EMD 630/MGMT 711
Modeling Infectious Diseases: Theory and Applications
Systems of Flow, with Application to the Stable Population Model

Edward H. Kaplan*

August 2003

1. Systems of Flow

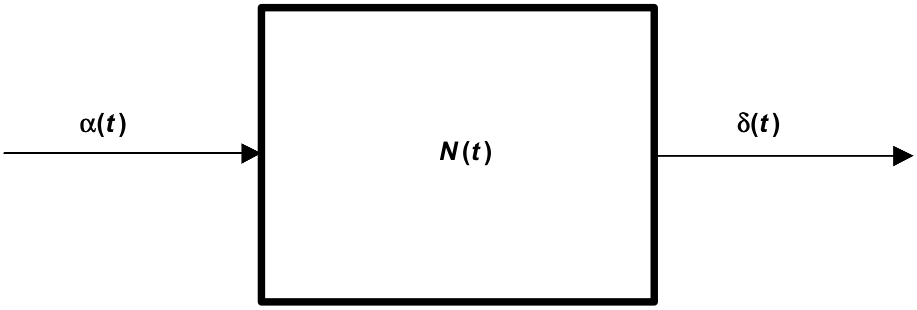
There are many systems, biological and other, that can be construed as systems of flow. Non-biological examples include river systems, transportation networks, computer and telecommunication networks, manufacturing supply chains, financial networks and even accounting systems (ever heard of a cash flow?). Biological systems include blood circulation, population dynamics (of animals, humans, bacteria, viruses etc.), and of course, disease transmission and progression, which is the subject of this class. While the modeling of infectious diseases is usually approached directly in terms of epidemiological specifics (and eventually we'll get there too), there are some basic principles governing systems of flow that are quite remarkable in their generality. Knowledge of these principles enables their application to *any* system of flow, including epidemiological systems of disease transmission and progression. We will therefore explore some of these ideas before jumping into what would be considered a more standard treatment of epidemic

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

modeling, and when we do take the required jumps, hopefully the material below will clarify matters.

1.1. Fundamental Equation of Flow

In the simplest system of flow, we focus on the movement of material (people, water, goods, money) into and out of a single "state" or "compartment" (see figure below). In economics or operations research, the quantity of material in the state is commonly referred to as "stock" (as in "the stock of inventory" or the "stock of money"), but we'll usually just talk about the "compartment size" in referring to the amount of material in the given state. Let $N(t)$ denote the compartment size at time t, $\alpha(t)$ denote the arrival rate at time t, and $\delta(t)$ denote the departure rate at time t.



In deterministic models, the amount of material that "flows" into the compartment between times t and $t + \Delta t$ is given by $\alpha(t)\Delta t$, while the amount that flows out of the compartment during this same time slice equals $\delta(t)\Delta t$. Arguing from first principles, we have a very simple and fundamental equation that describes the compartment size over time:

$$N(t + \Delta t) = N(t) + \alpha(t)\Delta t - \delta(t)\Delta t. \tag{1.1}$$

This *fundamental equation of flow* simply says that the compartment size in the next time instant equals the compartment size now, plus whatever shows up over the next time instant, minus whatever leaves. Think of all the things this equation describes. The number of persons infected with HIV tomorrow equals the number of persons infected today, plus the number of new HIV infections, minus the number of HIV-infected persons who die. As a less harrowing example, the amount of money in your checking account tomorrow equals the amount today, plus deposits, transfers in, interest etc. that “arrive” to your account, minus withdrawals, checks cleared, transfers out etc. that “depart.” My kid is taking a bath; how much water is in the tub in the next instant? Well, whatever is there now, plus whatever comes in from the faucet, minus whatever said kid splashes out (or drinks, or releases down the drain...). You get the idea.

In continuous time, it is easy to convert the fundamental equation of flow to differential form by taking the limit as $\Delta t \to 0$. That is, re-writing the equation as

$$\frac{N(t+\Delta t) - N(t)}{\Delta t} = \alpha(t) - \delta(t) \tag{1.2}$$

and letting $\Delta t \to 0$ yields

$$\frac{dN(t)}{dt} = \alpha(t) - \delta(t) \tag{1.3}$$

which has the general solution

$$N(t) = N(0) + \int_0^t [\alpha(u) - \delta(u)]\, du \tag{1.4}$$

where $N(0)$ is the compartment size at time 0, which itself can be arbitrary (e.g. Dec 16, 1955 which is my personal time 0, or the time at which an epidemic begins, etc.). Equation (1.4) is hardly a surprise, and really is just a re-statement of the fundamental equation of flow: what you’ve got at time t is whatever you started with at time 0, plus *everything* that arrived between 0 and t, minus *everything* that left.

So far the arguments have been in deterministic terms, but the ideas carry over to probabilistic (or stochastic) flow as well. In probabilistic models, $\alpha(t)\Delta t$ represents the probability of a single arrival between $t + \Delta t$, while $\delta(t)\Delta t$ is the probability of a single departure over the same time interval. Think of the flow system as representing the queue (including the person using the machine if applicable) at your local ATM machine. The probability a customer shows up and joins the queue in the next Δt time units is $\alpha(t)\Delta t$, while the probability that someone leaves (due to completing service, or getting impatient and just dropping out) equals $\delta(t)\Delta t$. For stochastic models, equation (1.4) models the *expected* compartment size over time (though we really should replace $N(t)$ by $E[N(t)]$ and similarly for $N(0)$ to recognize what’s fixed and what’s random).

In this class (and in most of the literature governing mathematical epidemiology), we technically will take the deterministic view. However, I will argue that in spite of this, opportunities for probabilistic interpretations abound when studying epidemiological systems of flow (i.e. disease transmission and progression), and in fact the ability to employ such interpretations not only makes the models easier to explain (and more convincing to the explainee!), it also sometimes leads to amazingly elegant shortcuts to the formulation and solution of models for infectious diseases.

