Math at the frontlines of the covid war

Pavel Etingof (MIT)

August 12, 2021

Contents

1. Introduction

Ever since the arrival of covid-19, we have been showered with various numbers from all sides. People have been using them to justify their views on the pandemic response. As this conversation has become unduly politicized, some people have been misusing numbers to promote all kinds of questionable theories about covid and misleading others and themselves.

- Covid is as bad as a bad flu, or just a tiny bit worse.
- Masks and social distancing don't really work.
- Lockdowns aren't effective.
- Vaccines are dangerous and don't help much.
- We are already at herd immunity.

And so on.

The numbers they quote are usually correct, but they often give them out of context, to support their preconceived opinions.

This mini-course is about learning how to use guesstimation and other kinds of math to interpret covid-related and other numbers you see in the news and social networks, and use them to correctly assess risks and make rational decisions.

The first part will use only very elementary math, and will be accessible to middle schoolers. The second part, which concerns the SIR model, will be a bit more advanced and involve some high school math, such as calculus.

Disclaimer 1: There will be **only math** and **no biology** (in fact, I don't know any!).

Disclaimer 2: We will not be able to actually debunk any theories by covid deniers or antivaxxers so that they would concede. Any math we do is necessarily based on some non-mathematical assumptions you have to believe, and they will question these assumptions rather than the math. There is nothing we can do about this as mathematicians; ultimately we have to develop our own views based on some beliefs, common sense and the math. In this mini-course we will only worry about getting the math right.

2. The effective contact rate, infectious rate, and basic reproduction ratio

The spread of infection during an epidemic is characterized by how various quantities, such as

- the fraction S of susceptible people,
- the fraction I of infectious people,
- the fraction B of people who have had the infection, etc.

change in time. This depends on several parameters.

2.1. **Effective contact rate.** The first parameter β is called the **effective contact rate**. It is the reciprocal

$$\beta = \frac{1}{t_c}$$

of the average time t_c between two "effective" contacts of an infectious person (i.e., ones which result in this person infecting someone else).

We have

$$\beta = \beta_* p,$$

where β_* is the suitably normalized number of contacts between people per unit of time and p is the probability that infection will be passed during a contact between an infectious person and a susceptible one, called the **secondary attack rate**.

2.2. **Infectious rate.** The second parameter γ is called the **infectious rate**. It is the reciprocal

$$\gamma = \frac{1}{t_r}$$

of the average time t_r an infected person remains infectious, i.e., the average time it takes an infectious person to be removed from the epidemic through recovery or death. So we assume that a recovered person acquires immunity and cannot get sick again.[1]

2.3. **Basic reproduction ratio.** It is convenient to rescale time by replacing t by the dimensionless quantity

$$T = \gamma t = \frac{t}{t_r}.$$

In other words, we take t_r as the natural unit of time. Then it turns out that the dynamics of the infection spread depends only on the ratio

$$R_0 := \frac{t_r}{t_c} = \frac{\beta}{\gamma} = \frac{\beta_* p}{\gamma} = \beta_* \cdot p \cdot t_r.$$

So R_0 is the average number of people infected by a sick person. It is the **basic reproduction ratio**, one of the main parameters in epidemiology.

[1]Note that this is not quite so for many infections, including covid-19.

2.4. **Exponential growth.** At the beginning of the epidemic, the infection grows exponentially with exponent $c := R_0 - 1$:

$$I(T) = I_0 e^{cT}, \tag{1}$$

although later the growth slows down as the epidemic "runs out of fuel". This is similar to computing **compound interest** on a bank account, with c playing the role of the interest rate.

In more detail, consider first the simplified discrete model, when "interest" is computed every unit of time. So at $T = 1$ the people infectious at $T = 0$ stop being infectious, but each of them is replaced, on average, by $R_0 = 1+c$ others. So we have $I(1) = I_0 R_0$, $I(2) = I_0 R_0^2$, etc., i.e.,

$$I(T) = I_0 R_0^T$$

for positive integer T.

However, this is only an approximation since new infections appear continuously, not only at integer values of T. We will get a better approximation if we compute the "interest" not with interval 1 but more often, with interval $1/2$. In time $1/2$, about half of the infectious people at $T = 0$ will stop being infectious, with each replaced by R_0 others. Thus

$$I(\tfrac{1}{2}) = \tfrac{I_0}{2} + I_0 \tfrac{R_0}{2} = I_0(1 + \tfrac{c}{2}).$$

So $I(1) = I_0(1 + \frac{c}{2})^2$ and in general

$$I(T) = I_0(1 + \tfrac{c}{2})^{2T},$$

where T is now a half-integer. More generally, if the "interest" is computed with intervals $1/k$ for a positive integer k, we get

$$I(\tfrac{1}{k}) = I_0(1 - \tfrac{1}{k}) + I_0 \tfrac{R_0}{k} = I_0(1 + \tfrac{c}{k}),$$

so

$$I(T) = I_0(1 + \tfrac{c}{k})^{kT},$$

where T is now a multiple of $1/k$. In the limit $k \to \infty$ (which corresponds to computing the "interest" continuously), we have

$$(1 + \tfrac{c}{k})^k \to e^c,$$

so we get (1).

Thus for infection to be able to spread widely and cause an epidemic, one must have $R_0 > 1$.

Example 1. For real infectious diseases R_0 normally varies from 1.2 (seasonal flu) to 18 (measles). The list of values of R_0 for various diseases can be found at `https://en.wikipedia.org/wiki/Basic_reproduction_number`. According to this list, the original covid strain had $R_0 \approx 3$, while for the delta variant R_0 is between 8 and 9 (according

to the latest CDC data). The only common diseases whose R_0 is higher are measles, mumps and chickenpox.

