This replaces section 2.3, The pitfalls of overlooking dependence. The notation is different from the book, I follow closely the notation used in the introductory lectures.

The sick nurse and the naive manager

A nurse in a hospital takes care of N patients. The patiens have private rooms so that they do not contact each other, whereas the nurse contacts each one of them at a rate γ. The nurse gets sick, but goes to work nevertheless. In each contact she makes while she is sick, she infects the patient with probability $\hat{p}$; hence she makes an infectious contact to each patient at a rate $\beta = \hat{p}\gamma$. Upon an infectious contact, the patient gets the infection if not already infected. The infection has a constant recovery rate ν. We are thus making the same assumptions about transmission and recovery as in the SIR model, except that there is only a finite number of patients (susceptibles to the nurse's infection) and they do not infect each other[1]. The question is, what is the probability $p(i)$ that the nurse infects i patients ($i = 0, 1, ..., N$) before she recovers.

The manager of the hospital department has attended some introductory lectures on the dynamics of infectious diseases, and never a better time to use what he has learned: he sets out to calculate $p(i)$ (hoping, of course, that $p(0)$ will turn out to be high, so that the nurse likely harms no-one). He focuses on one patient at a time. The nurse gives infectious contacts to the focal patient at a rate β, and she recovers at a rate ν. These are two 'competing' exponential decay processes. If the nurse gives an infectious contact to the focal patient before she recovers, then the patient gets sick; this happens with probability $\beta/(\beta + \nu)$. If the nurse recovers before infecting the focal patient, then the patient has escaped; this happens with probability $\nu/(\beta+\nu)$. The same goes of course for each patient.

So far, so good. Now digging out a book on probabilities, the probability that no patient is infected should be $p(0) = [\nu/(\beta + \nu)]^N$; and the probability that i patients are infected should be (good old binomial distribution!)

$$p(i) = \binom{N}{i} \left(\frac{\beta}{\beta + \nu}\right)^i \left(\frac{\nu}{\beta + \nu}\right)^{N-i} \tag{1}$$

But this is WRONG.

[1]This is the difference from the standard models of epidemics in finite populations, e.g. exercise 1.34 of the book.

What the manager has overlooked is that the fates of these patients are *not* independent of each other. If the nurse is lucky to recover soon (i.e., the length of her infection, T, happens to be short), then she likely does not infect anyone. Re-wording this from purely the patients' point of view, if the first patient escaped infection, then there is some chance that this was *because* the nurse recovered quickly and had no time to infect the first patient. This is good news to all other patients; they all enjoy a high probability of escaping the infection since they all have the same nurse with her quick recovery.

Overlooking dependence (through the shared nurse) is a problem because the formula $p(0) = [\nu/(\beta+\nu)]^N$ assumes independence between the N patients (and so does the general binomial formula for $p(i)$ in (1)). In our case, the first patient has indeed probability $\nu/(\beta+\nu)$ to escape the infection; but with dependence, the probability for the second (and third, etc.) patient to escape is higher than this, because the escape of the first patient may signal a fast recovery of the nurse.

In what follows, I first calculate the correct value of $p(i)$ and then show what this problem has to do with Jensen's inequality and taking averages properly (hence the model is in the chapter "The art of averaging").

To handle the problem of dependence, we do not consider patients one by one, but all N patients together. When all N are still suscpetible, then the nurse makes a new infection at a rate βN and recovers at a rate ν; hence $\beta N/(\beta N+\nu)$ is the probability that the first event is an infection. When the first infection is made, then there are only $N-1$ susceptibles left, so that the probability that the second event is also an infection is $\beta(N-1)/(\beta(N-1)+\nu)$. And so on; the probability that the first i events are infections and then recovery happens is

$$p(i) = \frac{\beta N}{\beta N+\nu}\cdot\frac{\beta(N-1)}{\beta(N-1)+\nu}\cdot\ldots\cdot\frac{\beta(N-i+1)}{\beta(N-i+1)+\nu}\cdot\frac{\nu}{\beta(N-i)+\nu} \qquad (2)$$

