9 Toward Realism II: Overdispersion and Superspreading

Overdispersion is the word given by statisticians to the presence of greater variability in the data than one would expect under some given statistical model. For example, a Poisson distribution with mean λ also has variance λ. This means that if we expect to see Poisson-distributed values (for instance, Poisson-distributed secondary infections) and instead we see something with variance greater than λ, we would classify this as overdispersion.

Superspreading is the word given by infectious disease epidemiologists to the presence of overdispersion in secondary infections relative to a Poisson model. When superspreading is an attribute of a person, we might call that person a **superspreader**, and when the superspreading is an attribute of a particular situation, we might call that a **superspreading event**. This week, we'll investigate both.

9.1 Secondary infection distributions

Last week we introduced the idea that each infection may lead to a varying number of secondary infections, and we suggested a few different distributions for the count of secondary infections. This week, we'll begin by asking: *why* might some individuals infect more than others?

9.1.1 Contacts

One reason that some individuals infect more that others is rates of contact. For instance, early in the HIV pandemic, May and Anderson[1] compiled various sources of unpublished data, shown in Fig. 1, on the number of sexual contacts (a) per month in London, 1986, (b) per 2-years in San Francisco, 1984-1985, and (c) per year in London, 1984 among a group who we would today refer to as men who have sex with men, MSM. They contrasted that with (d) heterosexuals 18-44 in England, 1986. A first-order observation of these data would note the contrast between the MSM and heterosexual populations. However, from the perspective of heterogeneity in secondary infections, note the substantial variation in sexual contacts *within* each chart–including panel (d).

Similar observations have been made about vector-borne diseases. For example, the Pareto Principle (also called an "80-20 rule") has been identified in the number of vectors per host (see, for example, Woolhouse et al. Figure 2) such that 80% of the mosquitos bite just 20% of the hosts.[2]

The key idea of these observations is that *individuals' variability in contacts can produce variability in secondary infections.*

[1] Sadly, the world lost Robert May in April 2020. May was a giant in complex systems and infectious disease modeling and ecology. Read an obituary from the Santa Fe Institute here. Together with his coauthor of many years, Roy Anderson, they published many highly influential studies in our field. Perhaps your Adopt A Pathogen Program work has in fact led you to an Anderson & May, or a May & Anderson publication!

[2] By the way, guess who the last author on that paper was?

