

# STOCHASTIC THINKING



(download slides and .py files from Stellar to follow along!)

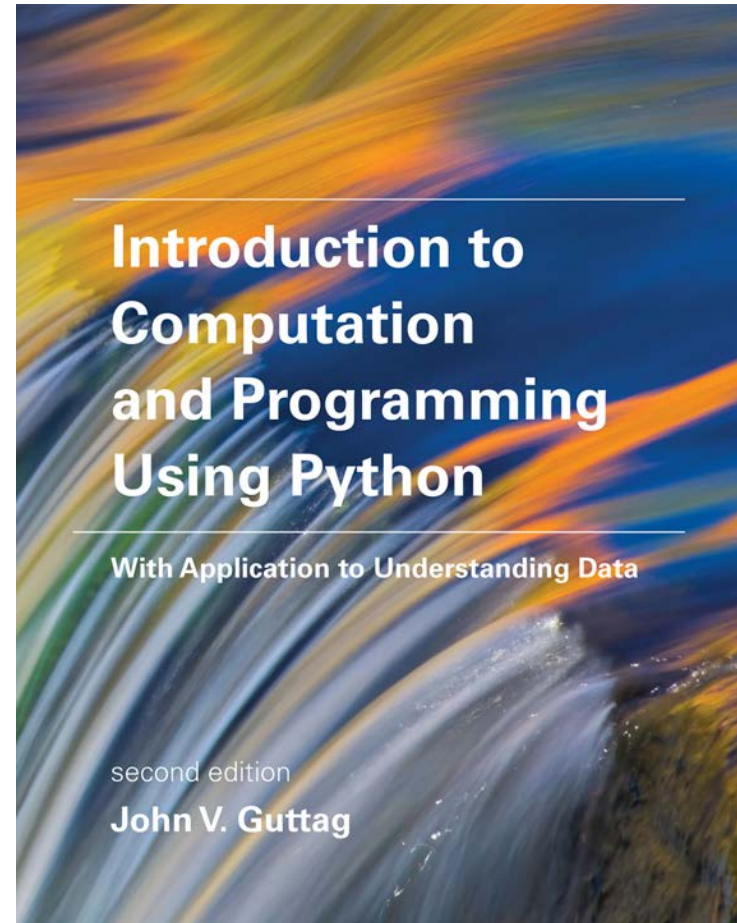
---

6.0002 LECTURE 4

# Assigned Reading

---

- Today:
  - Sections 15.1-15.3
  - Section 15.5
- Monday:
  - Chapter 14



# The Simple World of Newtonian Mechanics

---

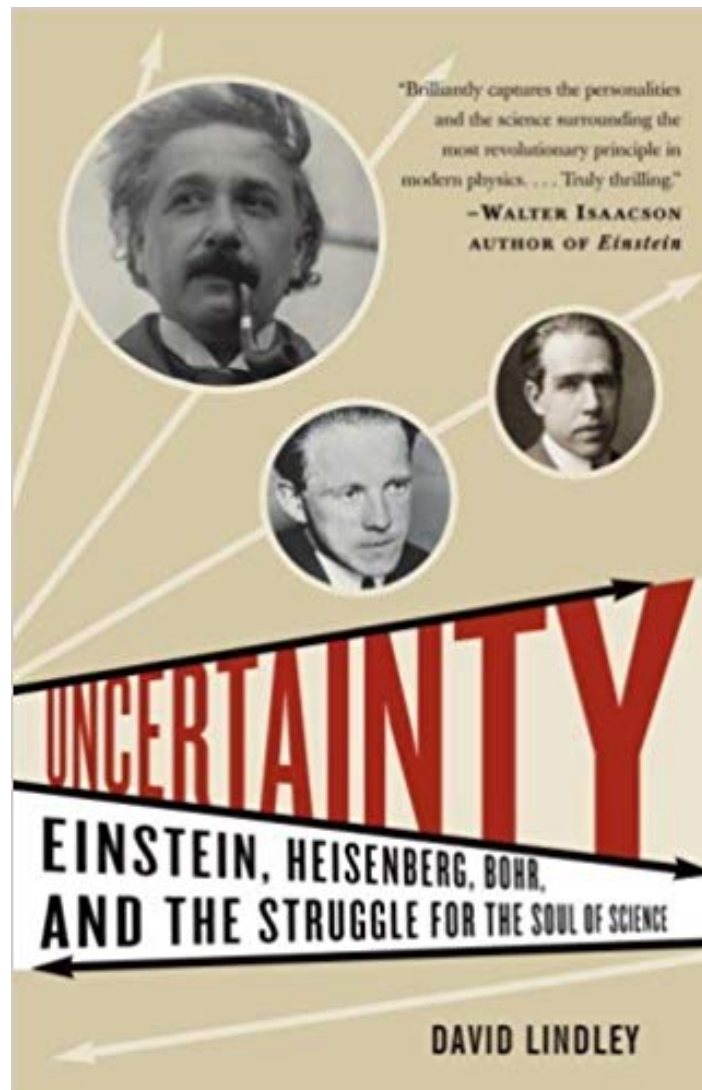
- Every effect has a cause
- The physical world can be understood causally
- 18<sup>th</sup> century mathematics
  - 18.01, 18.02
- 18<sup>th</sup>-19<sup>th</sup> century physics
  - 8.01



