Introduction to data science & artificial intelligence (INF7100)

Arthur Charpentier

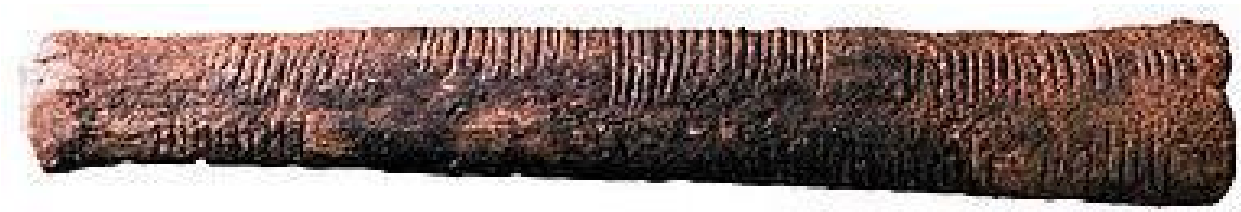
#132 Uncertainty and Randomness

été 2020

Uncertainty

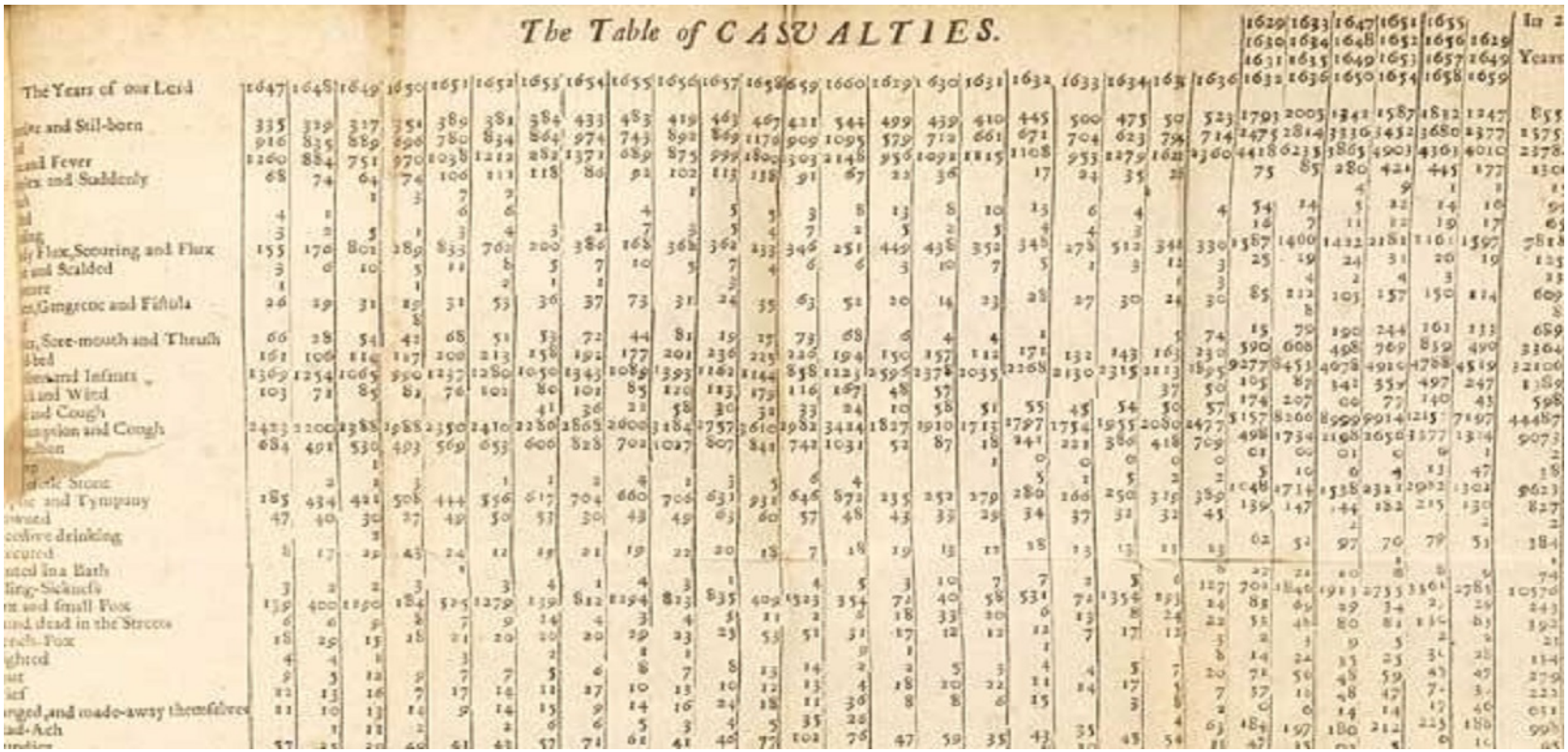
Data

Storing Data: Tally sticks, used starting in the Paleolithic area



A tally (or tally stick) was an ancient memory aid device used to record and document numbers, quantities, or even messages.

Data

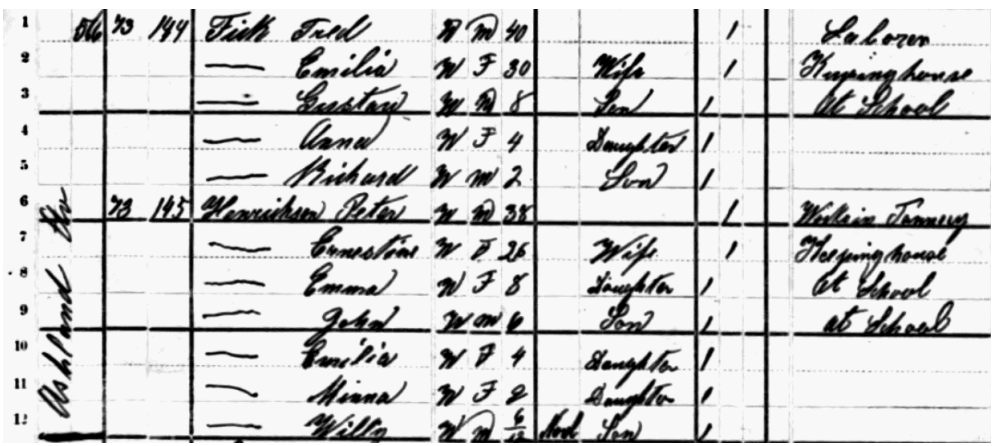


The Table of CASUALTIES.

Collecting Data: John Graunt conducted a statistical analysis to curb the spread of the plague, in Europe, in 1663

Data

Data Manipulation: Herman Hollerith created a Tabulating Machine that uses punch carts to reduce the workload of US Census, in 1881, see 1880 Census, n =50 million Americans.



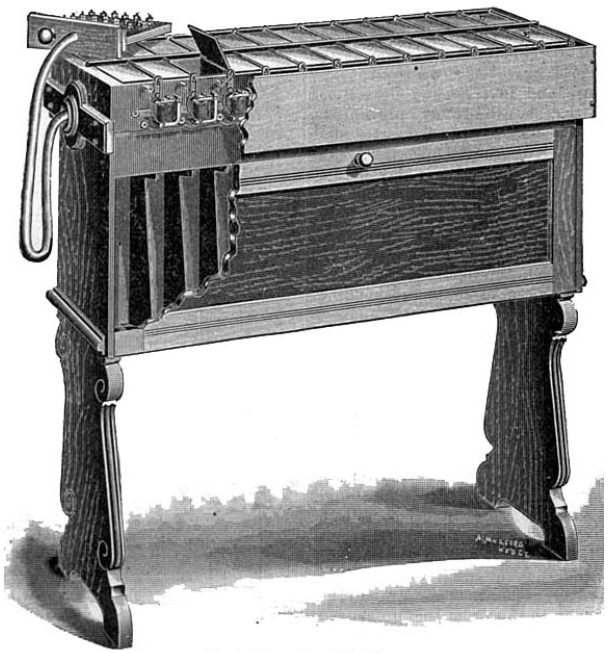


Fig. 3.—Sorting Machine.

