Application of Probability Generating Functions to Infectious disease modelling[1]

Joel C. Miller
La Trobe University and Institute for Disease Modeling

19 May 2019

[1] Largely based on *A primer on the use of probability generating functions in infectious disease modeling*, Infectious Disease Modelling 3 192–248, also Work in Progress: *Distribution of outbreak sizes for SIR disease in finite populations*

Summary

Infectious disease and probability distributions

In the setting of a stochastic epidemic, there are many contexts where we are interested in a probability distribution p_k on the non-negative integers.

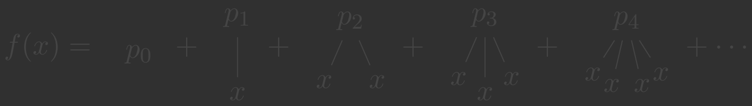
- p_k might represent the probability an infected individual will infect k individuals — the *offspring distribution*.
- p_k might represent the probability an outbreak infects exactly k individuals — the *final size distribution*.
- $p_k(g)$ or $p_k(t)$ might represent the probability of k infected individuals in generation g or time t — the *intermediate size distribution*

What is a Probability Generating Function (PGF)?

Consider a probability distribution on the non-negative integers, with p_k giving the probability of k.
Then

$$f(x) = \sum_{k=0}^{\infty} p_k x^k$$

is the *Probability Generating Function* for the distribution.
We can "visualize" the PGF as:

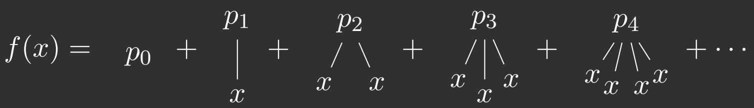


What is a Probability Generating Function (PGF)?

Consider a probability distribution on the non-negative integers, with p_k giving the probability of k.
Then

$$f(x) = \sum_{k=0}^{\infty} p_k x^k$$

is the *Probability Generating Function* for the distribution.
We can "visualize" the PGF as:



Examples

The PGFs of many common distributions have compact analytic forms.

- Poisson distribution: $p_k = \frac{e^{-\lambda}\lambda^k}{k!}$

$$f(x) = \sum_{k=0}^{\infty} e^{-\lambda}\frac{\lambda^k x^k}{k!} = e^{-\lambda}e^{\lambda x} = e^{\lambda(x-1)}$$

- Geometric distribution: $p_k = (1-r)r^k, \;\; 0 \leq r < 1$

$$f(x) = \sum_{k=0}^{\infty}(1-r)r^k x^k = \frac{1-r}{1-rx}$$

Examples

The PGFs of many common distributions have compact analytic forms.

- Poisson distribution: $p_k = \frac{e^{-\lambda}\lambda^k}{k!}$

$$f(x) = \sum_{k=0}^{\infty} e^{-\lambda}\frac{\lambda^k x^k}{k!} = e^{-\lambda}e^{\lambda x} = e^{\lambda(x-1)}$$

- Geometric distribution: $p_k = (1-r)r^k, \ 0 \leq r < 1$

$$f(x) = \sum_{k=0}^{\infty}(1-r)r^k x^k = \frac{1-r}{1-rx}$$

Examples

The PGFs of many common distributions have compact analytic forms.

- Poisson distribution: $p_k = \frac{e^{-\lambda}\lambda^k}{k!}$

$$f(x) = \sum_{k=0}^{\infty} e^{-\lambda}\frac{\lambda^k x^k}{k!} = e^{-\lambda}e^{\lambda x} = e^{\lambda(x-1)}$$

- Geometric distribution: $p_k = (1-r)r^k, \; 0 \leq r < 1$

$$f(x) = \sum_{k=0}^{\infty}(1-r)r^k x^k = \frac{1-r}{1-rx}$$

Properties

- $p_0 = \sum p_k 0^k = f(0)$. (important if interested in extinction of a disease).

$$p_0 = \quad p_0 \quad + \begin{array}{c} p_1 \\ | \\ 0 \end{array} + \begin{array}{c} p_2 \\ / \quad \backslash \\ 0 \quad\; 0 \end{array} + \begin{array}{c} p_3 \\ / \,|\, \backslash \\ 0 \;\; 0 \;\; 0 \end{array} + \begin{array}{c} p_4 \\ / / \backslash \backslash \\ 0 \; 0 \; 0 \; 0 \end{array} + \cdots$$