Virtually all models for infectious diseases can be viewed as repeated application of the fundamental equation of flow in various guises. Much of our attention will focus on creating biologically and/or epidemiologically plausible sub-models for the arrival and departure rates $\alpha(t)$ and $\delta(t)$ for epidemiologically (and sometimes demographically and in the case of intervention programs even logistically) meaningful compartments. Key to such sub-models is to recognize some additional relationships. For example, what is the relationship between the number of deaths per unit time in a population and the duration of life itself? More generally, how can one relate the time spent by an individual in a compartment to the aggregate rate with which individuals of all ages leave said compartment? Note that the compartment in question could represent just about anything: persons infected with some disease, or persons waiting to be vaccinated but otherwise uninfected, or persons placed in quarantine, or persons displaying symptoms, etc.

Since so much of epidemiology involves the *duration* of important periods (e.g. time to infection, incubation period, duration of infectiousness, time to recovery, time to death etc.), it is not surprising that biostatisticians have developed a family of models for durations data under the general heading of survival analysis. But, as with our more general view of systems of flow, these same tools are much more general. Reliability theory is a long-established sub-field in engineering that, among other things, pays a lot of attention to the time until systems fail with the idea of designing systems that postpone failure beyond some reasonable time with high probability (enabling preventative maintenance to occur before a catastrophic failure does). Of course things don't always work out nicely, as Three Mile Island, or the recent East Coast blackout (which, for convenience, we'll blame on Cleveland) demonstrate. Economists also have use for such models. Think about the labor market: what are the typical durations of employment, or unemployment for that matter? How about the duration of a strike?

I digress (it's happened before, why stop now?). The point is, we need to review some basic probability models for describing durations and connect them to epidemiology so that we have some basic tools and terminology to work with later. So here we go.

2. Modeling Durations: Basic Probability Models

Consider some state, and let the random variable T denote the total amount of time an individual spends in this state. Or alternatively, let T denote the time from some arbitrary starting point until some event (a person becomes infected, the strike ends, the Red Sox win the World Series) occurs. We will assume that T is finite (even for the Red Sox) and continuous. What are the most convenient ways to describe random variable T? There are three approaches that all convey the same information (in that mathematically you can always derive them from each other with no additional help), but have different advantages depending upon what you are trying to do. There, that was informative! The three functions used to describe T are the...

2.1. Survivor, Density, and Hazard Functions

Let's start with the survivor function $S(t)$. This is simply defined as the probability that the duration in question exceeds t, that is,

$$S(t) = \Pr\{T > t\}. \tag{2.1}$$

The survivor function, being a probability, always falls between 0 and 1. Also, the survivor function is non-increasing, which is to say that while it can remain constant over arbitrary swaths of time, when it changes value, it can only get smaller. This makes sense. Let T be the time to death (otherwise known as your age at death). The chance of living past the age of 10 cannot be smaller than the chance of living past the age of 100, especially since to make it past 100, you first have to make it past 10! That is,

$$\begin{aligned} \Pr\{T > 10\} &= \Pr\{10 < T \leq 100\} + \Pr\{T > 100\} \\ &\geq \Pr\{T > 100\}. \end{aligned} \tag{2.2}$$

Note also that T is a non-negative random variable (life starts at birth, that is, at time 0), and that (usually) $S(0) = 1$ and $S(\infty) = 0$, which means that everyone gets to play the game of life, but no one lives forever. If we are considering human mortality, this is not strictly true as of course there are stillbirths (or miscarriages or abortions etc.). On the other hand, focusing on all durations that actually start (live births in our example), the endpoints clearly make sense.

A different way to describe durations is directly in terms of a probability density $f(t)$, that is,

$$\Pr\{t < T \leq t + \Delta t\} = f(t)\Delta t. \tag{2.3}$$

The relationship between $S(t)$ and $f(t)$ follows directly from the laws of probability, whence

$$S(t) = \Pr\{T > t\} = \int_t^\infty f(u)du \tag{2.4}$$

or, working in reverse,

$$f(t) = -\frac{dS(t)}{dt}\cdot \tag{2.5}$$

Note that

$$\int_0^\infty f(t)dt = S(0) = 1 \tag{2.6}$$

which simply says that the probability that the duration equals *something* is equal to unity. Knowledge of the probability density leads definitionally to all sorts of interesting things, such as the mean and second moment of the survival time T, defined as

$$E(T) = \int_0^\infty tf(t)dt \tag{2.7}$$

and

$$E(T^2) = \int_0^\infty t^2 f(t)dt \tag{2.8}$$

from which one obtains the variance of T as

$$Var(T) = E(T^2) - [E(T)]^2 \tag{2.9}$$

as is well known. However, as we will see shortly, there are other formulas worth knowing for these same quantities that will simplify matters greatly.

The third and most physically or biologically motivated approach to describing the duration T is via the hazard function $h(t)$. The hazard function is the *instantaneous* probability that the duration ends in the next Δt time units, *conditional* on survival through time t. Applied to human lifetimes, the hazard rate becomes the age-specific mortality rate, usually denoted $\mu(a)$, familiar to demographers and others. In general, the hazard rate is defined by

$$\begin{aligned} h(t)\Delta t &= \Pr\{t < T \leq t + \Delta t | T > t\} \\ &= \frac{f(t)\Delta t}{S(t)} \end{aligned} \tag{2.10}$$

whence

$$h(t) = \frac{f(t)}{S(t)}\cdot \tag{2.11}$$

Now, to obtain the survivor function from the hazard, using equation (2.5) in equation (2.11) yields

$$-\frac{1}{S(t)}\frac{dS(t)}{dt} = h(t) \tag{2.12}$$

and upon integration (and remembering that $S(0) = 1$) we obtain the important result

$$\log(S(t)) = -\int_0^t h(u)du \tag{2.13}$$

whence

$$S(t) = e^{-\int_0^t h(u)du} \tag{2.14}$$

and, from equation (2.11),

$$f(t) = h(t)S(t) = h(t)e^{-\int_0^t h(u)du}. \tag{2.15}$$

2.2. Special Cases

In principle, almost any non-negative random variable can be used to model durations, as evidenced in your readings from *Probabilistic Reliability: An Engineering Approach.* We will have opportunities to use many different duration models (to describe, for example, incubation times for AIDS, anthrax, or smallpox). However, two special cases are worth considering at this time.