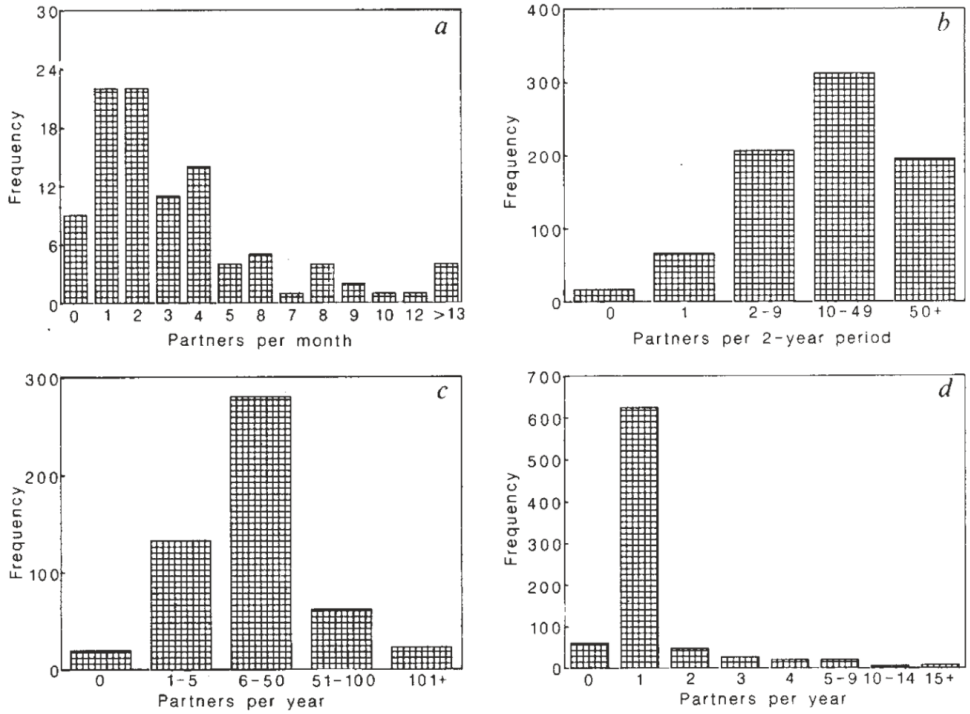

Figure 1: The number of sexual contacts (a) per month in London, 1986, (b) per 2-years in San Francisco, 1984-1985, and (c) per year in London, 1984 among a group who we would today refer to as men who have sex with men, MSM. These charts contrast the number of reported sexual partners for (d) heterosexuals 18-44 in England, 1986.(Fig. 4 from Transmission dynamics of HIV infection, 1987.

Before moving to other sources of variability in secondary infections, we need to talk about ethics. When we make observations that individuals' contacts vary substantially, this may cause us to pass some judgment on them, or blame them for spreading an infectious disease. Consider, for instance, whether you carry biases of your own when you look at Fig. 1, or when you reflect back on the COVID-19 pandemic and various behaviors around contacts.

Here, we should consider the cautionary tale of Gaëten Dugas, who was for some time referred to (and misdescribed!) as "Patient Zero" of the AIDS in the US. Dugas was a Quebecois flight attendant who was much maligned in the landmark book *And the Band Played On* for recklessly spreading HIV through hundreds of annual sexual contacts. However, since the 1980s, the ability to use phylogenetic trees to map the relationship between one infection and many others has shown that HIV from Dugas does not sit near the root of the tree, but instead in the middle. However, the die is cast: Dugas remains associated with the spread of HIV; we should instead remember him as a human whose name teaches us a cautionary tale about the ethics passing judgment or blame before all the facts are in. Stigma is a powerful force, leading many of those who died early in the AIDS crisis to die shunned and isolated from the wider society.

9.1.2 Duration of infection

Another reason that some individuals infect more that others is variability in the duration of infection. Perhaps one of the most easily recognized instances of variability in clearance time comes from *P. falciparum* which causes malaria. Parasites use a process of antigenic variation to prolong infection, but this process is stochastic meaning that some infections may last far longer than others. In fact, if left untreated, *P. falciparum* infections can last, at low levels for not just month, but years: the longest confirmed *P. falciparum* infection lasted 13 years.

Some of the best data we have on untreated *P. falciparum* infections comes from **malariotherapy**, the pre-antibiotic era procedure by which people were intentionally given malaria (by venous injection) with the goal of creating fever. After enough time with the tertian fever[3] quinine would be administered to cure the malaria, particularly if the person recovered from the infection being treated with fever. This was particularly applied to late-stage neurosyphilis patients, and in fact led to a 1927 Nobel Prize. To see some analysis of the malariotherapy data in a modeling context, see work by Lauren Childs and Caroline Buckee.

9.1.3 Infectiousness

Yet another reason that some individuals infect more that others is variability in infectiousness. There are no natural units for infectiousness, making it hard to measure, but proxies for infectiousness are common in the literature. One proxy for infectiousness discussed widely during the COVID-19 pandemic is **viral load**, which is a measurement of the number of copies of viral RNA (and typically one gene or a few genes

[3] A tertian fever is fever every two days, characteristic of the blood stage life cycle of *P. falciparum*. The name is weird because "tertian" has a root of 3, and refers to the fact that the fever reappears on the third day, i.e. on days 1, 3, 5, etc. A quartian fever occurs on days 1, 4, 7, etc.

