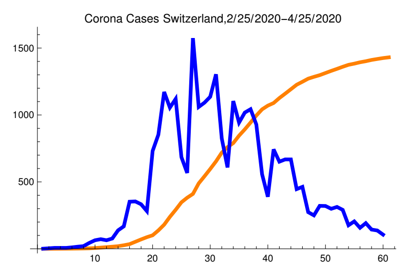
Unit 34: Data

Data

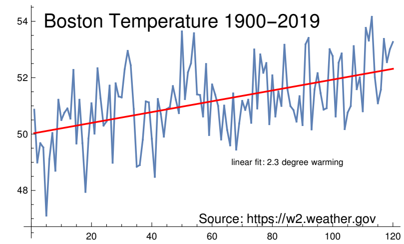
34.1. Information is stored in the form of **data**. Ultimately data can always be stored as numerical values $f(k)$, where k is a label. On a computer is is one function $f(1), \cdots, f(n)$ with $f(k)$ taking values in $\{0, \dots, 254\}$, where k is the **memory address** and where where n is the total number of **Bytes** it can store. Of course, these data are organized in a more convenient form like as a list of numbers, encoding a song or an array of numbers encoding a picture or an array of arrays of numbers encoding a movie. Here are the data of corona cases in Switzerland. They are given by an array $f(1) = 1, f(2) = 2, f(3) = 6..., f(61) = 28611$ of numbers: [1]



34.2. Anything which has been developed in calculus can be applied to data. For example, we can look for differences, sum up data, average data, produce distributions. Knowing multi-variable calculus and linear algebra and probability theory helps a lot to model, visualize and reduce data. You should consider to learn multi-variable calculus and linear algebra in the future. It turns out however that most of the insight we can gain from data come from functions of one variable. We want to know the value $S(n)$ of a stock prize, the temperature $T(n)$ or the number $I(n)$ of infected people by a disease on day n. Here is the development of the average year temperature in Boston over the last 120 years. [2]

[1] https://www.corona-data.ch/

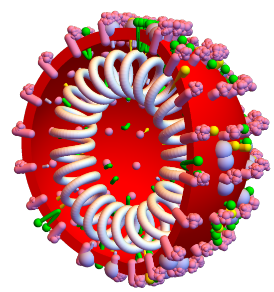
[2] https://w2.weather.gov



34.3. In some cases like illnesses, the data are estimates. For weather temperatures, they are quite accurate. A good model should allow to predict what the data should look like other times. But every model is just an application. A model can not only be either right or wrong, it can also be just irrelevant. Still, we can try build a model and then compare with the model. Both for pandemics as well as for climate change there are models. For climate we have a lot of accurate data, for pandemics, we do not.

Population dynamics

34.4. We are currently interested in the population dynamics of a virus "you know who". Neil Shubin mentioned last week in a public virtual lecture that there are more viruses exist than stars in the universe, actually millions more [3] and that some of our genoms contain viral DNA so that we really owe our existence to the construct of viruses [4] This is a bit of a different perspective in a time we all lose our minds.

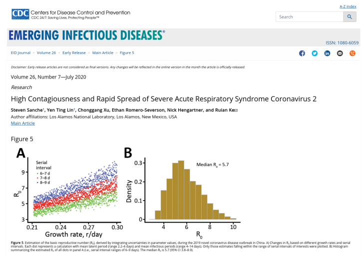


34.5. The simplest model for population dynamics is **exponential growth** given by the $f(t) = e^{ct}$ where c is a constant. In epidemiology, it is custom to define $f(t+h)/f(t) = R_0$, where R_0 is the **reproduction number** and h is an **infect period time**. This means with the notation introduced in the first week that $Df(t) = f(t+h) - f(t) = (R_0 - 1)f(t)$ so that $f(t) = (1+h)^{(R_0-1)t} f(0) = e^{ct} f(0)$ with $c = (R_0 - 1)\log(1+h)$. In the news, we often see R_0 values but not the **mean infection period** which is also important for estimating the growth. This spring one has tried through **increased hygiene** and **social distancing** the R_0 value down. You see in articles like in the figure that the R_0 value depends on the infection period h which is not yet precisely known. [5]