- $\sum k p_k = \sum k p_k 1^{k-1} = f'(1)$ is the expected value of the distribution (this could give $\mathcal{R}_0$).

$$\langle K \rangle = \quad 0p_0 \quad + \begin{array}{c} 1p_1 \\ | \\ 1 \end{array} + \begin{array}{c} 2p_2 \\ / \quad \backslash \\ 1 \quad\; 1 \end{array} + \begin{array}{c} 3p_3 \\ / \,|\, \backslash \\ 1 \;\; 1 \;\; 1 \end{array} + \begin{array}{c} 4p_4 \\ / / \backslash \backslash \\ 1 \; 1 \; 1 \; 1 \end{array} + \cdots$$

Properties

- $p_0 = \sum p_k 0^k = f(0)$. (important if interested in extinction of a disease).

$$p_0 = \quad p_0 \quad + \quad \begin{array}{c} p_1 \\ | \\ 0 \end{array} \quad + \quad \begin{array}{c} p_2 \\ / \quad \backslash \\ 0 \qquad 0 \end{array} \quad + \quad \begin{array}{c} p_3 \\ / | \backslash \\ 0 \; 0 \; 0 \end{array} \quad + \quad \begin{array}{c} p_4 \\ / / \backslash \backslash \\ 0 \; 0 \; 0 \; 0 \end{array} \quad + \cdots$$

- $\sum k p_k = \sum k p_k 1^{k-1} = f'(1)$ is the expected value of the distribution (this could give $\mathcal{R}_0$).

$$\langle K \rangle = \quad 0 p_0 \quad + \quad \begin{array}{c} 1p_1 \\ | \\ 1 \end{array} \quad + \quad \begin{array}{c} 2p_2 \\ / \quad \backslash \\ 1 \qquad 1 \end{array} \quad + \quad \begin{array}{c} 3p_3 \\ / | \backslash \\ 1 \; 1 \; 1 \end{array} \quad + \quad \begin{array}{c} 4p_4 \\ / / \backslash \backslash \\ 1 \; 1 \; 1 \; 1 \end{array} \quad + \cdots$$

Products/addition

The PGF for the sum of two distributions is the product of their PGFs.

	p_0	p_1 $\vert$ x	p_2 / \ x x	p_3 / $\vert$ \ x x x	$\cdots$
q_0	q_0p_0	q_0p_1 $\vert$ x	q_0p_2 / \ x x	q_0p_3 / $\vert$ \ x x x	$\cdots$
q_1 $\vert$ x	q_1p_0 $\vert$ x	q_1p_1 / \ x x	q_1p_2 / $\vert$ \ x x x	q_1p_3 / / \ \ x x x x	$\cdots$
q_2 / \ x x	q_2p_0 / \ x x	q_2p_1 / $\vert$ \ x x x	q_2p_2 / / \ \ x x x x	$\ddots$	
q_3 / $\vert$ \ x x x	q_3p_0 / $\vert$ \ x x x	q_3p_1 / / \ \ x x x x	$\ddots$	$\ddots$	
$\vdots$	$\vdots$	$\vdots$			

In particular, if we choose k numbers from a distribution with PGF $f(x)$, their sum has PGF $[f(x)]^k$ (the grandchild distribution, conditional on k).

Function composition

If we choose k from a distribution with PGF $\xi(x) = \sum p_k x^k$ and then choose k numbers $a_1, \ldots, a_k$ from another distribution with PGF $h(x) = \sum q_a x^a$. Then the sum $a_1 + \cdots + a_k$ has PGF

$$f(x) = \quad p_0 \quad + \quad \begin{array}{c} p_1 \\ | \\ h(x) \end{array} \quad + \quad \begin{array}{c} p_2 \\ / \quad \backslash \\ h(x) \quad h(x) \end{array} \quad + \quad \begin{array}{c} p_3 \\ / | \backslash \\ h(x) \; h(x) \; h(x) \end{array} \quad + \cdots$$

$$= \xi(h(x))$$

[If the "offspring distribution" has PGF $\mu(x)$, the number after g generations has PGF $\mu^{[g]}(x) = \mu(\mu(\cdots \mu(x) \cdots))$.]

