EMD 630/MGMT 711

Modeling Infectious Diseases: Theory and Applications

The Hazard Function Approach to SIR and Related Models

Edward H. Kaplan*

September 2003

1. Defective Duration Variables

Recall from the "system of flow" notes that a random duration T can be characterized in three different ways: via the survivor function $S(t) = \Pr\{T > t\}$, the probability density function $f(t)$, or the hazard function $h(t) = f(t)/S(t)$. Our discussion of duration variables focused on *proper* random variables, that is, those where $S(0) = 1$ *and* $S(\infty) = 0$ (equivalently, $\int_0^\infty f(t)dt = 1$, or equivalently, $\int_0^\infty h(t)dt$ is infinite – convince yourself that these conditions do amount to the same thing). These conditions are satisfied when the duration variable T is finite, as is appropriate for many durations (such as a human lifespan, the duration of infectiousness, or the length of a strike for that matter!). There are a few bizarre random variables where these conditions are satisfied yet the expected value of T is infinite, but we won't encounter them.

In the stable population model, we focused on the sorts of diseases that, essentially, would infect every individual in the population within their lifetimes (this

* Yale School of Management, and Department of Epidemiology and Public Health, Yale School of Medicine. edward.kaplan@yale.edu

is the implication of our "useful approximation" that $L >> T_S$ where L is disease-free lifetime, and T_S is the time spent susceptible, equivalently time to infection). We also assumed that, essentially, everyone recovered from the infection (in terms of the useful approximation, not only is $L >> T_S$, but also $L >> T_S + T_I$ where T_I is the duration of infectiousness, so everyone gets infected before they die, but everyone recovers too – and *then* they die). Before the advent of vaccination, these assumptions worked well for rubella, measles, mumps, etc.

Now our attention shifts to modeling true *epidemics*. Here, a new infection of some sort is introduced, people become infected, and over time we either see the epidemic "run its course," or we see the infection approach a steady (endemic) state. We will focus on the first situation where epidemics crest and fall. In such a situation, it will generally *not* remain true that everyone becomes infected (nor need it remain the case that virtually everyone recovers from disease).

If it is the case that not everyone becomes infected, then we can no longer assume that the time to infection is finite. If one lived forever, one might not become infected at all in an epidemic. Think of ebola fever virus, influenza, or other fast-moving (and sometimes fatal) infections that tend to "run their course" in a population following introduction (and for a variety of reasons to be explained as our course "runs its course").

Back to duration modeling. Interpreting the duration T as the time until some event occurs, we want a model that allows for the possibility that the event will *never* occur. Now, if T is the time until an event occurs, but it is possible that the event in question will *never* occur, then it must be possible for T to be infinite. If it is possible for T to be infinite, that is, there is some *probability* that the event never occurs, then we need a model where $\lim_{t\to\infty} \Pr\{T > t\} > 0$. That is, we will remove the constraint that $S(\infty) = 0$ (though we will keep $S(0) = 1$), and instead allow for the possibility that $\lim_{t\to\infty} S(t) = 1 - p > 0$. Note that with this possibility, we immediately have $\lim_{t\to\infty} \Pr\{T \leq t\} = p < 1$. What is the interpretation of p? Remember, if T is the time until an event occurs, then if T is infinite, the event in question simply *does not* happen, while if T is finite, then the event in question eventually *does* occur. Thus, p is simply the probability that the event in question actually happens. Applied to an epidemic, recall that T_S is the time spent susceptible, which can be interpreted as the time to infection. If a person is *never* infected during the epidemic, then T_S is infinite. Thus, p represents the probability of getting infected over the duration of the epidemic.

Now, there are several implications of the generalization discussed above. First, if $S(\infty) = 1 - p$, and $S(0) = 1$, then $\int_0^\infty f(t)dt = S(0) - S(\infty) = p$.

This is sensible – if we integrate over the probability density of the duration variable, instead of arriving at the conclusion that there must be *some* finite duration (which is true for proper random variables where the area under the probability density equals 1), we see that there is a finite duration (i.e. the event occurs) with probability p. This is consistent with the previous paragraph.

Here is a second implication. Recall that the relationship between the survivor function and the hazard function can be expressed as

$$S(t) = e^{-\int_0^t h(u)du}. \tag{1.1}$$

If the event in question never occurs, we have

$$\lim_{t\to\infty} S(t) = e^{-\int_0^\infty h(u)du} = 1 - p > 0 \tag{1.2}$$

which implies that the integral of the hazard function is finite, that is

$$\int_0^\infty h(u)du = -\log(1-p) < \infty. \tag{1.3}$$

This works in reverse too – if $\int_0^\infty h(u)du < \infty$, then $S(\infty) > 0$ and there is a probability p that the event in question occurs, and complementary probability $1-p$ that the event does not occur. A random duration variable with the properties described above (a probability p that T is finite, and a probability $1-p$ that T is infinite) is said to be *improper* or *defective*. **Equations (1.1) and (1.2) will prove critical to our approach to epidemics, as we will soon see.**

As an example of a defective duration, suppose the hazard function is given by

$$h(t) = e^{-t}. \tag{1.4}$$

Since the integral of this hazard function equals

$$\int_0^t h(u)du = \int_0^t e^{-u}du = 1 - e^{-t} \tag{1.5}$$

the associated survivor function $S(t)$ is given by

$$S(t) = e^{-\int_0^t h(u)du} = e^{-\left(1-e^{-t}\right)}. \tag{1.6}$$

Now, since $e^{-t} \to 0$ as $t \to \infty$ we see that

$$S(\infty) = e^{-1} = .3678 = 1 - p \tag{1.7}$$

and thus the event associated with the duration T, whatever that event is, occurs with probability $p = 1 - .3678 = .6322$ $(= 1 - 1/e)$.