Hollerith's Electric Sorting and Tabulating Machine.

Data

Allen-Scott Report

Data Center Plan Called Privacy Invasion

By ROBERT S. ALLEN
and PAUL SCOTT

WASHINGTON — A special White House task force is recommending the creation of a federal data center which eventually could have a comprehensive file on every man, woman and child in the country.

Now under study in inner administration circles, the still-secret report advocates the gradual transfer of all governmental records and statistics to magnetic computer tape, which would be turned over to a newly-created agency that would function as a general data center.

The computerized information would be available, at the push of a button, to a wide range of government authorities.

Estimated cost of the pro-

cial Security, census data, medical, credit and criminal reports.

"Comprehensive information of this kind, centralized in one agency," says Gallagher, "could constitute a highly dangerous dossier bank. Such an agency would be a distinct departure from our American tradition."

Subcommittee investigators have ascertained that the task force's report states that a vast accumulation of government records already is on computer tape and could be turned over to the proposed general data center immediately. Listed as among these available files are:

Internal Revenue Service — 74 2 million personal and corporate tax returns.

Defense Department — 14

the most intimate information, the investigators learned, are freely passed around among agencies. Graphically illustrative of this practice and its harsh consequences are the following two instances:

A teenager visiting Washington stayed with an uncle, at his mother's suggestion. During the night the boy was sexually assaulted by the uncle. Years later, as a Phi Beta Kappa graduate from a leading Eastern university, the boy applied for a job with the National Security Agency. During a required lie detector test he told about the assault. His frank admission cost him the desired job.

But that wasn't all. This affair, in which he was an inno-

Data Center: The US Government plans the world's first data center to store 742 million tax returns and 175 million sets of fingerprints, in 1965.

Data

Literary Digest Poll based on 2.4 million readers, Özbey (2018)
A. Landon: 57% vs. F.D. Roosevelt: 43%

George Gallup sample of about 50,000 people
A. Landon: 44% vs. F.D. Roosevelt: 56%

Actual results
A. Landon: 38% vs. F.D. Roosevelt: 62%

Today

AMERICA SPEAKS
THE NATIONAL WEEKLY POLL of PUBLIC OPINION

Next Sunday

Institute Forecasts the Re-election of Franklin D. Roosevelt, Gives Him 54% of Popular Vote, Minimum of 315 Electors

Major Party Percent Is 55.7; New York in F.D.R. 'Sure' Column

Election Forecast

1—The American Institute of Public Opinion predicts the re-election of Franklin D. Roosevelt and John N. Garner.

* * *

2—The Institute's latest presidential poll indicates that Roosevelt will receive approximately 56% of the major party vote (minor parties eliminated); to 44% for Alfred M. Landon and Frank Knox. In 1932 the President received 59.1% of the major party vote.

* * *

3—With minor parties included, President Roosevelt's percentage of the total popular vote will be approximately 54%, to 46% for Landon.

* * *

4—The President will receive a minimum of 315 electoral votes. The number necessary to win is 266. Should last-minute shifts in the group of states where the race is too-and-fork give this entire group to Roosevelt, he would receive more electoral votes than in 1932, when he polled 472.

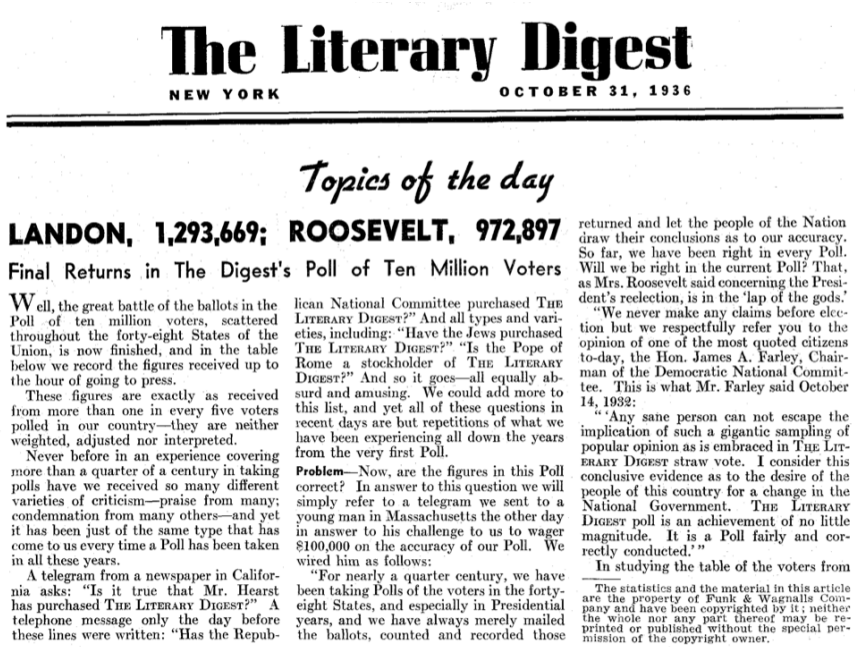
* * *

5—William Lemke, candidate of the Union Party, will poll fewer than 1,000,000 popular votes, and carry no state.

* * *

6—Norman Thomas, Socialist candidate, will poll about half as many votes as in 1932, when he received 884,000.

Election Will Test Clashing Poll Methods



The Literary Digest

NEW YORK — OCTOBER 31, 1936

Topics of the day

LANDON, 1,293,669; ROOSEVELT, 972,897

Final Returns in The Digest's Poll of Ten Million Voters

Well, the great battle of the ballots in the Poll of ten million voters, scattered throughout the forty-eight States of the Union, is now finished, and in the table below we record the figures received up to the hour of going to press.

These figures are exactly as received from more than one in every five voters polled in our country—they are neither weighted, adjusted nor interpreted.

Never before in an experience covering more than a quarter of a century in taking polls have we received so many different varieties of criticism—praise from many; condemnation from many others—and yet it has been just of the same type that has come to us every time a Poll has been taken in all these years.

A telegram from a newspaper in California asks: "Is it true that Mr. Hearst has purchased The Literary Digest?" A telephone message only the day before these lines were written: "Has the Republican National Committee purchased The Literary Digest?" And all types and varieties, including: "Have the Jews purchased The Literary Digest?" "Is the Pope of Rome a stockholder of The Literary Digest?" And so it goes—all equally absurd and amusing. We could add more to this list, and yet all of these questions in recent days are but repetitions of what we have been experiencing all down the years from the very first Poll.

Problem—Now, are the figures in this Poll correct? In answer to this question we will simply refer to a telegram we sent to a young man in Massachusetts the other day in answer to his challenge to us to wager $100,000 on the accuracy of our Poll. We wired him as follows:

"For nearly a quarter century, we have been taking Polls of the voters in the forty-eight States, and especially in Presidential years, and we have always merely mailed the ballots, counted and recorded those returned and let the people of the Nation draw their conclusions as to our accuracy. So far, we have been right in every Poll. Will we be right in the current Poll? That, as Mrs. Roosevelt said concerning the President's reelection, is in the 'lap of the gods.'