1643 – 1727

# Two Centuries Later

---



# Copenhagen Doctrine

- Heisenberg and Bohr argued that at its most fundamental level, the behavior of the physical world cannot be predicted
  - For example, cannot precisely measure position and momentum of a particle at the same time
  - Fine to make statements of the form “x is highly likely to occur,” but not of the form “x is certain to occur”

“The more precise the measurement of position, the more imprecise the measurement of momentum, and vice versa.”

“Those who are not shocked when they first come across quantum theory cannot possibly have understood it. ...If you think you can talk about quantum theory without feeling dizzy, you haven't understood the first thing about it.”



# Many Were Indeed Shocked

---

- Einstein and Schrodinger objected
  - “God does not play dice with the universe” – Albert Einstein



# DOES IT REALLY MATTER?

---



- Suppose we flip a coin twice
- Could we predict whether the flips would yield
  - 2 heads
  - 2 tails
  - 1 head and 1 tail?
- Need to know accurately:
  - weight distribution of coin
  - exact velocity and acceleration of finger
  - air flow
  - height above landing spot
  - other effects

# The Moral

- The world may or may not be inherently unpredictable
- But lack of knowledge doesn't allow accurate predictions
- Therefore, we might as well treat the world as inherently unpredictable



**Causal nondeterminism** – some events truly random

**Predictive nondeterminism** – in principle might be able to predict, but don't have enough information. There is chaos, but not randomness

# Why Should We Care about Stochastics?

---

- Want to model systems where there are choices in state transitions
  - Where will a molecule move when a fluid is heated?
- Want to model systems where measurement is uncertain
  - How tall are you?
- Want to model systems where we can't measure entire system
  - Will the Celtics win the NBA title this year?
- Etc.



# Stochastic Processes

---

- In general, a system defined by a set of state variables, and processes that determine the next set of values
- In a stochastic system, the process for determining next state might depend on both the previous states **and some random element**

# Two Specifications

---

```
def rollDie():  
    """returns an int between 1 and 6"""  
    return 3
```

```
import random
```

```
def rollDie():  
    """returns a random int between 1 and 6"""  
    return random.choice([1,2,3,4,5,6])
```

Any implementation that satisfies the second specification would also satisfy the first.

But one that satisfies the first specification might or might not satisfy the second

# Trying It

---

```
def rollDie():  
    """returns a random int between 1 and 6"""  
    return random.choice([1,2,3,4,5,6])  
  
def testRoll(n):  
    result = ''  
    for i in range(n):  
        result = result + str(rollDie())  
    return result  
  
for i in range(10):  
    print(testRoll(5))
```

Predictive or causal non-determinism?  
How probable is the output 11111?

# Probability is About Counting

---

- Count the number of possible events
- Count the number of events that have the property of interest
- Divide second by the first
- Probability of 11111?
  - All events: 11111, 11112, 11113, ..., 11121, 11122, ..., 66666
  - Ratio:  $1/(6^{*}5)$
  - $\sim 0.0001286$
- Probability of 12345?

