The Covid-19 pandemic

Statistics, statistical concepts and perspectives

With the world in the grip of a pandemic of coronavirus disease (Covid-19), newspapers, TV news broadcasts, websites and social media are flooded with numbers. There are daily reports of cases, treatments and deaths – plus the subsequent analyses – to digest, but much of the reporting also discusses concepts such as SIR models, case-fatality ratios, transmission rates and basic reproduction numbers – ideas that readers, viewers, listeners might be unfamiliar with.

It can all be, in a word, overwhelming.

As a publication committed to explaining statistical ideas and concepts, we set out to help readers understand what is going on, to make sense of the data they are confronted with, and to put important information into the appropriate context.

Contributors responded to our call “to explain the statistics of Covid-19”, and what follows is a selection of the many articles that have been published on our website – significancemagazine.com – since early April.

As one past contributor remarked recently, “This may be the most statistically relevant global crisis ever”. Many statisticians are doing important work to tackle the spread of the disease; others are fighting the spread of false information, which can be just as hazardous to health.

Our hope is that the articles that follow provide some clarity and insight at what is otherwise a confusing and concerning time.

Stay safe and keep well informed. ■

Brian Tarran

How do epidemiologists know how many people will get Covid-19?

Patrick Ball explains the SIR model

How many people are infected now with the novel coronavirus, SARS-CoV-2? How many will be infected tomorrow? These are hard questions because we cannot just see who is infected, and we do not know how each infected person is interacting with others, either infecting them or not. Here I will explain the basic framework used to estimate how many people are likely to become sick, and how many will recover or die. There is some notation, but I will present each equation in words. At this level, we are reasoning more with logic than with mathematics.

Patrick Ball is director of research at the Human Rights Data Analysis Group.

Why this is not just arithmetic

In the United States (as of 9 April 2020), there is a severe shortage of tests for SARS-CoV-2, the virus that causes people to become sick with Covid-19. As a consequence, even people in hospital with obvious and severe symptoms of Covid-19 are rarely tested for the virus. Furthermore, it is by now becoming clearer that many, perhaps most, people infected with SARS-CoV-2 have mild or no symptoms. The combination of these factors means that only a small fraction of SARS-CoV-2 cases are ever confirmed by a positive test. In practice, the confirmed case counts tell us more about the availability and distribution of tests than about the prevalence of infections. Consequently, the true number of SARS-CoV-2-positive

individuals is necessarily much higher than the reported counts. But how much higher?

The total size of the infected population determines how many people will need critical care, and how many will ultimately die. It is therefore important for health-care planning, economic policy, and public communication to estimate the population prevalence of SARS-CoV-2 infection.

The foundation of epidemiological modelling is the SIR model. This model enables us to estimate the progression, day by day, of the sizes of three sub-populations: those who are susceptible (*S*), infectious (*I*), and removed (*R*).

On each day, some of those who are susceptible get infected, and themselves become infectious to other people. The number of people who become infected is a combination of how contagious the disease is, as well as the effectiveness of social distancing, handwashing, and other practices that can limit transmission.

Similarly, on each day, some of those who are infected will recover or die, and these people are no longer at risk of transmitting the disease. We call these people “removed”, in the sense that they are no longer part of the infected population.

This model is *dynamic*, which means that it changes over time. To understand what is happening today, each component of the model depends on what was happening yesterday with the other components.

The notation and meaning

Scientists often convert word problems into notation to see connections among the pieces and then do calculations. Here is how the pieces of the SIR model fit together.

On any given day, the population can be divided into the people still susceptible to the virus, the people infected with the virus, and the people who have been removed, that is, either recovered or died. We will write this as $N = S_t + I_t + R_t$, where the subscript *t* means “this day”. Naturally, that means that the subscript *t* – 1 means “the day before this day”. Keep in mind that the *R* term includes both people no longer sick and those who have died.

Deaths

We can observe the number of deaths on each day: D_t . This is pretty much the only part of the pandemic we can measure without a lot of uncertainty. (There is still a little error because not all deaths due to Covid-19 are reported correctly, and some deaths that are not due to Covid-19 might be reported in error.)

FIGURE 1 Deaths, susceptible, infected, and removed in a hypothetical population. From Johndrow et al.¹

Susceptible (*S*)

“Susceptible” in this context means “available to become infected”. The number of people (in a population *N*) who are susceptible on day *t* equals the number susceptible yesterday minus the people who are newly infected today (v_t): $S_t = S_{t-1} - v_t$. Note that as long as people can only be infected once, the number of susceptible people can only go down.

Recovered

This is not usually defined as a separate term, but it is worth noticing that the number of people who recover and are no longer sick each day is some fraction (γ) of all the people infected up to and including yesterday, minus all the people who died up to and including yesterday: $\gamma(I_{t-1} - D_{t-1})$. The γ term is a proportion between 0 and 1, and it tells us how many people are recovering. Most SIR models lump together the recovered and dead, but I think it is useful to think about them as separate processes. I will use this definition of “recovered” in the next two definitions.

Infected (*I*)

The number of infected people on day *t* equals the number infected yesterday, plus the people newly infected today, minus the people who died yesterday, minus the people who recovered yesterday (note the recovered term at the end): $I_t = I_{t-1} + v_t - D_{t-1} - \gamma(I_{t-1} - D_{t-1})$. This number can go up as new infections occur and down as people recover or die.

Removed (*R*)

The number of people removed on day *t* equals the number removed yesterday, plus the people who died yesterday, plus the people who recovered yesterday (note the recovered term at the end): $R_t = R_{t-1} + D_{t-1} + \gamma(I_{t-1} - D_{t-1})$. Note that as long as people can only be infected once, the number of removed people can only go up.