2.2.1. Constant Survival Time (Type I Survival)

As argued by Anderson and May, the model of constant survival time provides an approximation for lifespans in the developed world. While it is not true that everyone literally dies at exactly the same age, the distribution of age at death is not that variable. For constant survival through death at age ℓ, we have

$$S(t) = \begin{cases} 1 & 0 < t \leq \ell \\ & \\ 0 & t > \ell \end{cases} . \tag{2.16}$$

The probability density and hazard functions are not particularly nice for Type I survival. The hazard function is defined for this case as

$$h(t) = \begin{cases} 0 & 0 < t < \ell \\ \infty & t = \ell \\ \text{who cares?} & t > \ell \end{cases} \tag{2.17}$$

while the density is given by

$$f(t) = \mathfrak{i}_\ell(t) \tag{2.18}$$

where $\mathfrak{i}_\ell(t)$ is the unit impulse function centered at ℓ, that is, $\mathfrak{i}_\ell(t) = 0$ for $t \neq \ell$, yet somehow $\lim_{\varepsilon\to 0} \int_{\ell-\varepsilon}^{\ell+\varepsilon} \mathfrak{i}_\ell(t)dt = 1$. Stick with the survivor function. Oh, and of course for Type I survival, $E(T) = \ell$ and $Var(T) = 0$.

2.2.2. Constant Hazard Rate (Type II Survival)

Mainly employed for mathematical convenience, the assumption of constant hazards for mortality and disease progression dominates in epidemiological models. Though this assumption is rarely justified on the basis of descriptive accuracy, the results one obtains regarding several (though not all) aspects of epidemics turn out to be surprisingly robust to departures from this assumption. Proceeding, if the hazard rate $h(t)$ is constant and equal to (say) $\mu > 0$, then direct application of equations (2.14) and (2.15) leads to the following results for the survivor function and probability density:

$$S(t) = e^{-\mu t} \text{ for } t > 0 \tag{2.19}$$

and

$$f(t) = \mu e^{-\mu t} \text{ for } t > 0. \tag{2.20}$$

Under Type II survival, random variable T follows the exponential distribution with mean $1/\mu$ (as is easily verified by Equation (2.7)). That $\mu = 1/E(T)$ is a special case of a more general result as we will discover.

2.3. Expected Survival Time

As mentioned in the discussion surrounding equation (2.7), one can obtain the mean duration $E(T)$ definitionally from the probability density. But for non-negative random variables such as durations, there is an even simpler formula called "integrating the tail" that involves only the survivor function. Here's how it works: suppose T refers to human lifetime, and focusing on a random live birth at time 0, define

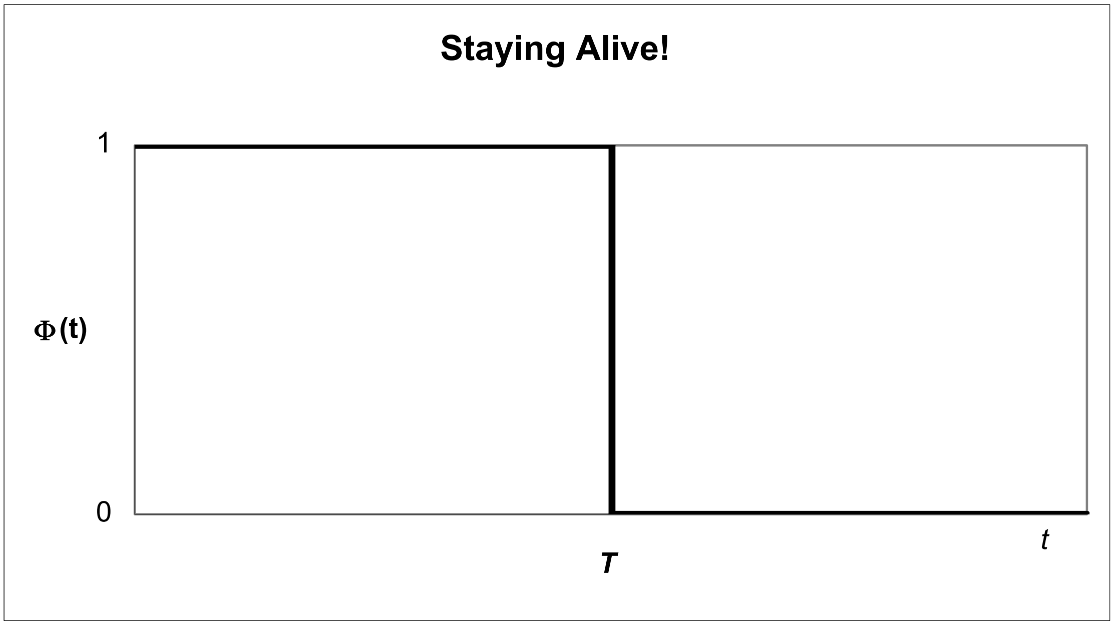
$$\Phi(t) = \begin{cases} 1 & \text{Still alive at time } t \\ \\ 0 & \text{Not alive at time } t \end{cases} . \tag{2.21}$$

Then the duration of life (equivalently, age at death), T, can be expressed cleanly as

$$T = \int_0^\infty \Phi(t) dt \tag{2.22}$$

as is clear from the figure below. Thus, the *expected* duration of life is given by

$$E(T) = E\left[\int_0^\infty \Phi(t) dt\right] = \int_0^\infty E\left[\Phi(t)\right] dt. \tag{2.23}$$



So we have changed the problem into figuring out the expected value of $\Phi(t)$. But wait, this is easy – since $\Phi(t)$ only takes on the values 0 or 1, we have

$$\begin{aligned} E[\Phi(t)] &= \Pr\{\Phi(t) = 1\} \times 1 + \Pr\{\Phi(t) = 0\} \times 0 \\ &= \Pr\{\Phi(t) = 1\}. \end{aligned} \tag{2.24}$$

Now, what is the probability that $\Phi(t) = 1$? This is exactly the same as the probability that someone born at time 0 is alive at time t, that is, if $T > t$ (since T after all is the time of death, so if your time of death is greater than t, you are alive at time t!). And as we all know, $\Pr\{T > t\} = S(t)$, the survivor function! We have thus shown that

$$E[\Phi(t)] = \Pr\{\Phi(t) = 1\} = \Pr\{T > t\} = S(t), \tag{2.25}$$

and as a consequence, the expected duration $E(T)$ is given by

$$E(T) = \int_0^\infty E[\Phi(t)]dt = \int_0^\infty S(t)dt. \tag{2.26}$$

This is what we mean by integrating the tail: the mean duration is equal to the integral of the survivor function.