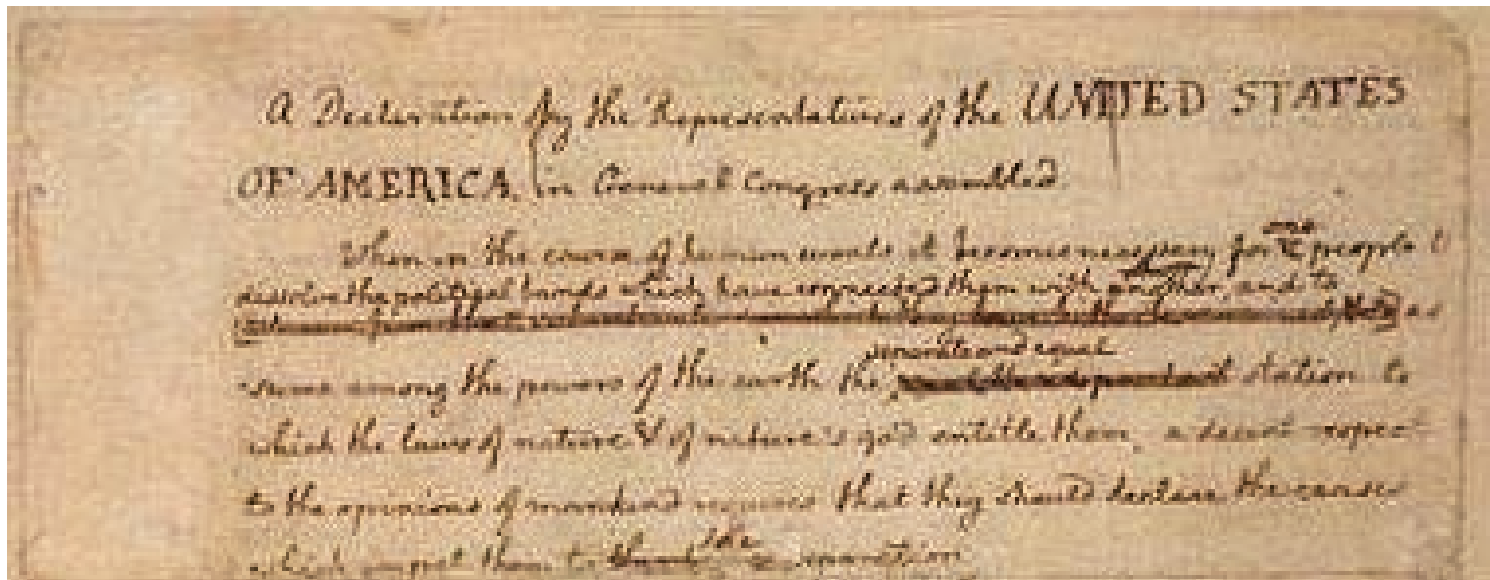
# Some Basic Facts about Probability

---

- Probabilities are always in the range **0 to 1**. 0 if impossible, and 1 if guaranteed
- If the probability of an event occurring is **p**, the probability of it not occurring must be **1-p**
- When events are **independent** of each other, the probability of all of the events occurring is equal to the **product** of the probabilities of each of the events occurring

# Independence

- Two events are **independent** if the outcome of one event has no influence on the outcome of the other
- Independence should not be taken for granted
  - Is your score on today's micro quiz independent of how well you did on the first problem set?



# Some Probability Algebra

---

- $P(A) = \frac{\text{number of events in } A}{\text{number of events in } U}$
  - $P(A \text{ and } B) = P(A)P(B)$ , if  $P(A)$  and  $P(B)$  independent
    - What is probability of flipping a head and rolling a 5 on a die?
    - $P(\text{head}) = 1/2$
    - $P(\text{roll } 5) = 1/6$
    - Product is  $1/12$
  - $P(A \text{ or } B) ?$
- $P(A) + P(B) ?$*

H1, H2, H3, H4, H5, H6  
T1, T2, T3, T4, T5, T6

# Some Probability Algebra

---

- $P(A) = \frac{\text{number of events in } A}{\text{number of events in } U}$
- $P(A \text{ and } B) = P(A)P(B)$  if independent
  - What is probability of flipping a head and rolling a 5 on a die?
  - $P(\text{head}) = 1/2$
  - $P(\text{roll } 5) = 1/6$
  - Product is  $1/12$
- $P(A \text{ or } B) ?$ 
  - $1 - P(\neg A \text{ and } \neg B)$

H1, H2, H3, H4, H5, H6  
T1, T2, T3, T4, T5, T6

$P(\text{head or rolling } 5)?$

$$1 - (1/2 * 5/6) = 7/12$$

# Independent vs. Dependent Probabilities

---

- Deal two cards at random from a standard deck of 52 cards
- What is the probability that the first card is a king and the second a king?
- **With replacement**
  - $4/52 * 4/52$
- **Without replacement?**



# Using Simulation to Estimate Probabilities

---


- First of many simulations we will see
- Talk more carefully about simulation in future lectures
- Simulation
  - Run many trials in which we select one value from the universe of possible values
  - For each trial, compute some properties of value
  - Report some statistics about the properties over the set of trials