The number R_0, however, is not an absolute constant for a given disease, and it is hard to define, let alone compute, precisely. It depends on many things, such as the density of population, people's lifestyle and habits in a given country, etc. The table gives a range of values of R_0 for each disease which characterize the spread of the infection **without any intervention.** In reality, once an epidemic begins, people change their behavior, trying to mitigate the spread of infection. The goal of **mitigation measures**, such as lockdowns, social distancing, mask mandates, isolation and quarantine is to reduce R_0. Namely, **lockdowns** reduce β_*, **masking and distancing** reduce p, and **isolation and quarantine** reduce t_r, i.e., increase γ.

2.5. **Effective reproduction ratio.** The number R_0 is the reproduction ratio of the infection at the beginning of the epidemic, assuming all the population is susceptible ($S = 1$). If a proportion of the population has immunity through having had the disease or through vaccination ($S < 1$) then R_0 is replaced by the **effective reproduction ratio** $R_e = R_0 S$ (i.e., R_0 is multiplied by the probability that a given contact is susceptible).

2.6. **Herd immunity.** In particular, when $S = \frac{1}{R_0}$, we get $R_e = 1$, so the infection can no longer expand, and for larger S it starts to wane. This happens when the fraction of immune people is

$$H = 1 - \frac{1}{R_0}.$$

This is called the **herd immunity threshold**.

Herd immunity can be reached by enough people getting infected, or through vaccination, or both. However, for high R_0, when H is close to 1, it is hard to reach by vaccination alone, since there are many people resisting vaccination, and since vaccines do not provide 100% immunity.

Example 2. For the seasonal flu $H \approx 1 - \frac{1}{1.2} = \frac{1}{6} \approx 17\%$, for the original covid strain $H \approx 1 - \frac{1}{3} = \frac{2}{3} \approx 67\%$, while for the delta variant $H \approx 1 - \frac{1}{8} = \frac{7}{8} = 87.5\%$. Thus, we may not be able to reach herd immunity from the delta variant through vaccination alone.

However, even after the herd immunity is reached, the infection continues to spread, albeit at a waning rate. So the proportion of people who will eventually get sick (without mitigation) is larger than H (although, curiously, it is less than 1!). We will see that when R_0 is

close to 1, the share of people who will ultimately get sick is about $2H$. For example, for seasonal flu it is about 32%, which for the US is about 100 million. The actual number of flu cases seen every year is smaller (about three times, on average), because of flu shots, weakening of the virus as the weather warms up, and other reasons.

Finally, an important caveat: when we talk about herd immunity, we assume that the geographic and social distribution of the immune portion of the population are uniform, which is often not the case. If in some parts of the country the percentage of immune population is lower than H, the infection will continue to spread in those areas, the faster the lower this percentage. The same applies if non-immune people tend to hang out with each other, e.g. because attitude to vaccination is correlated to the social stratum.

This is what is happening now in areas with low vaccination rates. For example, on August 5, 2021, the 7-day average daily case count in Florida reached 18,120, while in Massachusetts it was 906.[2]. But the Massachusetts' population is only 3 times less than Florida's. Thus the case count per capita in Florida on that day (averaged over the preceding week) was 6 times higher than in Massachusetts.

2.7. **Doubling time.** Finally, a good characteristic of the epidemic at the growth stage is the **doubling time** t_d of the infection. It solves the equation

$$e^{c\frac{t_d}{t_r}} = 2,$$

which yields

$$t_d = \frac{t_r \log 2}{c} \approx \frac{0.7 t_r}{c}.$$

For example, for the original strain of covid (R_0=3) and $t_r = 7$ days (as was in the first wave), we get

$$t_d \approx 2.5 \text{ days}.$$

This is what was observed in New York City during the exponential growth period.

Note that t_d is easy to measure from available data, and one can also estimate t_r (the time a sick person can keep infecting others). So we can use this equation to compute R_0 (or R_e):

$$R_0 \approx 1 + 0.7\frac{t_r}{t_d}.$$

[2]The 7-day average is a more robust quantity than the daily case count since on weekends a lot fewer cases are recorded.

For example, recent data from Florida shows that the 7-day average of the number of new infections doubled in 7 days (July 10-July 17, 2021). So we get $R_e = 1.7$. Now t_d has increased to 14 days, so we get $R_e \approx 1.35$. This is quite far from $R_0 \approx 8$ for the delta variant, but we should remember that there are many mitigation measures: masks, social distancing, testing, isolation, quarantine, many immune people, etc.

3. CFR and IFR

3.1. **The CFR.** The **case fatality rate** (CFR) of an infection is

$$\text{CFR} := \frac{\text{number of confirmed } \mathbf{deaths}}{\text{number of confirmed } \mathbf{cases}}.$$

The CFR is normally easy to compute because both of these numbers are readily available.

Example 3. For instance, for covid all this data is on the website `https://www.worldometers.info/coronavirus/`. According to this website, there are about 200 million confirmed covid cases in the world thus far (as of August 5, 2021), and about 4.3 million fatalities. So we get CFR=2.15%. And we get about the same number for individual countries. E.g., for the US, about 36 million cases and about 635,000 deaths, so CFR=1.75%, and for the UK about 6 million cases and 130,000 deaths, so CFR=2.2%.[3]

The CFR, however, is not very useful, since both numbers (confirmed cases and confirmed deaths) can deviate considerably from reality.[4]

Example 4. The CFR for covid in NYC during the first wave (March-April 2020) was about 10%, while in the second wave (December 2020-January 2021) it was under 2%. Did they get better at treating covid? Somewhat, but not nearly this much. Most of this effect is due to very limited testing during the first wave (since a covid case is registered on the basis of a positive PCR test). So we see that during the first wave, *at most* **one out of five** cases was reported (in fact, we will see that it was even less than that).

