**Superspreading**

Lloyd-Smith et al. (2005) show that the 2002-2004 SARS-CoV-1 epidemic was driven by a small number of events where one case directly infected a large number of secondary cases – a so called superspreading event. This means that for SARS-CoV-1 the distribution of how many secondary cases each primary case generates is heavy tailed. More specifically, the effective reproduction number describes the mean number of secondary cases a primary case generates during the outbreak, i.e. it is the mean of the offspring distribution. In order to address dispersion around this mean, Lloyd-Smith et al. (2005) use the [negative binomial distribution](https://en.wikipedia.org/wiki/Negative_binomial_distribution) with mean \(R(t)\) and over-dispersion parameter \(k\) as a probability model for the offspring distribution. The number of offspring that case \(i\), which got infected at time \(t\_i\), causes is given by \[  
Y\_{i} \sim \operatorname{NegBin}(R(t\_i), k),  
\] s.t. \(\operatorname{E}(Y\_{i}) = R(t\_i)\) and \(\operatorname{Var}(Y\_{i}) = R(t\_i) (1 + \frac{1}{k} R(t\_i))\). This parametrisation makes it easy to see that the negative binomial model has an additional factor \(1 + \frac{1}{k} R(t\_i)\) for the variance, which allows it to have excess variance (aka. over-dispersion) compared to the Poisson distribution, which has \(\operatorname{Var}(Y\_{i}) = R(t\_i)\). If \(k\rightarrow \infty\) we get the Poisson distribution and the closer \(k\) is to zero the larger the variance, i.e. the heterogeneity, in the distribution is. Note the deliberate use of the effective reproduction number \(R(t\_i)\) instead of the basic reproduction number \(R\_0\) (as done in Lloyd-Smith et al. (2005)) in the model. This is to highlight, that one is likely to observe clusters in the context of interventions and depletion of susceptibles.

That the dispersion parameter \(k\) is making epidemiological fame is a little surprising, because it is a parameter in a specific parametric model. A parametric model, which might be inadequate for the observed data. A secondary objective of this post is thus to focus more on describing the heterogeneity of the offspring distribution using classical statistical concepts such as the [**Gini coefficient**](https://en.wikipedia.org/wiki/Gini_coefficient).

**Negative binomial distributed number of secondary cases**

Let’s assume \(k=0.45\) as done in Adam et al. (2020). This is a slightly higher estimate than the \(k=0.1\) estimate by Endo et al. (2020)[2](http://staff.math.su.se/hoehle/blog/2020/05/31/superspreader.html#fn2) quoted in the Science article. We want to derive statements like “the x% most active spreaders infected y% of all cases” as a function of \(k\). The PMF of the offspring distribution with mean 2.5 and dispersion 0.45 looks as follows:

Rt <- 2.5

k <- 0.45

# Evaluate on a larger enough grid, so E(Y\_t) is determined accurate enough

# We also include -1 in the grid to get a point (0,0) needed for the Lorenz curve

df <- data.frame(x=-1:250) %>% mutate(pmf= dnbinom(x, mu=Rt, size=k))

So we observe that 43% of the cases never manage to infect a secondary case, whereas some cases manage to generate more than 10 new cases. The mean of the distribution is checked empirically to equal the specified \(R(t)\) of 2.5:

sum(df$x \* df$pmf)

## [1] 2.5

Lloyd-Smith et al. (2005) define a **superspreader** to be a primary case, which generates more secondary cases than the 99th quantile of the Poisson distribution with mean \(R(t)\). We use this to compute the proportion of superspreaders in our distribution:

(superspreader\_threshold <- qpois(0.99, lambda=Rt))

## [1] 7

(p\_superspreader <- pnbinom(superspreader\_threshold, mu=Rt, size=k, lower.tail=FALSE))

## [1] 0.09539277

So 10% of the cases will generate more than 7 new cases. To get to statements such as “10% generate 80% of the cases” we also need to know how many cases those 10% generate out of the 2.5 average.

# Compute proportion of the overall expected number of new cases

df <- df %>% mutate(cdf = pnbinom(x, mu=Rt, size=k),

expected\_cases=x\*pmf,

prop\_of\_Rt=expected\_cases/Rt,

cum\_prop\_of\_Rt = cumsum(prop\_of\_Rt))

# Summarise

info <- df %>% filter(x > superspreader\_threshold) %>%

summarise(expected\_cases = sum(expected\_cases), prop\_of\_Rt = sum(prop\_of\_Rt))

info

## expected\_cases prop\_of\_Rt

## 1 1.192786 0.4771144

In other words, the superspreaders generate (on average) 1.19 of the 2.5 new cases of a generation, i.e. 48%.

These statements can also be made without formulating a superspreader threshold by graphing the cumulative share of the distribution of primary cases against the cumulative share of secondary cases these generate. This is exactly what the [Lorenz curve](https://en.wikipedia.org/wiki/Lorenz_curve) is doing. However, for outbreak analysis it appears clearer to graph the cumulative distribution in decreasing order of the number of offspring, i.e. following Lloyd-Smith et al. (2005) we plot the cumulative share as \(P(Y\geq y)\) instead of \(P(Y \leq y)\). This is a variation of the Lorenz curve, but allows statements such as “the %x cases with highest number of offspring generate %y of the secondary cases”.

