The Birthday Paradox: A Practical Analysis

Abhishek Banerjee, Ian Klatzco, Robert Kaufman

Overview

- We “scraped” **43,960** birthdays of assorted individuals from the “Astro-Databank Project” and analyzed them for birthday problem type coincidences.
  - Collisions and Strong Collisions
  - Comparison with randomly-generated birthdays
  - “Broken Stick” Analysis
- We expected some birthdate bias going in

- To do this project, we used Python, a popular programming language.

Data Extraction

- We used Python package to extract data from webpages:
  - requests - getting webpages
  - lxml - webpage parsing
-
- The code went through approx. 130 pages each containing around 500 links to pages of people recorded in the “Astro-Databank Project” and placed them into a list.
-
- Then, we extracted and stored the birthdate for each of appx 46000 pages

ASTRO DATABANK maintained by www.astro.com

Special page

All pages

All pages | Previous page (1943 Frankford Junction derailment) | Next page (Adams, Rev. Joy)

Main page
All pages
New pages
Recent changes
Random page
Help
Help external to Astrodienst

Tools
Special pages
Printable version

All pages

Display pages starting at: 1943 Frankford Junction derailment

Display pages ending at:

Namespace: (Main) Hide redirects

Go

1943 Frankford Junction derailment	2015 Eckwersheim TGV derailment	*2015 Murders of Alison and Adam*
2015 Philadelphia train derailment	*2015 Thalys attack*	*2015 Tianjin explosion*
A, Dominique	A. D. G.	A. J. Croce
Aabel, Andreas	Aabel, Per	Aadland, Beverly
Aadland, Florence	Aafjes, Bertus	Aal-Pomares, Henri Francois
Aalberg, Andre	Aalberg, André	Aaliyah
Aantjes, Willem	Aaron, Dave	Aaron, Didier
Aaron, Hank	Aaron, Jean-Claude	Aarts, Johannes Josephus

Main page
All pages
New pages
Recent changes
Random page
Help
Help external
to Astrodienst

Tools
What links here
Related changes
Special pages
Printable version
Permanent link
Page information

Hamilton, Alexander

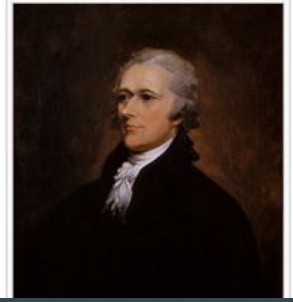
Name	Hamilton, Alexander	Gender: M
born on	11 January 1755	
Place	Charlestown, St Kitts-Nevis, 17n08, 62w37	
Timezone	LMT m62w37 (is local mean time)	
Data source	Date w/o time	Rodden Rating **X**
		Collector: Starkman
Astrology data	☉ ♑ 21°14' ☽ ♑	

add Alexander Hamilton to 'my astro'

Alexander Hamilton
natal chart (noon, no houses)
natal chart English style (noon, no houses)

Biography

Founding Father of the United States, chief of staff to General Washington, one of the most influential interpreters and promoters of the Constitution, the founder of the nation's financial system, and the founder of the first American political party.



As Secretary of the Treasury, Hamilton was the primary author of the economic policies of the George Washington administration, especially the funding of the state debts by the Federal government, the establishment of a national bank, a system of tariffs, and friendly trade relations with Britain. He became the leader of the Federalist Party, created largely in support of his views, and was opposed by the Democratic-Republican Party, led by Thomas Jefferson and James Madison.

Hamilton served in the American Revolutionary War. At the start of the war, he organized an artillery company and was chosen as its captain. He later became the senior aide-de-camp and confidant to General George Washington, the American commander-in-chief. He served again under Washington in the army raised to defeat the Whiskey Rebellion, a tax revolt of western farmers in 1794. In 1798, Hamilton called for mobilization against France after the XYZ Affair and secured an appointment as commander of a new army, which he

Parsing

- Post-extraction, the birthdates were not in a form favorable to analysis.