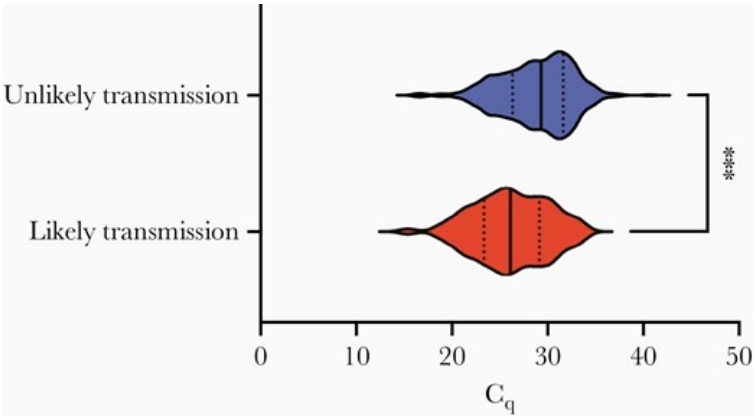

Figure 2: In the roommates transmitting SARS-CoV-2 (red), the viral loads are higher than among the roommates not transmitting (blue), from a study of roommates at CU Boulder. The horizontal axis units are C_q, the quantification cycle for the sample, such that lower C_q means a higher viral load in the person's saliva.

in particular) per ml of sample. The logic here is that more virus means a higher probability that a contact leads to infection.

One study implicating viral load in infectiousness came from our campus at CU Boulder. The study looked at discordant roommate pairs[4] and analyzed whether likelihood of transmission was associated with viral load, finding that lower viral loads were associated with lower odds of transmission (Fig. 2).

9.2 A model of individual variation — Lloyd-Smith et al 2005

Regardless of cause—contacts, duration, infectiousness—suppose that there exist some distribution of individual reproductive numbers, $Pr(\nu)$, whose expected value is R_0,

$$E[Pr(\nu)] = R_0 \ .$$

A person i will have individual reproductive number ν_i.

The value of ν_i tells us the *typical* number of infections caused by someone with that individual reproductive number. Our model will let the *actual* number of infections be drawn from a Poisson distribution with that mean. In other words, if Z_i is the number of secondary infections from individual i, then

$$Z_i \sim \text{Poisson}(\nu_i) \ .$$

[4] **Discordant** means that one roommate was infected while the other was not. This word often comes up in the analysis of couples in which one has a chronic or incurable infection, e.g. HSV or HIV, and the other does not.

This model represents a composition of two probability distributions, which is also called a **compound probability distribution** or a mixture distribution. A compound probability distribution comes up any time you assume that a random variable comes from some parametric distribution whose parameter, itself, has a distribution.

Lloyd-Smith et al. considered three cases, which we will also develop.

1. The easiest case is to assume that $\nu = R_0$, i.e. that the individual reproductive number has zero variance and everyone is equally transmitting.
2. The second case is to assume the dynamics of the SIR model with rates β and γ, another model which we analyzed in previous lectures. Here, ν is exponentially distributed with mean $R_0 + 1$.[5]
3. The third case is to assume that ν is Gamma distributed with mean R_0 and dispersion parameter k.[6]

Each of these distributions for ν, when "plugged in" to the Poisson-distributed number of actual secondary infections, produces a different distribution Z. It is important to understand how these distributions are derived, and so we will take each one at a time.

The general formula for computing a compound probability distribution is to recognize what we wish to compute: the probability that an individual i will create some number of secondary infections z. For example, say $z = 0$. The compound distribution must tell us the probability that we draw from Z and get $z = 0$. However, there are infinite possible ways to get $z = 0$ because there are infinite possible values that could have been drawn for ν_i. This means that we're going to use the law of total probability to say

$$Pr(Z = z) = \int_{\nu} Pr(Z = z \mid \nu)\, Pr(\nu)\, d\nu \,.$$

9.2.1 Case 1: Everyone has the same $\nu = R_0$

This case assumes that $Pr(\nu) = \delta_{R_0}(\nu)$. This *delta* function is a special probability distribution that is identically zero for all values *except* R_0, and is undefined at exactly R_0, but by definition integrates to exactly 1.[7] We therefore have

$$Pr(Z = z) = \int_0^{\infty} \frac{\nu^z}{z!} e^{-\nu}\, \delta_{R_0}(\nu)\, d\nu \,. \tag{1}$$

It turns out that integrating against δ functions is convenient, because the integrand pops out, evaluated at the point where the δ function is nonzero, i.e.

$$Pr(Z = z) = \frac{(R_0)^z}{z!} e^{-R_0} \,. \tag{2}$$

[5] Exercise: Can you derive this result?

[6] This is what is assumed in Homework 3, Question 3.