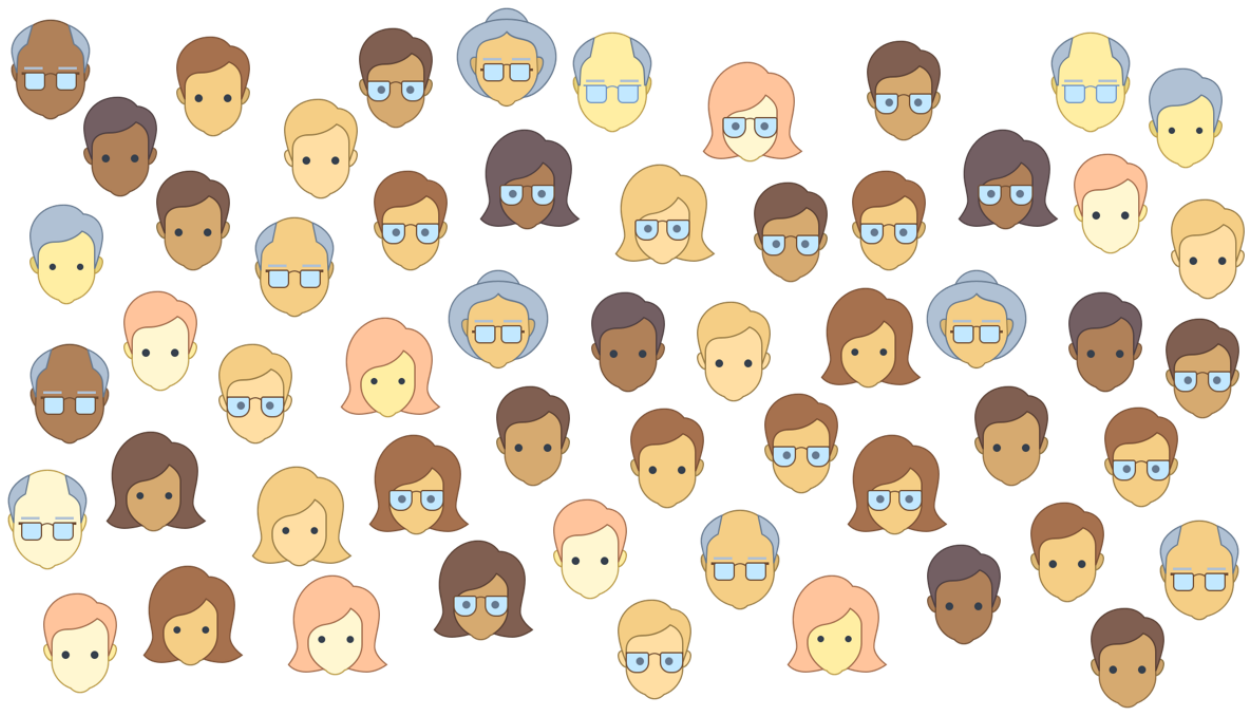
"We never make any claims before election but we respectfully refer you to the opinion of one of the most quoted citizens to-day, the Hon. James A. Farley, Chairman of the Democratic National Committee. This is what Mr. Farley said October 14, 1932:

"'Any sane person can not escape the implication of such a gigantic sampling of popular opinion as is embraced in The Literary Digest straw vote. I consider this conclusive evidence as to the desire of the people of this country for a change in the National Government. The Literary Digest poll is an achievement of no little magnitude. It is a Poll fairly and correctly conducted.'"

In studying the table of the voters from

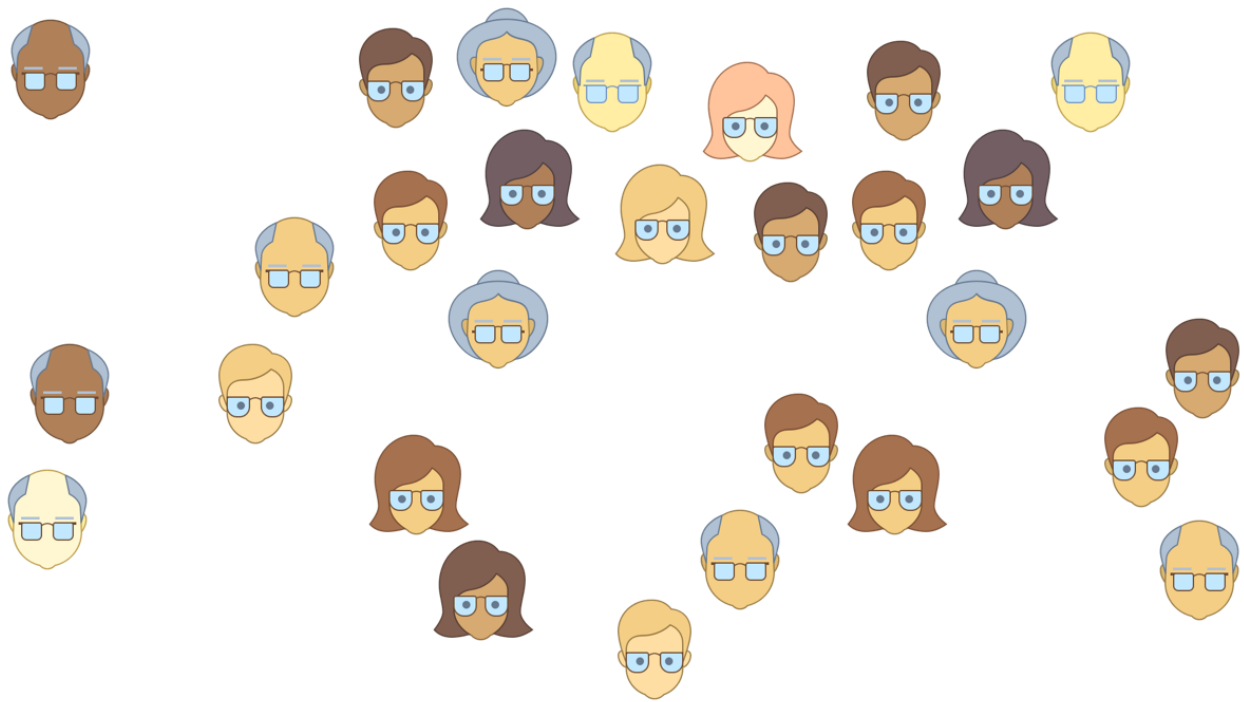
1936 Election & Sampling

A badly chosen big sample is much worse than a well-chosen small sample, see Why the 1936 Literary Digest Poll Failed or How biases can arise in sampling

