As shown in the steps leading to equation (2.6), it is possible to remove the variable I from the system by the division of equation (2.1) by equation (2.3), which (after integrating) gives an expression for S in terms of R. Bearing in mind that by definition $S+I+R=1$, we can rewrite equation (2.3):

$$\frac{dR}{dt} = \gamma(1-S-R).$$

After substituting for S from equation (2.6), this gives

$$\frac{dR}{dt} = \gamma(1-S(0)e^{-R_0 R}-R). \tag{2.9}$$

As it stands, this equation is not solvable. If, however, we assume that $R_0 R$ is small, we can Taylor expand the exponential term to obtain:

$$\frac{dR}{dt} = \gamma\left(1-S(0)+\left(S(0)R_0-1\right)R-\frac{S(0)R_0^2}{2}R^2\right). \tag{2.10}$$

It is messy but possible to solve this equation. Omitting the intermediate steps, we get

$$R(t) = \frac{1}{R_0^2 S(0)}\left(S(0)R_0-1+\alpha\tanh\left(\frac{1}{2}\alpha\gamma t-\phi\right)\right), \tag{2.11}$$

where

$$\alpha = \left[(S(0)R_0-1)^2+2S(0)I(0)R_0^2\right]^{\frac{1}{2}},$$

and

$$\phi = \tanh^{-1}\left[\frac{1}{\alpha}\left(S(0)R_0-1\right)\right].$$

To obtain the epidemic curve as a function of time, we need to differentiate equation (2.11) with respect to time, giving

$$\text{reported cases per unit time} \sim \frac{dR}{dt} = \frac{\gamma\alpha^2}{2S(0)R_0^2}\,\text{sech}^2\left(\frac{1}{2}\alpha\gamma t-\phi\right). \tag{2.12}$$

As usual, it is important to scrutinize the assumptions made while deriving this result. The key step was in going from equation (2.9) to equation (2.10) and it involved the assumption that $R_0 R$ is small. This condition is most likely to be met at the start of the epidemic (when $R \ll 1$) or if the infection has a very small R_0. Hence, the approximation will, in general, probably not be very accurate for highly infectious diseases such as measles, whooping cough, or rubella with estimated R_0 values of 10 or higher (see Table 2.1.). In addition, the ease with which these equations can be numerically integrated largely negates the need for such involved approximations on a regular basis.

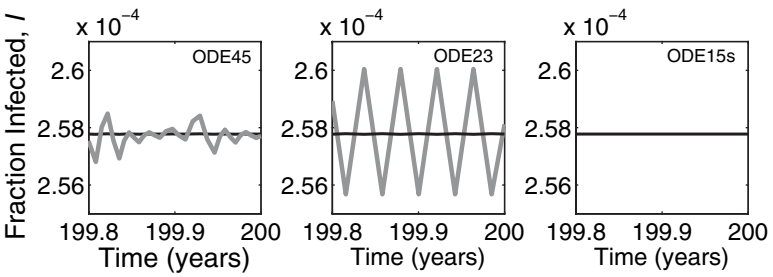
be estimated from the data, as shown in Figure 2.3. Note that the derivation of this equation requires $R_0 R$ to be small, which is unlikely to be the case for many infectious diseases, especially near the end of the epidemic when R typically approaches 1 (see Figure 2.2), but is a good approximation during the early stages often up to the peak of the epidemic. For most practical purposes, however, we need to numerically solve the SIR equations to calculate how variables such as X and Y change through time. The basic issues involved in such an endeavor are discussed in Box 2.3 and the associated figure.

Box 2.3 Numerical Integration

Given the nonlinearity in epidemiological models, it is typically not possible to derive an exact equation that predicts the evolution of model variables through time. In such cases, it is necessary to resort to numerical integration methods. By far the simplest algorithm is called Euler's method, due to the eighteenth century Swiss mathematician. It involves the translation of differential equations into discrete-time analogues. Simply, if we are interested in integrating the SIR equations, we consider the change in each variable during a very small time interval, δt. This is approximately given by the rate of change of that variable at time t multiplied by δt. For example, the equation describing the dynamics of the fraction of infectives is given by

$$\begin{aligned} I(t+\delta t) &\approx I(t)+\delta t \times \frac{dI}{dt} \\ &\approx I(t)+\delta t \times (\beta S(t) I(t) - (\mu+\gamma) I(t)). \end{aligned}$$

Using this scheme, we can simulate the dynamics of a system of ordinary differential equations (ODEs) through time. The problem with the method, however, is that it is rather crude and possesses low accuracy (the error in predicted trajectories scales with δt rather than a higher power). In extreme cases (with δt too large), the method has been known to generate spurious dynamics such as cyclic trajectories when it is possible to demonstrate analytically that all solutions converge to a stable equilibrium (see figure below).



This figure demonstrates $SEIR$ (see Section 2.5) dynamics predicted by different numerical integration schemes. The gray line depicts model predictions using untransformed variables, and the black line represents log-transformed variables (see below). The three different integration schemes demonstrated are the Runge-Kutta Dormand-Prince ("ODE45" in Matlab; left), Runge-Kutta Bogacki and Shampine pairs ("ODE23"; center), and Gear's method ("ODE15s"; right) (see Shampine and Reichelt 1997 for more details of these methods). As shown by the fluctuations in the left and center, numerical integration schemes can generate spurious dynamics, highlighting the importance of a thorough examination of model dynamics using alternative methods, as well as potentially substantial gains from log-transforming variables. (Parameter values are $\beta = 1250$ per year, $\mu = 0.02$ per year, $1/\sigma = 8$ days, and $1/\gamma = 5$ days.)

Such numerical issues can be easily overcome using more sophisticated integration methods, where the deviation from the true solution scales with higher powers of δt (see Press et al. (1988) for a review of different methods, as well as algorithms for their implementation).

However, no numerical scheme will ever be exact due to computational rounding error and therefore it is optimal to reformulate the equations so that such errors are minimized. Dietz (1976) suggested the use of log-transform variables ($\hat{x} = \log(S)$ and $\hat{y} = \log(I)$). The

transformed SIR equations become

$$e^{\hat{x}}\frac{d\hat{x}}{dt} = \mu - \beta e^{\hat{y}+\hat{x}} - \mu e^{\hat{x}}$$
$$e^{\hat{y}}\frac{d\hat{y}}{dt} = \beta e^{\hat{x}+\hat{y}} - (\mu+\gamma)e^{\hat{y}},$$

which after simplification become

$$\frac{d\hat{x}}{dt} = \mu e^{-\hat{x}} - \beta e^{\hat{y}} - \mu$$
$$\frac{d\hat{y}}{dt} = \beta e^{\hat{x}} - \mu - \gamma.$$

These new equations, although looking inherently more complex, are far less prone to numerical error. In the above figure, we show that although some integration schemes applied to the standard SIR equations can generate spurious fluctuating dynamics (gray lines), the use of log-transformation of variables (black lines) can overcome this problem—as can the use of more sophisticated integration methods.

2.1.1.3. Worked Example: Influenza in a Boarding School

An interesting example of an epidemic with no host demography comes from an outbreak of influenza in a British boarding school in early 1978 (Anon 1978; Murray 1989). Soon after the start of the Easter term, three boys were reported to the school infirmary with the typical symptoms of influenza. Over the next few days, a very large fraction of the 763 boys in the school had contracted the infection (represented by circles in Figure 2.4). Within two weeks, the infection had become extinguished, as predicted by the simple SIR model without host demography.

We can get an understanding into the epidemiology of this particular strain of influenza A virus (identified by laboratory tests to be A/USSR/90/77 (H1N1)) by estimating the parameters for the SIR model from these data. Using a simple least squares procedure (minimizing the difference between predicted and observed cases), we find the best fit parameters yield an estimated active infectious period ($1/\gamma$) of 2.2 days and a mean transmission rate (β) of 1.66 per day. Therefore, the estimated R_0 of this virus during this epidemic is $\beta/\gamma = 1.66 \times 2.2$, which is 3.65. As shown in Figure 2.4, model dynamics with these parameters is in good agreement with the data. Note, however, that as pointed out by Wearing et al. (2005), the precise value of R_0 estimated from these data is substantially affected by the assumed model structure (in Section 3.3 we deal with this issue in more detail).

2.1.2. The SIR Model With Demography

In the last section, we presented the basic framework for the SIR model given the assumption that the time scale of disease spread is sufficiently fast so as not to be affected by population births and deaths. Some important epidemiological lessons were learned from this model, but ultimately, the formalism ensured the eventual extinction of the pathogen. If we are interested in exploring the longer-term persistence and endemic dynamics of an infectious disease, then clearly demographic processes will be important.

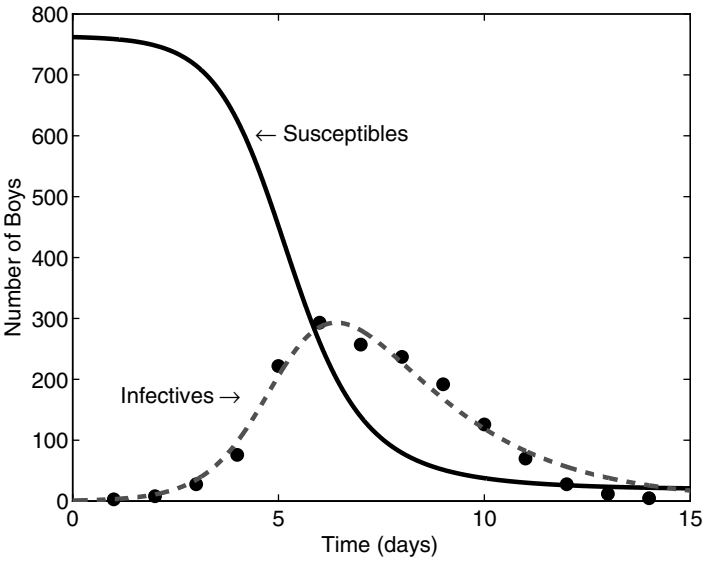


Figure 2.4. The SIR dynamics. The filled circles represent the number of boys with influenza in an English boarding school in 1978 (data from the March 4th edition of the *British Medical Journal*). The curves represent solutions from the SIR model fitted to the data using least squares. Estimated parameters are $\beta = 1.66$ per day and $1/\gamma = 2.2$ days, giving an R_0 of 3.65.

In particular, the most important ingredient necessary for endemicity in a population is the influx of new susceptibles through births.