- We needed to “parse” the data - write a script to make it look like:

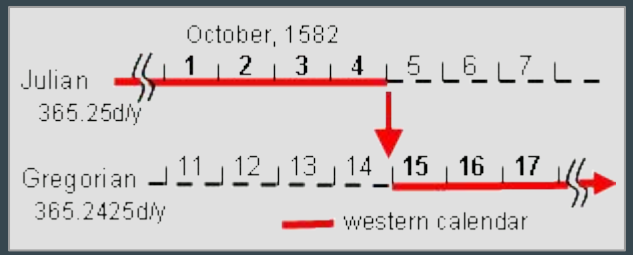
42/365, 93/365, 182/365, 12/365, 201/365

- We initially sorted it into two categories:
  - Only dates with birth times
  - All dates, without birth times

Parsing - Initial Problems

- Python libraries account for leap years, whereas we wanted to ignore them.
  - If the year was not a leap year, we kept the regular day.
  - If it was and it came after the leap day, we subtracted one from the day of the year

- The parser code also had to handle julian dates by converting them to gregorian



http://keisan.casio.com/keisan/lib/real/system/2006/1227757509/Calendar.gif

Parsing - More Problems

- In our parsing code, ~~Ian messed up~~ there was an error that created duplicates of Julian-converted birthdates back to back.

- In order to fix that ~~Rob un-messed up~~ we wrote a little code that removed a date if the same date had come just before it.

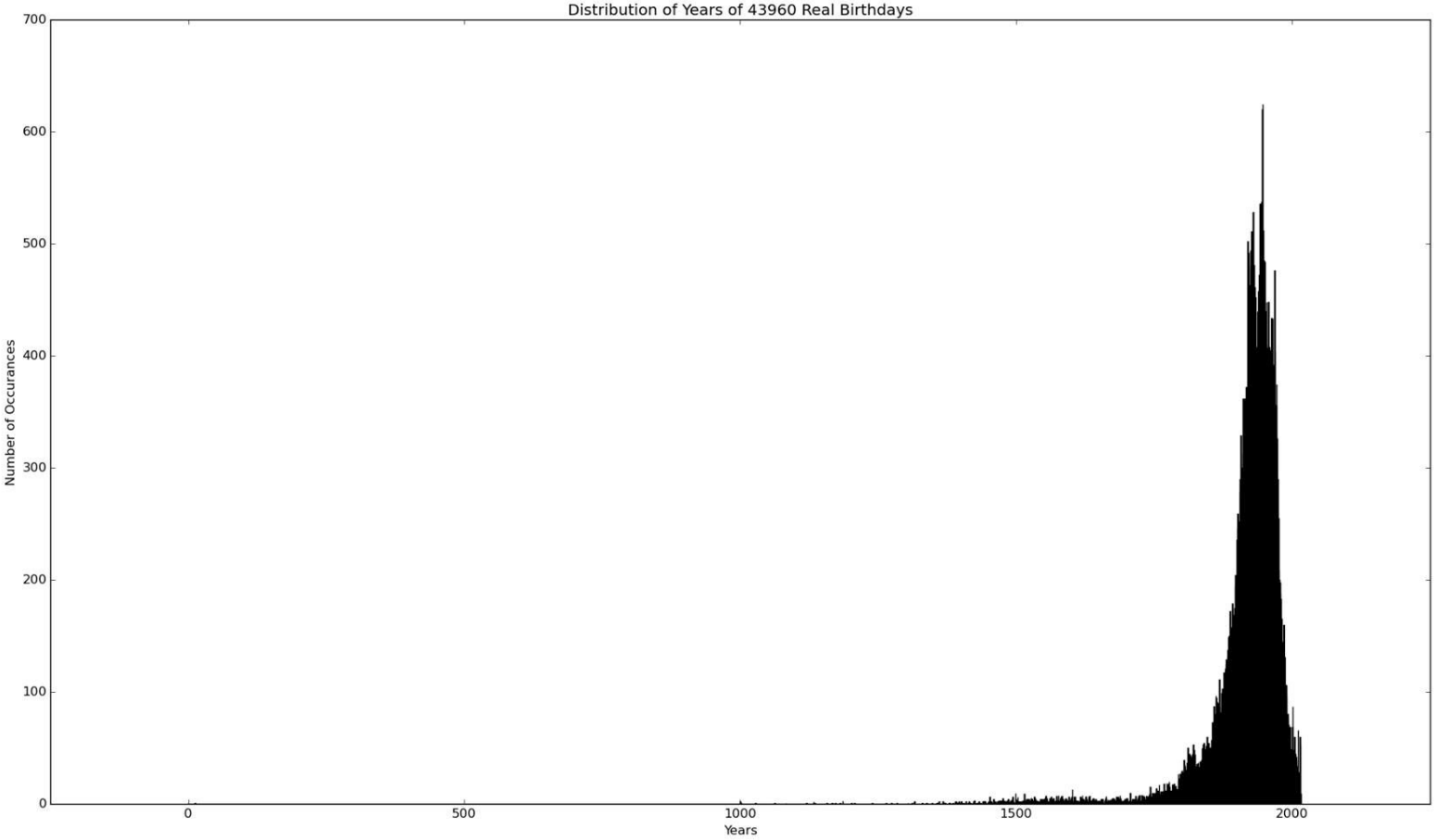
- It's possible that in doing this we introduced some error for two consecutive birthdates, but improbable.