# A Simulation of Die Rolling

---

```
def runSim(goal, numTrials):
    total = 0
    for i in range(numTrials):
        if i != 0 and i%1000 == 0:
            print('Starting trial', i)
        result = ''
        for j in range(len(goal)):
            result += str(rollDie())
        if result == goal:
            total += 1
    print('Actual probability of', goal, '=',
          round(1/(6**len(goal)), 8))
    estProbability = round(total/numTrials, 8)
    print('Estimated Probability of', goal, '=',
          round(estProbability, 8))

runSim('11111', 1000)
```



# Output of Simulation

---

- Actual probability = 0.0001286
- Estimated probability = 0.0
- Actual probability comes directly from math, but estimated probability?
- Why did simulation give me the **wrong** answer?
  - $6 \times 5 = 7776$  possibilities, so 1000 trials unlikely to observe one event in that universe with enough frequency to estimate accurately
  - Even if observed once, estimate would then be 0.001

Let's try 1,000,000 trials

# Morals

---

- Moral 1: it takes a lot of trials to get a good estimate of the frequency of occurrence of a **rare** event. We'll talk lots more in later lectures about how to **know** when we have enough trials to trust the estimate
- Moral 2: one should not confuse the **sample probability** with the actual probability
- Moral 3: there was really no need to do this by simulation, since there is a perfectly good closed form answer. We will see many examples where this is not true

# The Birthday Problem

---

- What is the probability of at least two people in a group having the same birthday?
- If are 30 people in a room, should you be surprised if two share a birthday?



# The Birthday Problem

---

- What if there are 367 people in the group?
  - Use the “pigeonhole principle”
- What about fewer people?



# The Birthday Problem

---

- If assume each birthdate equally likely, then use “inclusion/exclusion principle”
- Probability that  $n$  ( $< 367$ ) people each have different birthday

$$\frac{366}{366} * \frac{365}{366} * \frac{364}{366} * \dots * \frac{366 - n + 1}{366} = \frac{366!}{366^N * (366 - N)!}$$

Don't worry  
about why it's  
this formula

- Probability that at least two of  $N$  people have same birthday:

$$1 - \frac{366!}{366^N * (366 - N)!}$$

# The Birthday Problem

---

```
import math

def trueProb(numPeople):
    #assumes each birth date equally probable
    numerator = math.factorial(366)
    denom = (366**numPeople)*math.factorial(366-numPeople)
    return 1 - numerator/denom

for i in (2, 4, 8, 16, 32, 64, 128, 256):
    print('Probability of a shared birthday with', i,
          'people =', round(trueProb(i), 8))
```

# Approximation Using a Simulation

```
def sameDate(numPeople, numSame, possibleDates):
    birthdays = [0]*366
    for p in range(numPeople):
        birthDate = random.choice(possibleDates)
        birthdays[birthDate] += 1
    return max(birthdays) >= numSame

def birthdayProb(numPeople, numSame, numTrials,
                 possibleDates):
    numHits = 0
    for t in range(numTrials):
        if t != 0 and t%100000 == 0:
            print('Starting trial', t)
        if sameDate(numPeople, numSame, possibleDates):
            numHits += 1
    return numHits/numTrials

possibleDates = range(366)
for i in (2, 4, 8, 16, 32, 64, 128, 256):
    print('Est. probability of a shared birthday with', i,
          'people =', birthdayProb(i, 2, 1000, possibleDates))
```