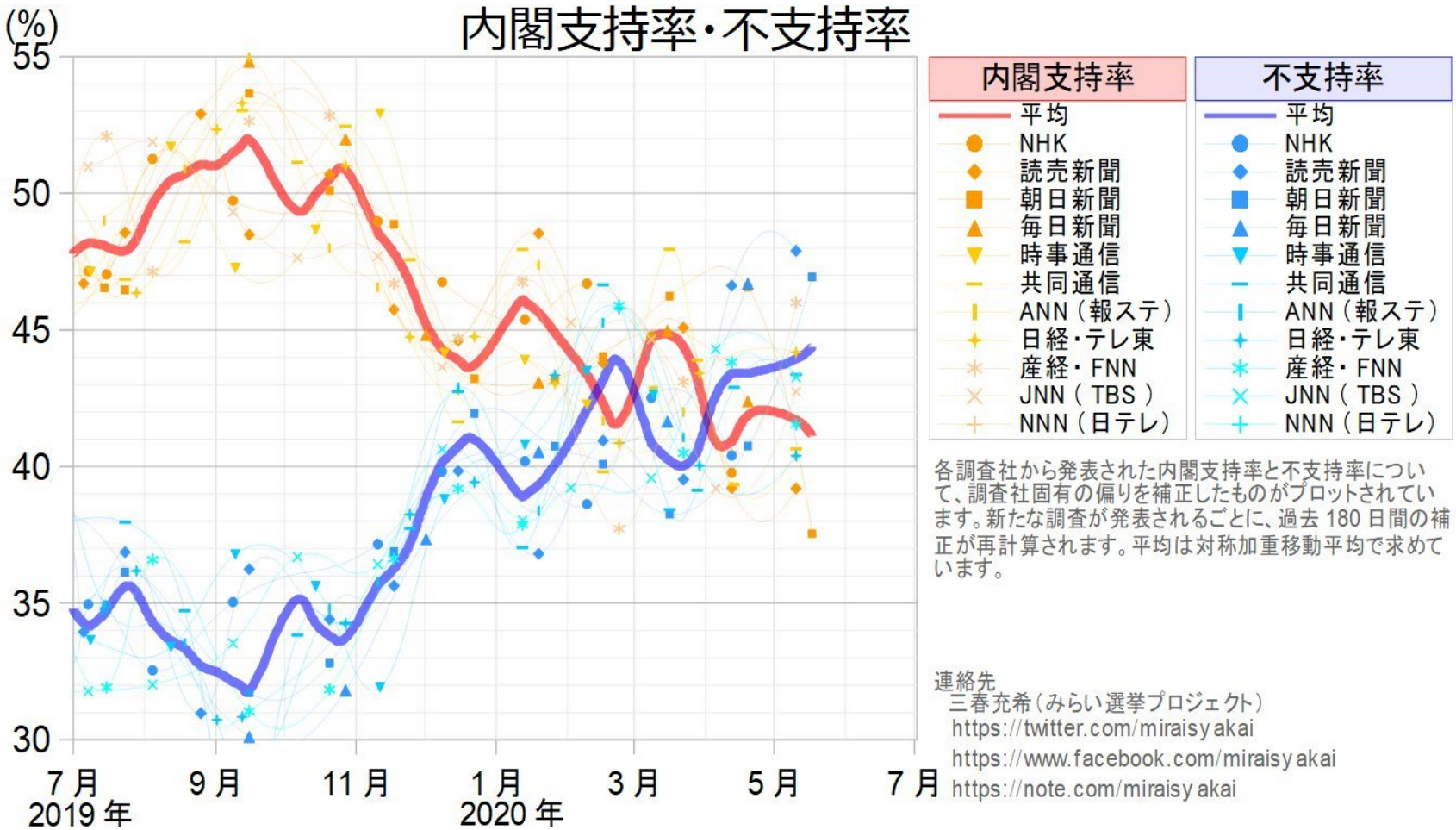


1936 Election & Sampling

A badly chosen big sample is much worse than a well-chosen small sample, see Why the 1936 Literary Digest Poll Failed or How biases can arise in sampling



Sampling Uncertainty



"Dans toute statistique, l'inexactitude du nombre est compensée par la précision des décimales" Alfred Sauvy (see #331)

Sampling Error and Uncertainty

Sample $\mathcal{X} = \{x_1, x_2, x_3, \cdots, x_{n-1}, x_n\}$

Consider some statistics $t(\mathcal{X})$, e.g. $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

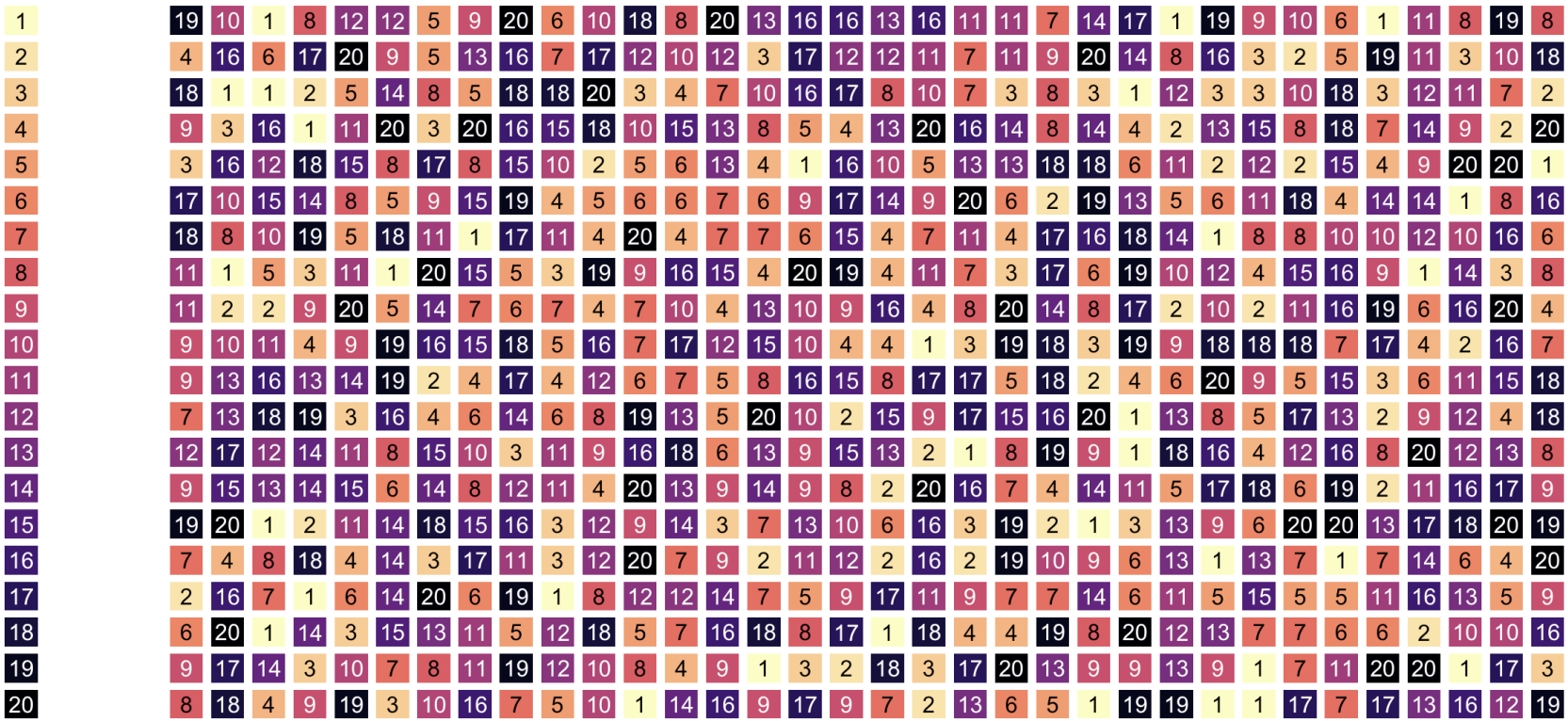
What if we had had another sample $\mathcal{X}'$?
How different $t(\mathcal{X}')$ could have been...?