Let's try this out for our two special cases. For Type I survival, where $S(t) = 1$ for $0 < t \leq \ell$ and $S(t) = 0$ otherwise, we have

$$E(T) = \int_0^\infty S(t)dt = \int_0^\ell 1dt + \int_\ell^\infty 0dt = \ell \tag{2.27}$$

(big surprise there – we assume everyone dies at age ℓ, and lo and behold, the average duration of survival exactly equals ℓ). For Type II survival, we get

$$E(T) = \int_0^\infty e^{-\mu t}dt = \frac{1}{\mu} \tag{2.28}$$

as claimed earlier.

Using the representation $T = \int_0^\infty \Phi(t)dt$ also enables the derivation of higher moments. For example, see if you can convince yourself that the second moment $E(T^2)$ is given by

$$E(T^2) = 2\int_0^\infty tS(t)dt, \tag{2.29}$$

a fact we will have use for below.

2.4. Disease Transmission/Progression and Duration Modeling

In considering infectious diseases, individuals pass through various phases or states. Each one of these phases can be thought of as a compartment subject to the fundamental equation of flow as individuals become infected, then infectious, then recover or die. Connecting up the various compartments leads to an epidemiological system of flow, while doing so mathematically leads to an epidemiological *model*.

Epidemiology texts and journals abound with disease-specific observations and descriptions of the relevant phases. A brief but good discussion can be found in Anderson and May's text, pp. 29-31 (Table 3.1 reports ranges for incubation, latent and infectious periods for several different infections), and also pp. 55-57. Here I wish simply to review some definitions and link them to duration variables.

The first duration variable is the disease-free lifespan, which I will denote by L (note that Anderson and May reserve L for the average time until death). All of the rules discussed in the previous section for the general duration variable T apply to L, that is, it has a survivor, density and hazard function representation, is finite, etc.

Let's consider a common, endemic (that is, stable prevalence), contagious infection in the population that makes people ill but doesn't kill them. Now, suspend reality and imagine living forever. Of interest is the time to infection, or equivalently the time spent susceptible. If one lived forever, one would become infected at some point. Let's denote the time to infection in this imaginary world by T_S for the time spent susceptible. Now, back to reality. From birth, either one becomes infected, or one dies for non-disease related reasons (so-called "natural" mortality). Thus, the actual time a person would remain susceptible in this circumstance would be the minimum of T_S and L. Either you die before becoming infected (in which case you are no longer susceptible to infection since you're dead), or you get infected (in which case you are no longer susceptible since you're infected). However, for many diseases (especially childhood diseases), $T_S << L$ so it is not so unreasonable to ignore natural mortality when computing time to infection (see Table 3.1 in Anderson and May). So, one can imagine a compartment of susceptibles in the population with arrival rates determined by births (and perhaps also by immigration) and departure rates related to the time to infection.

Upon infection, for most diseases, one enters a latent phase where, though infected, one is not infect*ious*. Having already grabbed the letter L to denote disease-free lifespan, denote the duration of the latent period by T_E (E for exposed). Similar definitions attend the time spent infectious T_I (or equivalently the time to recovery or other "removal" from the population), and the time spent recovered T_R.

One can imagine compartments for each of these different phases – time spent susceptible, latent, infectious and recovered – within the overall population. Linking the compartments together is the first step – this often involves nothing more than drawing a convincing diagram. Determining appropriate formulas for the various arrival and departure rates is the next step, and that can be more challenging. Only then can one begin to truly *analyze* the model created.

Anyway, the main point of import now is that infectious diseases can be represented as systems of flow with different compartments for different epidemiologically meaningful states, while the time spent in such states (e.g. susceptible, latent, infectious, recovered) can be described using duration modeling tools (such as survivor and hazard functions). But first, let's take a closer look at the one compartment model, for the insights derived will help in creating more complicated models for infectious diseases.

3. Stable Flow: The One Compartment Model

3.1. The Stable Population

In this section, we consider a model of stable flow through a one compartment system. For ease of discussion, we will refer to a steady-state (i.e. constant size and age-distributed) population of humans replenished by births but depleted by natural mortality. The overall arrival rate in this population is constant, that is, $\alpha(t) = \alpha$ independent of time (so the (expected) number of births over a period of arbitrary duration τ is given by $\alpha\tau$). For the population to remain constant, it must be that the aggregate death rate $\delta(t)$ is also constant and equal to the arrival rate, thus $\delta(t) = \delta = \alpha$.

The duration of life (equivalently age at death) in this population is described by the random variable L, which itself is characterized by the hazard rate (equivalently age-specific mortality rate) $\mu(a)$. We denote the density of persons in the population of age a by $N(a)$, which means that the actual *number* of persons with ages between a and $a + \Delta a$ is given by $N(a)\Delta a$. The total population size will simply be denoted by N and is found by tallying everyone alive at all ages, that is

$$N = \int_0^\infty N(a)da \tag{3.1}$$

One immediate equivalence is that the aggregate birth rate α must equal the population density evaluated at age 0, that is,

$$N(0) = \alpha. \tag{3.2}$$

Another is that the aggregate death rate δ must be consistent with random variable L, which after all is determined by the age-specific mortality rates $\mu(a)$. Thus, we must have

$$\delta = \int_0^\infty \mu(a)N(a)da = \alpha. \tag{3.3}$$

3.2. Age Distribution in the Stable Population

What is the distribution of the population by age? This turns out to be extremely easy to answer. The only people in the stable population now who are exactly aged a are those who arrived a time units ago and remain alive. Given that the arrival rate is constant, the number of persons who arrived between a and $a - \Delta t$ time units ago is equal to $\alpha \Delta a = N(0)\Delta a$, and of these, the fraction $S(a)$ survived until the present moment, and are thus aged a. The result couldn't be simpler:

$$N(a) = N(0)S(a). \tag{3.4}$$

Note that this immediately verifies equation (3.3), since

$$\int_0^\infty \mu(a)N(a)da = \int_0^\infty \mu(a)N(0)S(a)da = N(0) = \alpha \tag{3.5}$$

where we have used equation (2.15) to recognize that $\mu(a)S(a) = f(a)$, the probability density of age at death, and equation (2.6) which notes that probability densities integrate to unity.