Now, obviously the mean duration is infinite for a random duration that has a positive probability of being infinite. However, it could be of interest to focus on those with finite durations. The conditional survivor function given that the duration variable T is finite can be stated (see if you can understand why) as

$$\Pr\{T > t | T < \infty\} = \frac{S(t) - S(\infty)}{1 - S(\infty)} = \frac{S(t) - (1-p)}{p}. \tag{1.8}$$

To get a better feel for this, ask yourself what happens if you substitute $t = 0$ in the formula above (remember that $S(0) = 1$). Now try substituting $t \to \infty$. From this formula, one can derive the associated conditional density and hazard functions, given that the event of interest occurs. An application could be to consider the various properties of the time to infection *for those who get infected*. For example, if the duration given by the survivor function in equation (1.6) referred to the time to infection, then the expected time to infection *for those who get infected* would be equal to

$$\int_0^\infty \frac{e^{-\left(1-e^{-t}\right)} - e^{-1}}{1 - e^{-1}} dt = \frac{e^{-1}}{1 - e^{-1}} \int_0^\infty (e^{e^{-t}} - 1) dt \approx 0.77. \tag{1.9}$$

2. Incidence of Infection as the Hazard Function for the Time to Infection T_S

We now turn our attention to "fast epidemics." By fast, I mean that the epidemic runs its course over a period of time that is short relative to the normal human lifespan L. Note that this is not the same as our "useful approximation" $T_S + T_I << L$ from the stable population model. The import of assuming that an epidemic is fast is that we can ignore demographic changes in the population. The viewpoint is thus that of a cohort: at time 0, an infection is introduced to an otherwise susceptible population of size N, and we assume that the population size remains constant over the time required for the epidemic to run its course (though several members of this "constant" population could be dead by the time the epidemic is finished!). The key issues we wish to address are whether a newly introduced infection can lead to a serious epidemic (that is, will the infection "take off," or equivalently, will an epidemic even occur following introduction of the infection); the final size of the epidemic expressed as either the number in

or fraction of the population that becomes infected; the maximal incidence rate achieved over the duration of the epidemic, and epidemic dynamics such as the time required to infect a certain percentage of those who eventually will become infected, or the mean time to infection for those who get infected.

We will make one key assumption, which is that at any point in time, all susceptibles in the population face the same risk of infection at that point in time. With the introduction of infection at time 0, let $\lambda(t)$ denote the instantaneous risk of infection to any susceptible in the population. That is, the probability that a randomly selected person from the population would be infected between t and $t + \Delta t$, given that they have not become infected by time t, is equal to $\lambda(t)\Delta t$. In epidemiological terms, the function $\lambda(t)$ is the instantaneous *incidence rate of infection at time t*. Mathematically, $\lambda(t)$ is just the hazard function for the duration variable T_S, the time to infection. Now, since in general not everyone will get infected in the epidemic, it is possible for T_S to be infinite (for recall that if an event does not occur, then the time until that event occurs must be infinite!). Thus, the instantaneous incidence rate $\lambda(t)$ is the hazard rate for the defective duration T_S.

If we know the incidence rate, we can apply the results of the previous section. Let $p(t)$ denote the probability that an individual in the population becomes infected by time t after the infection is introduced at time 0. Using equation (1.1) we immediately conclude that

$$p(t) = \Pr\{T_S \leq t\} = 1 - e^{-\int_0^t \lambda(u)du} \tag{2.1}$$

and thus the final size of the epidemic (expressed as the fraction of the population that ultimately gets infected) can be expressed as

$$p \equiv \lim_{t\to\infty} p(t) = 1 - e^{-\int_0^\infty \lambda(u)du}. \tag{2.2}$$

These are fundamental equations.

Now let's consider some epidemic dynamics. Let t_α denote the time by which a fraction α of those who will ultimately get infected do so. To find t_α, we solve

$$\Pr\{T_S \leq t_\alpha | T_S < \infty\} = \frac{p(t_\alpha)}{p} = \alpha. \tag{2.3}$$

To determine the mean time until infection for those infected, we evaluate (see equation (1.8))

$$E(T_S|T_S < \infty) = \int_0^\infty \frac{S(t) - (1-p)}{p} dt = \int_0^\infty \frac{p - p(t)}{p} dt. \tag{2.4}$$

We defer discussion of whether an infection can "take off" for just a bit longer.

3. The Basic SIR Model: Classical Approach

Now let's very quickly restate the classical SIR model as described in most epidemic texts (though as we will soon see, this is the simplest case of the SIR model). The model to be described dates back to a 1927 paper by Kermack and McKendrick, and is often described as *the* Kermack-McKendrick model, but in fact Kermack and McKendrick's model was much more general, and what follows was presented as a special case.

Let the number of **S**usceptible, **I**nfected, and **R**ecovered individuals in the population (hence **SIR***)* be denoted by $X(t), Y(t)$ and $Z(t)$ respectively. Assuming that the population remains constant at size N we have the conservation equation

$$X(t) + Y(t) + Z(t) = N \text{ for } t \geq 0. \tag{3.1}$$

The model assumes free mixing, thus the incidence rate at time t is given by

$$\lambda(t) = \beta Y(t) \tag{3.2}$$

where β is the (assumed constant) transmission rate (with units per person per unit time). In the aggregate, since there are $X(t)$ susceptibles in the population at time t, each getting infected at rate $\lambda(t) = \beta Y(t)$, we can describe the dynamics of $X(t)$ with the differential equation

$$\frac{dX}{dt} = -X(t)\lambda(t) = -\beta X(t)Y(t). \tag{3.3}$$

Note that equation (3.3) is deterministic, so even though I will continue using the language of probability, what we are doing is using probabilistic methods to understand, formulate, solve and interpret what is in truth a completely deterministic model. Also, note that I refer to numbers of individuals, suggesting a discrete model, when in fact we are treating populations as continuous (so we are really looking at "flows" of infection and recovery).