But even if testing is readily available, many cases are not reported, because people just don't go to get tested. In fact, it is now

[3]These numbers have been going down and will continue to do so as more people are getting vaccinations. The CFR can also decrease as doctors get better at treating the disease.

[4]For example, if the same patient has two positive tests a few days apart, this may be recorded as two separate cases.

known that many cases of covid are asymptomatic, so people may not even know that they are sick! Estimates show that even now, when you can get tested for free in a pharmacy (although, importantly, with some wait time), every other covid case is probably missed, maybe even two out of three at the peaks: `https://www.npr.org/sections/health-shots/2021/02/06/964527835/why-the-pandemic-is-10-times-worse-than-you-think` (Exercise: why is the headline of this article misleading?). So during the first wave in NYC, it could be that as few as just **one out of ten** (or even fifteen) covid cases was reported.

The number of confirmed deaths is somewhat more reliable but not really solid either, because there can be different protocols of reporting a covid death.[5] Namely, since most covid deaths occur in very old and/or sick people, even in presence of a positive PCR test one may question if the patient died from covid or from other causes while being covid positive. This can also be affected by authorities who do not want people to see a high death toll (as happened in Russia where covid deaths have been underreported, likely for political reasons).[6]

Also, even though CFR for covid is currently about 2% in most countries, it varies from country to country.

Example 5. In Israel for covid CFR=0.73%, in United Arab Emirates 0.3%, while in the US it is 1.75% and in Italy 2.9%. Did the first two do so much better? Maybe somewhat, but not this much. To understand this discrepancy, it is important to remember that the vast majority of covid deaths occurs in people over 65. So we should look at the percentage of population in this age group. In the US we find 16.5%, in Italy 23.1%, while in Israel 12.4% and in UAE only 1.1%. This explains the discrepancy with UAE, especially given that UAE is at the very top in testing per capita (and it was so already early in the pandemic), so they have likely discovered a higher proportion of cases than other countries.

On the other hand, it still appears that Israel did better than the US and Italy. Maybe somewhat, but we should remember that Israel largely avoided the first wave of covid in March-April 2020 when tests were scarce, so they were able to detect a larger percentage of cases. That was not the case in the US in March-April 2020, and even less so

[5]E.g., according to `https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/burden.html`, only 1 in 1.3 actual covid deaths were reported in the US by May 31, 2021.

[6]We note that now in Russia there is other covid data coming from Rosstat which is more accurate.

in Italy, where there was a big wave early in the pandemic (February-March 2020) when very few tests were available. Only 1 in about 15 cases was detected in this wave.

3.2. **The IFR.** Thus, we see that CFR is indeed not a very good measure of severity of the disease. A much better one is the IFR, **infection fatality rate**,

$$\text{IFR} := \frac{\text{number of actual } \mathbf{deaths}}{\text{number of actual } \mathbf{cases}}.$$

This is still not very well defined though, since the definition of a covid case and covid death is subject to argument. Also, while it is not affected by availability of testing, by the same reason it is hard to compute, since at least the denominator is not easily available and has to be estimated.

This lack of information is widely used by covid deniers, who claim that the denominator is in fact much larger, and consequently IFR much smaller than people think. For the flu, the IFR is about 0.1% to 0.17%, so they claimed that covid is just as bad as the flu or just a tiny bit worse. As more and more data started to come out and it became increasingly difficult to defend this claim, they changed the narrative slightly and started saying that covid is about twice as bad as the flu, with IFR=0.3%. They used it as a basis for questioning mitigation measures, as we take no such drastic measures for the flu season.

This claim, however, has no merit for a number of reasons. One is that since covid is much more contagious than the flu (even the original strain, let alone the new variants, such as the delta variant), without mitigation a lot more people will get infected during the same period of time (in fact, in the US it is already more than the flu, even with pretty severe lockdowns, mitigation measures, etc.) But also this claim is based on two statements:

Statement 1. Covid cases are more underreported than is widely believed.

Statement 2. Covid deaths are overreported.

So how can one estimate the true numbers of cases and deaths and the IFR to verify these statements?

For the number of cases, covid deniers claimed that a lot more people actually got sick than reported. As we know, they are right about that, and the disagreement with mainstream experts is only to what extent it is so. One way to estimate that is **random antibody tests**.

Example 6. On April 23, 2020 the New York State announced that according to random antibody testing, 14% of its residents had antibodies for covid-19. By that day, NY had reported about 270,000 cases, for about 20 million population, i.e., the reported rate was about 1.35%. This means that slightly less than **one out of 10 cases** was reported (i.e., the **ascertainment bias** was about 10), and the true number of cases in NY state by that date estimates to 2.8M. On the other hand, the reported total number of covid deaths by that day was about 21,000. This gives IFR$\approx$ 0.75%, which is 4 to 7.5 times worse than the flu (not taking into account that it is also more contagious). This is about average among the numbers given in various studies (credible estimates vary from 0.5% to over 1%).

In fact, 14% may have been an overshoot since the antibody testing was held in stores, and people who go to stores are more likely to get sick. If so, the estimate for the IFR will be closer to 1%.

We can also try to guesstimate what percentage of population, say in NY state, have had covid so far. Our computation gives 2.8M until April 23, 2020. From April 23 to September 1, 2020 there were about 200,000 more reported cases, but the testing was ramped up, so let's say 1 out of 5 cases was found during this period. This gives another 1 million cases. Finally, since September 1, 2020, there were about 1.8 million registered cases, but the testing was adequate, so let us say about 1 out of 2 cases was found. This would give another 3.6 million. Thus we get a total of 7.4 million. The population of NY is roughly 20 million, so we get that about 42% of New Yorkers have had covid so far. With about 54,000 deaths so far in NY, we get IFR$\approx$ 0.73%, agreeing with our previous estimate of 0.75%.