3.3. Stable Population Size and Little's Theorem

Now that we know the age distribution of the population, we can find the total population size N by direct integration. We have

$$N = \int_0^\infty N(a)da = \int_0^\infty N(0)S(a)da = N(0)E(L). \tag{3.6}$$

This says something pretty important: the total population size equals the arrival rate multiplied by the expected duration of stay in the population. This is an instance of what is known as "Little's Theorem" from queueing theory, which states that the expected number of customers in a queueing system is equal to the product of the arrival rate and their average waiting time. The stable population can indeed be thought of as a queue: customers arrive (i.e. people are born) at rate $\alpha = N(0)$, and wait in the system (i.e. live!) on average for $E(L)$ units of time.

Now that we know the age distribution and the population size, we can divide to obtain the age-density of the population, that is, the fraction of the population of age a. Denoting this density by $g(a)$ we have

$$g(a) = \frac{N(a)}{N} = \frac{N(0)S(a)}{N(0)E(L)} = \frac{S(a)}{E(L)}. \tag{3.7}$$

This is also a special instance of a well-known formula from a branch of stochastic processes called "renewal theory." While the fraction of new arrivals to the population that will die at age a is given by $f(a) = -dS(a)/da$, the fraction of the stable population *currently* aged a is given by equation (3.7).

3.4. A Note Regarding the "Average" Hazard

Sometimes there is a need to produce an average death rate per person in the population. The obvious figure one should employ is simply the number of deaths per person per unit time. Equations (3.3) and (3.5) remind us that the number of deaths in total per unit time is simply $\delta = N(0)$, while equation (3.6) provides the total population size, so the number of deaths per person per unit time, that is, the per capita death rate, is given by

$$\frac{\delta}{N} = \frac{N(0)}{N(0)E(L)} = \frac{1}{E(L)}. \tag{3.8}$$

This generalizes the result for Type II survival (constant mortality rate independent of age) as described in the discussion following equation (2.20).

Another way to see this is to consider a population-weighted average of the age-specific mortality rates. Since $N(a)$ is the density in the population of age a, and thus subject to death at rate $\mu(a)$, it must be that overall average mortality is given by

$$\int_0^\infty \frac{N(a)}{N}\mu(a)da = \int_0^\infty \frac{S(a)}{E(L)}\mu(a)da = \frac{1}{E(L)} \tag{3.9}$$

where we have again recognized that $\mu(a)S(a) = f(a)$, and $\int_0^\infty f(a) = 1$.

3.5. Mean Age in the Stable Population

The mean age at death equals $E(L)$, but what is the mean age in the population? Given that the age density in the stable population is given by (3.7), applying first principles yields

$$\int_0^\infty ag(a)da = \int_0^\infty a\frac{S(a)}{E(L)}da = \frac{E(L^2)}{2E(L)} \tag{3.10}$$

as follows from equation (2.29).

3.6. Special Cases

We again consider the special cases of Type I and Type II survival. For Type I survival, the age-density $g(a)$ in the stable population is *uniform* between 0 and ℓ, that is,

$$g(a) = \begin{cases} \frac{1}{\ell} & 0 < a \leq \ell \\ \\ 0 & \text{all other } a \end{cases} \tag{3.11}$$

from which the mean age in the population is easily seen to equal $\ell/2$ (this also follows directly from equation (3.10) upon recognizing that $E(L) = \ell$). With Type II survival, recalling that $E(L) = 1/\mu$ and $S(a) = e^{-\mu a}$, we obtain

$$g(a) = \frac{S(a)}{E(L)} = \frac{e^{-\mu a}}{1/\mu} = \mu e^{-\mu a} = f(a) \text{ (!!)} \tag{3.12}$$

This is interesting, for it says that the fraction of the stable population at age a is exactly equal to the fraction of new arrivals who will expire at age a. In particular, the average age in the population is, somewhat oddly, equal to the average age at death averaged over new arrivals. This means that while a new arrival can expect to die at age $1/\mu$, the average age of those *currently alive* in the stable population equals this same value of $1/\mu$. This is another well-known implication of the "memoryless property" of the exponential distribution. It forgets – so not only is the average age in the population equal to $1/\mu$ – a living person in the population selected at random would, without regard to his or her current age, also expect to live on average another $1/\mu$ time units.

4. Disease Transmission and Progression in the Stable Population

We are now ready to contemplate disease transmission and progression in a stable population. The material to be covered leads to results that are essentially equivalent to what Anderson and May present in pp. 66-75 of their text, but instead of developing everything with (partial) differential equations, our approach will build on the methods discussed thus far. There are notational differences as shown in the table below, but hopefully after reading both sets of notes, you will have a much better foundation for modeling infectious diseases.

Some Differences In Notation

Function	Kaplan Notes	Anderson and May
Survivor Function	$S(a)$	$\ell(a)$
Arrival Rate	$\alpha = \delta = N(0)$	$B = N(0)$
Life Expectancy	$E(L) = \ell$	L

4.1. Three Compartments: Susceptible, Infected, Recovered

As in Anderson and May, we will assume a simple three stage model whereby arrivals to the population are uninfected but susceptible, that those susceptibles who become infected also immediately become infectious, and that those infectious either die of other causes while still infectious, or recover from infection – and *then* die of something else (you just can't win in this game).

4.2. Age Distribution of Number Susceptible, Infected and Recovered

We begin rather generally by assuming that susceptibles arrive to the population with rate α, that the duration of life is described by the non-negative random variable L which can have any distribution we wish, and that in the absence of natural deaths, new arrivals would remain susceptible for T_S time units, followed by an infectious period of duration T_I. Furthermore, we assume that the random variables L, T_S and T_I are mutually independent. Upon recovery from infection, persons simply go on with their lives until they die.

