Overdispersion – How Superspreading Drives the Pandemic

By James Sharpe for

COVID-19 Actuaries Response Group – Learn. Share. Educate. Influence.

Summary

There is much focus on efforts to reduce R below one, but little public recognition of the disproportionate impact of superspreading. This bulletin explores how superspreading is an important contributor to R.

Broad strategies can be successful at reducing R but may come at a high economic cost. Strategies which focus on reducing superspreading may – if successful – lead to suppression of the virus at a lower economic cost.

Introduction - The pandemic and R

A key metric used in media and government communications throughout the pandemic is "R" – a measure of the contagiousness of infectious diseases. R is the average number of people each infected person goes on to infect. Government scientists and politicians have been clear about the target to get R below one, so that case numbers reduce, with each infected person infecting less than one person on average.

R depends on both virus transmissibility and people's individual actions. Prior to people acting in response to the virus, its value appeared relatively stable in some countries.

To give an indication of R0 (the value of R at the beginning of an outbreak) for various diseases in recent history, values are shown below for a few different outbreaks[1]:

Disease	Outbreak	R0 estimates
COVID-19	Wuhan early 2020	1.4 – 5.7
MERS	Saudi Arabia 2014	0.5 – 3.9
Ebola	West Africa 2014	1.5 – 2.5
SARS	Hong Kong 2003	1.7 – 3.6
Influenza	US / Europe 1918	2.2 – 2.9
Measles	UK / US 20th century	12.0 – 18.0

A range of values is typically given due to challenges with estimating R. Some of these challenges depend on the method of estimation and have been well discussed elsewhere:

- Case-based – affected by the number of tests carried out as well as reporting delays
- Hospitalisations-based – subject to a delay in people being hospitalised, reporting delays, and possibly uncertainty about infection
- Deaths-based – subject to delays in occurrence and reporting delays, as well as difficulties identifying cause of death.

[1] Zimmer K, The Scientist, "Why R0 Is Problematic for Predicting COVID-19 Spread" [https://www.the-scientist.com/features/why-r0-is-problematic-for-predicting-covid-19-spread-67690]

This note focuses on a different challenge. R is the average (mean) number of infections per infected person, but the mean is just one feature of a complex probability distribution. Other features of this probability distribution are also important to understanding the nature of COVID-19 transmission, and those features are the focus of this note.

Jargon Buster

Some terms used in the rest of this note are defined here:

Probability distribution – gives the probability for a range of possible outcomes (e.g. it can be used to answer questions like what is the probability someone is taller than 2m). The different shapes of distributions are key – in particular what sort of 'extreme behaviour' is implied.

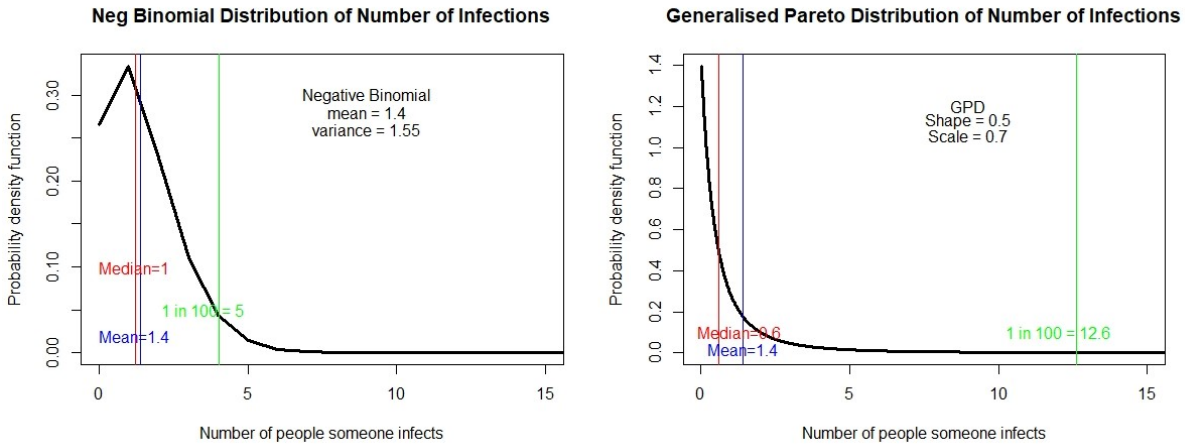
Mean – The average of the data i.e. the sum of the data, divided by the number of data points.