The simplest and most common way of introducing demography into the SIR model is to assume there is a natural host "lifespan", $1/\mu$ years. Then, the *rate* at which individuals (in any epidemiological class) suffer natural mortality is given by μ. It is important to emphasize that this factor is independent of the disease and is not intended to reflect the pathogenicity of the infectious agent. Some authors have made the alternative assumption that mortality acts only on the recovered class (see Bailey 1975; Keeling et al. 2001a; Brauer 2002), which makes manipulation easier but is generally less popular among epidemiologists. Historically, it has been assumed that μ also represents the population's crude birth rate, thus ensuring that total population size does not change through time ($\frac{dS}{dt} + \frac{dI}{dt} + \frac{dR}{dt} = 0$). This framework is very much geared toward the study of human infections in developed nations—our approach would be different if the host population exhibited its own "interesting" dynamics (as is often the case with wildlife populations; see Chapter 5). Putting all these assumptions together, we get the generalized SIR model:

$$\frac{dS}{dt} = \mu - \beta SI - \mu S, \tag{2.13}$$

$$\frac{dI}{dt} = \beta SI - \gamma I - \mu I, \tag{2.14}$$

$$\frac{dR}{dt} = \gamma I - \mu R. \tag{2.15}$$

This is online program 2.2

Note that for many diseases, such as measles, newborns may have passive immunity derived via the placental transfer of maternal antibodies (if the mother had experienced infection or had been vaccinated). Given that the average age at which this immunity to measles and other childhood infections is lost (approximately 6 months) is considerably smaller than the typical age at infection (4–5 years in developed nations; Anderson & May 1982), the assumption that all newborns enter the susceptible class is not unreasonable. In cases when the mean age at infection is very small—in developing nations, for example—maternally derived immunity needs to be explicitly incorporated into models (see McLean and Anderson 1988a,b; and Hethcote 2000).

Before proceeding further, it is useful to establish the expression for R_0 in this model. Starting with the basic definition of R_0 (number of secondary infectives per index case in a naive population of susceptibles), we can look closely at equation (2.14) to work it out. The parameter β represents the transmission *rate* per infective, and the negative terms in the equation tells us that each infectious individual spends an average $\frac{1}{\gamma+\mu}$ time units in this class—the infectious period is effectively reduced due to some individuals dying while infectious. Therefore, if we assume the entire population is susceptible ($S = 1$), then the average number of new infections per infectious individual is determined by the transmission rate multiplied by the infectious period:

$$R_0 = \frac{\beta}{\gamma + \mu}. \tag{2.16}$$

This value is generally similar to, but always smaller than, R_0 for a closed population because the natural mortality rate reduces the average time an individual is infectious. Having established the expression for the R_0, we can now explore some of the properties of the system. This model has proved very useful, primarily for (1) establishing disease prevalence at equilibrium, (2) determining the conditions necessary for endemic equilibrium stability, (3) identifying the underlying oscillatory dynamics, and (4) predicting the threshold level of vaccination necessary for eradication. We will successively explore these features and discuss some of the relevant mathematical underpinnings.

2.1.2.1. The Equilibrium State

The inclusion of host demographic dynamics may permit a disease to persist in a population in the long term. One of the most useful ways of thinking about what may happen *eventually* is to explore when the system is at equilibrium, with $\frac{dS}{dt} = \frac{dI}{dt} = \frac{dR}{dt} = 0$. We therefore set each equation in the system (equations (2.13)–(2.15)) to zero and work out the values of the variables (now denoted by S^*, I^*, and R^*) that satisfy this condition.

Without needing to do much work, the *disease-free equilibrium* is self-evident. This is the scenario where the pathogen has suffered extinction and, in the long run, everyone in the population is susceptible. Mathematically, this state is expressed as $(S^*, I^*, R^*) = (1, 0, 0)$. Below, we discuss the likelihood of observing this state in the system.

Establishing the *endemic equilibrium* requires slightly more work. Perhaps counterintuitively, we start by setting the equation for the infectives (equation (2.14)) to zero:

$$\beta SI - (\gamma + \mu)I = 0. \tag{2.17}$$

After factoring for I, we have

$$I(\beta S - (\gamma + \mu)) = 0, \tag{2.18}$$

which is satisfied whenever $I^* = 0$ or $S^* = \frac{(\gamma+\mu)}{\beta}$. The first condition is simply the disease-free equilibrium, so we concentrate on the second. The quantity on the right-hand side of the equality should look familiar: It is the inverse of the R_0. This leads to an important result:

In the SIR model with births and deaths, the endemic equilibrium is characterized by the fraction of susceptibles in the population being the inverse of R_0.

Having established that $S^* = \frac{1}{R_0}$, we substitute this into equation (2.13) and solve for I^*. Missing out a few lines of algebra, we eventually arrive at

$$I^* = \frac{\mu}{\gamma}\left(1 - \frac{1}{R_0}\right) = \frac{\mu}{\beta}(R_0 - 1). \tag{2.19}$$

One universal condition on population variables is that they cannot be negative. Hence the endemic equilibrium is biologically feasible only if $R_0 > 1$, which agrees with our earlier ideas about when an epidemic is possible. Now, utilizing $S^* + I^* + R^* = 1$, we can obtain an expression for R^*. The endemic equilibrium is, therefore, given by:

$$(S^*, I^*, R^*) = \left(\frac{1}{R_0}, \frac{\mu}{\beta}(R_0 - 1), 1 - \frac{1}{R_0} - \frac{\mu}{\beta}(R_0 - 1)\right).$$

2.1.2.2. Stability Properties

So far, we have derived expressions for the disease-free and endemic equilibrium points of the SIR system, and the restrictions on parameter values for these equilibria to be biologically meaningful. We now would like to know how likely we are to observe them. In mathematical terms, this calls for a "stability analysis" of each equilibrium point, which would provide conditions on the parameter values necessary for the equilibrium to be (asymptotically) stable to small perturbations. The basic idea behind stability analysis is explained in Box 2.4. When this technique is applied to our two equilibrium states, we find that for the endemic equilibrium to be stable, R_0 must be greater than one, otherwise the disease-free equilibrium is stable. This makes good sense because the pathogen cannot invade if each infected host passes on the infection to fewer than one other host (i.e. $R_0 < 1$). However, if successively larger numbers are infected ($R_0 > 1$), then the "topping up" of the susceptible pool by reproduction ensures disease persistence in the long term.

In the SIR model with births and deaths, the endemic equilibrium is stable if $R_0 > 1$, otherwise the disease-free equilibrium state is stable.

2.1.2.3. Oscillatory Dynamics

An important issue for any dynamical system concerns the *manner* in which a stable equilibrium is eventually approached. Do trajectories undergo oscillations as they approach the equilibrium state or do they tend to reach the steady state smoothly? The SIR system is an excellent example of a "damped oscillator," which means the inherent dynamics contain a strong oscillatory component, but the amplitude of these fluctuations declines over time as the system equilibrates (Figure 2.5).

Figure 2.6, shows how the period of oscillations (as determined by equation (2.23) in Box 2.4) changes with the transmission rate (β) and the infectious period ($1/\gamma$). We note

that (relative to the infectious period) the period of oscillations becomes longer as the reproductive ratio approaches one; this is also associated with a slower convergence toward the equilibrium.

Box 2.4 Equilibrium Analysis

Here, we outline the basic ideas and methods of determining the stability of equilibrium points. Although we are specifically concerned with the SIR model here, the description of the concepts will be quite general and could be applied to a range of models discussed in this book. Assume we have n variables of interest, N_i $(i = 1, 2, \ldots n)$. The dynamics of the system are governed by n coupled Ordinary Differential Equations (ODEs; in the SIR equations, n is three):

$$\frac{dN_i}{dt} = f_i(N_1, N_2, \ldots, N_n), \qquad i = 1, 2, \ldots, n \tag{2.20}$$

To explore the equilibrium dynamics, we must first determine the system's equilibrium state (or states). This is done by setting equations (2.20) to zero and solving for the solutions $N_1^*, N_2^*, \ldots, N_n^*$—noting that multiple solutions may exist. We know that if the system is at equilibrium, then it will remain at equilibrium (by definition). But what happens when the system is (inevitably) perturbed from this state? In mathematical jargon, we are interested in determining the consequences of small perturbations to the equilibrium state. This is achieved by looking at the rates of change of these variables when each variable is slightly shifted away from its equilibrium value. This is done by making the substitutions $N_i = N_i^* + \epsilon_i$ in equations (2.20) and exploring the growth or decline of the perturbation terms, ϵ_i, over time. In any specific case, we can carry out each of these steps, but there is a more generic methodology.

Mathematical results dating back some 200 years have established that for a series of equations such as those described by equations (2.20), the stability of an equilibrium point is determined by quantities known as *eigenvalues*, here represented by Λ_i. For a system of n ODEs, there will be n eigenvalues and stability is ensured if the real part of all eigenvalues are less than zero—these eigenvalues are often complex numbers. Having established the usefulness of eigenvalues, we need to explain how to calculate these terms. Before we can do that, a matrix, $\boldsymbol{J}$, known as the *Jacobian* must be introduced. It is given by:

$$\boldsymbol{J} = \begin{pmatrix} \frac{\partial f_1^*}{\partial N_1} & \frac{\partial f_1^*}{\partial N_2} & \cdots & \frac{\partial f_1^*}{\partial N_n} \\ \frac{\partial f_2^*}{\partial N_1} & \frac{\partial f_2^*}{\partial N_2} & \cdots & \frac{\partial f_2^*}{\partial N_n} \\ \vdots & & \ddots & \vdots \\ \frac{\partial f_n^*}{\partial N_1} & \frac{\partial f_n^*}{\partial N_2} & \cdots & \frac{\partial f_n^*}{\partial N_n} \end{pmatrix}.$$

The terms f_i^* refer to the functions $f_i(N_1, N_2, \ldots, N_n)$ calculated at equilibrium, i.e. $f_i(N_1^*, N_2^*, \ldots, N_n^*)$. The eigenvalues Λ_i $(i = 1, 2, \ldots, n)$ are the solutions of $\det(\boldsymbol{J} - \Lambda \boldsymbol{I}) = 0$; where $\boldsymbol{I}$ is the identity matrix. That is, we subtract Λ from each diagonal element of the Jacobian and then work out the determinant of the matrix. This will give rise to a polynomial in Λ of order n. This is called the *characteristic polynomial*, which, when set to zero and solved, gives rise to the eigenvalues $(\Lambda_1, \Lambda_2, \ldots, \Lambda_n)$.

Let us demonstrate these ideas by applying them to the SIR system of equations. Finding the two equilibrium states is described in the main text. So next, we work out the Jacobian:

$$\boldsymbol{J} = \begin{pmatrix} -\beta I^* - \mu & -\beta S^* & 0 \\ \beta I^* & \beta S^* - (\mu + \gamma) & 0 \\ 0 & \gamma & -\mu \end{pmatrix}.$$

To obtain the characteristic polynomial, we subtract Λ from the diagonal elements and calculate the determinant. This gives:

$$(\beta I^* - \mu - \Lambda)(\beta S^* - (\mu + \gamma) - \Lambda)(-\mu - \Lambda) + (\beta I^*)(\beta S^*)(-\mu - \Lambda) = 0.$$

Notice that $(-\mu - \Lambda)$ can be factored immediately, giving one eigenvalue ($\Lambda_1 = -\mu$) that is negative. The remaining two solutions $\Lambda_{2,3}$ are found by solving the following quadratic equation:

$$(\beta I^* - \mu - \Lambda)(\beta S^* - (\mu + \gamma) - \Lambda) + \beta I^* \beta S^* = 0. \tag{2.21}$$

Let us first consider the disease-free equilibrium. If we make the appropriate substitutions ($S^* = 1$ and $I^* = 0$), we have

$$(-\mu - \Lambda)(\beta - (\mu + \gamma) - \Lambda) = 0.$$

This clearly has two solutions, $\Lambda_2 = -\mu$ and $\Lambda_3 = \beta - (\mu + \gamma)$. For this equilibrium to be stable, we need to ensure all eigenvalues are negative, hence the stability criterion becomes $\beta < \mu + \gamma$, which translates into ensuring $R_0 < 1$.