Denote the density of persons in the population of current age a who are susceptible, infectious, or recovered by $X(a)$, $Y(a)$, and $Z(a)$. We can immediately write down formulas for these variables via direct probabilistic reasoning. We will do so, and then we will investigate their consequences.

Let's start with susceptibles of age a. To be susceptible and aged a, one must have arrived a time units ago, still be alive (so $L > a$), and still be susceptible (so $T_S > a$). We therefore immediately see that

$$X(a) = N(0) \Pr\{L > a\} \Pr\{T_S > a\} \tag{4.1}$$

where the multiplication of the probabilities follows from the independence of L and T_S.

Note that we have not made any assumptions about the nature of the infection process other than, whatever the age-specific infection rates are among susceptibles, the resulting time to infection (absent death from other causes) is given by the random variable T_S. To see how the infection process and T_S are linked, let $\lambda(a)$ denote the age-specific *incidence* of infection among susceptibles of the same age. Then $\lambda(a)$ is the hazard function associated with the duration T_S, which means that

$$\Pr\{T_S > a\} = e^{-\int_0^a \lambda(u)du} \tag{4.2}$$

as shown in equation (2.14). For example, if $\lambda(a) = \lambda$, a constant independent of age, then the time spent susceptible is exponentially distributed.

Now focus on the density of infectious persons of age a. To be alive and infectious at age a means that one must first be alive (so $L > a$), have become infected (so $T_S \le a$), but not to have recovered yet (so $T_S + T_I > a$). This implies that

$$Y(a) = N(0)\Pr\{L > a\}\Pr\{T_S \le a < T_S + T_I\}. \tag{4.3}$$

Again, this is a general result for the stable population.

The density of recovered individuals follows from similar reasoning. One must be alive ($L > a$) and have been infected but recovered ($T_S + T_I \le a$) in order to be both alive and recovered, whence

$$Z(a) = N(0)\Pr\{L > a\}\Pr\{T_S + T_I \le a\}. \tag{4.4}$$

The total population density at age a, $N(a)$, is simply the sum $X(a)+Y(a)+Z(a)$. Noting that

$$\Pr\{T_S > a\} + \Pr\{T_S \le a < T_S + T_I\} + \Pr\{T_S + T_I \le a\} = 1 \tag{4.5}$$

(why?), we see that as in the one compartment model,

$$N(a) = N(0)\Pr\{L > a\}. \tag{4.6}$$

4.3. Applications of Little's Theorem

4.3.1. The Total Number of Susceptibles

Recall from equation (4.1) that $X(a) = N(0)\Pr\{L > a\}\Pr\{T_S > a\}$. The product $\Pr\{L > a\}\Pr\{T_S > a\}$ is simply the survivor function for the *minimum* of the two random variables L and T_S. That is, if $M = \min(L, T_S)$, then $M > a$ if and only if both $L > a$ and $T_S > a$. We therefore see that the total number of susceptibles in the stable population is given by

$$X = \int_0^\infty X(a)da = N(0)E[\min(L, T_S)]. \tag{4.7}$$

This is nothing more than an application of Little's Theorem. The arrival rate is $N(0)$, while the expected duration of time spent susceptible (now including time spent susceptible by those who die prior to infection) is exactly $E[\min(L, T_S)]$.

4.3.2. Prevalence = Incidence × Duration

The total number of new infections per unit time in the stable population, that is, the *aggregate* incidence rate, can be represented in two general ways. First, noting that one becomes infected if $T_S \leq L$ (the time susceptible is less than the total lifetime), we immediately have

$$\text{Aggregate Incidence} = N(0)\Pr\{T_S \leq L\}. \tag{4.8}$$

The second approach is to note that those susceptibles of age a become infected with rate $\lambda(a)$, whence

$$\begin{aligned} \text{Aggregate Incidence} &= \int_0^\infty \lambda(a)X(a)da \\ &= \int_0^\infty \lambda(a)N(0)\Pr\{L > a\}\Pr\{T_S > a\}da \\ &= N(0)\Pr\{T_S \leq L\} \end{aligned} \tag{4.9}$$

with the last equality following by recognizing $\lambda(a)\Pr\{T_S > a\}$ as the probability density for the duration T_S.

Now, once infected, what is the expected time spent infected? Again one has to worry about disease-free mortality removing people from the population prior to recovery. Given that a person has just become infected, that is, that $T_S \leq L$, the remaining lifespan is simply $L - T_S$. The mortality-free duration of infection is by definition equal to T_I. The actual time spent infected will be the minimum of these, so given that infection has occurred, the expected time spent infected is equal to $E[\min(L - T_S, T_I)|T_S \leq L]$. From Little's Theorem, we find that the total number of infected persons in the stable population is given by

$$\underbrace{Y}_{\text{Prevalence}} = \int_0^\infty Y(a)da = \underbrace{N(0)\Pr\{T_S \leq L\}}_{\text{Incidence}} \times \underbrace{E[\min(L - T_S, T_I)|T_S \leq L]}_{\text{Duration}}. \tag{4.10}$$

This celebrated "prevalence = incidence x duration" result is true in the stable population for an endemic infection. Note that if we divide both sides of equation (4.10) by the population size N we obtain the per-capitized version of this result. That is, defining $\pi = Y/N$, $\iota = N(0)\Pr\{T_S \leq L\}/N$ and $d = E[\min(L - T_S, T_I)|T_S \leq L]$, we have

$$\pi = \iota \times d. \tag{4.11}$$

However, this formula is frequently applied incorrectly, owing to the common use of the word "incidence" to refer to the number of infections per *uninfected* person per unit time (which is often very close to, but not exactly the same as, the per-capita rate of infection ι).