[7] You can this of the δ function as the limit of a normal distribution whose mean stays fixed but whose variance $\sigma^2 \to 0$. It continues to integrate to one, but the peak of the distribution gets higher and higher, and approaches ∞.

This is nothing more than a Poisson distribution, and we therefore conclude that in the first case, the offspring distribution is $Z \sim \text{Poisson}(R_0)$.

9.2.2 Case 2: The values of ν are exponentially distributed

This case assumes $Pr(\nu) = \frac{1}{R_0} e^{-\frac{\nu}{R_0}}$. This exponential distribution means that the modal ν is 0, but that some ν are potentially very large. As before, the average ν must be R_0. Using the approach of our compound probability distribution,

$$Pr(Z = z) = \int_0^\infty \frac{\nu^z}{z!} e^{-\nu} \, \frac{1}{R_0} e^{-\frac{\nu}{R_0}} \, d\nu \, . \tag{3}$$

Attacking an equation like this will take courage! Let's first do all the simplifications that we can muster without any calculus.

$$Pr(Z = z) = \frac{1}{z!} \frac{1}{R_0} \int_0^\infty \nu^z e^{-\nu\left(1+\frac{1}{R_0}\right)} \, d\nu \, . \tag{4}$$

Good news! We've not only cleaned up our equation, but we can now see that the integrand is a polynomial in ν times an exponential in ν, which means it is susceptible to integration by parts. Now, these are class notes and not some kind of handwavy "it can be shown that..." sort of arrangement, so here are the details. Look away if you want to sort this one out on your own!

The integral has the form (written in the more comfortable x instead of ν),

$$\int_0^\infty x^z e^{-ax} \, dx \, ,$$

which means we can use the integration by parts formula repeatedly to knock x^z down, one integration-by-parts at a time,[8] and therefore,

$$\begin{aligned}
\int_0^\infty x^z e^{-ax} \, dx &= \frac{z}{a} \int_0^\infty x^{z-1} e^{-ax} \, dx \\
&= \frac{z(z-1)}{a^2} \int_0^\infty x^{z-2} e^{-ax} \, dx \\
&= \frac{z(z-1)(z-2)}{a^3} \int_0^\infty x^{z-3} e^{-ax} \, dx \\
\cdots &= \frac{z!}{a^z} \int_0^\infty e^{-ax} \, dx \\
&= \frac{z!}{a^{z+1}} \, .
\end{aligned} \tag{5}$$

[8] For your recollection, $\int_0^\infty u \, dv = u \, v \Big|_0^\infty - \int_0^\infty v \, du \, .$

Substituting ν for x, and $\left(1+\frac{1}{R_0}\right)$ for a, we can rewrite Eq. (4) as

$$Pr(Z=z) = \frac{1}{z!}\frac{1}{R_0}\left(\frac{z!}{\left(1+\frac{1}{R_0}\right)^{z+1}}\right)$$

which simplifies to

$$Pr(Z=z) = \frac{1}{R_0+1}\left(1-\frac{1}{R_0+1}\right)^{z}, \tag{6}$$

a geometric offspring distribution with parameter $p=\frac{1}{R_0+1}$, mean R_0, and starts at $z=0$.[9]

9.2.3 Case 3: The values of ν are Gamma distributed.

This case assumes that $Pr(\nu)$ is Gamma distributed with mean R_0 and dispersion parameter k. Because the offspring are Poisson distributed with rate ν, this is sometimes called a **Poisson-Gamma Mixture** in the broader literature. The derivation of the offspring distribution proceeds as follows, where we'll be making use of the gamma function $\Gamma(\alpha)$.[10] To keep the derivation simple, we'll parameterize the Gamma distribution with α and β, which are related to our parameters via $\beta=k$ and $\alpha=\frac{k}{R_0}$