Under the assumption that $x_i's$ are observed from some $\mathcal{N}(\mu, \sigma^2)$ distribution, we know (mathematical theory) that $\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$

- we need to know the distribution (here Gaussian)
- we need to know the parameters (here σ^2)

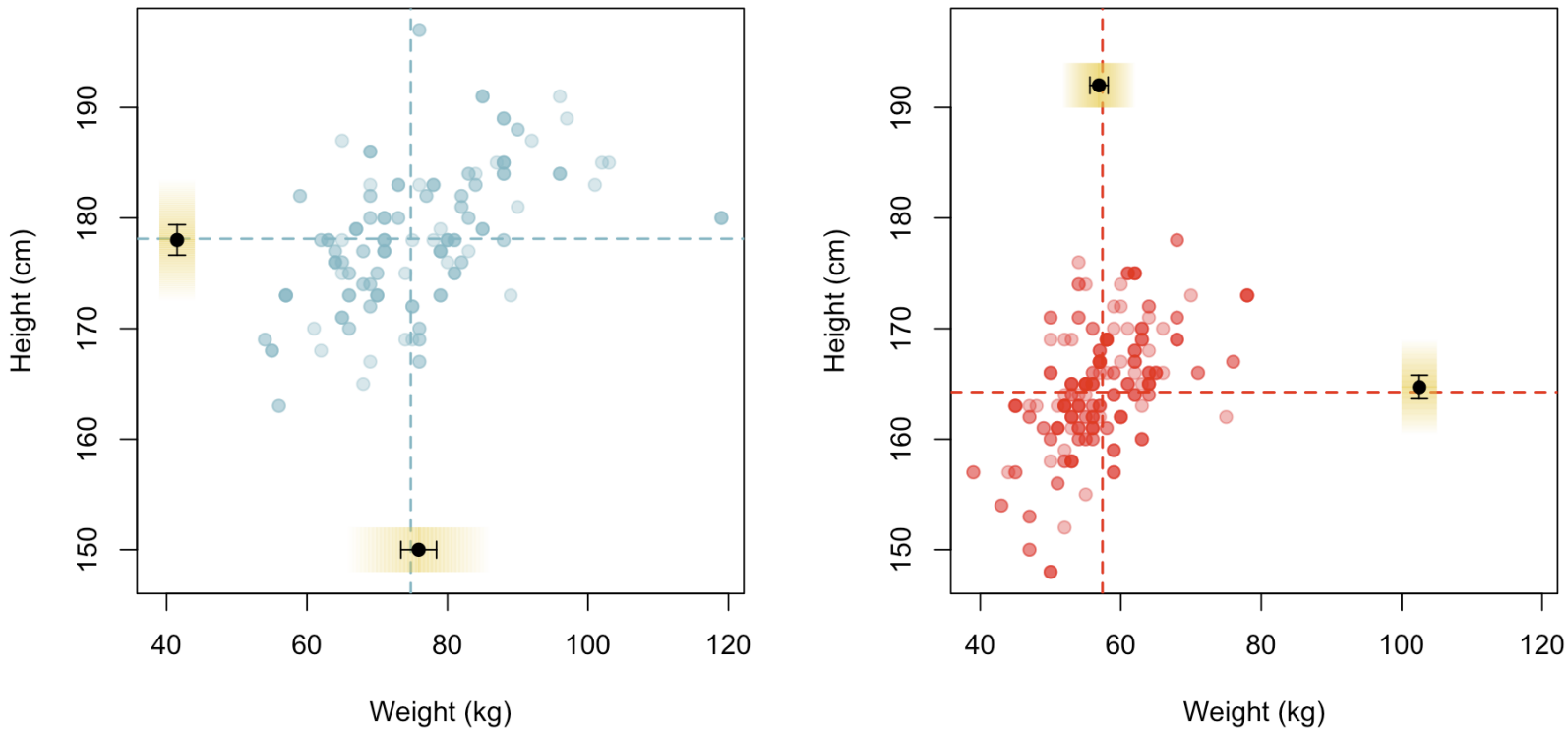
Sampling Error and Uncertainty

We can create some fake sample by resampling...



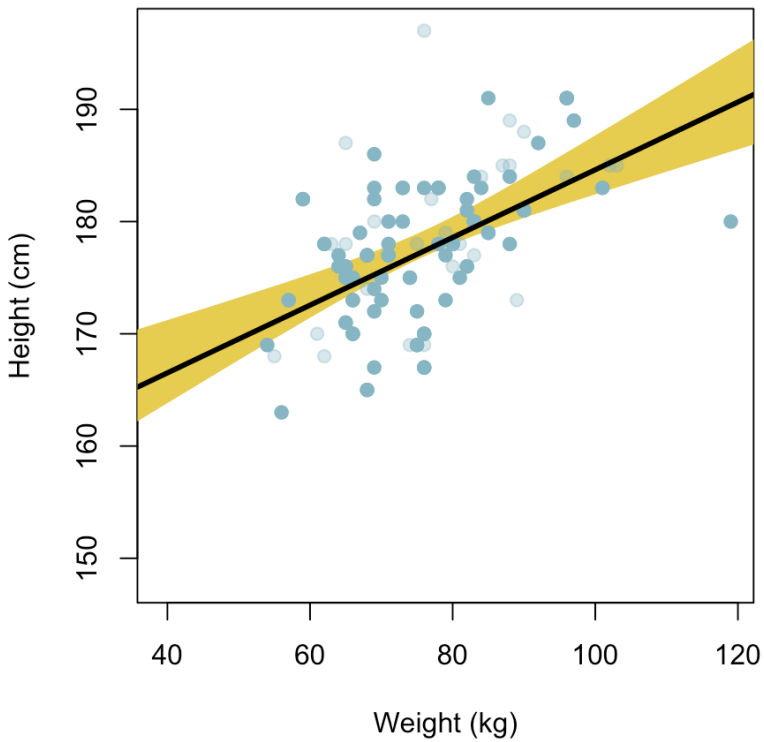
Resampling & Bootstrapping

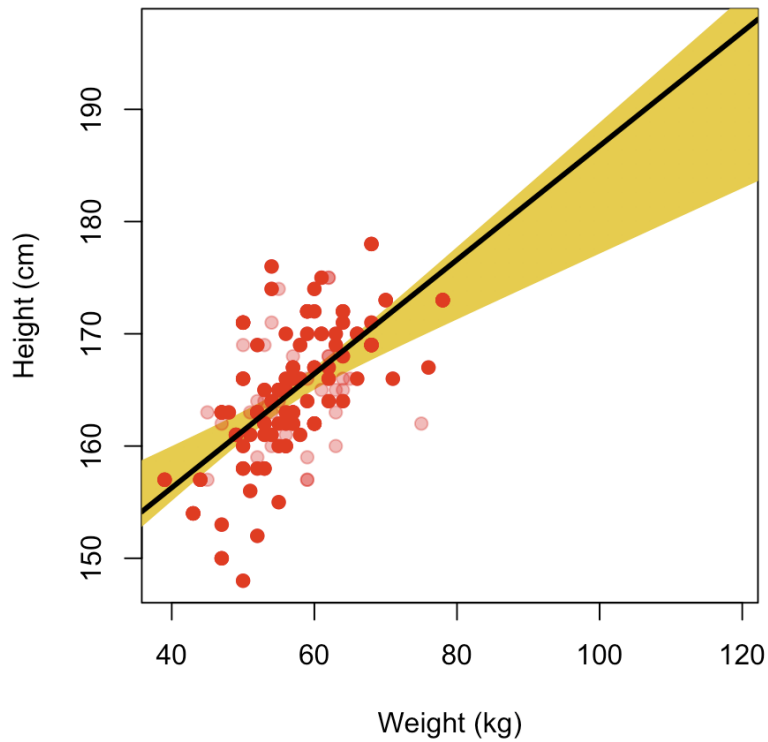
Students height and weight, of boys (left) and girls (right),