Data Analysis

- We analyzed our data using a Python plotting library.
  - mathplotlib

- Firstly, we plotted “Collisions” — when two people have identical birthdates. We also checked yearly and month birth frequency.

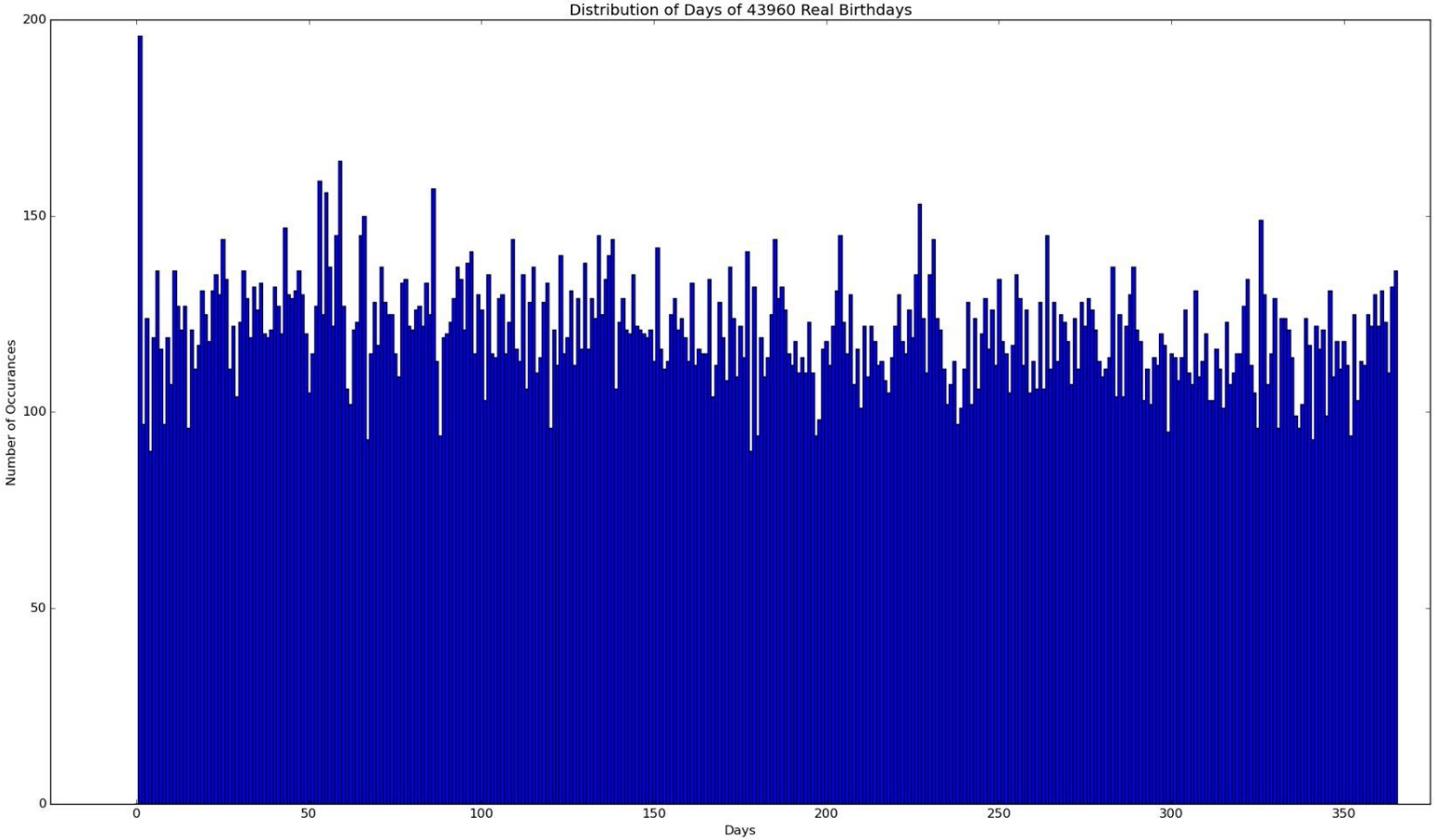
- Lastly we created a graph that summarized analysis of the birthday version of the broken stick problem.