$$\begin{aligned}
Pr(Z=z) &= \int_0^\infty \frac{\nu^z}{z!}e^{-\nu}\frac{\alpha^\beta}{\Gamma(\beta)}\nu^{\beta-1}e^{-\alpha\nu}\,d\nu \\
&= \int_0^\infty \frac{\alpha^\beta}{z!\,\Gamma(\beta)}\nu^{z+\beta-1}e^{-(\alpha+1)\nu}\,d\nu \\
&= \frac{\alpha^\beta}{z!\,\Gamma(\beta)}\frac{\Gamma(z+\beta)}{(\alpha+1)^{z+\beta}}\int_0^\infty \frac{(\alpha+1)^{z+\beta}}{\Gamma(z+\beta)}\nu^{z+\beta-1}e^{-(\alpha+1)\nu}\,d\nu \\
&= \frac{\alpha^\beta}{z!\,\Gamma(\beta)}\frac{\Gamma(z+\beta)}{(\alpha+1)^{z+\beta}} \\
&= \frac{\Gamma(z+\beta)}{\Gamma(z+1)\Gamma(\beta)}\left(\frac{\alpha}{\alpha+1}\right)^\beta\left(\frac{1}{\alpha+1}\right)^z \\
&= \binom{z+\beta-1}{z}\left(\frac{\alpha}{\alpha+1}\right)^\beta\left(\frac{1}{\alpha+1}\right)^z
\end{aligned} \tag{7}$$

[9] Recall that one can think of a geometric distribution in two ways: either we're counting the number of flips k till the first tails $p(1-p)^{k-1}$, which has mean $\frac{1}{p}$ and starts at $k=1$; or we're counting the number of heads k till the first tails $p(1-p)^k$ which has mean $\frac{p}{1-p}$ and starts at $k=0$. For our offspring distribution we have the latter, as $z=0$ offspring is not only plausible but common.

[10] By this point in your studies you are surely familiar with the factorial, i.e. $n! = n(n-1)(n-2)\ldots(2)(1)$. While the factorial is only defined at integer values of n, the Γ function is defined at all values of n, but it coincides with the factorial at the integers, offset by 1, i.e. $\Gamma(n)=(n-1)!$. Its definition is formally in terms of the integral $\Gamma(z)=\int_0^\infty t^{z-1}e^{-t}\,dt$.

Note that the term in blue is the integral of a Gamma's PDF over its entire support, and thus integrates conveniently to one. Finally, substituting in $\beta = k$, $\alpha = \frac{k}{R_0}$, we get

$$Pr(Z = z) = \binom{z + k - 1}{z} \left(\frac{k}{k + R_0}\right)^k \left(1 - \frac{k}{k + R_0}\right)^z , \tag{8}$$

which is a negative binomial distribution with dispersion parameter k, "coin flip bias" parameter $p = k/(k + R_0)$, and expected value R_0. Note that this distribution, too, begins at $z = 0$. Note also that we have also written the binomial coefficient in the parentheses typical of positive integer values of z and k to help spot the negative binomial more easily. However, we fully expect non-integer values of k and thus rely on the continuous extension of the binomial coefficient defined via gamma functions.[11]

Note that when $k \to \infty$, the negative binomial becomes a Poisson, and when $k \to 1$, the negative binomial becomes a geometric. This means that case 3 contains within it both special cases we have considered thus far, and then some: values of k between 1 and ∞ can interpolate between the models, but values of $k < 1$ can lead to more extreme overdispersion than even the geometric secondary case counts of the exponential model!

9.2.4 Summary

In summary, when $Pr(\nu)$ is a δ, $Pr(Z)$ is Poisson; when $Pr(\nu)$ is an exponential, $Pr(Z)$ is geometric; and when $Pr(\nu)$ is gamma, $Pr(Z)$ is negative binomial.

What good are these three models? To answer this question we need to introduce three important ingredients: data, maximum likelihood estimation, and model selection. The authors of this study collected detailed contact tracing data on multiple outbreaks of multiple diseases[12] This means that, assuming that contact tracing was performed adequately, each dataset represents a set of samples from Z for that particular disease in that particular environment.

The second key ingredient is **maximum likelihood estimation**. There are many possible ways that one might fit a model to data, and maximum likelihood is a common and powerful one. When we say "fit a model to data" what we really mean is "estimate the parameters of a model to best fit the observed data." The **maximum likelihood estimate** (or MLE) of a parameter is *the value of the parameter that maximizes the likelihood of observing the data, given the model.*

An example of MLE in action may help. You may know that if I flip a biased coin 100 times and end up with 30 heads and 70 tails, that a good estimate for the bias is $p = 0.3$. The MLE approach is to write down the likelihood of observing $h = 30$ heads and $t = 70$ tails, given a *variable* parameter for the coin's bias p.