Newly infected

The number of new infections each day is equal to the number of susceptible people yesterday times some fraction (β) of the infected proportion. The infected proportion is simply the proportion of the population who were infected yesterday:

$$v_t = S_{t-1} \beta \frac{I_{t-1}}{N}$$

The β term is a measure of infectiousness. β is mathematically related to the epidemiological term R_0 , which is the average number of people that each infected person newly infects. I have shown it here as β rather than R_0 so that the role of “infectiousness” can be clear in the mathematics relating the susceptible and infected populations.

Infected fatality rate

The fraction of the infected people who will eventually die equals the total deaths (summing all the daily totals) divided by the total people ever infected:

$$p = \frac{\sum_{t=0}^t D_t}{I_t + R_t}$$

This rate may change as health care is overloaded, or as new treatments are discovered. It is different for people of different ages and different “co-morbidities” such as diabetes, hypertension, and smoking. There is still considerable debate about the p number, which is usually expressed as a percentage. Most sources report that p seems to be between 0.5% and 2.0%, averaged across various studies, various age groups, and various co-morbidities.

Putting it all together

We can estimate the SIR values by using the relationships among them and information from clinical studies. The *New York Times* has an interactive online tool (nyti.ms/3cFqpNf) which allows a user to see immediately the effect of changing any of these parameters.

Covid-19 can be understood as a generic epidemic which has its own values of these parameters. Of course, policy and human behavior influence the parameters too, by reducing transmission through social distancing, reducing fatalities through better treatment, and ultimately, reducing susceptibility through a vaccine. The processes over time can be seen in Figure 1 (page 13).

The top graph shows the number of deaths each day, which is the only measure we can really observe. The middle graph shows the number of new infections each day. Note that it leads the deaths by a couple of weeks, and the length of this lead is another variable in the model which can only be observed through limited clinical studies. The bottom graph shows the SIR values: the susceptible population (green line) starts with everyone and declines over time. The infected group (orange line) rises for a while then slowly declines. The removed population (blue line) rises over time, eventually including everyone – the dead and the survivors.

Different modelling projects approach this framework differently. Which parts are assumed, measured, or modelled vary among different studies. Some models let the interactive user guess different values of R_0 (and thereby β) or p (and thereby γ), while other models incorporate measures from small clinical studies. Some models include an intervening term, *exposure*, between the susceptible population and infection (not all susceptible people are exposed, and the

unexposed people cannot become infected).

The mathematics that connects all the pieces is also different in different models.

In the long term, we will learn which models were best. However, time is too short for more than a tiny number of these models to be subject to formal peer review in time to be relevant. That means it is more important than ever that engaged laypeople (especially journalists) have at least a minimum sense of how to read these essential studies. ■

Reference

1. Johndrow, J., Lum, K. and Ball, P. (2020) Estimating the number of SARS-CoV-2 infections and the impact of social distancing in the United States. Preprint, [arXiv:2004.02605v1](https://arxiv.org/abs/2004.02605v1). bit.ly/2Yy1zLI

Further reading on the SIR model

- Kermack, W. and McKendrick, A. (1927) A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London Series A: Mathematical and Physical Sciences*, 115, 700–721.
- Anderson, R. M. and May, R.M. (1991) *Infectious Diseases of Humans*. Oxford: Oxford University Press.

Acknowledgements

I am interpreting these ideas from a paper by James Johndrow, Kristian Lum, and me.¹ I am grateful for their comments, as well as for suggestions from Maria Gargiulo, Megan Price, Tarak Shah, Stella Pierce, Danielle Fugere, Hope Howard, and Jacob Nelson.

Why we need more coronavirus tests than we think we need

James J. Cochran on the importance of testing a random sample

In the United States (as of 9 April 2020), President Donald Trump has said that testing for novel coronavirus infection will be limited to people who believe they may be infected. But if we only test people who believe they may be infected, we cannot understand how deep the virus has reached into the population. The only way this could work is if those who believe they may be infected are representative of the population with respect to novel coronavirus infection. Does anyone believe this is so?

The common characteristic of those who believe they may be infected is that they all show some outward symptoms of infection by the virus. In other words, people who are being tested for the novel coronavirus are disproportionately showing severe symptoms.

This would not be a problem if someone who is infected by the novel coronavirus immediately shows symptoms, but this is not the case. We have strong evidence that some people

develop mild cases, show no symptoms, and carry the virus without knowing it because they are asymptomatic. Thus, efforts to understand the virus’s penetration into the population must include observation of the asymptomatic.

The estimate of the proportion of the population who are infected can be calculated as:

$$\hat{p} = \frac{\text{number of symptomatic infections} + \text{number of asymptomatic infections}}{\text{number of symptomatic infections} + \text{number of asymptomatic infections} + \text{number not infected}}$$

So, we need data from a random sample of the entire population in order to gather data from infected people who are showing symptoms, infected people who are asymptomatic, and people who are not infected. All have some probability of being included in a true random sample of the population.

As of 23 April, leaders in Germany and New York State (see bit.ly/2Kp2iXd and dailym.ai/3bxZ5Au) had moved to implement random testing to assess how widespread the virus is, but there has been resistance from leaders elsewhere. This could be due to ignorance, disregard, or lack of appreciation of

James J. Cochran is associate dean for research, professor of applied statistics, and the Rogers-Spivey faculty fellow at the Culverhouse College of Business, University of Alabama. He is vice-chair of the *Significance* Editorial Board.