Continuing with the formulation, the number of infecteds clearly grows as new susceptibles are infected. Infecteds are assumed to become infectious immediately upon infection, and as stated earlier, transmit infections at a constant rate (as opposed to a rate which depends on how long they have been infected for, a realistic generalization we will attempt later). The duration of infectiousness T_I

facing each newly infected individual is assumed to be exponentially distributed with mean $1/\upsilon$, after which infected individuals "recover" and are no longer infectious. Note that "recovery" could be achieved by "dying!" Some recovery. Now, recall from earlier notes that exponentially distributed durations are equivalently characterized by constant hazard rates, and in the present situation, the constant hazard is the recovery rate υ. Thus, if there are $Y(t)$ infected and infectious individuals at time t, each recovering from infection with rate υ, then the aggregate recovery rate is simply $Y(t)\upsilon$. This leads to a differential equation for the number of infected (and infectious) persons in the population at time t as

$$\frac{dY}{dt} = \beta X(t)Y(t) - \upsilon Y(t). \tag{3.4}$$

Finally, whenever an infectious individual recovers, (s)he, uh, enters recovery! So at time t, the number of recovered individuals $Z(t)$ is augmented by new arrivals at rate $\upsilon Y(t)$. This leads to the equation

$$\frac{dZ}{dt} = \upsilon Y(t). \tag{3.5}$$

Note that if you sum equations (3.3)-(3.5) you get zero (why? check equation (3.1)). Equations (3.3)-(3.5) are commonly referred to as *the* SIR model (or *the* Kermack-McKendrick model). Again, what we really have is a special (indeed the simplest) case of the SIR model in these equations.

3.1. Mechanically Solving the Basic SIR Model

Suppose you want to generate the entire trajectory for a simple SIR model. You know the input parameters N (which could be 1 – any idea why you might want to do that?), β and υ, and you have initial conditions $X(0)$, $Y(0) > 0$ (or else you can never generate any infections – note that if $N = 1$, $0 < Y(0) < 1$ – again, why might you want to do things this way?), and $Z(0) = N - X(0) - Y(0)$ (usually one takes $Z(0) = 0$; why?). And, you have a spreadsheet program like Excel. How can you model an epidemic?

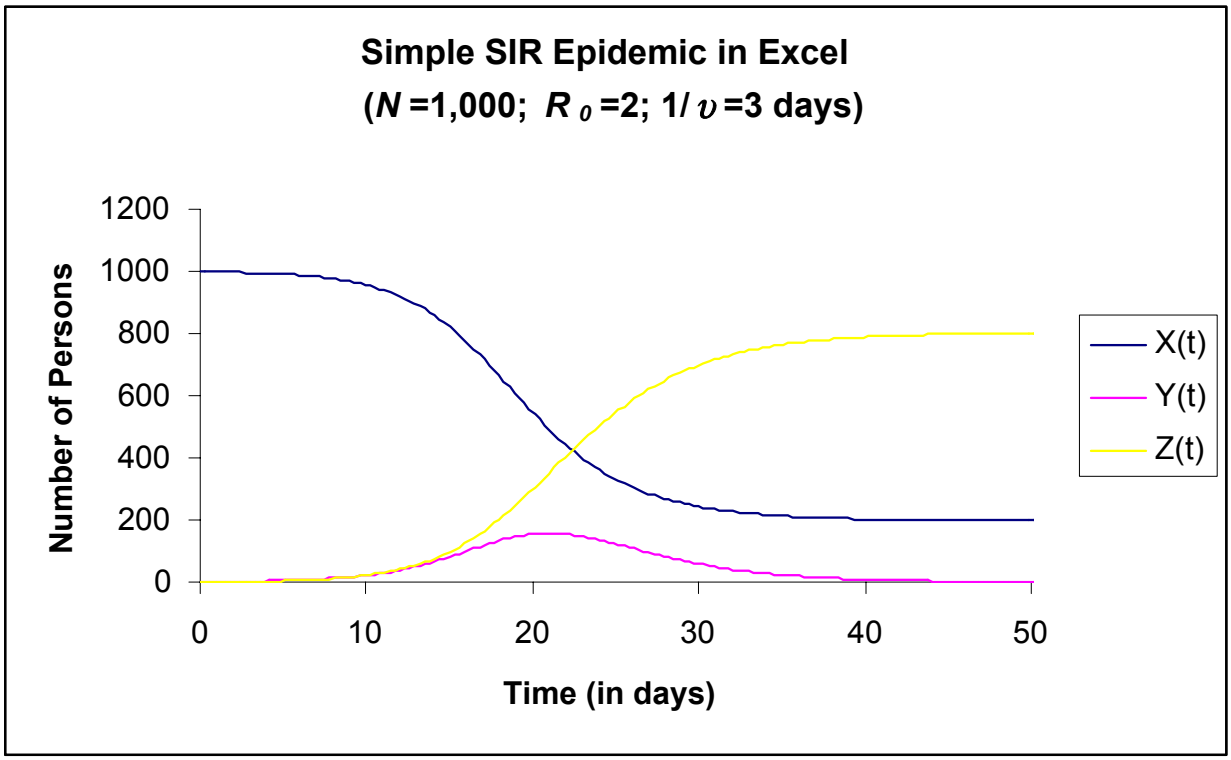
Simple – just do the entire thing in discrete time steps. Pick an appropriate time step ($0.1 \times 1/(\beta N - \upsilon)$ for example; see equation (3.8) for a clue underlying

this recommendation), call this time step Δt, and recursively solve:

$$\begin{aligned} X(t+\Delta t) &= X(t) - \beta X(t)Y(t)\Delta t \\ Y(t+\Delta t) &= Y(t) + [\beta X(t)Y(t) - \upsilon Y(t)]\,\Delta t \\ Z(t+\Delta t) &= Z(t) + \upsilon Y(t)\Delta t. \end{aligned} \tag{3.6}$$

Each of the rows in your spreadsheet will correspond to a point in time, while the columns will correspond to the different variables.

