INTRODUCTION TO CALCULUS

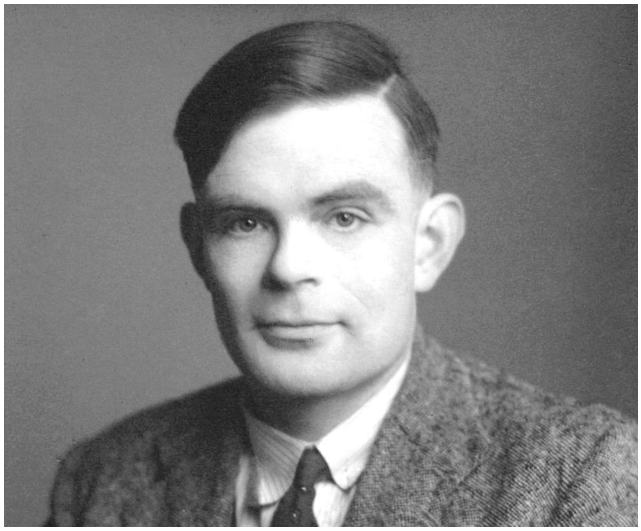
MATH 1A

Unit 34: More Data stories

34.1. In projects you have explored the derivative function of the prime number function $p(n)$ giving the $n'th$ prime, you have computed with functions describing polyhedra, you have seen unpredictability of simple deterministic systems and you have used Monte Carlo simulations to find areas. In this lecture we look at some more data stories.

Turing's picture

34.2. Information is stored in the form of **data**. Data can always be stored as numerical values $f(k)$, where k is a label. On a computer it is one function $f(1), \cdots, f(n)$ with $f(k)$ taking values in $\{0, \ldots, 254\}$ and k is the **memory address** and where where n is the total number of **Bytes** the computer can store. It was **Alan Turing** who illustrated best what **data** and what computing with data means. He coined the concept of a **Turing machine**.

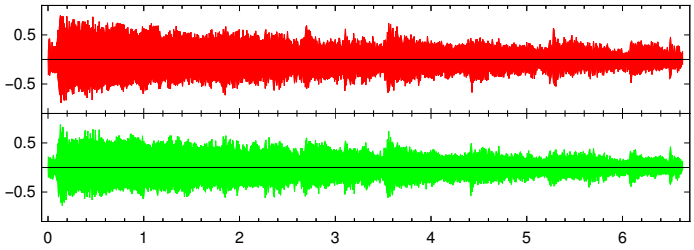


34.3. In that machine, all information is stored on data set $\{...x_{-m},\ x_0, x_n, ...\}$ on which only finitely many 1 appear. This **tape** contains the data. The machine has finitely many states $A_1, ..., A_k$. Depending on whether x_0 is 0 or 1 and what the state of the machine is, the machine now either replaces x_0 with some 0 or 1 or then moves the tape left or right or then changes the state. Turing demonstrated, that all computations we are capable to do can be done also by such a simple machine. The point is that all

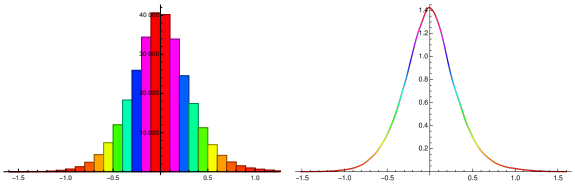
data we are ever can measure, process or produce can be described by a function $f(k)$ taking values 0 and 1 and having the property that $f(k) = 1$ only for finitely many k.

Audio data

34.4. Of course, these data are organized in a more convenient forms like a file representing a song or an array of numbers encoding a picture or an array of arrays of numbers that encode a movie. The following picture shows a few seconds of audio data of the song "Bohemian Rhapsody". It is the part "Mamaaah, Ohhhh". The command $\{f, g\} = AudioData[Import["queen.wav"]]$ gives you to two discrete functions of length $n = 292669$ corresponding to $n/44100 = 6.6$ seconds of sound. Looking up values like $f(4) = f[[4]]$ gives -0.091 which the amplitude of the left sound channel at time 4/44100 seconds. The function value $g(4) = g[[4]]$ gives 0.063507, the amplitude of the right sound channel. The function values range from -1 to 1.

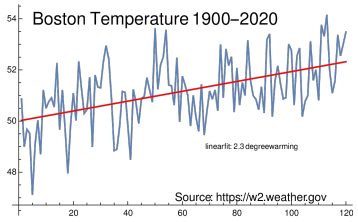


34.5. Anything which has been developed for calculus can be applied to data. For example, we can look for differences $f(k+1) - f(k)$. We can sum up data, average data, produce distributions. For the above sound clip, we can draw the histogram as well as the smooth histogram which is the distribution function of the sound data. What these histograms show is how many sound notes were in amplitude in a given interval. Naturally, very high amplitude sound data are rare.