Median – The mid-point of the data (50% probability of data being below or above this point).

Variance – A measure of dispersion, ie (sort of) the extent to which extremes can occur.

The infection rate probability distribution

As an example of the different features of a probability distribution that can affect transmission, the Negative Binomial distribution (NBD) and Generalised Pareto distributions (GPD) shown below both have the same mean (R=1.4). Whilst the means are the same, the dispersion is clearly very different.



For this NBD the mean and median are very similar. The variance (1.55) is just slightly larger than the mean. Towards the extreme, one person in 100 will infect five people, though this is not enormously different from the mean number of infections.

For the GPD presented, the mean is more than twice the median. Half the infected people infect 0.6 people or less; but one person in 100 will infect 12.6 people. This higher dispersion appears to better capture the super spreader events we see in practice for COVID-19[2].

One reason R is difficult to estimate is that the number of people each person transmits the virus to is difficult to track. However, it is possible to use "track and trace" data to get an estimate for the type

[2] Norman J, Bar-Yam Y, and Taleb N, Systemic risk of pandemic via novel pathogens – Coronavirus: A note, New England Complex Systems Institute [https://necsi.edu/systemic-risk-of-pandemic-via-novel-pathogens-coronavirus-a-note]

of distribution. Wang[3] investigated R and its dispersion using an NBD (mean of 1.23, variance of 8.31). Cave[4] gave some indication of the type of distribution that might be appropriate, mentioning the famous "Pareto principle" and indicating that 10% of cases are responsible for 80% of transmissions[5]. This type of transmission is consistent with the GPD, which encompasses distributions with infinite variance that are far more volatile.

The "Pareto principle" values have a direct link to the shape parameter used in the GPD[6]. They are each measures of "tail fatness" for probability distributions.

The GPD distribution is especially important in statistics as the tail of most probability distributions can be modelled by a GPD[7]. A 'fat' tail means we have a large number of extreme cases. The table below shows the GPD shape parameter that can be used to model the tail of a range of distributions.

GPD shape parameter	Example Distribution	Tail fatness
-1	Uniform	No tail
-1 to 0	Beta	Finite right end point (maximum is limited)
0	Normal, Gamma	Thin tail
0 to 1	Student T	Moderate to fat tail
1	Cauchy	Very fat tail

We can estimate the GPD shape parameter at approximately 0.3 to 0.6 using the Pareto Principle values quoted above. If the shape parameter is 0.5 or above the probability distribution has infinite variance, which can make the results much more complex and explain some of the counter-intuitive results seen in pandemics.

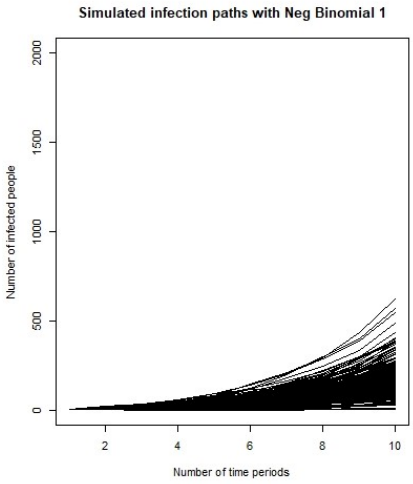
In the next section the GPD distribution is compared to two instances of the NBD to see the impact the differences would have on a pandemic outbreak.