As an example, the figure below reports an Excel-produced simulation using the equations above for an influenza outbreak ($\beta = 1/1500$, $1/\upsilon = 3$ days) in a population of $N = 1,000$ (maybe a high school). In the simulation, the time step used was $.1/(\beta N - \upsilon) = .1/(2/3 - 1/3) = .3$. Simulating for 50 days required $50/\Delta t = 50/.3 = 167$ rows in Excel. OK, back to the real stuff.



3.2. The Reproductive Number R_0 and the Basic SIR Model

Let's take a quick look at equation (3.4) and ask: what must be true for the infection to take off? Initially, the number of susceptibles comprises almost the entire population, so let's just approximate matters by assuming that $X(t) \approx N$ early in the epidemic. This approximation leads to the simpler, *linear* differential equation (indeed, what we are doing is referred to as linearization)

$$\frac{dY}{dt} = (\beta N - \upsilon)\, Y(t) \tag{3.7}$$

which has the direct solution

$$Y(t) = Y(0)e^{(\beta N - \upsilon)t}. \tag{3.8}$$

When will the number of infected people grow? The answer is clear: when the growth rate $\beta N - \upsilon > 0$, or equivalently when $\beta N > \upsilon$, or equivalently when

$$R_0 \equiv \frac{N\beta}{\upsilon} > 1. \tag{3.9}$$

R_0 is referred to as the basic reproductive number. It has a very intuitive meaning: if we stick a single newly-infectious person into a population of N susceptibles, this infectious person is (thanks to the free mixing assumption) spreading infections at rate $N\beta$ per unit time (recall that the units of the transmission rate β are per person per unit time, so multiplication by N gives us infections per unit time). How long does our index case continue spreading infections? On average for $E(T_I) = 1/\upsilon$ for the basic SIR model. Thus, the expected total number of infections directly transmitted by a newly-infectious individual early in the epidemic is equal to $N\beta/\upsilon$. By the way, recall equation (4.26), p. 24 from the "system of flow" notes – we have essentially the same formula for R_0 in the basic SIR model as in the stable population with an endemic infection!

The definition of R_0 as the expected number of infections generated by a newly-infectious individual early in the epidemic (essentially in a population that is 100% susceptible) survives for much more general models. Indeed, as we will see, it is often possible to write down a formula for R_0 by inspection once one understands whatever particular model assumptions are being made. A good and reasonably general version of R_0 to remember is

$$R_0 = N\beta E(T_I). \tag{3.10}$$

As an aside, some modelers prefer to focus upon what is called a "contact rate" and a corresponding probability of transmission *per contact*, as opposed to the more general transmission rate β. To see the equivalence, focus on a sexually transmitted disease like gonorrhea. Let c denote the average number of sexual contacts per person per unit time, and b denote the probability of infection per contact. A newly infected person would then have c sexual contacts with susceptibles in the population per unit time, a fraction b of which would lead to infection. Thus, the index case would be infecting others at rate cb per unit time, thus a total of $cbE(T_I)$ infections would be transmitted over the index's duration of infectiousness. This leads to an equivalent formula for R_0

$$R_0 = cbE(T_I) \tag{3.11}$$

and upon comparing equations (3.10) and (3.11), we see that it really doesn't matter which form we use providing that

$$cb = N\beta \tag{3.12}$$

or equivalently, that

$$\beta = \frac{cb}{N}. \tag{3.13}$$

For certain infectious diseases (STDs readily come to mind), the notion of a contact and per-contact probability of transmission are eminently sensible, but for others (e.g. influenza, smallpox) it is more difficult to structure things this way. In practice, as we will see, it is much more sensible to focus on R_0 and determine the transmission rate β via the equation

$$\beta = \frac{R_0}{NE(T_I)}. \tag{3.14}$$

3.3. Final Size of the Basic SIR Epidemic

If you like playing with differential equations, you'll like this. Divide equation (3.3) by equation (3.5) to obtain an equation for the number susceptible as a function of the number recovered (!), that is

$$\frac{dX}{dZ} = \frac{-\beta X(Z)}{\upsilon}. \tag{3.15}$$

This equation can be solved explicitly to yield

$$X(Z) = X(0)e^{-\frac{\beta}{\upsilon}Z}. \tag{3.16}$$

Now, $X(0) \approx N$, the population size. At the start of the epidemic, virtually everyone is susceptible, and certainly no one is recovered yet (see parenthetical remark on $Z(0)$, first paragraph of Section 3.1), so if $Z = 0$, we're at the start of the epidemic if $X(0) \approx N$. Remember, we're looking at X as a function of Z here, not of time! At the *end* of the epidemic, no one is infected (why???), and thus everyone is either susceptible (these folks never got infected), or recovered (guess what happened to everyone who got infected?). So jumping out to the end of the epidemic, it must be that Z accounts for *everyone* who was *ever* infected, that is, at the end of the epidemic we have

$$Z_\infty = Np \tag{3.17}$$

(where the subscript ∞ reminds us that we are at the end of the epidemic). Similarly, everyone who did not get infected by the end of the epidemic is still susceptible, thus it must be true that

$$X_\infty \equiv X(Z_\infty) = N(1-p). \tag{3.18}$$

Again reminding yourself that $X(0) \approx N$, equation (3.16) evaluated at Z_∞ leads to

$$X_\infty = N(1-p) = X(0)e^{-\frac{\beta}{v}Z_\infty} \approx Ne^{-\frac{\beta}{v}Np}. \tag{3.19}$$

Now, recall from equation (3.9) that $R_0 = N\beta/v$. We arrive at the beautiful result

$$1 - p \approx e^{-R_0 p} \tag{3.20}$$

or, as I like to write this,

$$p = 1 - e^{-R_0 p}. \tag{3.21}$$

So if you know R_0, you can find the fraction of the population ever infected, that is, the final size in an SIR epidemic, as the larger root of the equation above.