Resampling & Bootstrapping

Regression line of height against weight





Bootstrap & Bootstrapping

Algorithm 1: Bootstraping

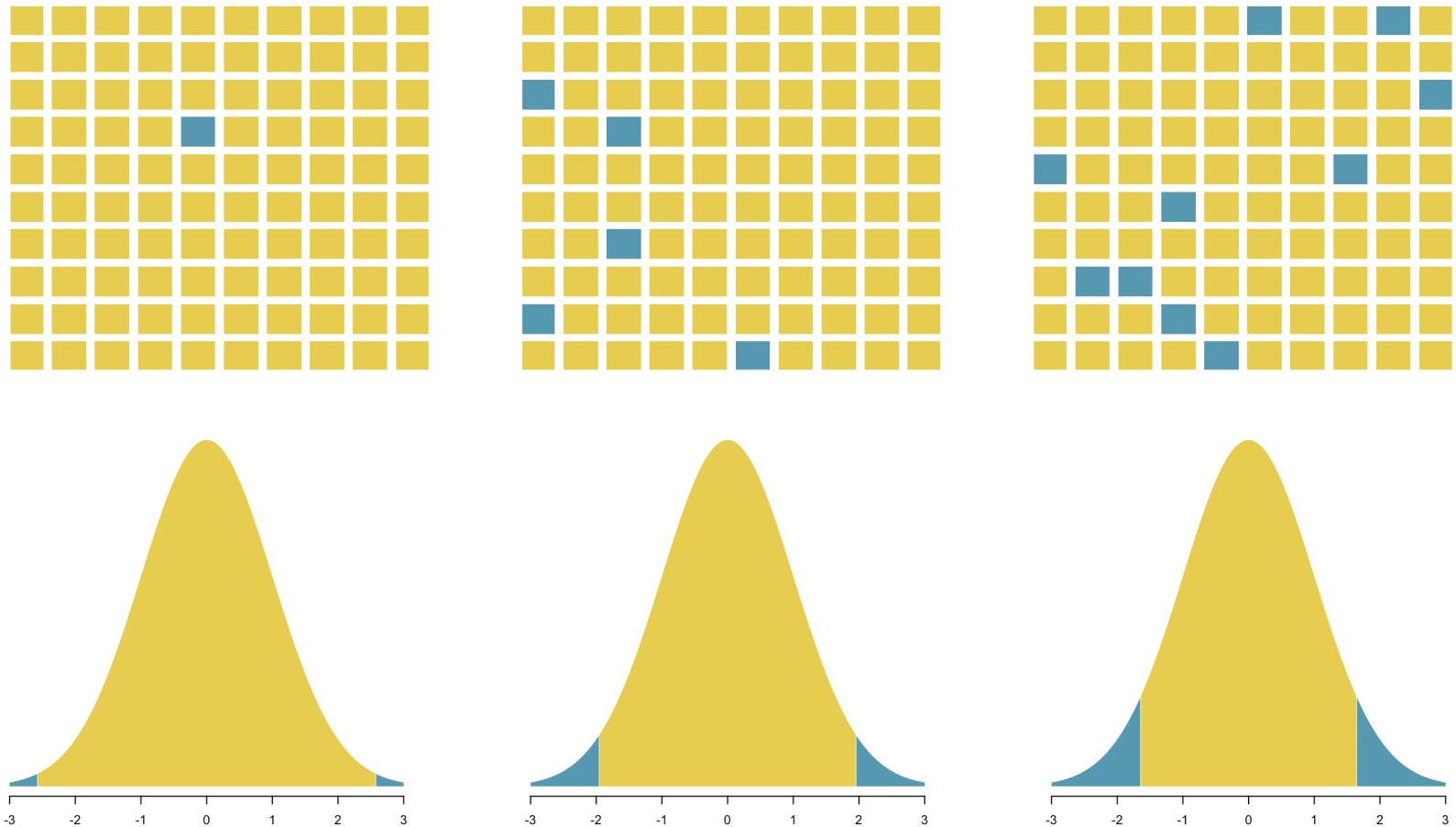
initialization : $\boldsymbol{x} = \{x_1, \cdots, x_n\}$;
for $b = 1, 2, ..., B$ **do**
    $\boldsymbol{x}_b \leftarrow$ resample from $\boldsymbol{x}$ (with replacement, same size);
    $t_b = T(\boldsymbol{x}_b)$

In python and R

```
> from sklearn.utils import resample
> x = [0.1, 0.2, 0.3, 0.4, 0.5, 0.6]
> xb = resample(x, replace=True)
> xb
[0.5, 0.3, 0.4, 0.5, 0.6, 0.5]
```

```
> x = c(0.1, 0.2, 0.3, 0.4, 0.5, 0.6)
> xb = sample(x,replace=TRUE)
> xb
[1] 0.4 0.4 0.5 0.1 0.5 0.6
```

1%, 5% or 10% chance



(Parametric) Probability Distribution

Is that the only way...?

What if we *know* the distribution of X_i's ?

We have here [for the Poisson distribution] a special case of the remarkable fact that there exist a few distributions of great universality which occur in a surprisingly great variety of problems. The three principal distributions, with ramifications throughout probability theory, are the binomial distribution, the normal distribution...and the Poisson distribution, William Feller, mentioned in Computational Models, Paul Humphreys,
(see 431, ... 434)

Monte Carlo

STAN ULAM, JOHN VON NEUMANN, *and the* MONTE CARLO METHOD

by Roger Eckhardt

The Monte Carlo method is a statistical sampling technique that over the years has been applied successfully to a vast number of scientific problems. Although the computer codes that implement Monte Carlo have grown ever more sophisticated, the essence of the method is captured in some unpublished remarks Stan made in 1983 about solitaire.

"The first thoughts and attempts I made to practice [the Monte Carlo method] were suggested by a question which occurred to me in 1946 as I was convalescing from an illness and playing solitaires. The question was what are the chances that a Canfield solitaire laid out with 52 cards will come out successfully? After spending a lot of time trying to estimate them by pure combinatorial calculations, I wondered whether a more practical method than "abstract thinking" might not be to lay it out say one hundred times and simply observe and count the number of successful plays. This was already possible to envisage with the beginning of the new era of fast computers, and I immediately thought of problems of neutron diffusion and other questions of mathematical physics, and more generally how to change processes described by certain differential equations into an equivalent form interpretable as a succession of random operations. Later... [in 1946, I] described the idea to John von Neumann and we began to plan actual calculations."

Von Neumann was intrigued. Statistical sampling was already well known in mathematics, but he was taken by the idea of doing such sampling using the newly developed electronic computing techniques. The approach seemed especially suitable for exploring the behavior of neutron chain reactions in fission devices. In particular, neutron multiplication rates could be estimated and used to predict the explosive behavior of the various fission weapons then being designed.