A similar computation nationwide gives about

$$1 \cdot 10 + 5 \cdot 5 + 30 \cdot 2 = 95 \text{ million} ,$$

or $95/330 \approx 29\%$ of the US population. With the death toll $635,000$, this gives IFR$\approx$ 0.67%.

CDC at `https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/burden.html` gives a somewhat higher figure for the estimated total number of infections nationwide: 120 million (i.e., their estimate for the percentage of unreported cases is a bit larger than ours). But they also say only 1 in 1.3 covid deaths is reported. So their estimate for the IFR is 0.68%.

3.3. **Hospitalization rates.** One can similarly define the **case hospitalization rate** CHR and the **infection hospitalization rate** IHR:

$$\text{CHR} := \frac{\text{number of } \textbf{hospitalizations}}{\text{number of confirmed } \textbf{cases}},$$

$$\text{IHR} := \frac{\text{number of } \textbf{hospitalizations}}{\text{number of actual } \textbf{cases}}.$$

They have similar behavior to CFR and IFR, and for covid tend to be about 3.25 times higher. E.g. the website `https://gis.cdc.gov/grasp/covidnet/COVID19_5.html` suggests that there have been slightly over 2 million reported covid hospitalizations nationwide, about 3.25 times higher than the death toll. So we have IHR$\approx 0.75 \cdot 3.25\% = 2.44\%$. Also since there have been about 36 million reported cases (until the beginning of the new wave), we get CHR$\approx 5.7\%$.[7]

3.4. **Health care system overload.** Now imagine that we have a poorly controlled covid surge in the UK in which only 10% of the population (=6.7 million) gets infected and which lasts 3 months (this actually happened in the winter of 2020/2021). This means that on average there would be 70,000 cases a day, so twice as many, i.e. 140,000 at the peak. About 2.5% of these will need hospitalization, that's about 3,500. If a covid patient is in the hospital for 1 week on average[8] (see `https://www.nuffieldtrust.org.uk/resource/chart-of-the-week-how-long-do-covid-19-patients-spend-in-hospital`) then about 25,000 covid patients would have to be in UK hospitals at a time. At least 1/3 of these patients (8,000) will require ICU care. So this would overwhelm the health care system, as there are only 6,500 ICU/critical care beds in the UK (`https://link.springer.com/article/10.1007/s00134-012-2627-8/tables/2`). Thus without any mitigation at all the health care system will be overwhelmed by a factor of 5 or higher.

[7]Unfortunately, the total number of hospitalizations due to covid is reported only by some states, so the total over the US is not known precisely. Based on the data from `https://covidtracking.com/data`, the ratio of hospitalizations to deaths varies widely, in the range 1 to 6, but for most states it is between 2.5 and 4, which roughly agrees with the estimate of 2 million nationwide. Our estimate 3.25 is in the middle of that range. On the other hand, the site `https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/burden.html` estimates that 1 in 1.8 covid hospitalizations is recorded, and estimates the actual number of covid hospitalizations in the US as 6.2 million, which for reported hospitalizations yields $6.2/1.8 \approx 3.4$ million, which is at odds with the above estimate of 2 million.

[8]We observe that the length of stay decreased during covid waves; this is because people were discharged prematurely due to shortage of beds.

4. Excess mortality

Covid deniers also claim (using that the vast majority of covid deaths occur among very old and sick people) that covid deaths are overreported, namely many deaths from old age or other illnesses are reported as being from covid, when there is a positive covid test (Statement 2). At the beginning of the pandemic, they said that this would be obvious once the **excess mortality** data for a long enough period (such as a year or two) come in. Let us see if they were really right.

Excess mortality is the number of additional deaths in a given state or country compared to the normal (=average) number during a similar period. According to CDC, excess mortality since Feb 1, 2020 in the US is about 760,000 (`https://www.cdc.gov/nchs/nvss/vsrr/covid19/excess_deaths.htm`). At the same time, the number of reported covid deaths is about 635,000 (including probable ones). So did 125,000 covid deaths go unreported? Some of them, but not all. Some of this is due to overloading the health care system, postponement of treatments, hesitancy to go to the hospital for fear of being infected, change in lifestyle due to lockdowns, etc. (This 125,000, while a large number, is about 4% of annual deaths in the US). The part of this 125,000 that should be directly attributed to covid depends on the exact definition of a covid death. In any case, it does not look like any over-reporting took place (more likely, covid deaths in the US were under-reported, if anything; indeed, as mentioned above, CDC estimates that only 1 out of 1.3 covid deaths in the US is reported). If we assume that half of these are under-reported covid deaths (probably most having occurred early when tests were scarce), i.e., 10% under-reporting, we will get 700,000 deaths from covid, and our estimate 0.75% for the IFR should be upgraded to 0.83%. Note that this is about 0.2% of the US population (330 million).[9]

On the other hand, if we look at Russia, they report about 135,000 deaths from covid by July 1, 2021, while the excess mortality by that date, according to the recent data of Rosstat, is about 557,000. So we see that covid deaths in Russia are severely underreported (at least 3 times). Also given that Russia's population is about 145 million, this is about 0.4% of the population, **twice as much as in the US**. This is due to low vaccination and rampant disregard for mitigation measures (such as mask wearing).[10]

[9]One should also take into account that covid measures reduced flu activity to very low levels, thereby reducing mortality. This makes covid mortality a bit higher.