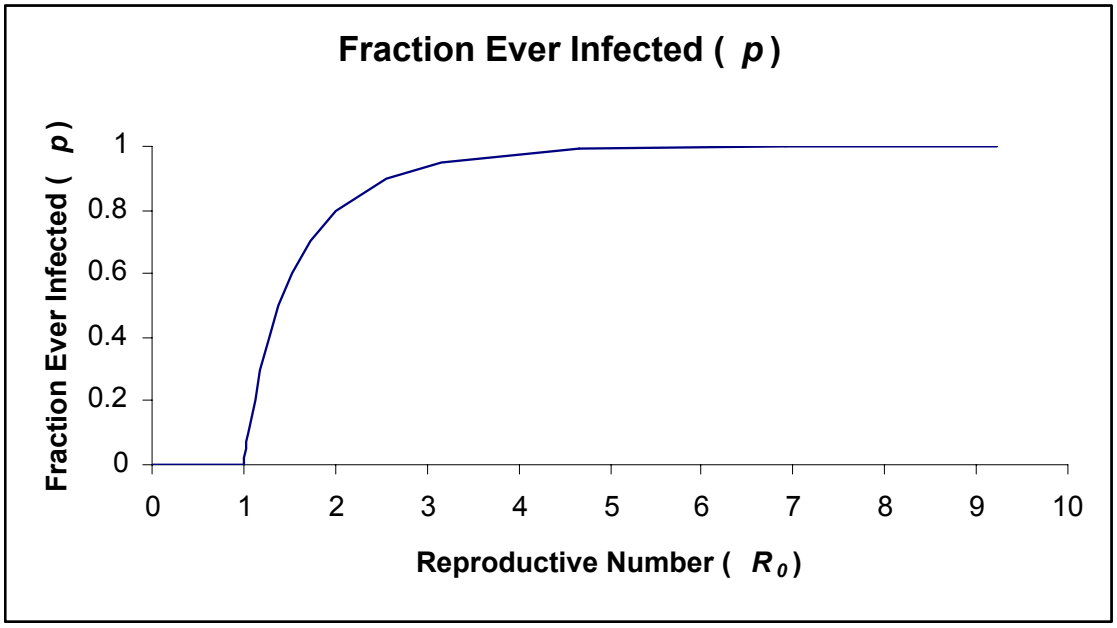
I said the larger root. Clearly zero is also always a root of this equation. When is it the correct answer? Suppose $R_0 \leq 1$ and consider both sides of equation (3.21) as a function of p. Note that the left-hand side (LHS) of this equation grows at rate 1 as a function of p ($dp/dp = 1$ after all), while the right-hand side (RHS) grows at rate $R_0 e^{-R_0 p}$ (this is the derivative of the RHS with respect to p). Now, if $R_0 \leq 1$, the derivative of the RHS is always less than the derivative of the LHS, which means that for all $p > 0$, the LHS is larger than RHS – and thus the *only* solution is $p = 0$ when $R_0 \leq 1$. Now, this is not strictly true thanks to the approximation $X(0) = N$ (which obviously ignores the initial number of infections

$Y(0)$ – it must be that $p \geq Y(0)/N$ no matter what R_0 is), but basically equation (3.21) gives us the right answer. You get a negligible epidemic if $R_0 \leq 1$.

Note that while p must be found numerically as the root of equation (3.21), one can write the reproductive number R_0 as an explicit function of the final size of the epidemic, that is,

$$R_0 = -\frac{\log(1-p)}{p}. \tag{3.22}$$

From this equation it is clear that $\lim_{p\to 0} R_0 = 1$ while $\lim_{p\to 1} R_0 = \infty$. And, as the fraction ever infected has to be even lower if $R_0 < 1$ than for $R_0 = 1$, we can use equation (3.22) and the limits just noted to produce the graph below.



Here is a similar trick you can play with this model. We can derive the number infected $Y(t)$ as a function of the number of susceptibles $X(t)$ by dividing equation (3.4) by equation (3.3) which yields

$$\frac{dY(X)}{dX} = \frac{\beta XY - \upsilon Y}{-\beta XY} = \frac{\upsilon}{\beta X} - 1. \tag{3.23}$$

This integrates to yield

$$Y(X) = \frac{\upsilon}{\beta}\log X - X + C. \tag{3.24}$$

To evaluate the constant of integration C, note that at the start of the epidemic when $X \approx N$, Y just equals the initial number infected, say Y_0. Thus,

$$Y(N) = Y_0 = \frac{\upsilon}{\beta} \log N - N + C \tag{3.25}$$

and returning to equation (3.24) we obtain

$$Y(X) = \frac{\upsilon}{\beta} \log X - X + Y_0 - \frac{\upsilon}{\beta} \log N + N. \tag{3.26}$$

Let $y = Y/N$ and $x = X/N$ denote the fractions of the population that are infected and susceptible respectively. Dividing both sides of equation (3.26) by N, taking $Y_0/N = y_0 \approx 0$ (since the fraction of the total population infected initially is negligible), and recalling that $R_0 = \beta N/\upsilon$, we obtain