Distribution of Years of 43960 Real Birthdays
Number of Occurances
Years
0
100
200
300
400
500
600
700
0
500
1000
1500
2000

Year Graph Analysis

- The previous graph contained the years of the birth of all of the people in our dataset.

- The maximum number of birthdays, 625, occurred in the year 1947.

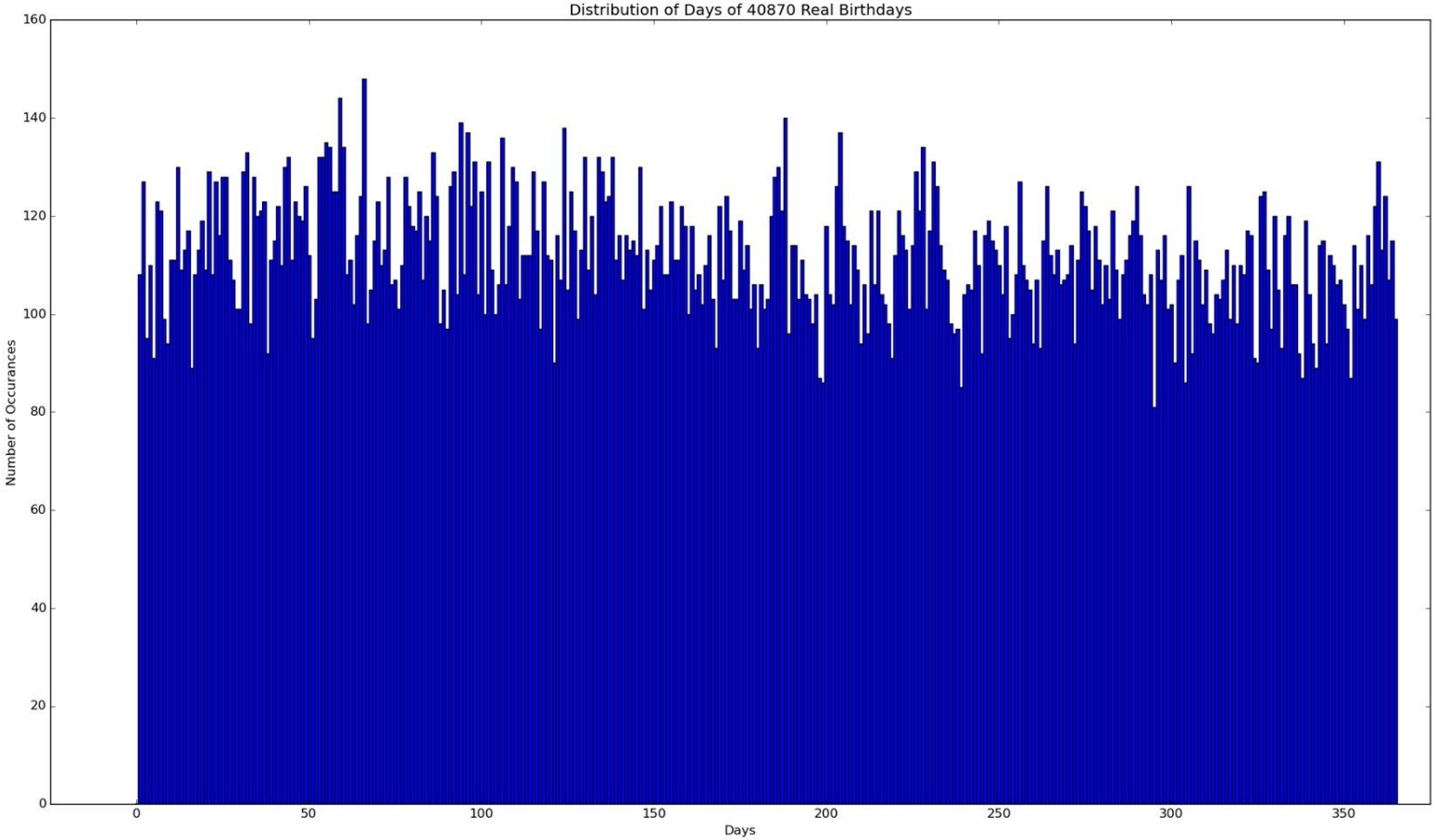

Distribution of Days of 43960 Real Birthdays
Number of Occurances
Days
0
50
100
150
200
250
300
350
200
150
100
50
0

Day Graph I Analysis

- The previous graph contained the days of birth of people with BOTH time of birth available and those without it.

- Our graph seems to have a fairly uniform distribution except for January 1st.

- January 1st has a disproportionate amount of births, 215, compared to average of 127/day, probably due to some sort of selection bias.