[10]Using the estimate of 0.75% for the IFR, we get that about 50% of the population in Russia have had covid-19. For Israel this computation gives about 13%,

What covid deniers should really do now to save their face is upgrade the "long term" for excess mortality from 1-2 years to 10 years or more. Hopefully the pandemic will subside by then, and since most covid fatalities are very old and sick people, they would likely not have made it to the end of this term anyway. So the effect of covid will be hard to see. In fact, by the same token, excess mortality may even go *negative* after the pandemic ends, so they may argue that covid is a good thing after all!

5. Risk estimation

You can use back-of-the-envelope calculations of this sort to estimate the risk of a given activity with regard to catching covid-19. For example, suppose you ride in a train car in Massachusetts on May 18, 2021. Suppose there are 100 people in this car. What is the chance that none of them was infectious with covid-19, assuming that a covid patient is contagious for 10 days?

Well, you can look up 7-day average the number of new cases in Massachusetts on that day. It is 620. So for 10 days it is about 6,200. These people are no problem since they likely won't be in the subway car (they are quarantined). However, we know that there were likely at least as many, possibly twice as many others who are not (yet) detected. Let us say, 9,000. This is about every 750-th person. So the probability is

$$p_1 \approx \left(1 - \frac{1}{750}\right)^{100} \approx \exp(-\tfrac{1}{7.5}) \approx 0.88.$$

So this is relatively safe (12% chance of contact).

On the other hand, on April 13, 2021 the 7-day average number of new cases was about 2,100, about 3.4 times higher. So the probability will now be

$$p_2 \approx (0.88)^{3.4} \approx 0.65.$$

So there is more than 1 in 3 chance that there is a contagious person in the car, between any two stops of the train, (35%).

Another question is whether this person will infect you, or not. This depends on how far away you are, how long you are exposed, and whether any of you or both wear a mask (which was required at that time). If you both do, this may well not happen.

Example 7. How safe is a 100-person camp starting August 14, 2021 (say, with all participants from Massachusetts, for simplicity)?

---

but we should correct for the demographics by multiplying by about 4/3, which gives 17%. For the US, as we have computed, this percentage is about 29%.

We look up the 7-day average of new cases on that day, it is about 1,200. So over 10 days it's 12,000. So we should assume that there are 18,000 contagious people walking around, about 1 in 400. So the probability of having no contagious person in the camp is

$$p_3 = (1 - \tfrac{1}{400})^{100} \approx 0.78,$$

i.e., there is a 22% chance of an infectious person in the camp. However, CDC says that a vaccinated person is 3 times less likely to be infected. So if everyone in the camp is vaccinated, we get

$$p_4 \approx 0.78^{1/3} \approx 0.92,$$

i.e., 8% chance of an infected person in the camp. So to make it safer, the camp may require PCR tests for participants, possibly also rapid tests at the entrance. This will lower the chance significantly (at least by a factor of 2), bringing the risk to 4% or less. And if participants of the camp are not random and mostly come from families of people who work from home and don't go to big gatherings and/or live in parts of the state with lower incidence of covid, then this will be further diminished.

6. Incubation period

There is, however, an aspect in which this calculation is too optimistic if a surge is just beginning. The issue is that covid, like many infectious diseases, has an **incubation period**, and also it takes time for a test to get through. For this reason, infection cases (those which do get discovered) are discovered with a delay. This is shown clearly on Chart 7 in `https://tomaspueyo.medium.com/coronavirus-act-today-or-people-will-die-f4d3d9cd99ca`, where the delay for covid is taken to be 8 days. We have seen that for $R_0 = 3$ and $t_r = 7$ days, the doubling time is 2.5 days, so during this delay period the number of infections grows more than 8-fold. Thus when we look at the number of new cases, we should multiply this number by 8. This means that during this period, **for every fatality we have about 1,000 sick people walking around** (most of them yet undiscovered!)

In the camp safety problem, the doubling time in Massachusetts is about 14 days, so we should multiply the number of actual cases only by about 1.5. This increases the risk roughly by the same factor.

7. Shelter-in-place orders

During the covid pandemic many cities and countries issued shelter-in-place orders. For example, in March 2020 Israel issued an order that

people should not go further from their house than 100 meters. Why is this helpful, and how much?

To understand this, we may use a very simple model. First suppose that people ride a subway in a big city. An infectious person remains infectious only one day, but during this day infects four neighbors on the subway car. Suppose one infectious person comes to the city. How many people will be infected in one week? In n days (when n is not too large) we get $N = 4^n$, so in a week $N = 4^7 \approx 16,000$. (Basically, we have the discrete time model with $R_0 = 4$ and $t_r = 1$).

On the other hand, assume that the city is under a shelter in place order. We will model it by people sitting at vertices of a square lattice (points with integer coordinates), and every day an infected person infects four neighbors at distance ≤ 5 from that person. (of course, if one of them is already infected, nothing happens). How many people will be infected in one week from one person at $(0, 0)$? In n days the infected people will be restricted to the disk of radius $5n$ around the origin, and the number of points in this disk is about $\pi \cdot (5n)^2$, which gives an upper bound for the number of people infected. In particular, in one week we get $\pi \cdot 35^2 < 4,000$.

We see that the infection spreads much slower under the shelter-in-place order (it grows quadratically, rather than exponentially). This is because after a while, many of the neighbors being infected at each step are already infected, so in reality each infected person infects less than 4 people (so $R_0 < 4$ and decreases with n, approaching 1 for large n).

8. Variants

Biologists explain that a virus mutates towards higher transmissibility (and this is what happened and continues to happen with the coronavirus). Why does this happen?

Mutation is a random process, and on a rare occasion it produces, purely by chance, a more transmissible variant. And if this variant manages to spread widely enough, it will beat the one in circulation and become prevalent. To understand why this happens and how quickly, let us consider the following simple model.