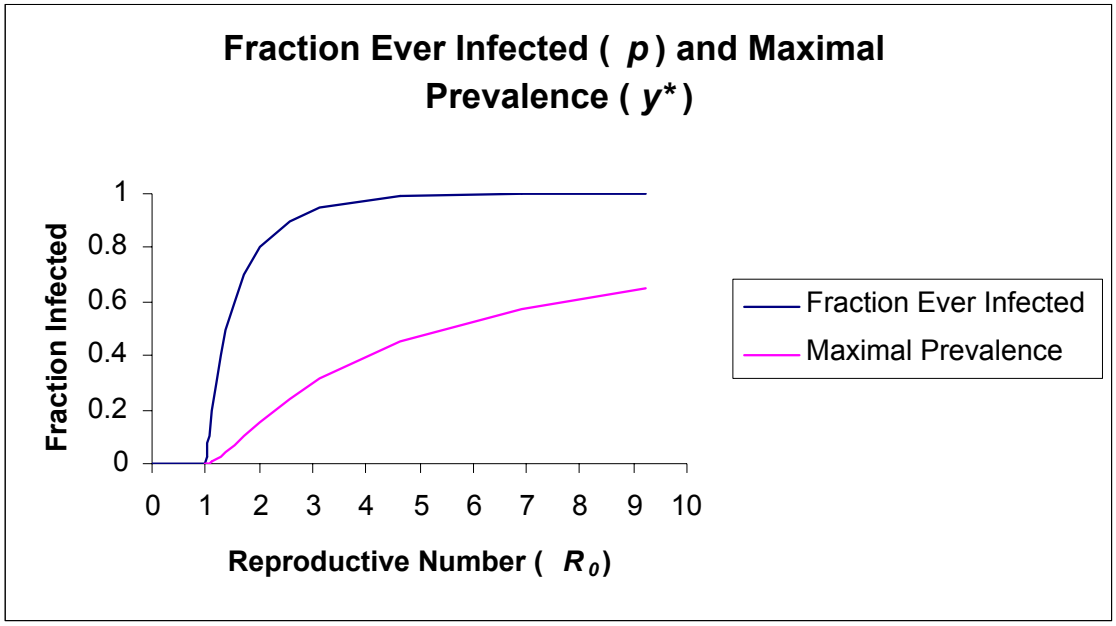
$$y(x) = \frac{\log(x)}{R_0} - x + 1. \tag{3.27}$$

This provides another route to the final size result, for at the end of the epidemic, everyone is either susceptible or recovered, so setting $y(x) = 0$ and rearranging the equation above (and recognizing that $p = 1 - x_\infty$ where x_∞ is the fraction susceptible at the end of the epidemic) yields equation (3.21).

There is another use of equation (3.27). Note from equation (3.4) that $dY/dt = 0$ when $\beta XY = \upsilon Y$, or when $X = \upsilon/\beta$. But since $Y(t)$ clearly rises and falls in an epidemic (look at the graph from our Excel simulation), $dY/dt = 0$ is exactly the condition for finding the maximum number infected at any point in time over the course of the epidemic. And, since $\lambda(t) = \beta Y(t)$, the maximum number infected leads directly to the maximum incidence rate. If $X = \upsilon/\beta$ at the time number infected is maximal, then the fraction susceptible at this same point in time must equal $x = X/N = \upsilon/(\beta N) = 1/R_0$. Sustituting into equation into equation (3.27), we see that the maximum prevalence of infection at any point in time over the course of the epidemic y^* is given by

$$\begin{aligned} y^* = y(\frac{1}{R_0}) &= \frac{\log(\frac{1}{R_0})}{R_0} - \frac{1}{R_0} + 1 \\ &= 1 - \frac{(1 + \log R_0)}{R_0}. \end{aligned} \tag{3.28}$$

The maximal incidence rate $\lambda^* = \beta N y^*$. A graph comparing the maximal prevalence to the final size as a function of R_0 appears below.



4. The Basic SIR Model: Final Size via the Hazard Function

The hazard function approach enables you to make probabilistic statements about a deterministic model! Of course, that requires knowing what the hazard function is. For the basic SIR model, we already saw from equation (3.2) that $\lambda(t) = \beta Y(t)$. The differential equations (3.3)-(3.5) (or their discrete analog in equation (3.6)) enable the actual computation of $Y(t)$, and thus the hazard $\lambda(t)$ is readily available for determining quantities such as t_α, $E(T_S|T_S < \infty)$ and so forth.

But the main advantage of the hazard approach is that it provides an *immediate* route to the final size result, and also enables generalization not only to more realistic SIR models, but also to models with non-random mixing (a subject down the road).

Let's focus on the final size result. From equation (2.2), we have directly that

$$p = 1 - e^{-\int_0^\infty \lambda(t)dt} = 1 - e^{-\int_0^\infty \beta Y(t)dt} \tag{4.1}$$

where we have substituted in the specific from of the hazard that corresponds to the basic SIR model. Focus on the integrated hazard $\int_0^\infty \beta Y(t)dt$, and ask yourself

what this is. Clearly $\int_0^\infty Y(t)dt$ is the total amount of "infectious person-time" spent in the population over the course of the epidemic. That is, $\int_0^\infty Y(t)dt$ is the *total* time spent infectious by *all* persons who were *ever* infectious.

But wait! The fraction of the population that was ever infectious is, by definition, equal to p! Thus, the number of persons who were ever infections must obviously equal Np. Furthermore, each infected person spends, on average, $E(T_I)$ time units being infectious, which means that the *total* time spent infectious by *all* persons who were *ever* infectious *must* be given by

$$\int_0^\infty Y(t)dt = NpE(T_I) \text{ (!!)} \tag{4.2}$$

This identity has to be true for *any* "fast" epidemic (where the population is constant and thus the epidemic can be viewed as a cohort), and not just for the basic SIR model. But, while we're still playing with the basic model, note that $E(T_I) = 1/\upsilon$ since the time to recovery (which is the same as the infectious period in this model) is exponentially distributed with mean $1/\upsilon$. Thus, for the basic SIR model, we directly arrive at the amazing result that

$$\int_0^\infty \beta Y(t)dt = \beta\frac{Np}{\upsilon} = \frac{N\beta}{\upsilon}p = R_0 p \text{ (!!)} \tag{4.3}$$

and thus, from equation (4.1), we instantly obtain the final size result

$$p = 1 - e^{-\int_0^\infty \beta Y(t)dt} = 1 - e^{-R_0 p} \text{ (!!)} \tag{4.4}$$

Furthermore, it seems like the substitution $E(T_I) = 1/\upsilon$ was a throwaway; really what we have is that

$$p = 1 - e^{-\beta N p E(T_I)} = 1 - e^{-R_0 p} \text{ (!!)} \tag{4.5}$$

where I have used the "general" definition of R_0 available from equation (3.10).