Function composition corresponds to looking at later generations.

More Properties

- If we know the analytic form of $\mu(x)$, we can numerically calculate $f(z) = \mu^{[g]}(z)$ for any value of z without knowing $f(z)$'s expansion *a priori*.
- Then we can find the coefficients of f by

$$p_k = \frac{1}{2\pi i}\oint_{|z|=1} \frac{f(z)}{z^{k+1}}\,\mathrm{d}z = \frac{1}{2\pi}\int_0^{2\pi} f\left(e^{2\pi i\theta}\right) e^{-ik\theta}\,\mathrm{d}\theta$$
$$\approx \frac{1}{M}\sum_{m=1}^{M} f\left(e^{2\pi i m/M}\right) e^{-2k\pi i m/M}$$

- We just evaluate f at the M locations once, and then we can do the summation for many values of k. It is very accurate if $M \gg k$, and reasonably accurate up to $k = M-1$.

More Properties

- If we know the analytic form of $\mu(x)$, we can numerically calculate $f(z) = \mu^{[g]}(z)$ for any value of z without knowing $f(z)$'s expansion *a priori*.
- Then we can find the coefficients of f by

$$p_k = \frac{1}{2\pi i}\oint_{|z|=1} \frac{f(z)}{z^{k+1}}\,\mathrm{d}z = \frac{1}{2\pi}\int_0^{2\pi} f\left(e^{2\pi i\theta}\right) e^{-ik\theta}\,\mathrm{d}\theta$$
$$\approx \frac{1}{M}\sum_{m=1}^{M} f\left(e^{2\pi i m/M}\right) e^{-2k\pi i m/M}$$

- We just evaluate f at the M locations once, and then we can do the summation for many values of k. It is very accurate if $M \gg k$, and reasonably accurate up to $k = M - 1$.

More Properties

- If we know the analytic form of $\mu(x)$, we can numerically calculate $f(z) = \mu^{[g]}(z)$ for any value of z without knowing $f(z)$'s expansion *a priori*.
- Then we can find the coefficients of f by

$$p_k = \frac{1}{2\pi i}\oint_{|z|=1} \frac{f(z)}{z^{k+1}}\,\mathrm{d}z = \frac{1}{2\pi}\int_0^{2\pi} f\left(e^{2\pi i\theta}\right) e^{-ik\theta}\,\mathrm{d}\theta$$

$$\approx \frac{1}{M}\sum_{m=1}^{M} f\left(e^{2\pi i m/M}\right) e^{-2k\pi i m/M}$$

- We just evaluate f at the M locations once, and then we can do the summation for many values of k. It is very accurate if $M \gg k$, and reasonably accurate up to $k = M - 1$.

Applications

PGFs have many applications:

- Calculating the probability an outbreak goes extinct in a large population.
- Determining the size distribution of an outbreak after a short time period.
- Predicting the final size distribution in a population (even if not small).
- Deterministic SIR dynamics in random networks.
- Deterministic SIR dynamics in well-mixed populations.
- and more!

Basic Disease assumptions

We assume:

- A single introduced infection.
- Discrete time: the offspring distribution has PGF $\mu(x)$.

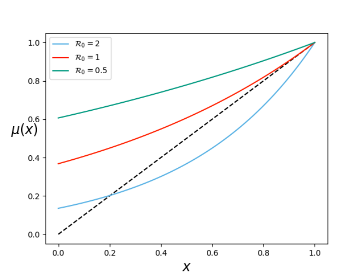
We can quickly conclude that

- $\mathcal{R}_0 = \mu'(1)$.
- The PGF for the number of infections in generation g is $\mu^{[g]}(x)$.

Extinction

Given that the offspring PGF is $\mu(y)$:

- It is reasonably well-known that the extinction probability α solves $y = \mu(y)$.
- $1 = \mu(1)$ is always a solution, but when $\mathcal{R}_0 > 1$, there is another solution in $[0, 1)$.
- We can find the other solution through iteration.