Suppose the existing variant has $R_e = 1.5$, while the new one has $R_e = 2$. Suppose initially the new variant comprises only 1% of cases. Suppose both variants grow exponentially. How fast will the new variant become prevalent?

After (rescaled) time T, the first variant grows by $e^{0.5T}$, while the second one by e^T. So the proportion of the second variant in time T

will be approximately

$$P = \frac{0.01e^T}{0.01e^T + 0.99e^{0.5T}} = \frac{1}{1 + 99e^{-0.5T}}.$$

So the value of T when the second variant becomes prevalent is a solution of the equation

$$99e^{-0.5T} = 1,$$

i.e.,

$$T = 2\log 99 \approx 9.2.$$

If $t_r = 7$ days, we get that the second variant becomes 50% prevalent in about 9 weeks, i.e., about 2 months.

What about 90% prevalence? This happens when

$$99e^{-0.5T} = \frac{1}{0.9} - 1 = \frac{1}{9}$$

so

$$T = 2\log 891 \approx 13.6.$$

So this happens in another month.

9. Vaccine safety

Adverse effects of vaccination can be reported to VAERS (Vaccine Adverse Effects Reporting System). Since covid vaccinations started, this system has recorded a sharp increase in the number of reported serious complications and deaths. In particular, the system has recorded 1,736 deaths in the US on days 0,1,2 after vaccination. Antivaxxers are eager to point out that this is as much as from all other vaccines ever since 1990! (see `https://www.ronjohnson.senate.gov/services/files/A4A76F9A-9B29-4CF9-B987-F9097A3F4CB7`).

Does it mean covid vaccines are dangerous, more so than other vaccines administered previously, as is claimed by antivaxxers?

Let us compute. One thing important to remember is that deaths reported to VAERS are deaths *after vaccination*, not necessarily *because of vaccination*. Now, covid vaccines have been given to more than 2/3 of elderly population in the US. Every year in the US there are about 3 million deaths, most of them elderly people. So about 2 million of them are vaccinated. This is about 6,000 deaths per day, or 18,000 per three days. 10 times as much as people dying on day 0,1,2 after vaccination in the VAERS system! So even if they don't vaccinate people who are "actively dying", this easily accounts for the deaths reported at VAERS (since more than 10% of deaths happen suddenly).

But what about this figure being as big as all the deaths from vaccines in the previous 30 years? Is this alarming? Let us compute again. Remember, the covid vaccination campaign of 2021 was unprecedented, by far largest in human history. How many times are you vaccinated in your life? Just a few times. So the number of covid vaccinations is at least, say, 1/3 of other vaccinations in the last 30 years. And these are done when you are a kid or a young person, much less likely to die![11]

We see that the VAERS data does not, in fact, give any grounds for questioning the safety of covid vaccines.

10. Vaccine efficacy

Vaccine efficacy, or efficiency, is said to be p if it reduces the probability to get infected, symptomatic, become seriously ill, or die of the disease by a factor $1 - p$. It is usually given as a percentage. For example, efficacy of 40% against infection means that the probability of getting infected is multiplied by 0.6.

Vaccine efficacy can be estimated by looking at **breakthrough cases**, i.e. covid cases among fully vaccinated people. CDC has been keeping track only of the serious cases, which led to hospitalization or death. There are about 8,000 such cases for the period May 1-August 9, 2021. What does it tell us about vaccine efficiency against serious disease?

Let us compute. There have been 3.5 million registered covid cases in the US during this period. On average, during this period, 40% of the population had been vaccinated (in fact, more among older people, but we will neglect that). So we may assume that at least 1.4 million of these cases were among the vaccinated. So we obtain that CHR among the vaccinated is $\leq \frac{8}{1400} = 0.57\%$. For comparison, as we computed, the CHR among general population is about 5.7%, which is 10 times more. So we see that vaccines are at least 90% effective against serious disease.

The number of breakthrough fatalities during the same period was about 1,500. This gives CFR among the vaccinated of about 0.1% and IFR about 0.05%, twice less than seasonal flu. This is 15 times less than for unvaccinated (0.75%), giving efficacy of at least about 94% at preventing death from covid-19. In fact, these figures are even bigger

[11]Here we excluded flu shots, which are performed in large quantities every year. However, since they have been around for many years, deaths "from a flu shot" are rarely reported to VAERS.

since the vaccination rate among the older population was higher than 40%.[12]

Also during this period there were about 40,000 covid deaths, so only 3.7% of covid deaths occurred among vaccinated people. Given that 2/3 of the older population was vaccinated, this needs to be divided by 2, which shows that in fact vaccination decreases the probability of dying from covid by a factor of 30. In other words, there is 97% efficacy at preventing death from covid-19.

For the delta variant, vaccines are less effective at preventing the infection itself (in mild form). Israel reports 40% for Pfizer. So the vaccine replaces R_0 with $\widetilde{R}_0 = 0.6R_0$ if all people are vaccinated. This does not seem like much, so what's the point of organizations (such as MIT, for instance) to require vaccination? Well, let us compute. Suppose initially $R_0 = 2$. Then $\widetilde{R}_0 = 1.2$, so this increases the doubling time 5-fold! Adding regular testing, one may further decrease R_0 to a value below 1, making it impossible for the infection to spread altogether.