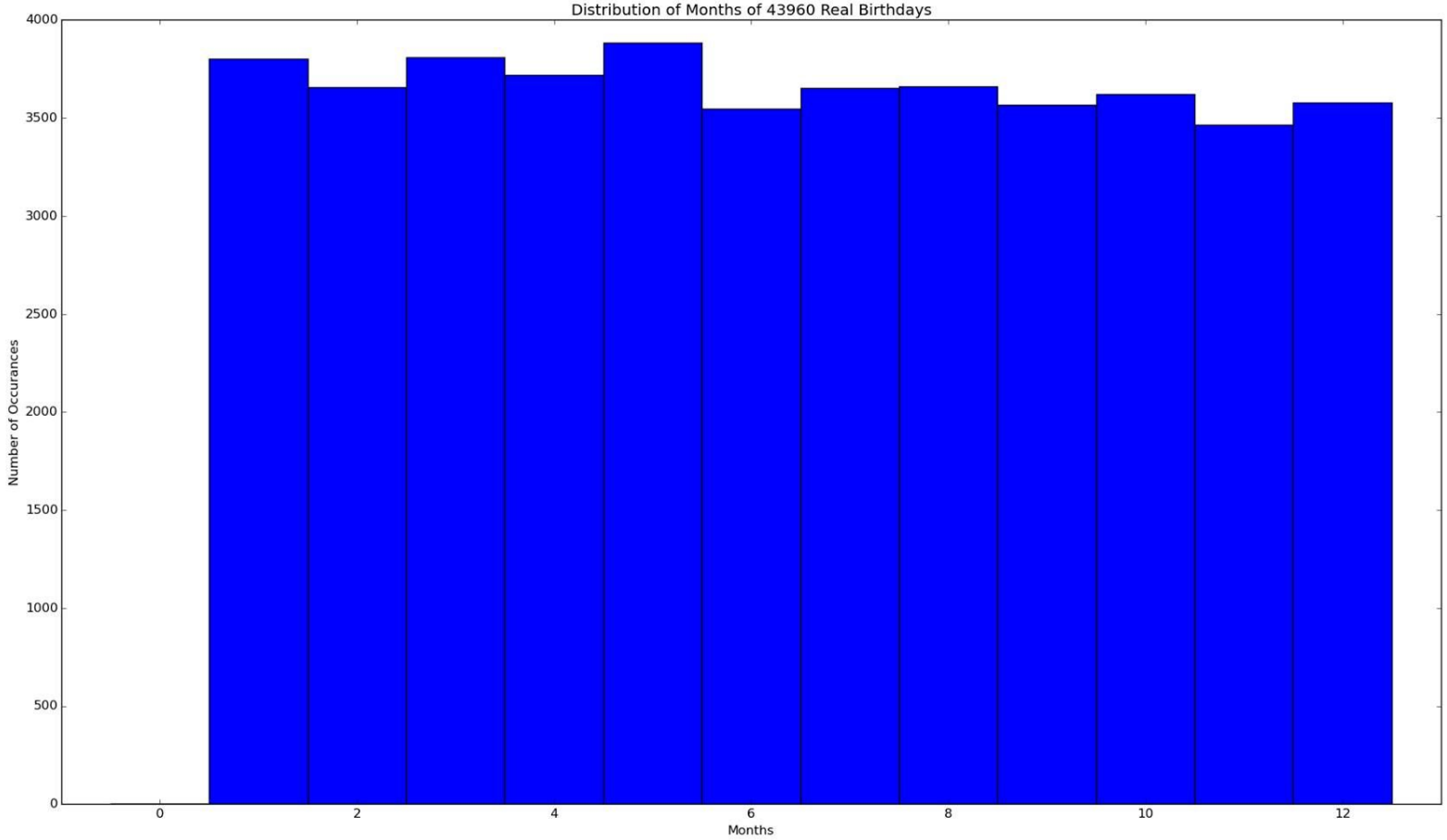

Distribution of Days of 40870 Real Birthdays
Number of Occurances
Days
0
20
40
60
80
100
120
140
160
0
50
100
150
200
250
300
350

Day Graph II Analysis