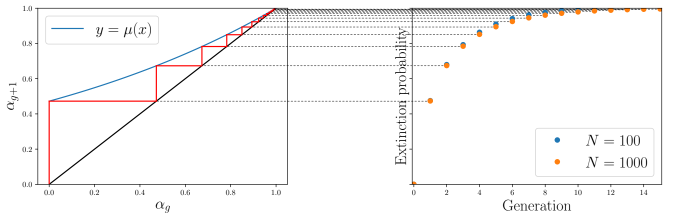


Extinction and Cobweb diagrams

- $\mu^{[g]}(y) = \mu(\mu(\cdots\mu(y)\cdots))$ is the PGF for the number of infections in generation g.
- Then $\alpha_g = \mu^{[g]}(0)$ is the probability the outbreak is extinct by generation g.
- How does this prediction compare with stochastic simulation?

Extinction and Cobweb diagrams

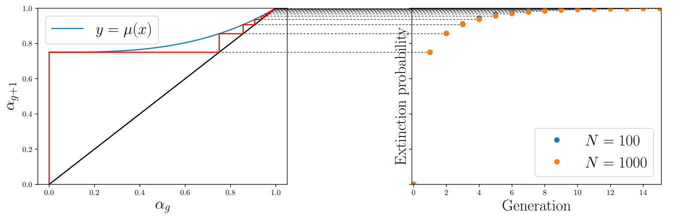
- $\mu^{[g]}(y) = \mu(\mu(\cdots \mu(y)\cdots))$ is the PGF for the number of infections in generation g.
- Then $\alpha_g = \mu^{[g]}(0)$ is the probability the outbreak is extinct by generation g.
- How does this prediction compare with stochastic simulation?



Poisson distribution, $\mu(y) = e^{0.75(y-1)}$, $\mathcal{R}_0 = 0.75$.

Extinction and Cobweb diagrams

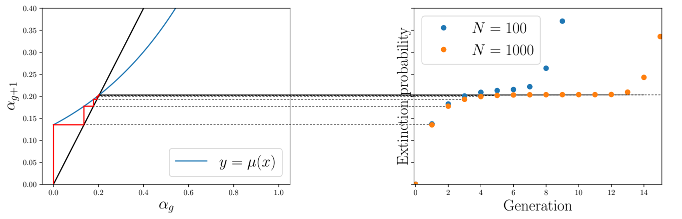
- $\mu^{[g]}(y) = \mu(\mu(\cdots\mu(y)\cdots))$ is the PGF for the number of infections in generation g.
- Then $\alpha_g = \mu^{[g]}(0)$ is the probability the outbreak is extinct by generation g.
- How does this prediction compare with stochastic simulation?



Bimodal distribution, $\mu(y) = (3+y^3)/4$, $\mathcal{R}_0 = 0.75$.

Extinction and Cobweb diagrams

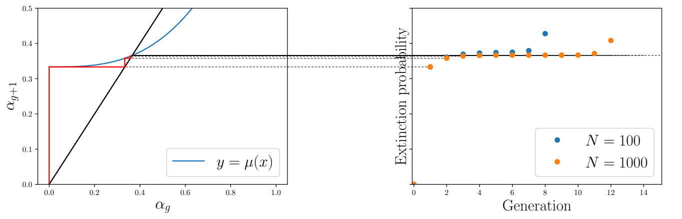
- $\mu^{[g]}(y) = \mu(\mu(\cdots\mu(y)\cdots))$ is the PGF for the number of infections in generation g.
- Then $\alpha_g = \mu^{[g]}(0)$ is the probability the outbreak is extinct by generation g.
- How does this prediction compare with stochastic simulation?



Poisson distribution, $\mu(y) = e^{2(y-1)}$, $\mathcal{R}_0 = 2$.

Extinction and Cobweb diagrams

- $\mu^{[g]}(y) = \mu(\mu(\cdots\mu(y)\cdots))$ is the PGF for the number of infections in generation g.
- Then $\alpha_g = \mu^{[g]}(0)$ is the probability the outbreak is extinct by generation g.
- How does this prediction compare with stochastic simulation?



Bimodal distribution, $\mu(y) = (1+2y^3)/3$, $\mathcal{R}_0 = 2$.