[11] Specifically, $\binom{z+k-1}{z} = \frac{\Gamma(z+k)}{\Gamma(z+1)\Gamma(k)}$.

[12] SARS (2 outbreaks), Measles (2), Smallpox (5), and Monkeypox, Pneumonic plagues, Hantavirus, Ebola, and Rubella (1 each). Read more in the authors' notes here.

If we assume that each of the flips is independent,

$$\mathcal{L}(h,t \mid p) \equiv Pr(h,t \mid p) = \binom{h+t}{h} \left[\prod_{i=1}^{h} p\right] \left[\prod_{i=1}^{t} (1-p)\right] = \binom{h+t}{h} p^h (1-p)^t \; .$$

Here, the fancy $\mathcal{L}$[13] means likelihood. Specifically, we are writing down *the likelihood of observing the data, given the parameter*. Our goal now is to determine the value of p that maximizes that likelihood. To do so, we'll first take a $\log$, and then take a derivative with respect to p. Recall that $\log$ is monotonic, which means that the locations of any extrema of $f(x)$ are at the same locations x as the extrema of $\log[f(x)]$. This means that taking a $\log$ and then taking a derivative will give us the same answer as just taking the derivative, but may be considerably easier. The **log likelihood** is

$$\log \mathcal{L}(h,t \mid p) = \log \binom{h+t}{h} + h \log p + t \log(1-p) \; .$$

Taking a derivative with respect to p, we get

$$\frac{\log \mathcal{L}}{dp} = \frac{h}{p} - \frac{t}{1-p} \; ,$$

and setting this equal to zero (to find the p that maximizes) we get

$$\hat{p} = \frac{h}{h+t} \; .$$

In other words, not only is the MLE for p the intuitive estimate, but it's also a statistically principled estimate. Note the hat on $\hat{p}$ which is a traditional way of showing that a parameter has been estimated.

In this example, we used a coin-flipping (binomial) model to estimate p from flip outcomes. In the outbreak data from Lloyd-Smith et al., the authors use each of their three models, in turn, to estimate each model's parameter(s) from offspring count data.

The third key ingredient is **model selection**. The idea of model selection is, as it sounds, to select or choose a model to explain the data. When we have multiple models, each of which could explain the data, model selection allows us to ask which model does the best job, but it also recognizes that a more complex model will more easily be able to explain the nuances of a dataset. Model selection, which may take various forms, simply penalizes models that are more complex so that they don't get an unfair advantage. You could think of model selection as a mathematical version of the Principle of Simplicity:

Everything should be made as simple as possible, but not simpler.

When Lloyd-Smith et al., analyzed their various outbreak data by fitting each model by MLE and using model selection to ask which of the models (Poisson, Geometric, Negative Binomial) was the best explanation for SARS data, they found that

[13] The LaTeX for $\mathcal{L}$ is mathcal{L}

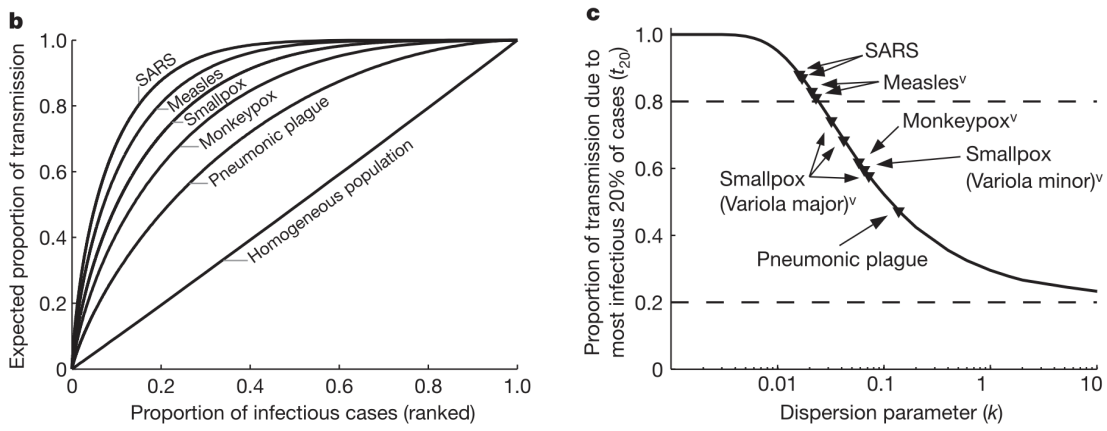

Figure 3: Panels b (left) and c (right) from Lloyd-Smith et al., Figure 1.