# Add information for plotting the modified Lorenz curve

df <- df %>%

mutate(cdf\_decreasing = pnbinom(x-1, mu=Rt, size=k, lower.tail=FALSE)) %>%

arrange(desc(x)) %>%

mutate(cum\_prop\_of\_Rt\_decreasing = cumsum(prop\_of\_Rt))

# Plot the modified Lorenz curve as in Fig 1b of Lloyd-Smith et al. (2005)

ggplot(df, aes(x=cdf\_decreasing, y=cum\_prop\_of\_Rt\_decreasing)) + geom\_line() +

coord\_cartesian(xlim=c(0,1)) +

xlab("Proportion of the infectious cases (cases with most secondary cases first)") +

ylab("Proportion of the secondary cases") +

scale\_x\_continuous(labels=scales::percent, breaks=seq(0,1,length=6)) +

scale\_y\_continuous(labels=scales::percent, breaks=seq(0,1,length=6)) +

geom\_line(data=data.frame(x=seq(0,1,length=100)) %>% mutate(y=x), aes(x=x, y=y), lty=2, col="gray") + ggtitle(str\_c("Scenario: R(t) = ", Rt, ", k = ", k))

Using the standard formulas to compute the [Gini coefficient](https://en.wikipedia.org/wiki/Gini_coefficient#Discrete_probability_distribution) for a discrete distribution with support on the non-negative integers, i.e.  \[  
G = \frac{1}{2\mu} \sum\_{y=0}^\infty \sum\_{z=0}^\infty f(y) f(z) |y-z|,  
\] where \(f(y)\), \(y=0,1,\ldots\) denotes the PMF of the distribution and \(\mu=\sum\_{y=0}^\infty y f(y)\) is the mean of the distribution. In our case \(\mu=R(t)\). From this we get

# Gini index for a discrete probability distribution

gini\_coeff <- function(df) {

mu <- sum(df$x \* df$pmf)

sum <- 0

for (i in 1:nrow(df)) {

for (j in 1:nrow(df)) {

sum <- sum + df$pmf[i] \* df$pmf[j] \* abs(df$x[i] - df$x[j])

}

}

return(sum/(2\*mu))

}

gini\_coeff(df)

## [1] 0.704049

A plot of the relationship between the dispersion parameter and the Gini index, given a fixed value of \(R(t)=2.5\), looks as follows

We see that the Gini index converges from above to the Gini index of the Poisson distribution with mean \(R(t)\). In our case this limit is

gini\_coeff( data.frame(x=0:250) %>% mutate(pmf = dpois(x, lambda=Rt)))

## [1] 0.3475131

**Red Marble Toy Example**

The described hypothetical scenario is translated to an offspring distribution, where a primary case either generates 1 (with probability 9/10) or 10 (with probability 1/10) secondary cases:

# Offspring distribution

df\_toyoffspring <- data.frame( x=c(1,10), pmf=c(9/10, 1/10))

# Hypothetical outbreak with 10000 cases from this offspring distribution

y\_obs <- sample(df\_toyoffspring$x, size=10000, replace=TRUE, prob=df\_toyoffspring$pmf)

# Fit the negative binomial distribution to the observed offspring distribution

# Note It would be better to fit the PMF directly instead of to the hypothetical

# outbreak data

(fit <- MASS::fitdistr(y\_obs, "negative binomial"))

## size mu

## 1.69483494 1.90263640

## (0.03724779) (0.02009563)

# Note: different parametrisation of the k parameter

(k.hat <- 1/fit$estimate["size"])

## size

## 0.590028

In other words, when fitting a negative binomial distribution to these data (probably not a good idea) we get a dispersion parameter of 0.59.

The Gini coefficient allows for a more sensible description for offspring distributions, which are clearly not negative-binomial.

gini\_coeff(df\_toyoffspring)

## [1] 0.4263158

**Discussion**

The effect of superspreaders underlines the stochastic nature of the dynamics of an person-to-person transmitted disease in a population. The dispersion parameter \(k\) is conditional on the assumption of a given parametric model for the offspring distribution (negative binomial). The Gini index is an alternative characterisation to measure heterogeneity. However, in both cases the parameters are to be interpreted together with the expectation of the distribution. Estimation of the dispersion parameter is orthogonal to the mean in the negative binomial and its straightforward to also get confidence intervals for it. This is less straightforward for the Gini index.

A heavy tailed offspring distribution can make the disease easier to control by targeting intervention measures to restrict superspreading (Lloyd-Smith et al. 2005). The hope is that such interventions are “cheaper” than interventions which target the entire population of infectious contacts. However, the success of such a targeted strategy also depends on how large the contribution of superspreaders really is. Hence, some effort is needed to quantify the effect of superspreaders. Furthermore, the above treatment also underlines that heterogeneity can be a helpful feature to exploit when trying to control a disease. Another aspect of such heterogeneity, namely its influence on the threshold of herd immunity, has recently been invested by my colleagues at Stockholm University (Britton, Ball, and Trapman 2020).