Intermediate Size

Given a single introduced infection in an infinite population and an offspring distribution with PGF $\mu(y) = \sum p_k y^k$:

- The number of active infections in generation $g+1$ has PGF $\Phi_{g+1}(y) = \mu(\Phi_g(y))$, with $\Phi_0(y) = y \;\Rightarrow\; \Phi_g(y) = \mu^{[g]}(y)$.

$$\Phi_{g+1}(y) = p_0 + \begin{array}{c} p_1 \\ | \\ \Phi_g(y) \end{array} + \begin{array}{c} p_2 \\ / \ \backslash \\ \Phi_g(y)\,\Phi_g(y) \end{array} + \begin{array}{c} p_3 \\ / | \backslash \\ \Phi_g(y) \ \Phi_g(y) \ \Phi_g(y) \end{array} + \cdots$$

- The number of completed infections has PGF $\Omega_{g+1}(z) = z\mu(\Omega_g(z))$ with $\Omega_0(z) = z^0 = 1$:

$$\Omega_{g+1}(z) = z\left(p_0 + \begin{array}{c} p_1 \\ | \\ \Omega_g(z) \end{array} + \begin{array}{c} p_2 \\ / \ \backslash \\ \Omega_g(z)\ \Omega_g(z) \end{array} + \begin{array}{c} p_3 \\ / | \backslash \\ \Omega_g(z) \ \Omega_g(z) \ \Omega_g(z) \end{array} + \cdots \right)$$

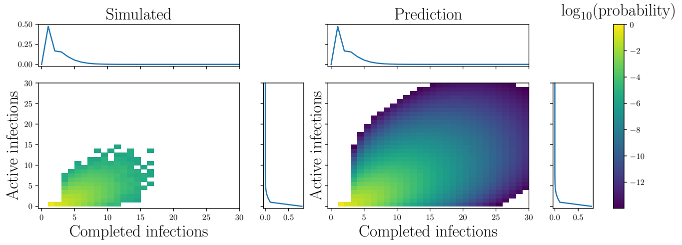
Intermediate size

- The joint distribution of completed infections and active infections has PGF $\Pi_{g+1}(y,z) = z\mu(\Pi_g(y,z))$ with $\Pi_0(y,z) = y$.

Intermediate size

- The joint distribution of completed infections and active infections has PGF $\Pi_{g+1}(y,z) = z\mu(\Pi_g(y,z))$ with $\Pi_0(y,z) = y$.

Generation 3, 5×10^5 simulations $N = 1000$ vs predictions

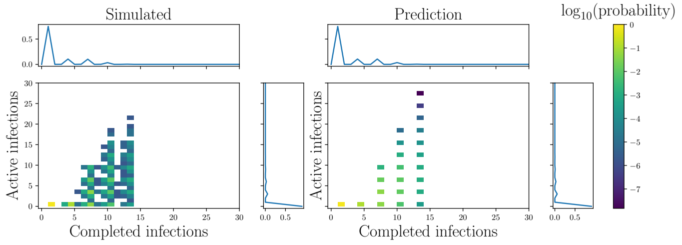


Poisson, $\mu(x) = e^{0.75(x-1)}$, $\mathcal{R}_0 = 0.75$.

Intermediate size

- The joint distribution of completed infections and active infections has PGF $\Pi_{g+1}(y,z) = z\mu(\Pi_g(y,z))$ with $\Pi_0(y,z) = y$.

Generation 3, 5×10^5 simulations $N = 1000$ vs predictions

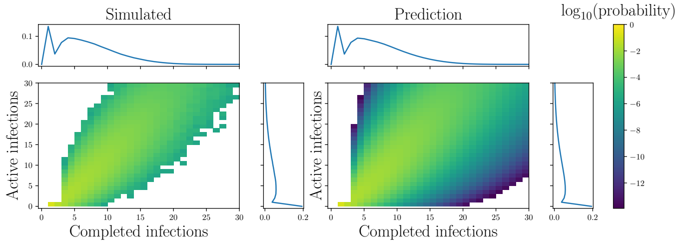


Bimodal, $\mu(x) = (3 + x^3)/4$, $\mathcal{R}_0 = 0.75$.

Intermediate size

- The joint distribution of completed infections and active infections has PGF $\Pi_{g+1}(y,z) = z\mu(\Pi_g(y,z))$ with $\Pi_0(y,z) = y$.