To explore the endemic equilibrium, again we substitute the expressions for S^* and I^* into equation (2.21) and explore the condition required for the remaining two eigenvalues to be negative. After making some simplifications, we arrive at the following quadratic equation,

$$\Lambda^2 + \mu R_0 \Lambda + (\mu + \gamma)\mu(R_0 - 1) = 0, \tag{2.22}$$

the solutions of which can be obtained by the standard formula, giving:

$$\Lambda_{2,3} = -\frac{\mu R_0}{2} \pm \frac{\sqrt{(\mu R_0)^2 - \frac{4}{AG}}}{2},$$

where the term $A = \frac{1}{\mu(R_0 - 1)}$ denotes the mean age at infection (see Section 2.1.2.4) and $G = \frac{1}{\mu + \gamma}$ determines the typical period of a host's infectivity.

To make further progress with this equation, we notice that often $(\mu R_0)^2$ is small enough to ignore and hence we can approximate the above solutions to

$$\Lambda_{2,3} \approx -\frac{\mu R_0}{2} \pm \frac{i}{\sqrt{AG}},$$

where $i = \sqrt{-1}$. Therefore, the endemic equilibrium is feasible only when R_0 is greater than one, but it is always stable. The fact that the largest ("dominant") eigenvalues are complex conjugates (they are of the form $\Lambda_{2,3} = x \pm iy$) tells us that the equilibrium is approached via oscillatory dynamics. The period of these damped oscillations, T, is determined by the inverse of the complex part of the eigenvalues multiplied by 2π:

$$T \sim 2\pi\sqrt{AG}. \tag{2.23}$$

2.1.2.4. Mean Age at Infection

When dealing with infectious diseases in the real world, an important indicator of prevalence is the host's mean age at infection, A (if it possible to contract the infection

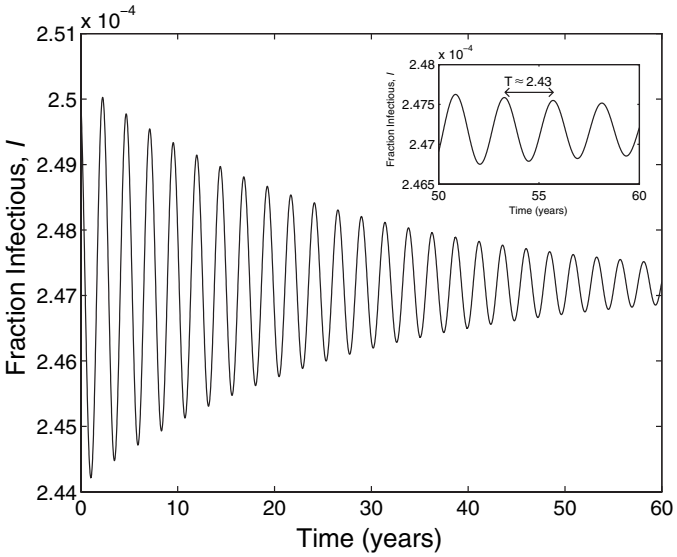


Figure 2.5. The SIR model's damped oscillations. The main figure shows how the fraction of infectives oscillates with decreasing amplitude as it settles toward the equilibrium. The inset shows a slice of the time-series with the period of fluctuations as predicted by equation (2.23). The figure is plotted assuming $1/\mu = 70$ years, $\beta = 520$ per year, and $1/\gamma = 7$ days, giving $R_0 = 10$. Initial conditions were $S(0) = 0.1$ and $I(0) = 2.5 \times 10^{-4}$.

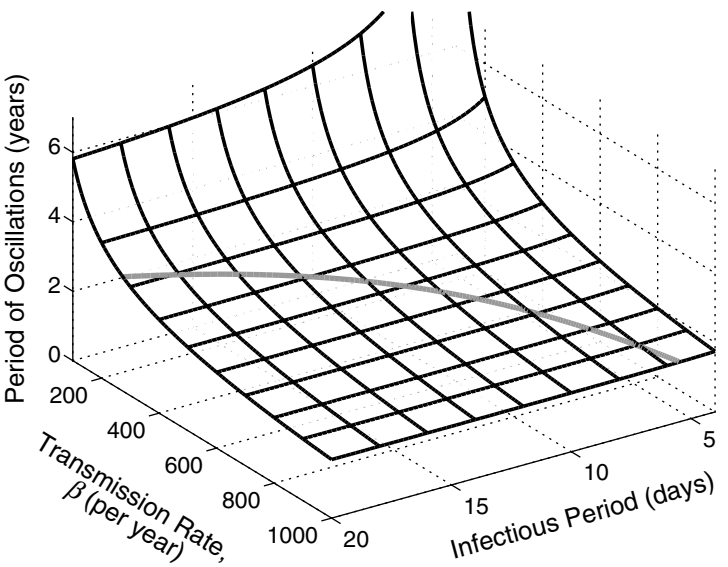


Figure 2.6. The period of damped oscillations in the SIR model as the transmission rate (β) and infectious period ($1/\gamma$) are varied. The gray line on the surface depicts the oscillatory period whenever the parameters combine to yield an R_0 of 15. The figure is plotted assuming $1/\mu = 70$ years.

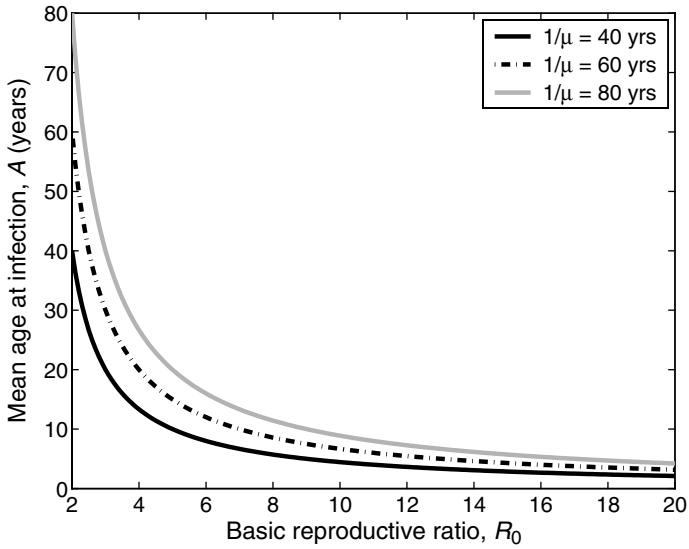


Figure 2.7. The mean age at infection for the SIR model as a function of the basic reproductive ratio (R_0) and the average life expectancy ($1/\mu$). Using equation (2.24), we see that as R_0 increases, there is a dramatic decline in the mean age at infection, whereas changes in μ are not as influential.

multiple times, then we would be interested in the mean age at first infection). This quantity can be measured in a straightforward way for many human or animal infections by analyzing age-specific serological data, which detect the presence of antibodies specific to pathogen antigens. The *rate* at which individuals seroconvert (from negative to positive) provides information on the force of infection (see Box 2.1). Age-stratified serological surveys, when conducted randomly through the population, can yield a relatively unbiased population level estimate of the mean age at infection. Typically, we may also have accurate independent estimates for some of the model parameters such as the host life expectancy (from demographic data) and the infectious period (from clinical epidemiology data). Thus, once we have an estimate of the mean age at infection from serological surveys, we can estimate the transmission rate β as long as we have an expression for A derived from the model (see Figure 2.7).

How do we calculate the average age at which susceptibles are infected, especially given a nonage-structured model? We can approach this question by taking equation (2.13) at equilibrium and calculating the mean time an individual remains susceptible—that is, the mean time from birth (or loss of maternally derived immunity) to infection. Ignoring the small, disease-independent mortality term, the average period spent in the susceptible class is approximated by the inverse of the force of infection, namely $\frac{1}{\beta I^*}$. Upon substituting for I^* from equation (2.19), we obtain an expression for the mean age at infection:

$$A \approx \frac{1}{\mu(R_0 - 1)}. \tag{2.24}$$

This equation can be rephrased as $R_0 - 1 \approx \frac{L}{A}$, where L is the host's life expectancy. The above step has proven historically very important in establishing a robust link between model parameters and population level quantities such as L and A. In their classic work,

Anderson and May (1982; 1991) used the estimates of A and L to calculate R_0 for numerous infections in different geographical regions and eras (see Table 2.1).

The mean age of (first) infection is equal to the average life expectancy of an individual divided by $R_0 - 1$.

2.2. INFECTION-INDUCED MORTALITY AND SI MODELS

The models described in the previous section have implicitly assumed that the infection is essentially benign. Transmission results in a period of illness, which is followed by recovery and lifelong immunity. This scenario is reasonable for largely harmless infections such as the common cold or chickenpox. However, numerous infectious diseases are associated with a substantial mortality risk. Examples include malaria, measles, whooping cough, SARS, and dengue fever, among others. How do we explore the consequences of infection-induced mortality? Specifically, how do we incorporate a mortality *probability* into the SIR equations? The obvious approach would be to add a term such as $-mI$ to equation (2.14), where m is a per capita disease-induced mortality *rate* for infected individuals. However, this may be tricky to interpret biologically or estimate from data. Instead, it is preferable to think about the probability, ρ, of an individual in the I class dying from the infection before either recovering or dying from natural causes. This is the quantity most likely estimated from clinical studies or case observations. Mathematically, this translates into the following equation:

$$\frac{dI}{dt} = \beta SI - (\gamma + \mu)I - \frac{\rho}{1-\rho}(\gamma + \mu)I, \tag{2.25}$$

where ρ represents the per capita *probability* of dying from the infection and takes values from zero to one. (The S equation remains as before.) Therefore, in order to convert this to a mortality rate, we should set $m = \frac{\rho}{1-\rho}(\gamma + \mu)$. This equation for the infection dynamics can be tidied up to give

$$\frac{dI}{dt} = \beta SI - \frac{(\gamma + \mu)}{1-\rho} I. \tag{2.26}$$

Note that as ρ approaches unity, new infectives die almost instantaneously and R_0 drops to zero; for such diseases it may be more appropriate to further subdivide the infectious period and allow mortality only in the later stages of infection (see Section 3.3).