4.3.3. The Number of Recovered Individuals

Finally, we come to the number of recovered individuals, $Z = \int_0^\infty Z(a)da$. We can again use Little's Theorem, note that the total number of new recoveries per unit time in the population must equal $N(0)\Pr\{T_S + T_I \leq L\}$, and also argue that the expected time spent in recovery is exactly the remaining lifetime $L - T_S - T_I$, given that the person has lived to both become infected and recover. This logic yields

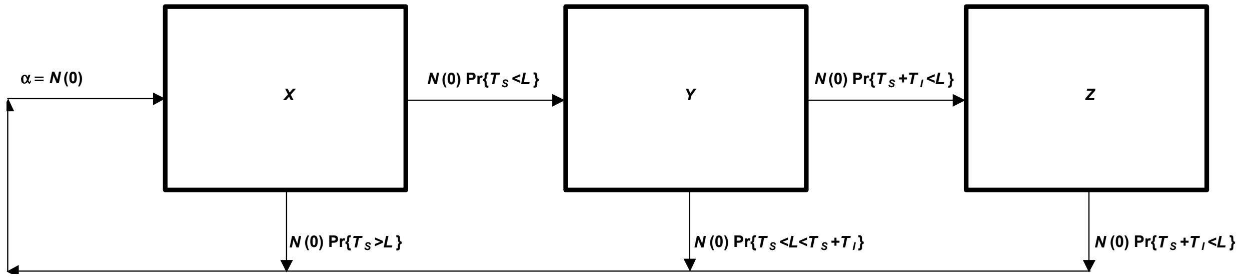
$$Z = N(0)\Pr\{T_S + T_I \leq L\} \times E(L - T_S - T_I|T_S + T_I \leq L). \tag{4.12}$$

But of course, there is an easier way out – since the population is stable, we must have the conservation $N = X + Y + Z$ (since everyone in the population is either susceptible, infected or recovered), and thus we can simply set $Z = N - X - Y$.

The equations for the total population quantities were derived directly and sometimes with amazing ease, but they are also very general and sometimes a bit clumsy. We can simplify matters considerably with a very useful approximation, but first, let's review the epidemiological system of flow we have spent so much time developing.

4.4. The Epidemiological System of Flow in the Stable Population

The age-aggregated epidemiological system of flow for the stable population is shown in the figure below. Note that for each of the three compartments, the total flow in always equals the total flow out. For example, the total flow out of the susceptible compartment (X) splits the arrival rate $\alpha = N(0)$ in two: those who become infected (with probability $\Pr\{T_S < L\}$) and those who die before becoming infected (with probability $\Pr\{T_S \geq L\}$). Similarly, the infected compartment (Y) splits the entering flow ($N(0)\Pr\{T_S < L\}$) into those who recover and those who die before recovering in a manner that conserves the total flow (for $\Pr\{T_S < L\} = \Pr\{T_S < L \leq T_S + T_I\} + \Pr\{T_S + T_I < L\}$). And, as assumed throughout, the total departures from the population exactly balance the total arrivals; convince yourself that the sum of the departure rates from each compartment exactly equals the arrival rate $N(0)$.



4.5. Useful Approximation: $T_S + T_I << L$

The model described could be used to model those non-lethal infections encountered in childhood, adolescence, or even early adulthood that virtually everyone experiences, and indeed if everyone eventually gets infected (just not at the same time!), it seems reasonable to presume that $T_S + T_I << L$, which means that $\Pr\{L > T_S\}$ and $\Pr\{L > T_S + T_I\}$ are both well approximated by 1. This simplifies life considerably! First, everyone gets infected, thus the aggregate incidence rate equals the arrival rate $N(0)$. The time spent in the susceptible state now becomes T_S (with mean $E(T_S)$) instead of $\min(L, T_S)$, and the age-specific density of susceptibles $X(a)$ becomes simply $N(0)\Pr\{T_S > a\}$. Similarly, the time spent infectious now becomes T_I (with mean $E(T_I)$ instead of the more general $E[\min(L - T_S, T_I)|T_S \leq L]$). Prevalence = Incidence x Duration is especially simple:

$$\underbrace{Y}_{\text{Prevalence}} = \underbrace{N(0)}_{\text{Incidence}} \times \underbrace{E(T_I)}_{\text{Duration}} \quad (\text{when } T_S + T_I << L) \tag{4.13}$$

Meanwhile, the time spent in the recovered state is just $L - T_S - T_I$ under this approximation. We will have many uses for this approximation, for it will often be the case that infections move very quickly when contrasted against lifetimes not afflicted by disease.

4.6. Mean Age at Infection

What is the age at infection? That is, of all those who become infected, what is the time at which they do so? One might expect that the answer is given by T_S, the time spent susceptible. In general, however, one needs to worry about non-disease mortality, which as usual clutters up the formulas. Let $\eta(a)$ denote the probability density for age at infection. Since the rate with which persons of age a become infected is given by $\lambda(a)X(a)$, while the total infection rate is given by $\int_0^\infty \lambda(a)X(a) = N(0)\Pr\{T_S < L\}$, the age density among those just infected is given by

$$\eta(a) = \frac{\lambda(a)X(a)}{N(0)\Pr\{T_S < L\}} \tag{4.14}$$

and thus the mean age at infection is given definitionally by

$$\text{Mean Age at Infection} = \int_0^\infty a\eta(a)da. \tag{4.15}$$

Again, considerable insight (and utility!) is gained by making use of the useful approximation $T_S + T_I << L$, for then $\eta(a)$ simplifies to $\lambda(a)\Pr\{T_S > a\}$ and as desired, we have

$$\text{Mean Age at Infection} = \int_0^\infty a\lambda(a)\Pr\{T_S > a\}da = E(T_S) \text{ (when } T_S + T_I << L). \tag{4.16}$$

4.7. The Anderson and May Stable Population Model (pp. 66-75)

To reproduce the Anderson and May stable population model requires recognizing the specific assumptions they have made. Here they are: disease incidence is age-independent (that is, the hazard rate $\lambda(a) = \lambda$, and hence T_S follows an exponential distribution with mean $1/\lambda$, that is, $\Pr\{T_S > a\} = e^{-\lambda a}$), while infected persons recover with constant, age-independent hazard rate υ (and thus T_I is exponentially distributed with mean $1/\upsilon$). Results are presented for Type I and Type II survival (constant lifetime, and constant hazard respectively). See if you can reproduce some of their results using the equations above (and noting the notational differences stated earlier).