Generation 3, 5×10^5 simulations $N = 1000$ vs predictions

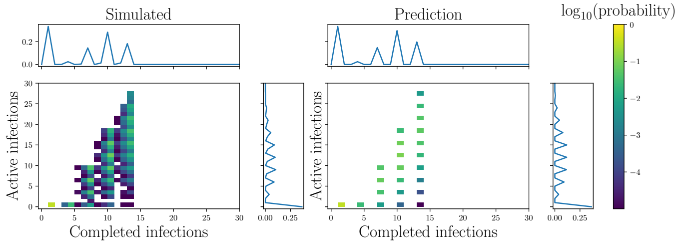


Poisson, $\mu(x) = e^{2(x-1)}$, $\mathcal{R}_0 = 2$.

Intermediate size

- The joint distribution of completed infections and active infections has PGF $\Pi_{g+1}(y,z) = z\mu(\Pi_g(y,z))$ with $\Pi_0(y,z) = y$.

Generation 3, 5×10^5 simulations $N = 1000$ vs predictions



Bimodal, $\mu(x) = (1+2x^3)/3$, $\mathcal{R}_0 = 2$.

Final size distribution

In an infinite population, the final size distribution is given by $\Omega(z) = \lim_{g\to\infty} \Omega_g(z)$ and is a solution to

$$\Omega(z) = z\mu(\Omega(z))$$

There is a theorem:

> *Given an offspring distribution with PGF $\mu(y)$, the probability of exactly $j < \infty$ infections is the coefficient of y^{j-1} in $[\mu(y)]^j$.*

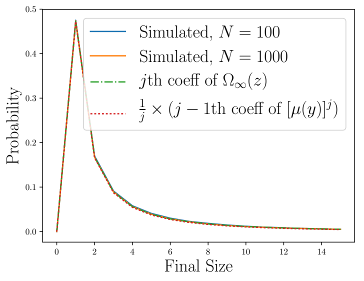
Final size distribution

In an infinite population, the final size distribution is given by $\Omega(z) = \lim_{g\to\infty} \Omega_g(z)$ and is a solution to

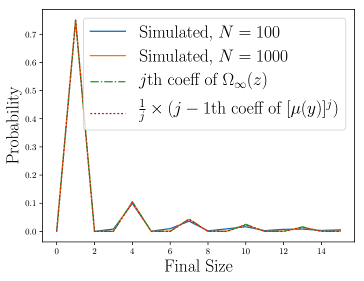
$$\Omega(z) = z\mu(\Omega(z))$$

There is a theorem:

> *Given an offspring distribution with PGF $\mu(y)$, the probability of exactly $j < \infty$ infections is the coefficient of y^{j-1} in $[\mu(y)]^j$.*



Poisson, $\mathcal{R}_0 = 0.75$



Bimodal, $\mathcal{R}_0 = 0.75$

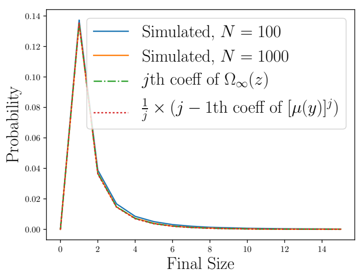
Final size distribution

In an infinite population, the final size distribution is given by $\Omega(z) = \lim_{g\to\infty} \Omega_g(z)$ and is a solution to

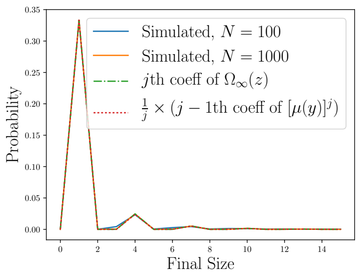
$$\Omega(z) = z\mu(\Omega(z))$$

There is a theorem:

Given an offspring distribution with PGF $\mu(y)$, the probability of exactly $j < \infty$ infections is the coefficient of y^{j-1} in $[\mu(y)]^j$.