2.2.1. Mortality Throughout Infection

Because infection actively removes individuals from the population, we can no longer implicitly assume that the population size is fixed—disease-induced mortality could lead to an ever-declining population size. One way around this is to incorporate a *fixed* birth rate (ν) into the susceptible equation (2.13), independent of the population size:

$$\frac{dS}{dt} = \nu - \beta SI - \mu S. \tag{2.27}$$

However, the fact that the population size N can vary means that we need to consider the transmission term βSI in much more detail. Until now, it has made little difference

whether density or frequency-dependent transmission was assumed, because the N term was constant and could be absorbed by rescaling the population size or rescaling the transmission rate β. However, because N is now variable, this is no longer possible and the choice of transmission mechanism can profoundly affect the dynamics. Given that this problem illustrates an important (and often confusing) issue in epidemiological modeling, we will deal with the two cases separately, paying particular attention to the underlying assumptions. To make matters explicit, we will deal with numbers rather than proportions throughout this section.

2.2.1.1. Density-Dependent Transmission

For density-dependent (pseudo mass action) transmission, we specifically consider the case where as the total population size N decreases, due to disease-induced mortality, there is reduced interaction between hosts. In this density-dependent formulation, the rate at which new cases are produced is βXY, which scales linearly with both the number of susceptibles and the number of infectious individuals.We start by considering the values of ν and μ and the dynamics of the disease-free state. In the absence of disease, we find that

This is online program 2.3

$$\frac{dN}{dt} = \nu - \mu N \qquad \Rightarrow \qquad N \to \frac{\nu}{\mu} \tag{2.28}$$

Hence, we can equate $\frac{\nu}{\mu}$ with the ecological concept of a carrying capacity for the population.

We can now repeat the kinds of analyses that were carried out on the generalized SIR model in Section 2.1.1.2. Once again, we find the system possesses two equilibrium points: one endemic (X^*, Y^*, Z^*) and one disease free ($\frac{\nu}{\mu}$, 0, 0). Missing out a few lines of algebra, we find the endemic equilibrium to be:

$$X^* = \frac{\mu+\gamma}{\beta(1-\rho)} = \frac{\nu}{\mu R_0}, \tag{2.29}$$

$$Y^* = \frac{\mu}{\beta}(R_0 - 1), \tag{2.30}$$

$$Z^* = \frac{\gamma}{\beta}(R_0 - 1). \tag{2.31}$$

$$\Rightarrow N^* = \frac{\nu}{\mu R_0}[1 + (1-\rho)(R_0 - 1)]. \tag{2.32}$$

In this case, R_0 ($= \frac{\beta(1-\rho)\nu}{(\mu+\gamma)\mu}$) contains a correction term $(1-\rho)$ that takes into account the reduced period of infectivity due to disease-induced mortality, as well as a term that takes into account the population size at the disease-free equilibrium. The condition necessary to ensure the feasibility of the endemic equilibrium (and hence the instability of the disease-free steady state) is found by ensuring $Y^* > 0$, which translates to $R_0 > 1$, as we have previously seen. This means that if $\rho > 0$, then the pathogen has to have a higher transmission rate per unit of infectious period ($\frac{\beta}{\mu+\gamma}$) in order to remain endemic, compared to a similar infectious disease that is benign—due to the fact that the effective infectious period is reduced by disease-induced mortality.

The stability properties of this system are similar to the basic SIR model. Indeed, stability analysis reveals the endemic equilibrium to be always locally asymptotically

stable. The approach to the equilibrium is also via damped oscillations, the (natural) period of which are determined by the following equation:

$$T \sim 2\pi\sqrt{AG}. \tag{2.33}$$

This is identical to equation (2.23), though note that here the terms A (the average age of infection) and G (the average generation length of infection), respectively, both contain a correction term due to taking infection-induced mortality into account.

When disease-induced mortality is added to the SIR model with density-dependent transmission, the equilibrium and stability properties simply reflect a change in parameters.

2.2.1.2. Frequency-Dependent Transmission

For frequency-dependent (mass action) transmission, the calculation is somewhat more involved. Recall that if X and Y are the number of susceptible and infectious hosts, then the mass-action assumption means that the transmission rate is given by $\beta XY/N$. Previously, we rescaled the variables such that S and I represented fractions of the population, which removed the N term in the transmission rate, leaving us with more elegant-looking equations. However, in this scenario N is varying and that trick is no longer appropriate. Instead, we retain the notion that X and Y are numbers of hosts. Our equation for the number of infectious individuals is therefore:

$$\frac{dY}{dt} = \beta XY/N - \frac{(\gamma+\mu)}{1-\rho}Y. \tag{2.34}$$

This is online program 2.4

We note that the mass-action assumption means that even when the population size is reduced, each individual still has the same average number of contacts.

It will not be a surprise that two equilibria exist: the endemic (X^*, Y^*, Z^*) and the disease-free ($\frac{\nu}{\mu}$, 0, 0). The endemic equilibrium can again be found by setting the rates of change equal to zero:

$$\begin{aligned} X^* &= \frac{\nu(1-\rho)(\gamma+\mu)}{\mu(\beta(1-\rho)-\mu\rho-\gamma\rho)} = \frac{N}{R_0}, \quad \Rightarrow \quad S^* = \frac{(\gamma+\mu)}{\beta(1-\rho)} = \frac{1}{R_0}, \\ Y^* &= \frac{\nu\beta(1-\rho)^2-\nu(\mu+\gamma)(1-\rho)}{(\mu+\gamma)(\beta(1-\rho)-\mu\rho-\gamma\rho)}, \quad \Rightarrow \quad I^* = \frac{\mu}{\beta(1-\rho)}(R_0-1), \end{aligned} \tag{2.35}$$

$$N^* = \frac{\beta\nu(1-\rho)^2}{\mu(\beta(1-\rho)-\mu\rho-\gamma\rho)} = \frac{\nu}{\mu}\left(\frac{R_0(1-\rho)}{(R_0-\rho)}\right).$$

As the expressions on the right-hand side demonstrate, equilibrium values are easier to understand intuitively if we deal with the proportion of the population in each state, rather than the absolute number. Again, we note that $R_0 > 1$ is necessary and sufficient for the disease to invade the population and for the equilibrium point to be feasible and stable.

For both the frequency- and density-dependent models, the population size can be reduced from its carrying capacity of $\frac{\nu}{\mu}$ by the impact of the disease. Not surprisingly, infectious diseases with the highest mortalities (ρ close to one) and largest R_0 have the greatest impact on the population (Figure 2.8). Although for low mortality levels both mixing assumptions lead to similar results, when the mortality is high the

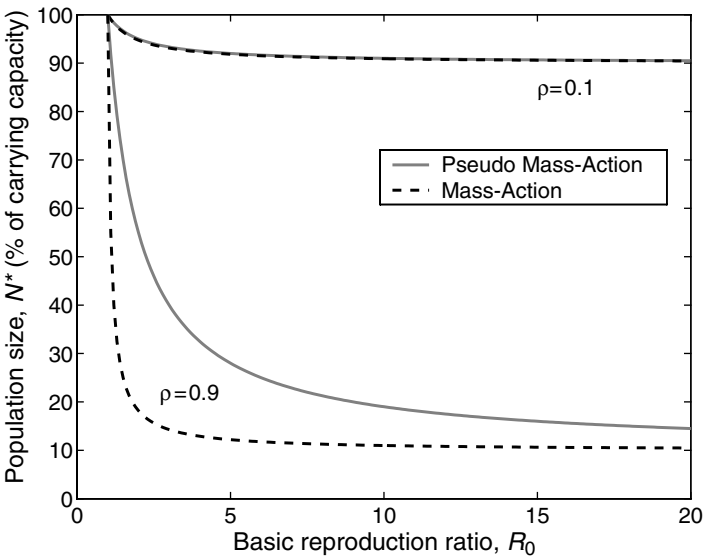


Figure 2.8. The population size, N^*, for diseases that are associated with mortality. Two different disease mortality probabilities are considered, $\rho = 0.1$ and $\rho = 0.9$, and the two mixing assumptions of density dependence (βXY, solid line) and frequency dependence ($\beta XY/N$, dashed line) are shown.

frequency-dependent (mass-action) assumption leads to the largest drop in the total population size. This is because pseudo mass-action mixing places a natural damping on transmission, such that as the population size decreases so does the contact rate between individuals, limiting disease spread and reducing disease-induced mortality. From this relatively simple example it is clear that when population sizes change, our assumption about mixing behavior can have profound effects on the dynamics.

When disease-induced mortality is added to the SIR model with frequency-dependent transmission, the equilibrium and stability properties can change substantially, especially if the probability of mortality is high.

2.2.2. Mortality Late in Infection

One difficulty with equation (2.26) is that when the mortality rate is very high, the infectious period is substantially reduced. In some cases, we may wish to consider a disease where mortality generally occurs at (or toward) the end of the infectious period. This is often plausible because the onset of disease (and symptoms) may be significantly delayed from the onset of infection. In such cases, the following model would be appropriate:

$$\frac{dS}{dt} = \nu - \beta SI - \mu S, \tag{2.36}$$

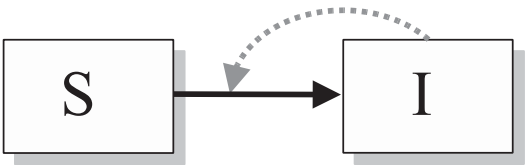
$$\frac{dI}{dt} = \beta SI - (\gamma + \mu)I, \tag{2.37}$$

$$\frac{dR}{dt} = (1-\rho)\gamma I - \mu R. \tag{2.38}$$

For this form of model mortality, we again need to consider the precise form of transmission. For frequency-dependent transmission, the reduction in the recovered population has no effect on the dynamics and hence this model has the same properties as the standard SIR model. However, for density-dependent transmission, we find that $X^* = \frac{\nu(\mu+\gamma(1-\rho))}{(\beta-\gamma\rho)\mu}$ and $Y^* = \frac{\nu(\beta-\gamma-\mu)}{(\beta-\gamma\rho)(\gamma+\mu)}$, although again this is stable and feasible if $R_0 = \frac{\beta\nu}{\mu(\gamma+\mu)} > 1$.

The equilibrium levels of diseases that cause mortality are critically dependent upon whether frequency- or density-dependent transmission is assumed, due to the changes in the total population size that occur. However, we generally find that the endemic equilibrium is feasible and stable as long as $R_0 > 1$.

2.2.3. Fatal Infections



The models of Sections 2.2.1 and 2.2.2 represent cases where infection does not always kill. There are, however, numerous examples of animal and plant pathogens that are always fatal (Feline Infectious Peritonitis (FIP), Spongiform Encephalopathy (BSE), Leishmaniasis, Rabbit Haemorrhagic Disease, and Highly Pathogenic Avian Influenza (H5N1)). In this situation, we can simplify the SIR equations by removing the recovered class, which leads to a family of models known as SI models. Here, infecteds are assumed to remain infectious for an average period of time $(1/\gamma)$, after which they succumb to the infection. Assuming frequency-dependent transmission, the equations describing the SI model are simply:

$$\frac{dX}{dt} = \nu - \beta XY/N - \mu X, \tag{2.39}$$

$$\frac{dY}{dt} = \beta XY/N - (\gamma+\mu)Y. \tag{2.40}$$

It is straightforward to demonstrate that the endemic equilibrium ($X^* = \frac{\nu}{\beta-\gamma}$, $Y^* = \frac{\nu(\beta-\gamma-\mu)}{(\beta-\gamma)(\gamma+\mu)}$) is feasible as long as $R_0 = \frac{\beta}{(\mu+\gamma)} > 1$ and is always locally stable.