Example 8. On August 15, 2021, in Israel there were 214 unvaccinated and 301 fully vaccinated hospitalized covid patients (`https://www.covid-datascience.com/post/israeli-data-how-can-efficacy-vs-severe-disease-be-strong-when-60-of-hospitalized-are-vaccinated`). Does this mean the vaccine does not work against serious disease? Let us compute. Here is the age distribution of these patients, with vaccination rates given in green:

Age	Population (%)		Severe cases		Efficacy
	Not Vax %	Fully Vax %	Not Vax per 100k	Fully Vax per 100k	vs. severe disease
All ages	1,302,912 18.2%	5,634,634 78.7%	214 16.4	301 5.3	**67.5%**
<50	1,116,834 23.3%	3,501,118 73.0%	43 3.9	11 0.3	**91.8%**
>50	186,078 7.9%	2,133,516 90.4%	171 91.9	290 13.6	**85.2%**

Thus, computing over the whole population, we get about 67% efficacy, while if we compute more carefully for age groups, we get efficacy of about 92% in the population under 50 and 85% in the population over 50, both much bigger than 67%! This is an instance of the so-called **Simpson's paradox**: `https://en.wikipedia.org/wiki/Simpson%27s_paradox`.

[12] On the other hand, not all breakthrough cases may be reported.

11. The SIR epidemiological model

Let us now discuss the most basic model of infection spread – the **SIR model**, https://en.wikipedia.org/wiki/Compartmental_models_in_epidemiology.

11.1. **The SIR differential equations.** The SIR model has two unknown functions – the fraction of **susceptible** population $S = S(t)$ and the fraction of **infected** population, $I = I(t)$. We also have the fraction of **removed** population, $R = 1 - I - S$, all the remaining people who are removed from the epidemic through recovery or death (this explains the abbreviation SIR). Thus the fraction of people infected by the time t is $N(t) = 1 - S(t)$.

Infections are passed from the infected to the susceptible, so the number of new infections appearing per unit of time is proportional to IS, and the coefficient is the *effective contact rate* $\beta = t_c^{-1}$. Thus for the rate of change of S we get the **first equation of the SIR model**:

$$\frac{dS}{dt} = -\beta IS.$$

The second equation of the SIR model describes the rate of change of I. This change comes from two sources: (1) susceptible people become infected (the rate of this, according to the first equation, is βIS), and (2) infected people are removed (recover or die). The rate of such removal is, of course, proportional to I, namely, it equals γI where $\gamma = t_r^{-1}$ is the *infectious rate*. So the **second equation of the SIR model** is

$$\frac{dI}{dt} = \beta IS - \gamma I.$$

Recall that we rescaled time by setting $T := \gamma t$. We then get

$$\frac{dS}{dT} = -R_0 IS, \ \frac{dI}{dT} = R_0 IS - I,$$

where $R_0 = \frac{\beta}{\gamma} = \frac{t_r}{t_c}$ is the *basic reproduction ratio*.

The initial conditions are $I(0) = I_0$, $S(0) = S_0 = 1 - I_0$ for some small $I_0 > 0$ (the epidemic begins from just a handful of cases, which is a tiny fraction of the population). By Picard's theorem, there is a unique solution with such initial conditions, which describes the spread of the infection.

Note that when T is small, I is small and S is close to 1, so the second equation of the SIR system is well approximated by the linear equation

$$I' = cI, \ c := R_0 - 1,$$

giving

$$I(T) \approx I_0 e^{cT},$$

as we already discussed above. This is the initial period of *exponential growth*.

The second equation also tells us at which point the infection (i.e., the number of active cases $I(T)$) starts to wane. This happens when $\frac{dI}{dT} = 0$, i.e., $R_0 S = 1$, which yields $S = 1/R_0$. Thus the total number of people who have been infected by this time is

$$H = 1 - S = 1 - \frac{1}{R_0},$$

which is the *herd immunity threshold*.

11.2. **Solving the SIR equations.** To find the solution of the SIR equations, note that we can eliminate I using the first equation:

$$I = -\frac{S'}{R_0 S} = -\frac{(\log S)'}{R_0}.$$

Then the second equation takes the form

$$-(\log S)'' = -R_0 S' + (\log S)'.$$

Thus, integrating, we get

$$-(\log S)' = C - R_0 S + \log S.$$

The initial condition tells us that

$$R_0 I_0 = R_0(1 - S_0) = C - R_0 S_0 + \log S_0.$$

So

$$C = R_0 - \log S_0.$$

Thus we get

$$-(\log S)' = R_0 - R_0 S + \log \frac{S}{S_0},$$

i.e.,

$$S' = S(R_0(S-1) - \log(S/S_0)). \tag{2}$$

with initial condition $S(0) = S_0$.

11.3. **How many people will never get sick?** This, in particular, tells us what fraction of the population S_∞ will never get sick (curiously, it is not zero!). Namely, it is determined by the condition that $S' = 0$, which leads to the transcendental equation

$$R_0(S_\infty - 1) = \log(S_\infty/S_0).$$

Exponentiating, we get

$$e^{-R_0} e^{R_0 S_\infty} = S_0^{-1} S_\infty.$$

So setting $Q := R_0 S_\infty$, we get

$$Qe^{-Q} = S_0 R_0 e^{-R_0}.$$

Thus introducing the Lambert function

$$W(x) = \sum_{n=1}^{\infty} \frac{n^{n-1}}{n!} x^n$$

which is inverse to $f(y) = ye^{-y}$ (as follows from the Lagrange inversion theorem), we obtain

$$Q = W(S_0 R_0 e^{-R_0}),$$

i.e.,

$$S_\infty = \frac{W(S_0 R_0 e^{-R_0})}{R_0} \approx \frac{W(R_0 e^{-R_0})}{R_0}$$

(as S_0 is close to 1). Note that $W(R_0 e^{-R_0})$ is the second solution of the equation $xe^{-x} = R_0 e^{-R_0}$ (the first solution is R_0). For $R_0 = 1$ both solutions are 1 (maximum point for the function $y = xe^{-x}$). Thus if $R_0 = 1 + c$ for small c then $W(R_0 e^{-R_0}) \approx 1 - c$. So for R_0 close to 1 we have

$$S_\infty \approx \frac{1-c}{1+c} \approx 1 - 2c.$$

So the fraction of people who will get sick is about $2c$, or about $2H$, where H is the herd immunity threshold. In other words, the herd immunity threshold is reached approximately in the middle of the epidemic.