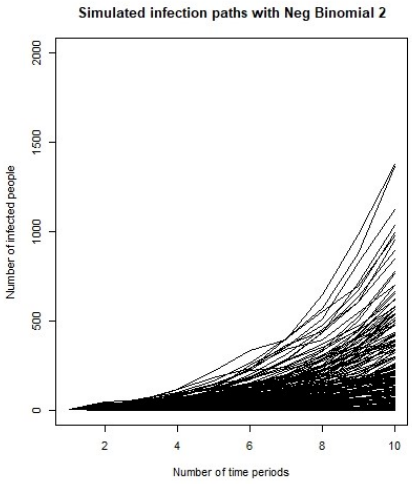
Simulation examples

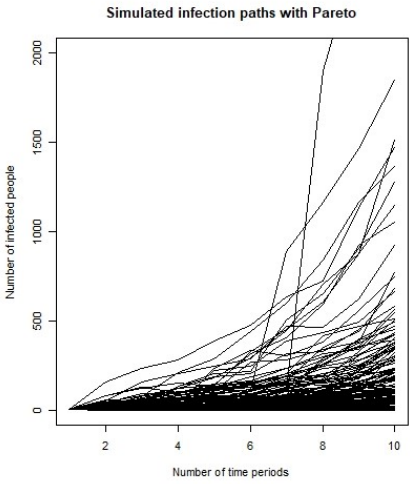
To examine the impact of different distribution shapes on an outbreak, with the same R value of 1.4, a simulation study is produced using both probability distributions described in the previous section[8] as well as another NBD (labelled "Neg Binomial 2") similar to that used by Wang[9].

Each simulation is a different outbreak, and 200 simulations are considered for each probability distribution type. The numbers of infected people are captured over 10 periods in the plots below:

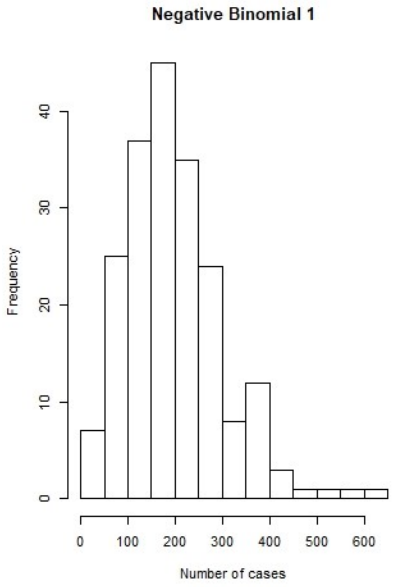
[3] Wang, L., Didelot, X., Yang, J. et al. Inference of person-to-person transmission of COVID-19 reveals hidden super-spreading events during the early outbreak phase. [https://www.nature.com/articles/s41467-020-18836-4]
[4] Cave E, Nature Public Health Emergency Collection, "COVID-19 Super-spreaders: Definitional Quandaries and Implications" [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7229875/]
[5] The same paper quotes 20% of cases being responsible for 80% of transmissions as a more general rule.
[6] Sharpe J, Juarez M, arXiv.org "Calibration of the Pareto and related distributions—a reference-intrinsic approach" [https://arxiv.org/pdf/1911.10117.pdf]
[7] McNeil A, Frey R, Embrechts P "Quantitative Risk Management"
[8] Negative Binomial with mean 1.4 and variance 1.55; GPD shape = 0.5 and scale = 0.7 (infinite variance)
[9] Negative Binomial with mean 1.4 and variance 8.2