Alternatively, if we assume pseudo mass-action transmission, such that the contact rate scales with density, we obtain:

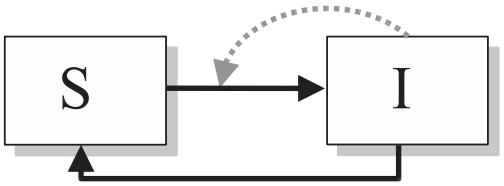
$$\frac{dX}{dt} = \nu - \beta XY - \mu X, \tag{2.41}$$

$$\frac{dY}{dt} = \beta XY - (\gamma + \mu)Y. \tag{2.42}$$

Similarly, for this system the endemic equilibrium ($X^* = \frac{\gamma+\mu}{\beta}$, $Y^* = \frac{\nu}{(\gamma+\mu)} - \frac{\mu}{\beta}$) is feasible as long as $R_0 = \frac{\beta\nu}{(\mu+\gamma)\mu} > 1$ and is again always locally stable.

2.3. WITHOUT IMMUNITY: THE SIS MODEL

The SI and SIR models both capture the dynamics of acute infections that either kill or confer lifelong immunity once recovered. However, numerous infectious diseases confer no long-lasting immunity, such as rotaviruses, sexually transmitted infections, and many bacterial infections. For these diseases, a individuals can be infected multiple times throughout their lives, with no apparent immunity. Here, we concentrate briefly on this class of models, called SIS because recovery from infection is followed by an instant return to the susceptible pool.



Simply, these SIS models are described by a pair of coupled ordinary differential equations:

$$\frac{dS}{dt} = \gamma I - \beta IS, \tag{2.43}$$

$$\frac{dI}{dt} = \beta SI - \gamma I. \tag{2.44}$$

This is online program 2.5

The parameters are as defined in the previous section, but with $S + I = 1$. In this simple example, demography (births and deaths) has been ignored. Despite this lack of susceptible births, the disease can still persist because the recovery of infectious individuals replenishes the susceptible pool. We can, therefore, substitute $S = 1 - I$ into equation (2.44) and simplify to get

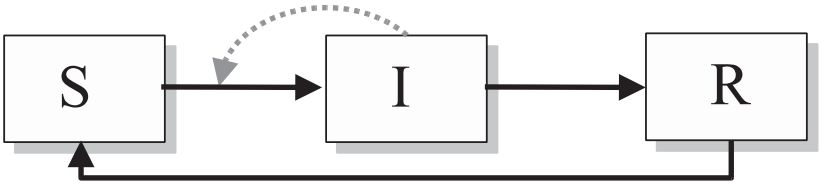
$$\frac{dI}{dt} = (\beta - \beta I - \gamma)I = \beta I((1 - 1/R_0) - I), \tag{2.45}$$

where, as usual, $R_0 = \beta/\gamma$. This equation is the equivalent of the logistic equation used to describe density-dependent population growth in ecology (Murray 1989). The equilibrium number of infectives in this population is obtained by setting equation (2.45) to zero and solving for I^*. This gives $I^* = (1 - 1/R_0)$, hence yet again $S^* = 1/R_0$, and the equilibrium will be stable as long as $R_0 > 1$. However, for this class of model, convergence to the equilibrium is monotonic with no oscillatory behavior. The dynamics of SIS models and sexually transmitted diseases is described in more detail in Chapter 3.

For an infectious disease that does not give rise to long-term immunity, as long as the infection is able to invade the population, loss of immunity will guarantee its long-term persistence.

2.4. WANING IMMUNITY: THE SIR MODEL

The SIR and SIS models are two behavioral extremes where immunity is either lifelong or simply does not occur. An intermediate assumption is that immunity lasts for a limited period before waning such that the individual is once again susceptible.



This translates into the following model:

$$\frac{dS}{dt} = \mu + wR - \beta SI - \mu S \tag{2.46}$$

$$\frac{dI}{dt} = \beta SI - \gamma I - \mu I \tag{2.47}$$

$$\frac{dR}{dt} = \gamma I - wR - \mu R, \tag{2.48}$$

where w is the rate at which immunity is lost and recovered individuals move into the susceptible class. Not surprisingly, the dynamics of this model provide a smooth transition between the SIR framework (when $w = 0$) and the SIS model (when $w \to \infty$). As before, $R_0 = \frac{\beta}{\gamma+\mu}$ and we require $R_0 > 1$ to obtain a plausible stable endemic solution. Once again, it is possible to obtain the oscillatory damping toward endemic equilibrium that was seen in the SIR model; the period of these oscillations is given by:

$$T = \frac{4\pi}{\sqrt{4(R_0 - 1)\frac{1}{G_I}\frac{1}{G_R} - \left(\frac{1}{G_R} - \frac{1}{A}\right)^2}},$$

where $A = \frac{w+\mu+\gamma}{(w+\mu)(\beta-\gamma-\mu)}$ is again the average age at first infection, whereas $G_I = 1/(\gamma + \mu)$ is the average period spent in the infectious class, and $G_R = 1/(w + \mu)$ is the average time spent in the recovered class.

Figure 2.9 shows the dramatic effects of changing the level of waning immunity: As the period of immunity ($1/w$) is reduced (and the dynamics more closely resemble those of the SIS model), we observe a dramatic increase in the prevalence of infectious disease together with a marked drop in the period of the damped oscillations. This waning immunity result was used by Grassly et al. (2005) to explain the 9–10-year cycle observed in syphilis cases in the United States. It was postulated that these cycles were determined by the natural oscillations of syphilis infection if waning immunity was assumed. Grassly et al. (2005) found that oscillations consistent with those observed could be generated by a model with $R_0 = 1.5$, $1/w = 10$ years, $1/\gamma = 2$ months, and $1/\mu = 33$ years—this high birth rate accounts for individuals entering and leaving the pool of sexually active individuals. We note that although the deterministic model predicts that eventually such oscillations

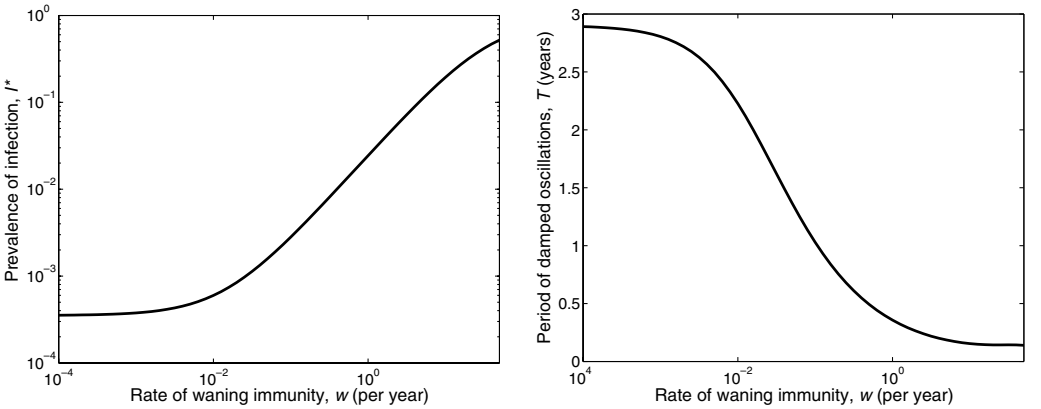
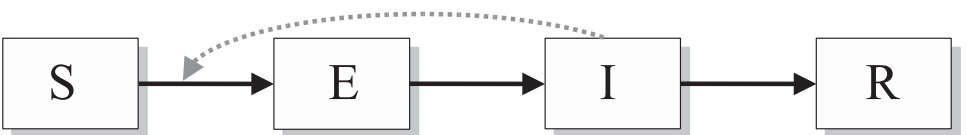


Figure 2.9. The effects of waning immunity on the equilibrium infection prevalence (left graph) and the period of the damped oscillations (right graph). ($R_0 = 10$, $1/\mu = 70$ years, $1/\gamma = 10$ days.)

will decay to zero, the chance nature of transmission can lead to sustained oscillations generally close to the natural period, T (see Chapter 6).

2.5. ADDING A LATENT PERIOD: THE $SEIR$ MODEL

We briefly introduce a refinement to the SIR model to take into account the latent period. The process of transmission often occurs due to an initial inoculation with a very small number of pathogen units (e.g., a few bacterial cells or virions). A period of time then ensues during which the pathogen reproduces rapidly within the host, relatively unchallenged by the immune system. During this stage, pathogen abundance is too low for active transmission to other susceptible hosts, and yet the pathogen is present. Hence, the host cannot be categorized as susceptible, infectious, or recovered; we need to introduce a new category for these individuals who are *infected* but not yet *infectious*. These individuals are referred to as Exposed and are represented by the variable E in $SEIR$ models.



Assuming the average duration of the latent period is given by $1/\sigma$, the $SEIR$ equations are:

$$\frac{dS}{dt} = \mu - (\beta I + \mu)S, \tag{2.49}$$

$$\frac{dE}{dt} = \beta SI - (\mu + \sigma)E, \tag{2.50}$$

$$\frac{dI}{dt} = \sigma E - (\mu + \gamma)I, \tag{2.51}$$

$$\frac{dR}{dt} = \gamma I - \mu R. \tag{2.52}$$

This is online program 2.6

As before, we typically assume $S + E + I + R = 1$ and hence equation (2.52) is redundant. The addition of a latent period is essentially akin to introducing a slight time delay into the system and we may expect that such a feature may act to destabilize and slow the system. As we will demonstrate, the dynamic properties of the $SEIR$ model are qualitatively similar to those of the SIR system.

Standard equilibrium stability analysis proceeds by finding the steady states of the system and determining the criteria for their stability. As with previous disease models, the $SEIR$ model also possesses both an endemic (S^*, E^*, I^*, R^*) and a disease-free (1, 0, 0, 0) equilibrium solution. As usual, the endemic fixed point is of greater interest and is given by

$$S^* = \frac{(\mu+\gamma)(\mu+\sigma)}{\beta\sigma} = \frac{1}{R_0}, \tag{2.53}$$

$$E^* = \frac{\mu(\mu+\gamma)}{\beta\sigma}(R_0 - 1). \tag{2.54}$$

$$I^* = \frac{\mu}{\beta}(R_0 - 1), \tag{2.55}$$

with $R^* = 1 - S^* + E^* + I^*$. The expression for R_0 is now slightly different, due to the death of some individuals in the exposed class who do not contribute to the chain of transmission. However, this difference is often negligible because typically $\sigma/(\mu+\sigma) \sim 1$ as the latent is far smaller that the expected lifespan. As expected, if the latent period is infinitesimally small (i.e., $\sigma \to \infty$), then we recover the same expression for R_0 as for the SIR model ($R_0 = \beta/(\gamma+\mu)$).