On the other hand, for large R_0

$$S_\infty \approx e^{-R_0},$$

i.e., decays exponentially fast with R_0.

Example 9. If $R_0 = 1.2$ (seasonal flu) we have $S_\infty \approx 0.68$, i.e. about 32 % will get sick (although the herd immunity threshold is $H = 1 - \frac{1}{1.2} \approx 16.6\%$, about half as much, as expected). If $R_0 = 2$ (optimistic estimate for the original covid strain) then $S_\infty \approx 0.2$ so 80% will get sick (while $H = 1 - \frac{1}{2} = 50\%$). But if $R_0 = 8$ (the delta

variant) then $S_\infty \approx 0.00033$, i.e. only 3 people in $10,000$ will remain uninfected. Virtually everyone will get sick! (However, in real life the epidemic will usually be mitigated, so R_0 will quickly drop to a much smaller value).

11.4. **Logistic growth.** This tells us the behavior of $S(T)$ at the end of the epidemic. Namely, as $T \to \infty$, the function $S(T)$ approaches S_∞, and the derivative of the function $S(R_0(S-1)-\log(S/S_0))$ at S_∞ is

$$-D = 2R_0S_\infty - R_0 - \log(S_\infty/S_0) - 1 \approx 2R_0S_\infty - R_0 - \log(S_\infty) - 1.$$

So for R_0 close to 1 we have $D \approx R_0 - 1$, while for large R_0 we have $D \approx 1$.

Thus for large T we have

$$S(T) \approx S_\infty(1 + e^{-D(T-T_1)}).$$

for some $T_1 \in \mathbb{R}$. So

$$I(T) = \frac{D}{R_0(1+e^{D(T-T_1)})} \approx \frac{D}{R_0}e^{-D(T-T_1)}.$$

This is the range of **logistic growth**.

11.5. **The exact solution of the SIR equations.** Let us now come back to solving the SIR equations. Separating variables, we have

$$\frac{dS}{S(R_0(S-1)-\log(S/S_0))} = dT.$$

So

$$T = \int_S^{S_0} \frac{dx}{x(R_0(1-x)+\log(x/S_0))} = \int_{R_0+\log(S/S_0)}^{R_0} \frac{dv}{v - Ke^v},$$

where

$$K := S_0R_0e^{-R_0}$$

(setting $v := R_0 + \log(x/S_0)$). So define the special function

$$V_K(u) := \int_1^u \frac{dv}{v - Ke^v}$$

(it cannot be expressed more explicitly). Then we get

$$T = V_K(R_0) - V_K(R_0 + \log(S/S_0)).$$

So we obtain the exact solution in implicit form:

$$S(T) = R_0^{-1}Ke^{V_K^{-1}(T_0-T)},$$

where $T_0 := V_K(R_0)$. So we see that as $S_0 \to 1$ (i.e., $I_0 \to 0$), the function $S(T)$ does not depend on S_0 very much, except it gets shifted

by $V_K(R_0)$. Moreover, the exponential form of the solution for small T implies that this time shift is

$$T_0 \approx -\frac{\log I_0}{c}.$$

11.6. **The approximate solution.** When S is close to 1, we have an approximation

$$\log(S) \approx S - 1.$$

Thus at the beginning of the epidemic the differential equation for S can be approximated by

$$S' = cS(S-1).$$

Moreover, $c := R_0 - 1$ is small then his approximation is good for all T since we know that in this case S will never get below a threshold approximately equal to $1 - 2c$.

This equation can be solved explicitly:

$$S(T) = \frac{1}{1 + (S_0^{-1} - 1)e^{cT}}.$$

So the total share of infected people by the time T is

$$B(T) = 1 - S(T) = \frac{(1-S_0)e^{cT}}{S_0 + (1-S_0)e^{cT}} = \frac{1}{1 + e^{-c(T-T_0)}},$$

where

$$T_0 = \frac{\log \frac{S_0}{1-S_0}}{c}.$$

So what does this approximate solution look like? The rate of growth $B'(T)$ (i.e., the renormalized number of new infections) is

$$B'(T) = -S'(T) = \frac{\frac{S_0 c}{1-S_0} e^{-cT}}{(1 + \frac{S_0}{1-S_0} e^{-cT})^2} = \frac{c}{4\cosh^2 \frac{c(T-T_0)}{2}}.$$

This has the shape of a **solitary wave** that we often see on the news.

11.7. **Duration of the surge.** Thus

$$T = T_0 - \frac{1}{c} \log\left(\frac{1-B}{B}\right).$$

So if $B = 0.05$ then

$$T \approx T_0 - \frac{3}{c},$$

while if $B = 0.95$ then

$$T \approx T_0 + \frac{3}{c}.$$

So if we define the duration of the epidemic as the period between $B = 0.05$ and $B = 0.95$, we will get the value

$$\mathbf{T} = \frac{6}{c}.$$

In other words, the actual duration is

$$\mathbf{t} = \frac{6t_r}{c}.$$

For example, assume that $R_0 = 1.5$ and $t_r = 7$ days (slightly mitigated original strain of covid). Then $\mathbf{t} = 84$ days. This is close to the duration of the first covid wave.

Acknowledgements. This mini-course was given at CampCapeCod in August 2021. I am grateful to the participants for many interesting questions.