5. Finding the Final Size for Generalizations of the Basic SIR Model

5.1. Non-exponential (*arbitrary*) Duration of Infectiousness T_I

Suppose that the duration of infectiousness in an SIR epidemic has an arbitrary, non-exponential duration. The (proper!) duration T_I can be described by allowing

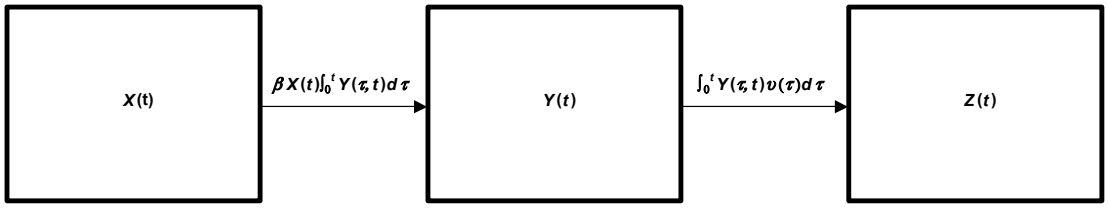
its hazard function $\upsilon(t)$ to be arbitrary. The total number of persons who are infectious at time t consists of persons who have been infected for different lengths of time and have yet to recover. Let $Y(\tau, t)$ denote the density of infectious persons at time t who have been infected for τ units of time. The total number of persons infected at time t, $Y(t)$, is then found by integrating $Y(\tau, t)$ over the time spent infected, that is,

$$Y(t) = \int_{\tau=0}^{t} Y(\tau, t) d\tau. \tag{5.1}$$

Note that $Y(0, t)$ is the total arrival rate of new infections at time t. To get a sense of the flows in this more general SIR model, review the figure below. We have retained the assumption of free mixing, so the incidence rate (equivalently hazard rate for infection) at time t is equal to

$$\lambda(t) = \beta Y(t) = \beta \int_{\tau=0}^{t} Y(\tau, t) d\tau \tag{5.2}$$

while the aggregate recovery rate can no longer be written as $\upsilon Y(t)$ since the duration of infectiousness is no longer exponential, but instead must be written as $\int_{\tau=0}^{t} Y(\tau, t)\upsilon(\tau)d\tau$. Make sure you understand this formulation.



To formulate this model with differential equations would require the use of *partial* differential equations (or PDE's). You will sometimes need to do this to obtain a complete trajectory of the epidemic, but we'll leave that to another day. However, we can immediately obtain the final size equation via our hazard function approach.

First, note that the reproductive number for this model is exactly given by equation (3.10). The expected number of infections generated by a newly infectious individual early in the epidemic would indeed be equal to

$$R_0 = N\beta E(T_I) \tag{5.3}$$

where we recognize that

$$E(T_I) = \int_0^{\infty} e^{-\int_0^t \upsilon(u)du} dt \tag{5.4}$$

as follows from recognizing the survivor function for T_I and integrating the tail. Second, from equations (2.2) and (5.2), we can write down the final size equation as

$$p = 1 - e^{-\int_0^\infty \lambda(t)dt} = 1 - e^{-\int_{t=0}^\infty \beta \int_{\tau=0}^t Y(\tau,t)d\tau dt}. \tag{5.5}$$

Now, as with the simple SIR model, note that $\int_{t=0}^\infty \int_{\tau=0}^t Y(\tau,t)d\tau dt$ is clearly the *total* time spent infectious by *all* persons who were *ever* infectious. As before, we arrive at the identity

$$\int_{t=0}^\infty \int_{\tau=0}^t Y(\tau,t)d\tau dt = NpE(T_I) \text{ (!!)} \tag{5.6}$$

Substituting this result back into equation (5.5) and comparing with equation (5.3), we obtain

$$p = 1 - e^{-\beta NpE(T_I)} = 1 - e^{-R_0 p}. \tag{5.7}$$

What we have just learned is that for a given value of R_0, it doesn't matter what the distribution of the duration of infectiousness equals in an SIR epidemic for purposes of determining the final size. The fraction of (and hence number in) the population that is ultimately infected only depends on the mean duration of infectiousness via the reproductive number R_0, and is otherwise *independent* of the probability distribution of the infectious period!

5.2. Stages of Infection with Stage-Dependent Infectiousness (Transmission)

Thus far, we have assumed that infectiousness remains constant over the course of the infectious period (as indicated by the fact that the transmission parameter β stays constant), and also that infectious individuals become infectious immediately upon infection. Most infections don't work this way. As discussed at length in the book by Anderson and May (Chapter 3), it is typical that a latent period during which those infected are *not* yet infectious must pass, followed by an infectious period. For some diseases, there are other identifiable stages during which infectiousness differs. For example, smallpox progressed through a latent period to a *prodrome* characterized by a high fever and flu-like symptoms before reaching the overt rash stage (I say progressed because smallpox is the one infectious disease that has been eradicated, though there are fears that it could potentially be re-introduced by terrorists). There is considerable argument regarding the infectiousness of persons in this prodrome relative to the overt rash stage, but the only important point here is that the infectiousness differs by stage of disease.

This is an easy generalization to incorporate into the basic SIR framework. Instead of having a single stage of infection (and associated duration of infectiousness), we now assume that upon infection, individuals pass through k serial stages of infection, each with their own infectiousness characterized by the transmission rate β_j in stage j, and duration of time T_{I_j} spent in stage j of infection, $j = 1, 2, ..., k$. Note that if stage j is a latent stage, then $\beta_j = 0$. Models with exactly one latent stage followed by one infectious stage followed by recovery are known as **SEIR** models (where the latent stage is represented by **E** for **E**xposed; we have already used the symbol L to represent the duration of a human lifespan, and besides, who would want to work with **SLIR** models, I mean how much **SLIR** could we get?).