For the endemic equilibrium to be feasible and stable (and the disease-free equilibrium to be unstable), equation (2.55) once again requires that the basic reproductive ratio exceed one ($R_0 > 1$). By exploring the Jacobian of the system (equations (2.49)–(2.51)) in the usual way, we obtain a quartic in the eigenvalues Λ. As for the SIR model (Box 2.4), $\Lambda = -\mu$ is an obvious solution, leaving us with a cubic equation:

$$\Lambda^3 + (\mu R_0 + 2\mu + \sigma + \gamma)\Lambda^2 + \mu R_0(2\mu + \sigma + \gamma)\Lambda + \mu(R_0 - 1)(\mu+\sigma)(\mu+\gamma) = 0.$$

Unfortunately, there is no obvious solution to this equation. As pointed out by Anderson and May (1991), however, in many cases σ and γ will be much larger than μ and μR_0. If so, then one solution to this equation is approximately $\Lambda \sim -(\sigma+\gamma)$, which leaves us with a quadratic for the two remaining eigenvalues:

$$\Lambda^2 + \mu R_0 \Lambda + \frac{\gamma\sigma}{\sigma+\gamma}\mu(R_0 - 1) \approx 0.$$

Clearly, this is similar to the analogous equation for the SIR model (equation (2.22)), with a slight correction term due to the death of exposed individuals. By working out the expression for these eigenvalues, the endemic equilibrium is stable given $R_0 > 1$ with perturbations dying out in an oscillatory manner. Again, we can find an expression for the natural period of oscillations in this case, which is given by $T \sim 2\pi\sqrt{AG}$, where G (the "ecological generation length" of the infection) is slightly modified to include the latent period: $G = \frac{1}{\mu+\gamma} + \frac{1}{\mu+\sigma}$.

From this analysis, one may be tempted to conclude that the $SEIR$ model is an unnecessary complication of the SIR model. Given a small death rate ($\mu \ll \gamma, \sigma$), the SIR

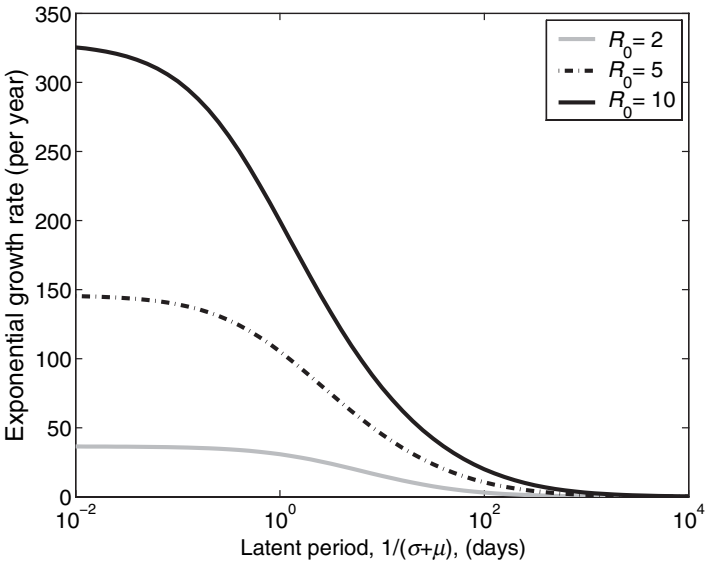


Figure 2.10. The effects of an exposed period on the initial growth rate of an infection in a totally susceptible population. Although a small exposed period leads to dynamics close to those predicted by the SIR model, a long exposed period can dramatically slow the spread of infection, or even prevent the spread if too many individuals die before becoming infectious. ($1/\mu = 70$ years, $1/\gamma = 10$ days.)

and $SEIR$ models behave similarly at equilibrium as long as the basic reproductive ratio ($\beta_{SIR}/\gamma_{SIR} = \beta_{SEIR}/\gamma_{SEIR}$) and average infected period ($1/\gamma_{SIR} = 1/\gamma_{SEIR} + 1/\sigma_{SEIR}$) are identical. However, the two models behave very differently at invasion, with the presence of an exposed class slowing the dynamics. Examining the eigenvalues at the disease-free equilibrium allows us to describe the increase in prevalence during the invasion phase:

$$I_{SEIR}(t) \approx I(0)\exp\left(\frac{1}{2}\left[\sqrt{4(R_0-1)\sigma\gamma+(\sigma+\gamma)^2}-(\sigma+\gamma)\right]t\right),$$

$$\left\{\approx I(0)\exp\left([(\sqrt{R_0}-1)\gamma]t\right) \qquad \text{if } \sigma=\gamma\right\},$$

whereas without an exposed class the dynamics are given by:

$$I_{SIR}(t) \approx I(0)\exp([(R_0-1)\gamma]t),$$

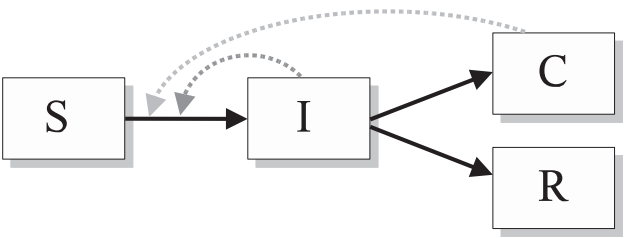
where natural mortality has been ignored to simplify the equations. This behavior is exemplified in Figure 2.10, showing how long exposed periods can slow or even prevent the spread of infection. Therefore, if large fluctuations in the prevalence of infection are of interest, or we wish to consider both invasion and equilibrium properties, the exposed class must be realistically modeled.

Although the SIR and $SEIR$ models behave similarly at equilibrium (when the parameters are suitably rescaled), the $SEIR$ model has a slower growth rate after pathogen invasion due to individuals needing to pass through the exposed class before they can contribute to the transmission process.

2.6. INFECTIONS WITH A CARRIER STATE

Although the SIR and $SEIR$ model paradigms are a good approximation to the epidemiological characteristics of many infectious diseases, such as measles or influenza, other infections have a more complex natural history. As an example of how such complexities can be accommodated in the model, will we consider infections such as hepatitis B or herpes, where a proportion of infected individuals may become chronic carriers, transmitting infection at a low rate for many years. The greater biological complexity of these systems can be readily incorporated into our current modeling framework, although accurate parameterization becomes more complex. Here we focus on the inclusion of a single carrier class, using hepatitis B as our prototypic infectious disease.

This general framework of building compartmental models can be readily extended to infections with more complex biology.



For diseases with carrier states, susceptible individuals can be infected by either carriers or acutely infectious individuals. It is generally assumed that the progress of infection within an individual is independent of their source of infection; that is, those infected by acutely infectious individuals and those infected by carriers are indistinguishable. A recently infected individual is acutely (highly) infectious for a given period and then either recovers completely or moves into the carrier class. Such dynamics lead to the following model:

$$\frac{dS}{dt} = \mu - (\beta I + \varepsilon\beta C)S - \mu S, \tag{2.56}$$

$$\frac{dI}{dt} = (\beta I + \varepsilon\beta C)S - \gamma I - \mu I, \tag{2.57}$$

$$\frac{dC}{dt} = \gamma q I - \Gamma C - \mu C, \tag{2.58}$$

$$\frac{dR}{dt} = \gamma(1-q)I + \Gamma C - \mu R. \tag{2.59}$$

This is online program 2.7

Here, C captures the proportion of carriers in the population, ε is the reduced transmission rate from chronic carriers compared to acute infectious individuals, q is the proportion of acute infections that become carriers while a fraction $(1-q)$ simply recover, and Γ is the rate at which individuals leave the carrier class—all other parameters have their standard meanings.

As with any new epidemiological model, it is important to understand when an epidemic can occur and hence to calculate the basic reproductive ratio, R_0. For infections with

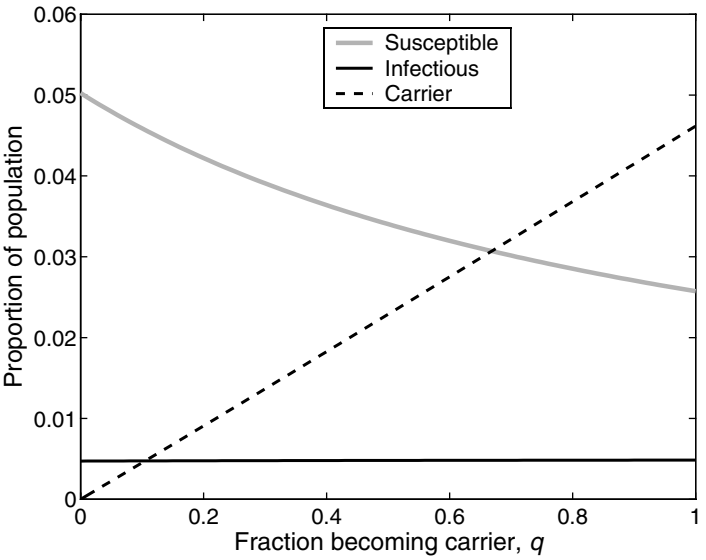


Figure 2.11. The equilibrium values of a model with a carrier class. ($\beta = 73$ per year, $1/\gamma = 100$ days, $1/\Gamma = 1000$ days, $\varepsilon = 0.1$, $1/\mu = 50$ years.) Using these parameters R_0 is relatively large (from 19.9 to 38.8), hence the level of susceptibles is always low and therefore the variation in the level of acute infections is negligible.

a carrier state, R_0 has two components: One comes from acutely infectious individuals, which follow the standard calculation for R_0 that has been given throughout this chapter; the other comes from infections caused while in the carrier state and must take into account the fraction of infecteds becoming carriers:

$$R_0 = \frac{\beta}{\gamma + \mu} + \frac{q\gamma}{(\gamma + \mu)} \frac{\varepsilon\beta}{(\Gamma + \mu)},$$

where the term $\frac{q\gamma}{\gamma+\mu}$ accounts for those individuals that do not die in the infectious class and go on to be come carriers. Therefore, as expected, the fact that infected individuals can enter an infectious carrier state rather than simply recovering increases the value of R_0.

The value of R_0 is the sum of separate components from the acutely infected and chronic carriers.

With only a little more work than usual, the equilibrium values of the model can be found:

$$S^* = \frac{\gamma + \mu}{\beta + \frac{q\gamma\varepsilon\beta}{\Gamma+\mu}} = \frac{1}{R_0},$$

$$I^* = \frac{\mu(1 - S^*)}{\gamma + \mu}, \qquad C^* = \frac{\gamma q \mu(1 - S^*)}{(\gamma + \mu)(\Gamma + \mu)},$$

Figure 2.11 shows how these equilibrium values change as q, the fraction of infected individuals that become carriers, varies. Due to their much longer "infectious" period, carriers can easily outnumber acutely infected individuals.