Computing $p(i)$ from the above formula is somewhat cumbersome, so we take only an example for illustration: the special case of $\beta=\nu$. There is of course no biological reason whatsoever why the two rates should equal, but this special assumption greatly simplifies the math. Now the rates simply cancel in (2), yielding

$$\begin{aligned} p(i) &= \frac{N}{N+1}\cdot\frac{N-1}{N}\cdot\frac{N-2}{N-1}\cdot\ldots\cdot\frac{N-i+1}{N-i+2}\cdot\frac{1}{N-i+1} = \\ &= \frac{1}{N+1} \end{aligned} \qquad (3)$$

It may look surprising that $p(0), p(1), \ldots$ are all equal; recall, however, that this is true only for the specific choice of parameters $\beta = \nu$. The point is that this result is obviously different from the naive manager's binomial distribution in the wrong equation (1).

To see the above as a problem of averaging, suppose that the length of the nurse's infection is known and equals T. When T is fixed, then there is no dependence beween the patients; if the first patient escapes infection, this does not signal anything to the fate of the others. With fixed T, we do get a binomial distribution

$$p_c(i; T) = \binom{N}{i} q^i (1-q)^{N-i} \tag{4}$$

where q is the probability that the nurse infects a patient; since infecting one patient is an exponential decay process during T, the probability of *not* infecting by the time the nurse recovers is $1 - q = e^{-\beta T}$, and therefore

$$q = 1 - e^{-\beta T} \tag{5}$$

The index c in the notation $p_c(i; T)$ reminds us that this probability is conditional on the fact that the length of the nurse's infection is T.

In reality, we do not know how long it will take for the nurse to recover (T), so that to compute $p(i)$, we need to average the probabilities in (4) over the distribution of T. In other words, now we imagine infinitely many realizations (nurses and hospitals). $p(i)$ is the probability that in a randomly chosen realization there are i patients infected. (The randomly chosen realization is the nurse and hospital of real life, but the same story could play out in many other hospitals, of which our hospital is but one.) Since the recovery rate ν is constant, the probability that a random nurse (and therefore the real nurse) recovers in time T is exponentially distributed with probability density function

$$f(T) = \nu e^{-\nu T} \tag{6}$$

With probability $f(T)dT$, the recovery time is between T and $T + dt$, and in this case the probability of i patients getting infected is $p_c(i; T)$. Integrating for all possible values of T, we get

$$p(i) = \int_0^\infty p_c(i; T) f(T) dt = \int_0^\infty \binom{N}{i} (1 - e^{-\beta T})^i (e^{-\beta T})^{N-i} \nu e^{-\nu T} dt \tag{7}$$

This is a nasty integral. We thus take again the special case $\beta = \nu$ for illustration; with this, some patience, and the hint in exercise 2.15 in the book (giving a formula for a complicated sum that arises on the way), one arrives at $p(i) = 1/(N+1)$ as in (3).

What we did was to compute the binomial probability $p_c(i;T)$ conditional on T and then average over T. What the naive manager did (in effect) was to take the average of q over T, which (from (5) and (6)) is

$$\begin{aligned}
\int_0^\infty (1-e^{-\beta T}) f(T) dt &= \int_0^\infty (1-e^{-\beta T}) \nu e^{-\nu T} dt = \\
&= \nu \left[\int_0^\infty e^{-\nu T} dt - \int_0^\infty e^{-(\beta+\nu)T} dt \right] = \\
&= \nu \left[\frac{1}{\nu} - \frac{1}{\beta+\nu} \right] = \\
&= \frac{\beta}{\beta+\nu}
\end{aligned}$$

and then to insert this average q into the binomial formula to arrive at the wrong result in (1). Thus in the end, the problem boils down to Jensen's inequality; because the binomial formula is a nonlinear function of its probability q, it matters whether we evaluate the binomial formula at the average q (wrong) or evaluate the formula at each possible value of q (which here we get from considering each possible value of T in (5)) and average the results as in equation (7) (correct).

You would not believe how often this is done in the wrong way.