Whether data

34.6. While multi-variable calculus and linear algebra and probability theory help more effectively to visualize and reduce data, much insight can be gained from data depending on one variable only: we want to know the value $S(n)$ of a stock prize, the temperature $T(n)$ on day n. Here is the development of the average year temperature in Boston over the last 121 years from 1990 to 2020 [1]

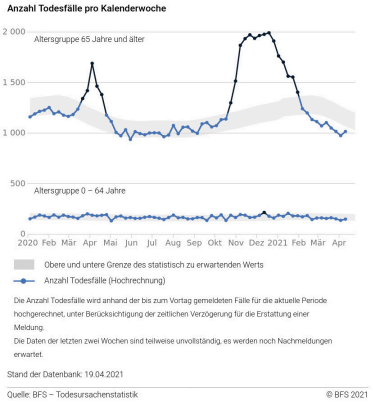


[1] https://w2.weather.gov

Health data

34.7. A good model needs to be able to predict what the data should look like other times. If we build a model we then compare the data with the model. Weather data are very reliable. Everybody can measure the temperature and compare it with the data which are published. Health data are much less reliable. While we have pretty good and accurate data about birth and death of the population, we have much less reliable data about the cause of death.

34.8. Switzerland has quite good and reliable statistical data. The BFS (Bundesamt für Statistik) is among the finest in the world. The bureau publishes also excess mortality. They are below average at the moment also for 65 and older. As a matter of coincidence, the New York times from today has just put a graph of the excess death rates from 2018-2021 on the front page.



Population growth

34.9. The simplest model for population dynamics is **exponential growth model** 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.

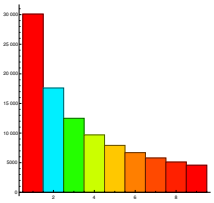
34.10. The function $f(t) = e^{ct}$ satisfies the differential equation $f'(t) = cf(t)$. We are not going more into differential equation but just want to point out that models usually only work in specific situations. Exponential growth models for example fail always after some time, simply because resources are finite. A population of rabbits will grow exponentially fast at first, but then, after some saturation has been reached, the exponential model (famously illustrated by the Fibonacci rabbits) can not be sustained and a better model will be needed.

34.11. 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 $f(t)(1-f(t))$. The logistic model is much better. Do you remember discrete version, the **logistic equation**. It turns out that differential equations $x'(t) = F(x(t))$ in one

dimension can not exhibit complicated behavior but that maps $x(t+1) = F(x(t))$ can do. The case $x(t+1) = 4x(t)(1-x(x))$ has illustrated this.

Benford's law

34.12. If you look at random data and make a statistics of the **first significant digit**, there is a surprise, as Simon Newcomb and Frank Benford realized in 1881 or 1938 respectively. The digit 1 appears more frequent than the number 2. One can look at stock market values, lengths of rivers or car prizes, one sees similar data. Here is an example, where one can prove the distribution to be true [2]



Homework

Problem 33.1: In the movie **Fermat's room**, a string of 169 zeros and one's are given. You are a mathematician in that movie and need to figure out what the numbers mean, otherwise, the room gets crushed. Can you figure it out? $\{0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,1,1,1,1,1,1,1,1,0,0,1,$
$1,1,1,1,1,1,1,1,1,1,0,0,1,1,0,0,0,1,0,0,0,1,1,0,0,1,1,0,$
$0,0,1,0,0,0,1,1,0,0,1,1,1,1,1,1,0,1,1,1,1,1,0,0,1,1,1,1,$
$0,0,0,1,1,1,1,1,0,0,0,1,1,1,1,1,1,1,1,1,1,0,0,0,0,0,1,0,1,0,$
$1,0,1,0,0,0,0,0,0,1,1,0,1,0,1,1,0,0,0,0,0,0,1,1,1,1,1,$
$1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0\}$.

Problem 33.2: Verify that the function $\frac{e^{4t}}{e^{4t}+1}$ satisfies the logistic equation $f'(t) = 4f(t)(1-f(t))$.

Problem 33.3: Make a list of the first 30 Fibonacci numbers and make a statistics about the first significant digit.

Problem 33.4: Look up the population of the numbers of at least 20 largest cities on Wikipedia.

Problem 33.5: Now make a statistics about the list the first significant digits with the data you have collected in 33.4.

Oliver Knill, knill@math.harvard.edu, Math 1a, Harvard College, Spring 2021

[2] This example appears in V. Arnold's, Mathematical methods of Mechanics