We assume the random variables T_{I_j} are mutually independent, proper duration variables that can have any probability distribution we wish. Otherwise, we retain the free mixing assumption, which again means that at any point in time, all susceptibles share the same risk of acquiring infection. The instantaneous incidence rate (hazard function) $\lambda(t)$ for this model is given by

$$\lambda(t) = \sum_{j=1}^{k} \beta_j Y_j(t) \tag{5.8}$$

where $Y_j(t)$, the number of infected persons in stage j of infection, is given by integrating over the density of time spent infectious in stage j, that is,

$$Y_j(t) = \int_{\tau=0}^{t} Y_j(\tau, t) d\tau \tag{5.9}$$

as in equation (5.1). Note that the arrival rate to the first stage of infection is equal to the aggregate rate of new infections, that is,

$$Y_1(0, t) = X(t)\lambda(t) \tag{5.10}$$

where as usual $X(t)$ is the number of susceptibles in the population at time t. For stages 2 through k, the arrival rate to stage j is exactly equal to the departure rate from stage $j - 1$, thus we have the flow conservation

$$Y_j(0, t) = \int_{\tau=0}^{t} Y_{j-1}(\tau, t)\upsilon_{j-1}(\tau) d\tau \tag{5.11}$$

where $\upsilon_j(t)$ is the hazard function associated with T_{I_j}, the j^{th} stage of infection.

The reproductive number R_0 for this model is still defined as the total number of infections transmitted by a single newly-infected individual over the course of the total duration of infectiousness, but we now need to break this up by stage of infection. This results in the appealing formula

$$R_0 = N \sum_{j=1}^{k} \beta_j E(T_{I_j}) = N\beta E(T_I) \tag{5.12}$$

where we recognize

$$E(T_I) = \sum_{j=1}^{k} E(T_{I_j}) \tag{5.13}$$

as the expected *total* duration of infectiousness, and

$$\beta = \frac{\sum_{j=1}^{k} \beta_j E(T_{I_j})}{\sum_{j=1}^{k} E(T_{I_j})} = \frac{\sum_{j=1}^{k} \beta_j E(T_{I_j})}{E(T_I)} \tag{5.14}$$

as the *average* transmission rate (where the average is weighted by the respective mean durations of infectiousness across stages).

To race straight to the final size of such an epidemic, we note that

$$\begin{aligned}
\int_0^\infty \lambda(t)dt &= \int_0^\infty \sum_{j=1}^{k} \beta_j Y_j(t)dt \text{ (eq 5.8)} \\
&= \sum_{j=1}^{k} \beta_j \int_0^\infty Y_j(t)dt \\
&= \sum_{j=1}^{k} \beta_j NpE(T_{I_j}) \text{ (total time infectious in stage } j) \qquad (5.15) \\
&= pN \sum_{j=1}^{k} \beta_j E(T_{I_j}) \\
&= pR_0 \text{ (!!) (eq 5.12).}
\end{aligned}$$

Once again we have, thanks to equation (2.2), the familiar final size equation

$$p = 1 - e^{-pR_0}. \tag{5.16}$$

This really helps, because all you need to get the final size of the epidemic is R_0. The specific combinations of infectious stage (including latent stage) durations, their distributions, and their stage-specific infectiousness don't matter. All that does is R_0.

5.3. Duration-Dependent Transmission and Arbitrary Duration of Infectiousness

One can go even further. Some would quibble that even a staged model of disease progression is an inadequate approximation, and that what is really required is a model where infectiousness varies continuously with the time from infection. Thus, if a person has been infected for some time τ, then transmission occurs at rate $\beta(\tau)$.

As Amanda Bynes often states on my daughter's favorite TV show (I know, I'm letting her watch too much junk), "Not a problem!!!" First, letting T_I denote the duration of infectiousness, we immediately can write down the formula for R_0 as

$$R_0 = N \int_{\tau=0}^{\infty} \beta(\tau) \Pr\{T_I > \tau\} d\tau = N\beta E(T_I) \tag{5.17}$$

where now we have defined the mean infectiousness β as

$$\beta = \frac{\int_{\tau=0}^{\infty} \beta(\tau) \Pr\{T_I > \tau\} d\tau}{\int_{\tau=0}^{\infty} \Pr\{T_I > \tau\} d\tau} = \frac{\int_{\tau=0}^{\infty} \beta(\tau) \Pr\{T_I > \tau\} d\tau}{E(T_I)}. \tag{5.18}$$

To understand equation (5.17), note that

$$\int_{\tau=0}^{\infty} \beta(\tau) \Pr\{T_I > \tau\} d\tau = \int_{u=0}^{\infty} f_{T_I}(u) \int_{\tau=0}^{u} \beta(\tau) d\tau du \tag{5.19}$$

where $f_{T_I}(u)$ is the probability density of the infectious period T_I. For someone who has been infectious for exactly u units of time, the total transmission potential is $\int_{\tau=0}^{u} \beta(\tau) d\tau$; note that if $\beta(\tau)$ was a constant β_0, this integral would just equal βu, and substitution into equation (5.19) would yield $\beta_0 E(T_I)$ (which is the case of arbitrary duration of infectiousness but with constant transmission studied two sections ago).

If the density of persons who have been infected for τ units of time at time t is denoted by $Y(\tau, t)$, then the incidence (hazard) can be written as

$$\lambda(t) = \int_{\tau=0}^{t} Y(\tau, t) \beta(\tau) d\tau. \tag{5.20}$$

One can repeat the by now familiar steps in our analysis (sounds like a good homework problem), and arrive at the usual result that

$$\int_0^{\infty} \lambda(t) dt = Np\beta E(T_I) = R_0 p \tag{5.21}$$

and that, one last time with feeling, we have

$$p = 1 - e^{-R_0 p}. \tag{5.22}$$

And this concludes our supplemental notes on the hazard function approach to SIR-type models.