- The previous graph contained the birth dates of those that ONLY had the day available and not the time.

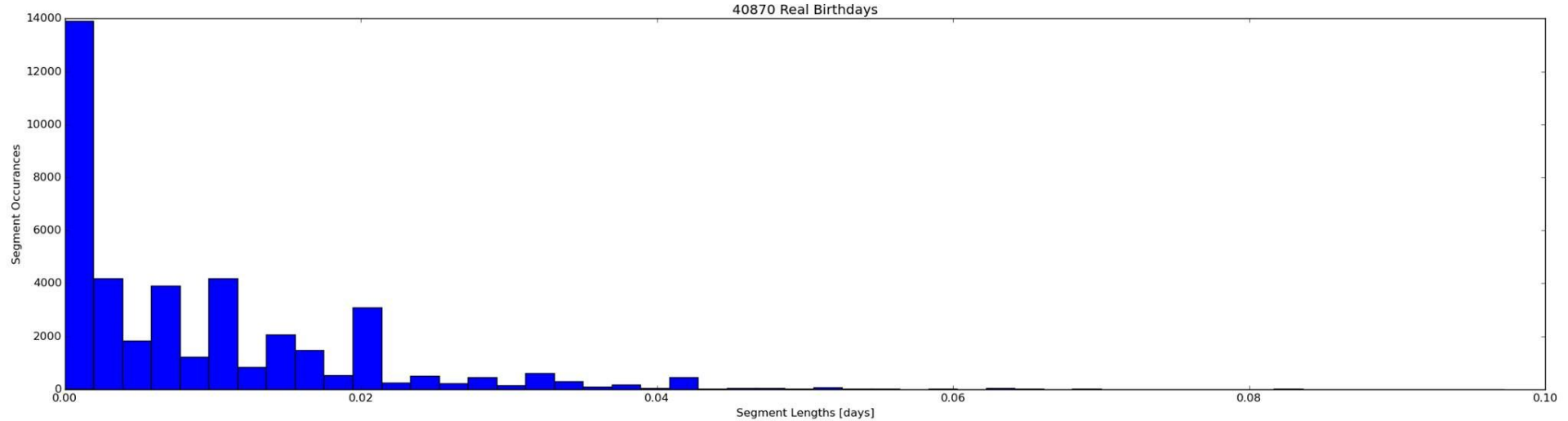
- The graph seems to be fairly uniformly distributed.