In practice, the epidemiology of diseases such as hepatitis B virus (HBV) is more complex, because the risk of becoming a carrier is age-dependent, with infected children more likely to become carriers. This can potentially lead to some surprising behavior with two stable endemic situations for a given set of parameters due to positive feedbacks between R_0 and the number of carriers (Medley et al. 2001). When R_0 is high, the average age of infection is low, which in turn leads to many carriers and hence a higher R_0; in contrast, a low R_0 means that the average age of infection is high and so few carriers are produced and R_0 remains low. In this manner, and using an age-structured formalism (Chapter 3), it is possible to observe "endemic stability" where both high and low R_0 solutions are stable and the equilibrium prevalence of infection depends on the initial starting conditions.

2.7. DISCRETE-TIME MODELS

This chapter has concentrated exclusively on epidemiological models written as differential equations. This is partly because the vast majority of models in the literature are based on this framework. The inherent assumption has been that the processes of disease transmission occur in real time and that variability in factors such as the infectious period may be dynamically important. Some have argued that if the latent and infectious periods are relatively constant, it is reasonable to construct models phrased in discrete time. Such models were first developed by Reed and Frost in 1928 (Abbey 1952) as probabilistic entities, assuming transmission probabilities are binomial (giving rise to "chain binomial" models; see Bailey 1975 and Daley and Gani 1999 for more details). Here, we introduce these models within a deterministic setting and refer interested readers to Chapter 6 for a discussion of analogous stochastic models.

An important issue that arises in modeling epidemics in discrete time is precisely what a time increment represents. Ideally, units of time should represent the "generation" length of the infection through a host, though in some cases this can lead to some difficulty especially if latent and infectious periods differ markedly. To demonstrate this approach, consider a disease with latent and infectious periods of exactly a week. We are therefore interested in determining the future changes in the fraction of individuals in the population who are susceptible (S_t), exposed (E_t), or infectious (I_t) in week t. The following *difference* equations can be used to represent such a scenario:

$$S_{t+1} = \mu - S_t e^{-\beta I_t}, \tag{2.60}$$

$$E_{t+1} = S_t(1 - e^{-\beta I_t}), \tag{2.61}$$

$$I_{t+1} = E_t, \tag{2.62}$$

where μ now represents the *weekly* per capita births. The exponential term in equation (2.60) represents the per capita probability of *not* contracting the infection given I_t infectives with transmission β (Box 2.1). This formalism inherently assumes that transmission probability per susceptible follows a Poisson distribution, with mean βI_t. Hence, the probability of escaping infection is given by the zero term of this distribution. The transmission parameter (β) is now analogous to the maximum reproductive potential—or R_0—of the infection. Consider the situation where everyone in the population is susceptible and a single infectious individual is introduced. Then, the initial spread of the

disease occurs at rate β, very much analogous to R_0 in the continuous-time models. Thus, as before, for the infection to invade, we require $\beta > 1$.

The introduction of discrete time does little to change the analytical approaches used to explore model dynamics. For example, we can easily perform equilibrium analysis on these equations (2.60)–(2.62) by solving for $S^* = S_{t+1} = S_t$, $E^* = E_{t+1} = E_t$, and so forth. Such an exercise yields the following equilibrium solutions:

$$S^* = \frac{\mu}{1 - e^{-\beta\mu}}, \tag{2.63}$$

$$E^* = \mu, \tag{2.64}$$

$$I^* = \mu. \tag{2.65}$$

To establish the criterion for endemicity, we insist that at equilibrium the proportion of susceptibles is less than one ($S^* < 1$). From equation (2.63), this translates into requiring that

$$\beta > \frac{-\log(1-\mu)}{\mu}.$$

This relationship is nearly always satisfied because (1) $\log|1-\mu| \sim \mu$ (for $\mu \ll 1$) and (2) for the infection to successfully invade, we require $\beta > 1$.

To establish the stability of this equilibrium, we follow a similar procedure to that outlined in Box 2.4. The Jacobian for this system is simply a matrix whose elements represent the partial derivatives of equations (2.60)–(2.62) with respect to S_t, E_t, and I_t, evaluated at equilibrium (for more details, see Murray 1989):

$$\boldsymbol{J} = \begin{pmatrix} e^{-\beta I^*} & 0 & -\beta S^* e^{-\beta I^*} \\ 1 - e^{-\beta I^*} & 0 & \beta S^* e^{-\beta I^*} \\ 0 & 1 & 0 \end{pmatrix}.$$

To obtain a polynomial for the eigenvalues, we subtract Λ from the diagonal elements and work out the determinant for the Jacobian (i.e. set $\det|\boldsymbol{J} - \Lambda\boldsymbol{I}| = 0$). As for the SIR equations, because we have three variables, we obtain a cubic equation in the eigenvalues. In this instance, because of the discrete time nature of the model, stability of the equilibrium solution requires the dominant eigenvalue to have magnitude less than one (May 1973; Nisbet and Gurney 1982; Murray 1989). For the system described by equations (2.60)–(2.62), this reduces to requiring all roots of the following equation to lie in the unit circle:

$$\Lambda^3 - e^{-\beta I^*}\Lambda^2 - \beta S^* e^{-\beta I^*}\Lambda + \beta S^* e^{-\beta I^*} = 0.$$

Although an analytical solution to this equation is not possible, we can use any number of commands in mathematical software packages to find its roots. In Figure 2.12, we have numerically calculated the dominant eigenvalues for a range of parameter combinations and show that as long as the endemic equilibrium is feasible, it is also stable. Note, however, that the absolute values of Λ are very close to unity, suggesting a weakly stable equilibrium: perturbations eventually decay to zero, but this may happen over a long time.

Recently, some authors (Finkenstädt and Grenfell 2000; Bjørnstad et al. 2002; Xia et al. 2004) have argued that when infection prevalence is relatively small, we may conveniently

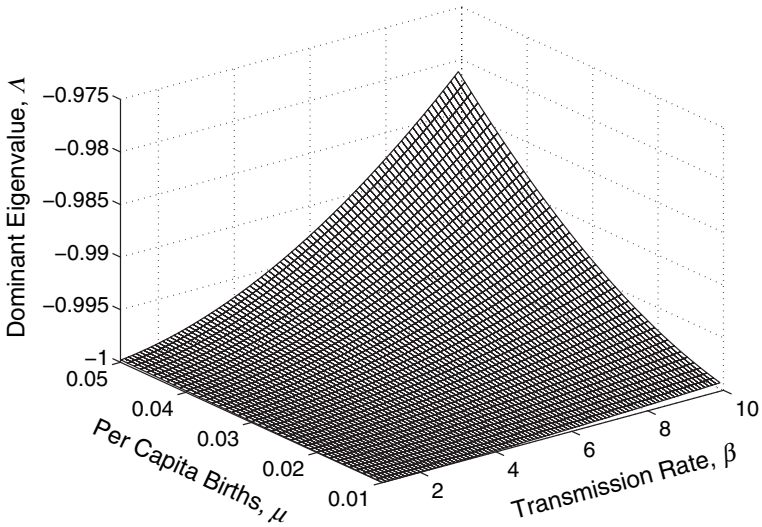


Figure 2.12. The stability of the endemic equilibrium for the SIR model phrased in discrete-time. The magnitude of the dominant eigenvalue is always less than one, but not massively so, suggesting very weakly stable dynamics; the dominant eigenvalue is also real and negative, indicating a 2-week oscillatory cycle in the approach to equilibrium.

rewrite equation (2.60) using a bilinear term to represent the transmission term:

$$S_{t+1} = S_t + \mu - \beta S_t I_t.$$

This is essentially the same as expanding the exponential term in equation (2.60) and ignoring the $\beta^2 I_t^2$ and higher-order terms. (For an ecological example, contrast the host-parasitoid models of Hassell (1978) and Neubert et al. (1995)). If we proceed with an equilibrium analysis, as above, we find that the dominant eigenvalue is *always* equal to unity. In this system, the endemic equilibrium is neutrally stable, with small perturbations to the equilibrium neither decaying nor growing. This is an undesirable and unrealistic property of the model and results largely from the above simplification. To overcome this undesirable structural instability of the model, Finkenstädt and Grenfell (2000) incorporated an exponent into the S_t and I_t terms:

$$S_{t+1} = S_t + \mu - \beta S_t^{\alpha} I_t^{\phi}.$$

When α and ϕ are less than one, the endemic equilibrium is stabilized.

The major benefit of formulating a model in discrete time is that it is much easier to parameterize using discrete time data than the associate differential equation counterparts (Bailey 1975; Mollison and Ud-Din 1993; Finkenstädt and Grenfell 2000). Their major drawback, however, is their demonstrated mathematical fragility (Glass et al. 2003).

2.8. PARAMETERIZATION

Although models in isolation can provide us with a deep understanding of the transmission and control of general infectious diseases, if we wish to examine the epidemiology of

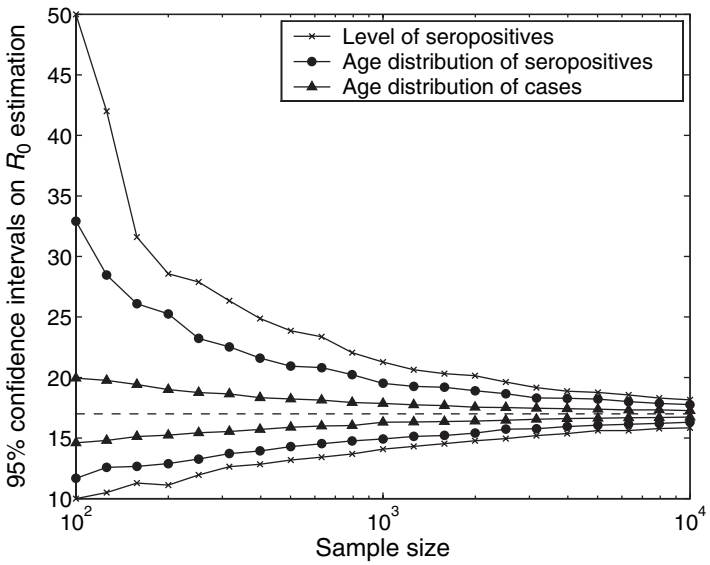


Figure 2.13. The effect of sample size on the estimated value of R_0. The true value of R_0 is 17, and the lines show the 95% confidence intervals associated with various estimation methods. For the serological estimates, the sample size refers to the number of sampled individuals (irrespective of disease status); for the case-based estimates (triangles), the sample size refers to the number of recorded cases. We have assumed no age bias in the reporting of cases.

a specific pathogen, we need to parameterize our models accordingly. Although generic models provide an intuitive explanation, it is only though detailed parameterization that useful public health guidance can be generated. From a modeling perspective, only the parameterization differentiates models of smallpox and measles, or models of chlamydia and HIV.

The standard SIR model has four parameters: the birth rate, the natural death rate, the average infectious period (or, conversely, the rate of recovery), and the transmission rate. In an ideal world, robust statistical inference would provide information on all parameters. In reality, this approach is hampered both by limited data, as well as underdeveloped appropriate inferential methdologies, though a substantial body of ongoing work redresses this shortcoming (see King and Lele, in prep, Ionedes et al. 2006, Vasco and Rohani, in prep). As a result, we often use other relevant information to assist with parameterization. A good understanding of the host's biology, for example, will provide accurate estimates of the birth and death rates (which we have generally assumed to be equal). The infectious period can usually be estimated independently via clinical monitoring of infecteds, either by observation of transmission events or by more detailed techniques measuring the amount of pathogen excreted. This leaves us with a single parameter—the transmission rate, β—to estimate. In practice, researchers generally focus attention on estimating the basic reproductive ratio, R_0, from which the transmission rate can be derived. There are multiple approaches to finding the value of R_0, which if appropriate will depend on the type of data that is available (Figure 2.13).