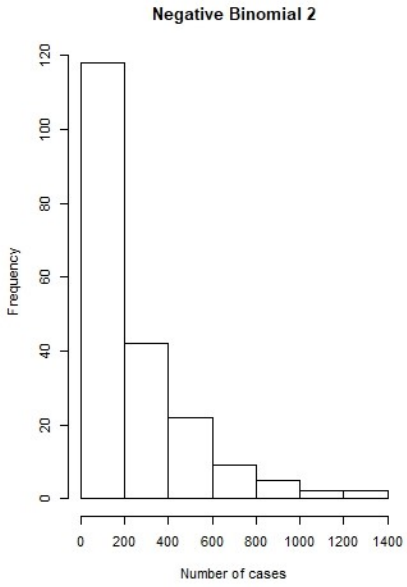


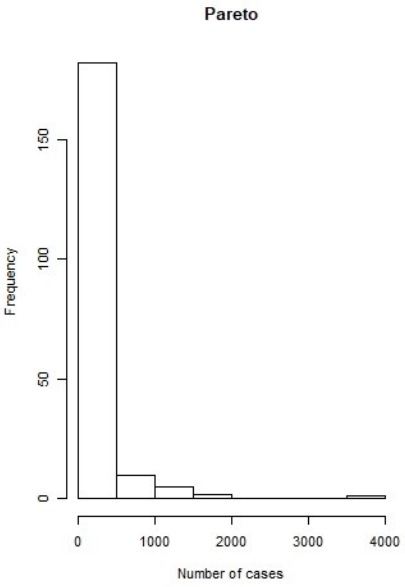




The distributions of the number of people infected at the end of each simulation are shown below. Note the different scales necessary for illustrating the number of cases.







The NBD cases on the left are relatively stable around a central value; none of the outbreaks go to zero, but very few spike high. The fatter tailed NBD in the middle includes more varied behaviour with some significant spikes as well as some cases disappearing. The GPD data has many outbreaks that do go to zero, with a small number that explode to very high levels.

In practice the behaviour seen with the higher variance NBD and the GPD is in line with experience. Outbreaks can quickly disappear in many cases, whilst elsewhere explode extremely fast (for example the contrasting experience of Lombardy and regions in the south of Italy). These counter intuitive results are often a source of confusion when analysing results, as it can be difficult to understand why some outbreaks explode and others disappear.

Stopping super spreaders – bringing R below 1

The key feature that causes the immense difference in results seen in the simulations above is the presence of super spreaders and super spreader events.

There are many reasons for the existence of super spreaders:

- Some people have close contacts with many people; hospital porters go from room to room in a hospital; delivery drivers have contact with many households; shop assistants see dozens of customers, etc.
- Some people may shed more virus and so infect a greater proportion of the people that they come into contact with.
- Some situations may give rise to superspreading events; for example, noisy indoor venues with low ventilation, or meat packing factories with high density production lines and cold temperatures.

These risk factors for superspreading may also operate in combination.

Cutting out superspreading could be disproportionately effective in our efforts to contain the pandemic. In statistical terms we are cutting off the right tail of the GPD, not just getting R below 1, but getting the GPD shape parameter below zero.

Given the multiple drivers of superspreading, multiple strategies can be combined to counter it. Identify key roles in society where the risk of super spreading is higher, and ensure these roles follow strict distancing and hygiene standards. Identify situations where superspreading occurs, and focus restrictions (or mitigations against transmission) on these. Evolve the tracing strategy to identify and focus on those who have passed on the virus (and have a high chance of being super spreaders) and trace their contacts.

Some of these strategies are already key parts of the global responses to COVID-19, but a greater focus on strategies which combat superspreading may enable a better balance between COVID-19 suppression and economic damage limitation.

Summary and conclusions

The distribution shape for the number of people each person infects is extremely fat tailed. This statistical property has real-world implications for the management of disease outbreaks, helping our understanding of how best to reduce R below one (i.e. targeting super spreaders).

Actuaries and statisticians modelling the pandemic would benefit from the literature on fat-tailed distributions and their application to pandemics.[10] [11] [12]

[10] Embrechts P, Kluppelberg C, Mikosch T "Modelling Extremal Events"
[11] Taleb N "Statistical consequences of fat tails"
[12] Cirillo, P., Taleb, N.N. "Tail risk of contagious diseases". [https://www.nature.com/articles/s41567-020-0921-x]