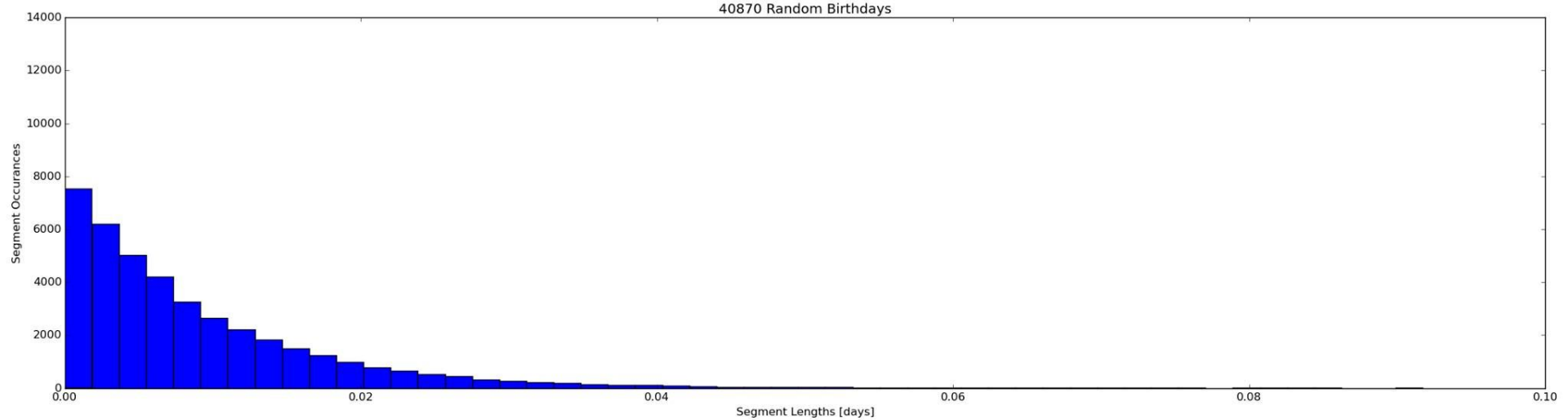
- This means that when we remove the birthdays with times, the amount of people born on each day is roughly equal.


Distribution of Months of 43960 Real Birthdays
Number of Occurances
Months
0
500
1000
1500
2000
2500
3000
3500
4000
0
2
4
6
8
10
12