2.8.1. Estimating R_0 from Reported Cases

An initial response to estimating transmission, β, or the basic reproductive ratio, R_0, would be to record the total number of reported cases of infection within a given community or area. However, if we examine the equilibrium prevalence (equation (2.19)), we find that:

$$I^* = \frac{\mu}{\beta}(R_0 - 1) = \frac{\mu}{\gamma + \mu}\left[1 - \frac{1}{R_0}\right].$$

Therefore, if R_0 is substantially larger than one, then estimation is going to be very difficult because the effect of R_0 on prevalence is relatively small. In addition, we are unlikely to record every case within a population as many infections will go unreported. Even for measles in England and Wales (for which we possess one of the most accurately recorded long-term epidemiological data sets), the reporting rate is only 60% (Clarkson and Fine 1985; Finkenstadt and Grenfell 2000), which swamps the small effects of R_0 that we wish to detect.

An alternative approach is to examine the early behavior of an epidemic. The SIR model predicts that the early growth of an epidemic will be exponential:

$$I(t) \sim I(0)\exp([R_0 - 1](\gamma + \mu)t)$$

Fitting to this type of behavior is possible, and during the early stages of a novel infection (such as the 2001 foot-and-mouth epidemic in the United Kingdom or the 2003 SARS epidemic in China) is the only way of estimating transmission levels. However, in general, three difficulties may exist: (1) this method can work only if there is an epidemic to be observed, and cannot provide information on endemic diseases; (2) in the early stages of an epidemic, due to the low number of cases, the dynamics may be highly stochastic and influenced by large fluctuations (see Chapter 6), which will cause the estimates of R_0 to also fluctuate. By the time stochastic fluctuations become negligible, it is likely that the epidemic behavior is nonlinear and therefore the exponential approximation is no longer valid. (3) The final difficulty with this approach is that unless the pathogen is novel to the population, some individuals are likely to be immune. Although this partial susceptibility can be incorporated into the equations leading to $I(t) \sim I(0)\ \exp([S(0)R_0 - 1](\gamma + \mu)t)$, it is impossible to separate its effects from the value of R_0. One plausible solution is to fit to the entire epidemic, finding which values of R_0, the initial level of susceptible, $S(0)$, and case identification probability lead to a model epidemic that most closely matches the recorded profile of cases (see Gani and Leach 2001 for an example of this method applied to smallpox data). Additional aspects affecting the use of this approach in estimating R_0 are discussed in Chapter 3 and Wearing et al. (2005).

At this stage, it is important to distinguish between the variable $Y(t)$ in models and the kinds of longitudinal epidemiological data typically available. The plague case notification data plotted in Figure 2.3, for example, represent the total number of *new* cases diagnosed in a given week—the incidence of infection. This is distinct from $Y(t)$, which is the number of individuals infectious at time t—the prevelance of infection. If the infectious period of a disease happens to be approximately the same as the resolution of the data, then the cases and Y are comparable (Ferrari et al. 2005). Otherwise, we need to introduce a new model variable that can be used in any estimation procedure. In general, it is reasonable to assume that case reports take place once an individual leaves the infectious class. This may be because diagnosis occurs most readily once an individual is symptomatic, at which point either the infection has already been cleared or the individual remains infectious but

has reduced transmission opportunities due to quarantining and convalescence. Therefore, $K(T)$, the number of new cases reported at time point T, may be represented by

$$K(T) = \int_{T-1}^{T} \gamma I \, dt.$$

This formulation assumes no reporting error, which may be inappropriate and should be addressed before estimation to ensure bias-free parameter values (see, for example, Finkenstädt and Grenfell 2000; Clarke 2004). The implementation of reporting error is discussed in more detail in Chapter 6.

Finally, one of the most powerful approaches using case reports it to identify the average age at infection in an endemic situation. We have already seen (equation (2.24)) that the average age of infection is given by $A \approx 1/[\mu(R_0 - 1)]$, and this approximation is most reliable when R_0 is large. The average age of infection is generally estimated by simply finding the average age of all reported cases. Figure 2.13, triangles, shows how successful this estimation procedure is at finding the value of R_0 as the number of sampled cases varies. Two potential difficulties exist with this methodology: (1) the age of patients may not be recorded as a matter of course, and therefore it may be impossible to analyze historical data with this technique; and (2), the researcher generally has little control over which individuals report infection, and therefore age-related reporting biases can influence the results.

2.8.2. Estimating R_0 from Seroprevelance Data

Estimating R_0 from case report data is problematic in humans because reporting is often patchy and biased because not all infected individuals seek medical advice. For wildlife diseases, obtaining data on individual cases is even more difficult. An alternative approach is to use standard molecular techniques to detect the presence of an antigen against a particular pathogen. In this way, we can differentiate the population into those who have not experienced infection (susceptibles) and those who have (either recovered or currently infected). The primary advantage of working with such serological data is that the researcher has full control over the sampling of the population. In contrast, the advantage of working with infectious cases (as detailed above) is that a larger amount of data can be collected during normal medical practices.

The simplest way of utilizing serological information is via the relationship that $S = 1/R_0$; this is shown by the crosses in Figure 2.13. Here the concept is relatively simple: R_0 is estimated as the inverse of the proportion of our sample that are seronegative (susceptible). The complication with this approach is that we need to make sure that our sample represents of the entire population, because the level of seroprevalence (proportion of individuals recovered from infection) is expected to increase with age.

An alternative method that can take far better advantage of the information from a serological survey is to once again utilize the age-dependent nature of the likelihood of being susceptible (Anderson and May 1991). For an individual of age a, the standard SIR model predicts that the probability an individual is still susceptible is given by $P(a) \approx \exp(-a\mu(R_0 - 1))$. Therefore, if we know the ages of our serological sample, we can construct the likelihood that the data comes from a disease with a particular R_0 value, and then find the R_0 that maximizes this likelihood. If we have n individuals who are susceptible (seronegative) of ages $a_1, \ldots, a_n$, and m individuals who are seropositive

(ages $b_1, \ldots, b_m$), then the likelihood is:

$$L(R_0) = \prod_{i=1}^{n} \exp(-a_i \mu (R_0 - 1)) \prod_{i=1}^{m} \left[1 - \exp(-b_i \mu (R_0 - 1))\right].$$

This use of additional age-related information together with serology provides a much better prediction of the value of R_0 (Figure 2.13, circles). However, if we are free to choose our serological sample from any members of the population, then we should preferentially select individuals of an age close to (what we initially believe to be) the average age of infection—sampling many older individuals will provide little information because they are all likely to be seropositive. In Figure 2.13, where $R_0 = 17$, then sampling individuals just in the 3- to 7-year-old age range produces estimates of similar accuracy to those from the case-based average age of infection assuming unbiased reporting (triangles).

2.8.3. Estimating Parameters in General

The above likelihood argument provides a general framework in which models can be successfully parameterized from data. Ideally, as much information (and parameterization) should be gained from individual-level observations; often this provides a very accurate parameterization of individual-level characteristics such as average infectious and latent periods, but is generally unreliable for transmission characteristics. For the missing parameters (for example, in Section 2.6, equations (2.56)–(2.59) require a transmission rate for both the infectious and carrier class), a maximum likelihood approach is often the most suitable. For a given set of parameters, we can determine the dynamics predicted by the model and then calculate the likelihood that the observed data came from such dynamics. By finding the set of parameters that maximize this likelihood, we are selecting a model that is in closest agreement with the available data. In the above example, the dynamics were the long-term equilibrium age-distribution of susceptibility, but there is no reason why more dynamic behavior could not be utilized. Finally, this maximum likelihood approach is not confined to parameter selection and can even allow us to distinguish between models, providing us with a means of selecting the most appropriate framework.

The ensuing chapters return to this question of parameterization, where appropriate, and show how the extra levels of heterogeneity can be characterized from the type of observation and available survey data.

2.9. SUMMARY

In this chapter, we have introduced the simplest models of epidemics, whose structures have been determined by the biology of the aetiological agent and its effects on the host. We have met (1) the SI model, which represents infections that are fatal; (2) the SIS model for infections that do not illicit a long-lasting immune response; (3) the $SIRS$ when immunity is not permanent; and (4) the SIR and $SEIR$ models that describe host-pathogen systems with lifelong protection following infection.

This chapter also introduced the epidemiologically important measure of a pathogen's reproductive potential, R_0. It summarizses the number of new individuals infected by a single infectious individual in a wholly susceptible population. Irrespective of model structure, the infection will experience deterministic extinction unless $R_0 > 1$, which is intuitively appealing (and obvious).

By concentrating on these simple models, a number of important epidemiological principles can be demonstrated:

➤ For an infection to invade a population, the initial fraction of the population in the susceptible class has to exceed $\frac{1}{R_0}$.

➤ Initially, the infection grows at a rate proportional to $R_0 - 1$.

➤ In the absence of host demography, strongly immunizing infections will always go extinct eventually. After the infection has died out, some susceptibles remain in the population. Thus, the chain of transmission is broken due to too few infectives, not a lack of susceptibles.

➤ To maintain an immunizing infection, the susceptible pool must be replenished via recruitment.

➤ In the SIR model with host demography, the endemic equilibrium is feasible if $R_0 > 1$ and is always stable. In this system, trajectories converge onto the asymptotic state via damped oscillations, the period of which can be determined as a function of epidemiologically simple measures such as the mean age at infection and the effective infectious period of the disease. This is often referred to as the "natural period" of the system.

➤ For diseases that fail to elicit long-term immunity to subsequent reinfection, the SIS model is appropriate, which is identical to the standard logistic model population growth. The key feature of that model is the presence of an exponentially stable equilibrium point. This means the infection will always be feasible and stable if $R_0 > 1$ and the approach to the equilibrium is not oscillatory.

➤ If we assume that infectious and latent periods are exactly fixed, then it is possible to formulate models in discrete time. The take-home message from these models is qualitatively similar to the standard SIR models, with the endemic equilibrium feasible and stable as long as $R_0 > 1$.

➤ Not surprisingly, discrete-time models are much less stable than their continuous-time counterparts. The endemic equilibria are very weakly stable, with perturbations decaying over long periods.

This chapter has focused on model simplicity, often sweeping known epidemiological and demographic complications under the carpet. In particular, we have focused exclusively on deterministic approaches to understanding disease dynamics. Implicit in this approach is an assumption that chance events (in the transmission process or host population demography) are either unimportant or can be "averaged out," to reveal the true underlying system traits. Chapter 6, relaxes this assumption and explores the consequences of stochasticity in SIR and SIS systems.