In March of 1947, he wrote to Robert Richtmyer, at that time the Theoretical Division Leader at Los Alamos (Fig. 1). He had concluded that "the statistical approach is very well suited to a digital treatment," and he outlined in some detail how this method could be used to solve neutron diffusion and multiplication problems in fission devices for the case "of 'inert' criticality" (that is, approximated as momentarily static config-

Los Alamos Science Special Issue 1987

131

Generating Random Numbers

predicting without understanding

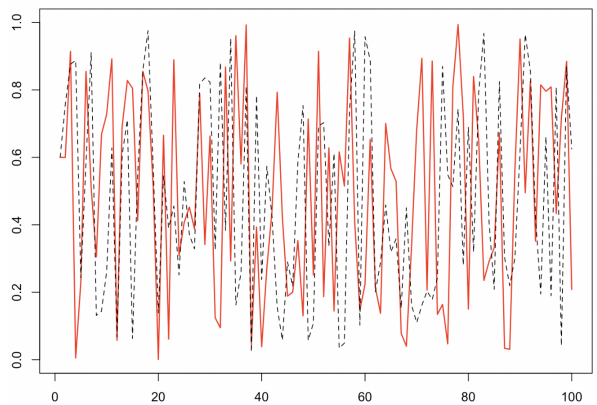
or

understanding without predicting

Quelle responsabilité pour les algorithmes ?

Sequences generated with Sedgewick algorithm, $u_n = x_n/m$ where

$$x_n = (ax_{n-1} + c) \text{ modulo } m$$



$a = 1013904223$, $c = 1664525$ and $m = 2^{32}$ with $u_1 = 0.6$ and $u_1 = 0.60001$.

See Histoire du hasard et de la simulation for further discussion

Generating Random Numbers

A Million Random Digits with 100,000 Normal Deviates, RAND, 1955

★☆☆☆☆ **Errors Throughout**
Reviewed in the United States on December 30, 2013

They sure don't come up with random numbers like they used to. If you look closely, you will note that every tenth digit or so is just a repeat of the last digit and every hundredth or so is a just the same digit repeated three times. How sloppy!

A sampling of this "work":

Page 36 - Line 6 - 15 characters in should be 5, not 4.
Page 99 - Line 18 - first three characters should be "453" not "345".
Page 145 - Line 2 - 7th and 19th characters transposed.
Page 190 - Whole line of numbers omitted betwen 6th and 7th lines.
Pages 210 and 211 - Two sections appear quasi-randomized, instead of randomized.

★★★☆☆ **Predictable**
Reviewed in the United States on October 30, 2013

It seemed like about 10% of the time I was able to predict which number was next. It was still better than Life of Pi which, aside from being irrational, included no estimations of Pi at all.

★☆☆☆☆ **Not really random**
Reviewed in the United States on September 26, 2012

I bought two copies of this book. I find that the first copy perfectly predicts what the numbers will be in the second copy. I feel cheated.

★★★★★ **great yet lacking....**
Reviewed in the United States on May 1, 2013

So far, it's great, but what I REALLY want is an audio version that I can listen to at the beach or the gym. Maybe narrated by Morgan Freeman???

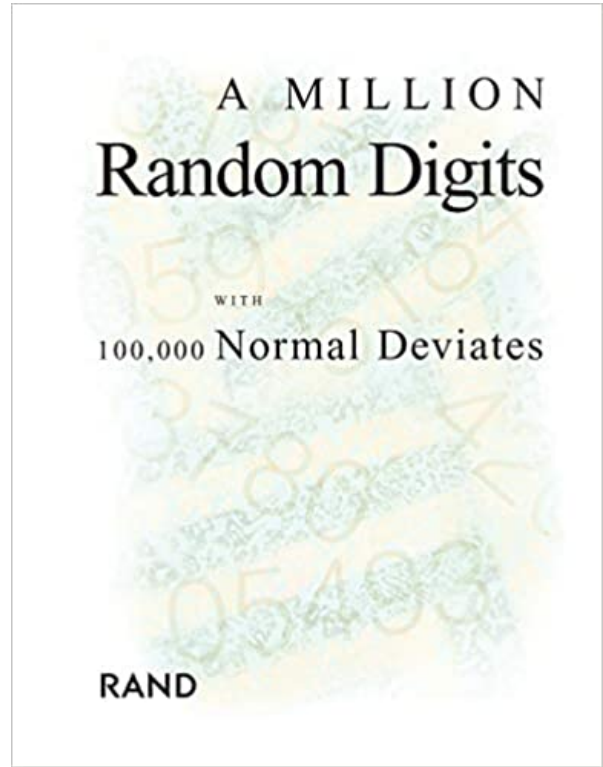


TABLE OF RANDOM DIGITS 1

00000	10097	32533	76520	13586	34673	54876	80959	09117	39292	74945
00001	37542	04805	64894	74296	24805	24037	20636	10402	00822	91665
00002	08422	68953	19645	09303	23209	02560	15953	34764	35080	33606
00003	99019	02529	09376	70715	38311	31165	88676	74397	04436	27659
00004	12807	99970	80157	36147	64032	36653	98951	16877	12171	76833
00005	66065	74717	34072	76850	36697	36170	65813	39885	11199	29170
00006	31060	10805	45571	82406	35303	42614	86799	07439	23403	09732
00007	85269	77602	02051	65692	68665	74818	73053	85247	18623	88579
00008	63573	32135	05325	47048	90553	57548	28468	28709	83491	25624
00009	73796	45753	03529	64778	35808	34282	60935	20344	35273	88435
00010	98520	17767	14905	68607	22109	40558	60970	93433	50500	73998
00011	11805	05431	39808	27732	50725	68248	29405	24201	52775	67851
00012	83452	99634	06288	98083	13746	70078	18475	40610	68711	77817
00013	88685	40200	86507	58401	36766	67951	90364	76493	29609	11062
00014	99594	67348	87517	64969	91826	08928	93785	61368	23478	34113
00015	65481	17674	17468	50950	58047	76974	73039	57186	40218	16544
00016	80124	35635	17727	08015	45318	22374	21115	78253	14385	53763
00017	74350	99817	77402	77214	43236	00210	45521	64237	96286	02655
00018	69916	26803	66252	29148	36936	87203	76621	13990	94400	56418
00019	09893	20505	14225	68514	46427	56788	96297	78822	54382	14598
00020	91499	14523	68479	27686	46162	83554	94750	89923	37089	20048
00021	80336	94598	26940	36858	70297	34135	53140	33340	42050	82341
00022	44104	81949	85157	47954	32979	26575	57600	40881	22222	06413
00023	12550	73742	11100	02040	12860	74697	96644	89439	28707	25815
00024	63606	49329	16505	34484	40219	52563	43651	77082	07207	31790
00025	61196	90446	26457	47774	51924	33729	65394	59593	42582	60527