4.8. The Reproductive Number R_0 and "Weak" Homogeneous Mixing

The reproductive number R_0 figures prominently in infectious disease epidemiology. It is essentially the number of infections a single infected person could transmit in a fully susceptible population. In the stable population, of course, it is not the case that 100% of all alive are susceptible. Rather, only the fraction X/N are susceptible. Thus it seems a stretch to imagine that an infected person could transmit R_0 infections in the stable population.

In fact, assuming that new infections are coming from those currently infectious, for the population to remain stable, it must be that the populations of each of the compartments are also stable (and they are as we have shown – we have formulas for X, Y and Z after all). This means that in the stable population, each infected person must transmit not R_0 infections on average, but rather 1 infection on average. Each infected person must infect one other to replenish him or herself.

We now introduce the assumption of "weak" homogeneous mixing, which simply says that the rate of new infections in the population is proportional to the fraction of the population that is susceptible. The implication of this assumption from an infectious person's point of view should be clear: if everyone is susceptible, then on average the infected person is able to generate R_0 infections, so it must be that R_0 is the proportionality constant defining weak mixing. That is, if each infected person would generate R_0 infections were all susceptible, but only generates a single infection when X out of N persons in the population are susceptible, R_0 must satisfy the equation

$$R_0 \times \frac{X}{N} = 1. \tag{4.17}$$

Since $N = N(0)E(L)$ (Little's Theorem!) and $X = N(0)E[\min(L, T_S)]$ (uh...Little's Theorem!), we have the following result:

$$R_0 = \frac{N}{X} = \frac{E(L)}{E[\min(L, T_S)]} \approx \frac{E(L)}{E(T_S)} \text{ (when } T_S + T_I << L). \tag{4.18}$$

This is pretty interesting – it says that we can find R_0 by looking at the ratio of the time spent in the population (which is $E(L)$) to the time spent susceptible (which is $E[\min(L, T_S)]$), and under our useful approximation, we get an even cleaner estimate of R_0 as the ratio of life expectancy to mean age at infection.

Note that if you substitute in the assumption of age-independent incidence (T_S is exponentially distributed with mean $1/\lambda$) and either Type I or Type II survival, you can replicate Anderson and May's results (equations 4.20 and 4.25). Starting with Type I survival, we have $E(L) = \ell$ and $E[\min(L, T_S)] = \int_0^\ell e^{-\lambda a} da = \frac{1-e^{-\lambda \ell}}{\lambda}$, whence

$$R_0 = \frac{E(L)}{E[\min(L, T_S)]} = \frac{\ell}{\frac{1-e^{-\lambda\ell}}{\lambda}} = \frac{\lambda\ell}{1-e^{-\lambda\ell}} \approx \lambda\ell. \tag{4.19}$$

For Type II survival, life is even simpler: $E(L) = 1/\mu$, and $E[\min(L, T_S)] = \int_0^\infty e^{-\mu a} e^{-\lambda a} da = \frac{1}{\mu+\lambda}$, thus for Type II survival we have

$$R_0 = \frac{E(L)}{E[\min(L, T_S)]} = \frac{1/\mu}{1/(\mu+\lambda)} = 1 + \frac{\lambda}{\mu} \approx \frac{\lambda}{\mu}\cdot \tag{4.20}$$

4.9. Strong Homogeneous ("Free") Mixing, the Transmission Rate β, and R_0

We have one remaining task to attempt, and that is to make the transmission of infection truly endogenous. Thus far, we have made no specific assumptions regarding *how* susceptibles in the stable population become infected. Rather, we have only assumed that they are infected at some rate. The assumption of strong homogeneous (or free) mixing is one way to close the loop. We do so by assuming that the hazard for infection experienced by susceptible individuals is proportional to the total number of infected persons in the population, that is,

$$\lambda = \beta \int_0^\infty Y(a) da = \beta Y \tag{4.21}$$

where β is referred to as the transmission rate.

Using this expression for λ enables explicit formulae for the reproductive number R_0 in terms of more fundamental parameters. Recall equation (4.10) for Y, repeated below:

$$\underbrace{Y}_{\text{Prevalence}} = \int_0^\infty Y(a) da = \underbrace{N(0) \Pr\{T_S \le L\}}_{\text{Incidence}} \times \underbrace{E[\min(L - T_S, T_I) | T_S \le L]}_{\text{Duration}}. \tag{4.22}$$

As in Anderson and May, assume that the time to recovery T_I is exponentially distributed with mean $1/\upsilon$. Applying the equation above to the case of Type II survival, we immediately obtain

$$\begin{aligned} Y &= N(0) \times \frac{\lambda}{\lambda+\mu} \times \frac{1}{\upsilon+\mu} \\ &= \frac{N\mu\lambda}{(\lambda+\mu)(\upsilon+\mu)} \end{aligned} \tag{4.23}$$

where we have used the flow balance $N(0) = \mu N$. Substitution into equation (4.21) yields

$$\lambda = \beta Y = \beta N\mu \times \frac{\lambda}{\lambda + \mu} \times \frac{1}{\upsilon + \mu} \tag{4.24}$$

which after cancelling the λ's and rearranging can be written as

$$\lambda = \frac{\beta N\mu}{\upsilon + \mu} - \mu. \tag{4.25}$$

Substitution in equation (4.20) finally yields

$$R_0 = 1 + \frac{\lambda}{\mu} = \frac{\beta N}{\upsilon + \mu} \approx \frac{\beta N}{\upsilon} \tag{4.26}$$

which is a well-known formula for R_0 under free mixing.

For Type I survival, the general results are messier, but if we accept the approximation $T_S + T_I << L$, then things simplify immediately since under the approximation,

$$Y \approx N(0) \times E(T_I) = \frac{N}{\ell\upsilon} \tag{4.27}$$

where we have used Little's Theorem in the form $N = N(0)\ell$ to express $N(0)$ as N/ℓ. From this we obtain

$$\lambda = \beta Y \approx \frac{\beta N}{\ell\upsilon}\cdot \tag{4.28}$$

Placing this result into equation (4.19) yields

$$R_0 \approx \lambda\ell \approx \frac{\beta N}{\upsilon} \tag{4.29}$$

in agreement with equation (4.26). Even though these results were derived for the stable population, it turns out they will apply to transient epidemics as well. This concludes our tour of the epidemiological system of flow in the stable population.