- For SARS outbreaks, the Negative Binomial model is superior to the Poisson and Geometric.
- The MLE of the dispersion parameter $\hat{k} = 0.16$ suggests *overdispersion* even relative to the geometric distribution (which would be $k = 1$).
- The corresponding $Gamma$ distribution for individual reproductive numbers would suggest that 73% of individuals have $\nu < 1$, but that 6% have $\nu > 8$.

This is precisely the heterogeneity that we expect with superspreading!

For other directly transmitted infectious diseases, the authors found that the Poisson model is essentially always ruled out, but that the geometric model may be appropriate. Interestingly, we know that the geometric model cannot possibly be justified using the assumptions we used to derive the exponentially-distributed ν: that would require that we assume exponentially-distributed infection durations, which are implausible. Therefore, when model selection favors the geometric distribution, it is a strong hint that variability in environment, contacts, and infectiousness must be implicated, as we cannot justify the exponential/geometric distribution from infectious duration alone!

There are two convenient ways of visualizing the results from this analysis of superspreading. One is what we call a **Lorenz curve** which is commonly used in economics in the analysis of inequalities in production or income. The Lorenz curve sorts individuals by their contribution to some total (e.g. their contribution to future infections) and then plots how the inclusion of a proportion x of individuals can account for a proportion y of the total. If everyone is identical, then each additional person contributes the same amount to the total, producing the $y = x$ equality line. Deviations above that line will be more extreme as variability (i.e. superspreading potential) increases (Fig. 3, left panel)

Another convenient way to visualize superspreading potential is to ask what transmission is driven by the

most infectious 20% of cases, which the authors call t_{20}. Here, "infectious" does not mean infectious in the absence of environment and contacts, but is instead inclusive of it. This may be done by showing how t_{20} changes as the value of k changes.[14] Because k can take on values across many orders of magnitude, the authors choose to use a log-scale (Fig. 3, right panel). When viewing these results, note how even pneumonic plague is plotted with $\hat{k} \approx 0.1$, far from even the geometric distribution.

9.3 Closing thoughts

Superspreading and overdispersion produce some counterintuitive predictions. One is that, for the same value of R_0, higher superspreading potential (lower k) will make it harder for an outbreak to grow large. This means that, even when R_0 is large, a small value of k will result in either (i) large outbreaks that grow rapidly, or (ii) extremely short outbreaks that die before a superspreading event occurs. In other words, with superspreading, finite-size outbreaks die out extremely quickly.

Superspreading may also lead to peculiar observations and difficulty in messaging around infectious diseases. When individual reproductive numbers vary substantially, we should expect infrequent but explosive epidemics after the introduction of a single case. Indeed, this is precisely what was observed for SARS in 2003, with similar results for SARS-CoV-2 beginning in 2019. Such real-world observations are nearly impossible to explain with $k = 1$ and especially with $k \to \infty$, and truly impossible to explain with a deterministic model from the SIR-ODE family.

Finally, while the definition of superspreading was relatively difficult to pin down prior to the Lloyd-Smith et al. paper, the authors do contribute a useful way to define superspreading events (SSEs):

1. Estimate R_e, the effective reproductive number for the outbreak in question, inclusive of behavior changes or interventions.
2. Construct Poisson(R_e) as a null model for transmissions Z.
3. Define an SSE as any infected individual who infects more than $Z^{(n)}$ others, where $Z^{(n)}$ represents the nth percentile of the distribution.

This approach implicitly treats the Poisson as a null model—deviations from which can be used to quantify and define SSEs.

[14] Note that, because t_{20} is essentially a calculation of the 80th percentile of these distributions, only the value of k matters; R_0 is irrelevant.