Generating Random Numbers

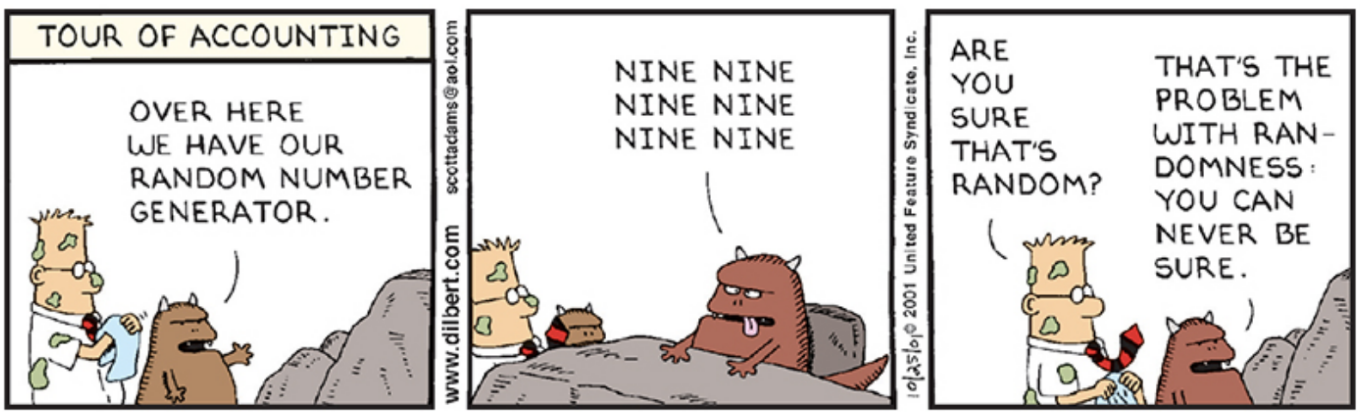
One can easily draw 'random' numbers from any distribution
most popular : uniform on $[0, 1]$ and Gaussian $\mathcal{N}(0, 1)$
In python and R

```
> import random
> random.uniform(0, 1)
0.15473258779393262
> import numpy as np
> x = np.random.normal(0, 1, 10)
> x
array([ 1.3136,  0.0829,  1.1754, -0.7242,  0.5681,
        0.0701, -0.0567,  0.3344, -1.1355, 1.1510])
```

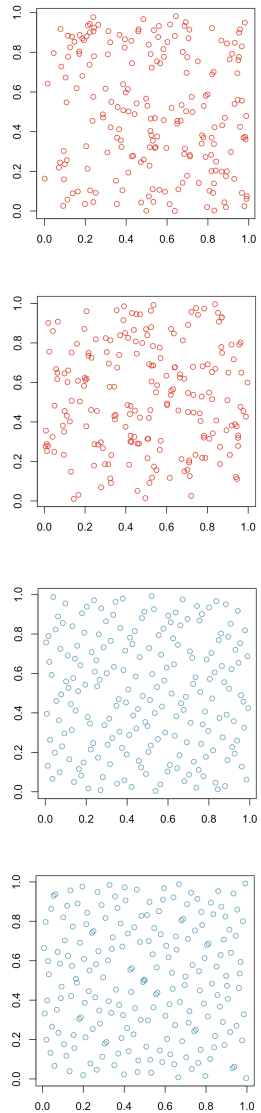
```
> runif(10)
 [1] 0.1118923 0.1714409 0.3878056 0.4111353 0.3819566
 [6] 0.0145239 0.2615349 0.8561546 0.9593199 0.2472850
 > rnorm(10)
 [1] -0.234491 -0.891424 -1.357470 -0.5840280  1.66809
 [6]  1.085743  0.557528 -1.636832 -0.271538  1.091798
```

Generating Random Numbers

- draw $n =$ one million values
  and plot the histogram to check
- values should be "independent"
  (between n and $n + 1$)

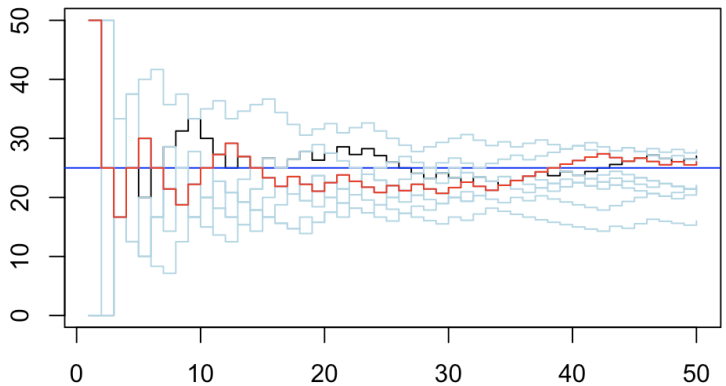


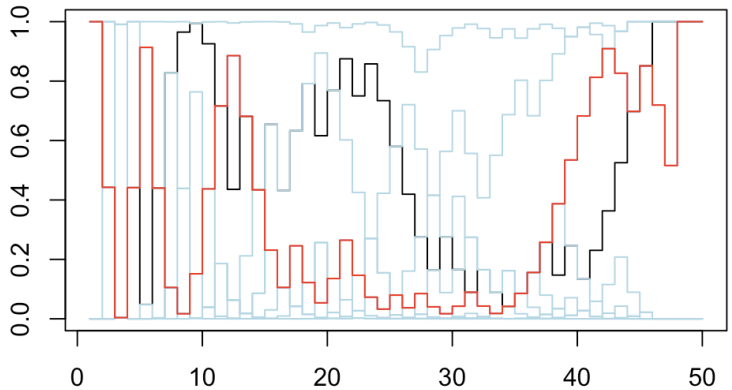
via Dilbert

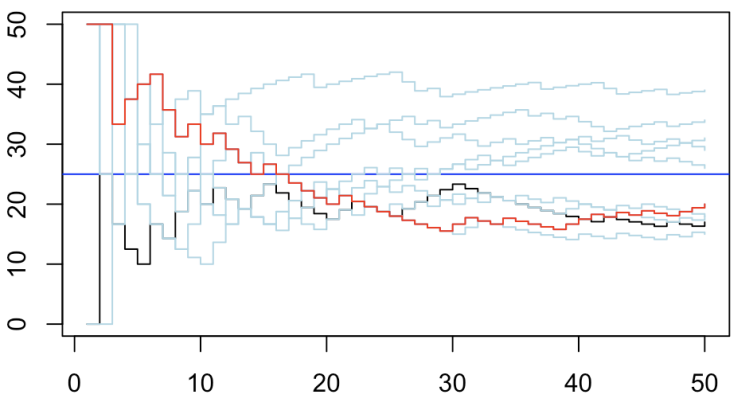


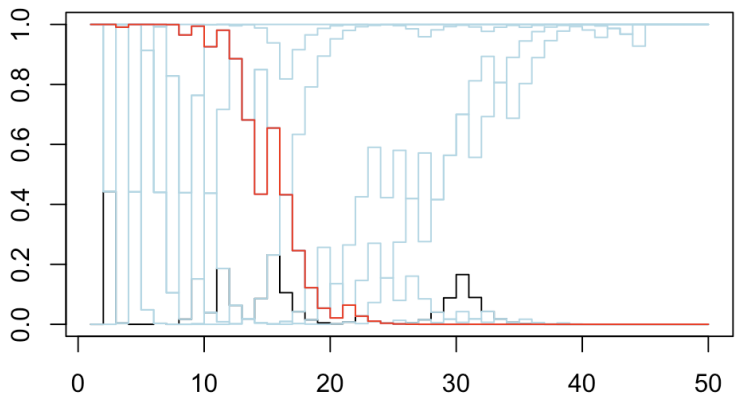
Forecasting Elections

Evolution of $n \mapsto \widehat{p}_n = \overline{x}_n$ (left)
and $n \mapsto \mathbb{P}[\overline{X}_N > 1/2 | x_1, \cdots, x_n]$ (right)









Forecasting Mortality

Important to add confidence bands on predictions
See recently @Alex_Blanchet's tweet on mortality in Québec