Month Graph Analysis

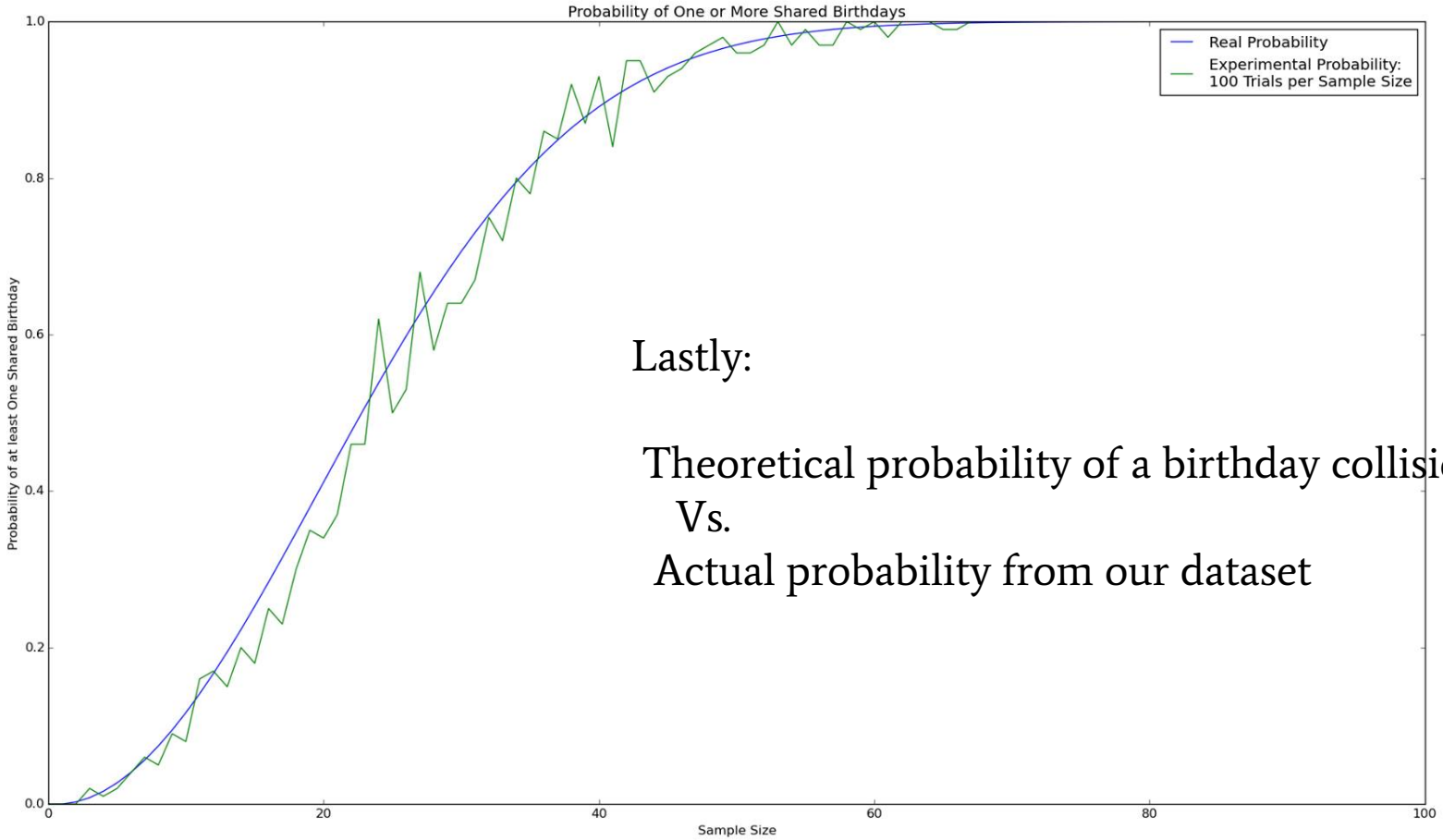
- The previous graph contained the distribution of people born in each month. This graph included both time and no time birth dates.
-
- The graph seems to also be fairly uniformly distributed.
  - Max: May
  - Min: November
-
- This means that in our large dataset, a roughly equal amount of people were born in each month.


40870 Real Birthdays
Segment Occurances
Segment Lengths [days]
0
2000
4000
6000
8000
10000
12000
14000
0.00
0.02
0.04
0.06
0.08
0.10


40870 Random Birthdays
Segment Occurances
Segment Lengths [days]
0
2000
4000
6000
8000
10000
12000
14000
0.00
0.02
0.04
0.06
0.08
0.10

Broken Stick Graph Analysis

- The x-axis of the graph represents the length between two points on a stick represented by days. (Fractions are smaller as 10,000s of points were placed on a graph with only 366 days.)
- The y-axis is the frequency of occurrences of a given segment length.
- The bottom graph is from a randomly generated set of 40,870 numbers from 1 to 365.
- We can see that the bottom graph is normally distributed while the birth date data has a disproportionate amount of data points very close to each other.



Lastly:

Theoretical probability of a birthday collision
Vs.
Actual probability from our dataset

Conclusions Drawn

- Data work is **hard**
  - Getting good data is **hard**

- Real-life birthdays have some deviation from the expected values
  - Eg,. Jan 1 bias

- The birthday paradox's answer is, in fact, backed by actual data

Works Cited

Dasgupta, Anirban. "The Matching, Birthday and the Strong Birthday Problem: A Contemporary Review." *Journal of Statistical Planning and Inference* 130.1-2 (2005): 377-89. ELSEVIER. Web. 11 Apr. 2016.

Rodden, Lois. "Main Page." *Astrology: Birth Data and Horoscope of 20000 Celebrities, Horoscopes for Astrological Research*. N.p., n.d. Web. 11 Apr. 2016.

Code available for your perusal on github.com/ianklatzco/astroDownloader.