[3] ttps://www.airspacemag.com/daily-planet/there-are-more-viruses-earth-there-are-stars-universe-180974433

[4] https://www.nytimes.com/2017/10/04/science/ancient-viruses-dna-genome.html

[5] Emerging infectious diseases, Volume 26, Number 7, July 2020 (early release)



34.6. Once a certain saturation has been reached, the exponential growth slows down to a logistic growth. This happens as faster as larger the R_0 value is. The reason for the slow down is the decreasing number of new hosts assuming **immunity** of subjects which have overcome the disease. A logistic type growth is inevitable in a finite population and infection numbers of Corona are a textbook example for that. Especially politicians like to explain the slow-down also through social distancing. Others have pointed out that the R_0 value was actually smaller than now than before social distancing measures have come into effect and that simple measures like washing hands were already very effective. More radical ideas are trying to increase R_0 so that "herd immunity" kicks in earlier, minimizing the **overall harm** of a economic shut-down which also leads to life casualties, as other illnesses can no more be treated effectively, through stress etc. Lock-downs had also beneficial impacts like improved air quality due to less travel, good things for battling climate change.

Population growth

34.7. A follow up course to Math 1a is Math 1b, where one can learn more about series and differential equations. For example, one can look at the differential equation

$$f'(t) = rf(t)$$

which is an equation for an unknown quantity $f(t)$ like the number of infected people during the early stage of a pandemic. The rule tells that the rate of change (the increase of the number of infected people). It leads to **exponential growth**

$$f(t) = e^{rt} .$$

You can check that the derivative $f'(t)$ agrees with $rf(t)$.

34.8. If a population is finite, this exponential growth will slow down. The text book model is the **logistic growth** like $f'(t) = f(t)(1 - f(t))$ which is a **differential equation**. You can check that the function $f(t) = \frac{e^t}{e^t+1}$ solves this equation. Can you do that? You have to check that the derivative $f'(t)$ agrees with $f(t)(1 - f(t))$.

35. Lessons

35.1. The recent weeks have taught us various lessons. Here are few:

Data have to come with sources.

If a source does not do that and even government sites like CDC do not do that often, one has to complain. Unfortunately, most news outlets fail to do that. There should be a general rule that any statement comes with complete data sets.

35.2.

Accurate data are important for decisions.

35.3. If not enough data are available, one is flying blind and fear takes over. Fear can be a blessing as it leads to action and caution, it can also paralyze or lead to over-reactions.

Data can be interpreted in various ways.

35.4. There is a classical book by Darrell Huff from 1954 called "How to Lie with Statistics". The recent months have given plenty of more examples.

It is possible to manipulate with pictures.

35.5. The book **"How to lie with statistics"** from the 1950ies which explains some manipulation tools. You have to be aware that manipulation is not always intended. Some authors want to justify a model, want to validate their data, want to influence people (for example to do trigger behavior in a crisis).

Also look critically statements of experts.

35.6. There are various forms of bias possible. People judge because of race, religion, gender, institution, title or appearance. We will look at examples like the **Dr Fox experiment**.

Look up the data yourself. How were the data obtained? How accurate are they?

35.7. Health data are difficult. On the CDC website, the annual burden of flue is the US estimated (April 26, 2020) with 12'000 to 61'000 deaths, 140'000-810'000 hospitalisations and 9.3 -45 million illnesses. Look at the large error margins. For example, hospitals are not always required to report influenza. With Covid 19, the problems have been discussed a lot. One does not know how many are infected, how many had it, nor how many deaths are really attributed to it as many victims had pre-conditions. So, unlike the temperature in Boston, the virus data can be interpreted in various ways.

35.8.

Even when highly complex multi-dimensional data exist, one dimensional visualizations are great.

So, even if you go beyond this course to other mathematical topics, single variable calculus will remain important!

Oliver Knill, knill@math.harvard.edu, Math 1a, Harvard College, Spring 2020