Poisson, $\mathcal{R}_0 = 2$



Bimodal, $\mathcal{R}_0 = 2$

Final size distribution

In a finite population of size N, the probability q_M of exactly M infections occurring given that number of transmissions an individual causes has PGF $\mu(x)$ is found by solving

$$\mathsf{C}\vec{q} = \vec{1}$$

where the lower triangular matrix C has

$$c_{\ell,M} = \begin{cases} 0 & \ell > M \\ \left[\mu\left(\frac{M-1}{N-1}\right)\right]^{-\ell} \prod_{j=1}^{\ell-1} \frac{M-\ell}{N-\ell} & \ell \leq M \end{cases}$$

Unpublished work in progress

Final size distribution

In a finite population of size N, the probability q_M of exactly M infections occurring given that number of transmissions an individual causes has PGF $\mu(x)$ is found by solving

$$\mathsf{C}\vec{q} = \vec{1}$$

where the lower triangular matrix C has

$$c_{\ell,M} = \begin{cases} 0 & \ell > M \\ \left[\mu\left(\frac{M-1}{N-1}\right)\right]^{-\ell} \prod_{j=1}^{\ell-1} \frac{M-\ell}{N-\ell} & \ell \leq M \end{cases}$$

Unpublished work in progress

Basic deterministic SIR model

We now consider deterministic SIR models:

$$\dot{S} = -\beta I S$$
$$\dot{I} = \beta I S - \gamma I$$
$$\dot{R} = \gamma I$$

with $S + I + R = 1$.

Use an integrating factor $\theta^{-1} = e^{\beta \int I \mathrm{d}t}$ on the $\dot{S}$ equation:

$$\dot{S} + \beta I S = 0 \quad \Rightarrow \quad S = S(0)\theta$$

where $\dot{\theta} = -\beta I \theta$ and $\theta(0) = 1$

Basic deterministic SIR model

We now consider deterministic SIR models:

$$\dot{S} = -\beta I S$$
$$\dot{I} = \beta I S - \gamma I$$
$$\dot{R} = \gamma I$$

with $S + I + R = 1$.
Use an integrating factor $\theta^{-1} = e^{\beta \int I \mathrm{d}t}$ on the $\dot{S}$ equation:

$$\dot{S} + \beta I S = 0 \quad \Rightarrow \quad S = S(0)\theta$$

where $\dot{\theta} = -\beta I \theta$ and $\theta(0) = 1$

Basic deterministic SIR model

With some more work we have

$$\dot{\theta} = -\beta\theta I$$
$$S = S(0)\theta$$
$$I = 1 - S(0)\theta - R(0) - \frac{\gamma}{\beta}\ln\theta$$
$$R = R(0) + \frac{\gamma}{\beta}\ln\theta$$

A single ODE!
We can derive this directly by treating θ as the probability a susceptible person hasn't received any transmissions.

This also works for models with multiple subpopulations

Basic deterministic SIR model

With some more work we have

$$
\begin{aligned}
\dot{\theta} &= -\beta\theta I \\
S &= S(0)\theta \\
I &= 1 - S(0)\theta - R(0) - \frac{\gamma}{\beta}\ln\theta \\
R &= R(0) + \frac{\gamma}{\beta}\ln\theta
\end{aligned}
$$

A single ODE!
We can derive this directly by treating θ as the probability a susceptible person hasn't received any transmissions.

This also works for models with multiple subpopulations

Heterogeneous susceptibility

Assuming that the susceptible individuals each have a k such that they become infected as a Poisson process with rate $k\beta I/\langle K\rangle$, we have

$$S = S(0)\psi(\theta)$$
$$I = \left(1 - S(0)\psi(\theta) + \frac{\gamma\langle K\rangle}{\beta}\ln\theta\right)$$
$$R = -\frac{\gamma\langle K\rangle}{\beta}\ln\theta$$

where $\psi(x) = \sum_k P(k)x^k$ and the system is governed by a single ODE

$$\dot{\theta} = \frac{-\beta\theta\left(1 - S(0)\psi(\theta) + \frac{\gamma\langle K\rangle}{\beta}\ln\theta\right)}{\langle K\rangle}$$

and initial condition

$$\theta(0) = 1\,.$$

Discussion

- Probability Generating Functions have many applications to infectious disease modeling.
- They provide efficient ways to calculate:
  - Epidemic probability
  - Outbreak size distribution
  - Deterministic SIR Epidemics

Acknowledgments

Thanks to many people at the Institute for Disease Modeling, including

- Mike Famulare
- Edward Wenger
- Hao Hu