One trial

Set of trials

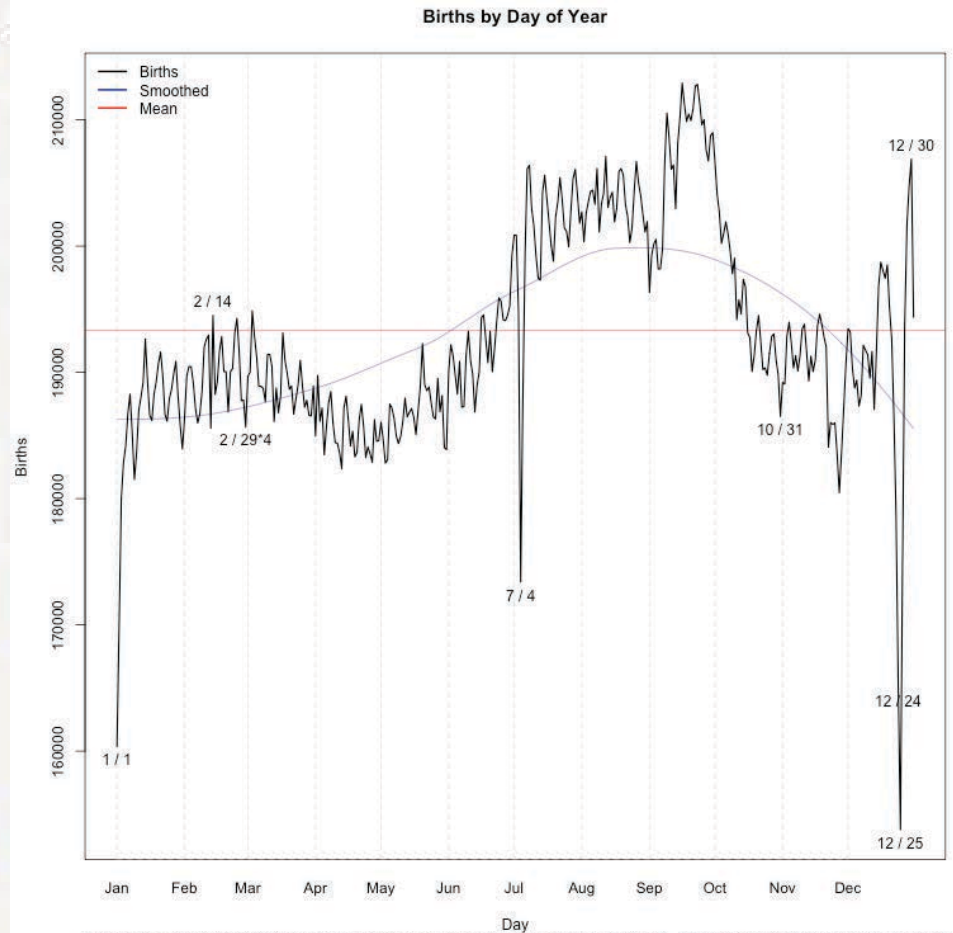
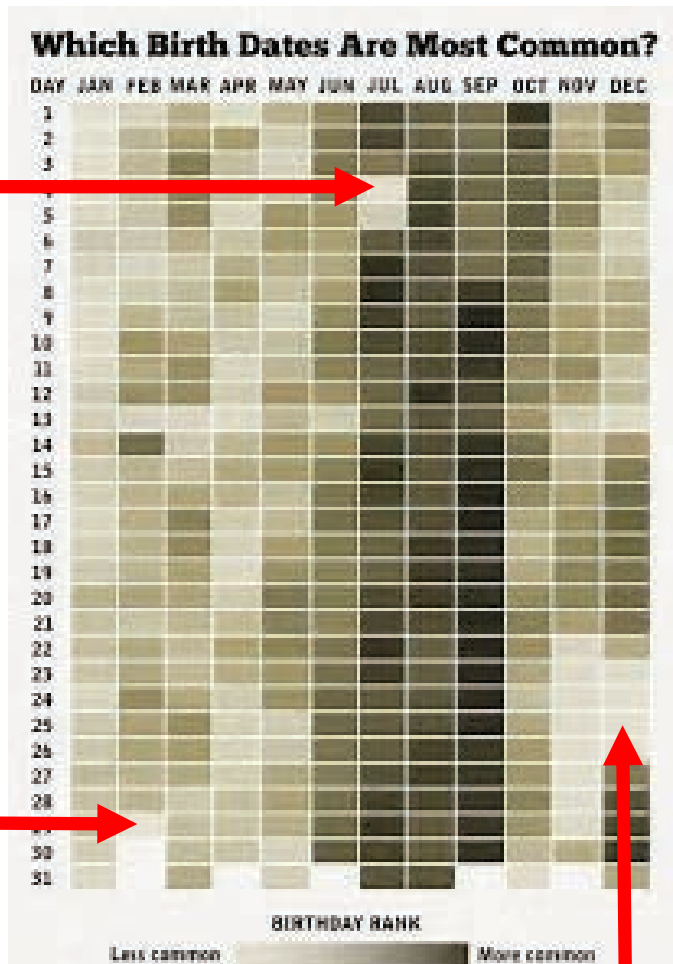
Suppose we want the probability of 3 people sharing a birthday?

# Why 3 Is Much Harder Analytically

---

- For 2 people, the complementary problem is “all birthdays are distinct”
- For 3 people, the complementary problem is a complicated disjunction
  - All birthdays distinct or
  - One pair and the rest distinct or
  - Two pairs and the rest distinct or
  - ...
- But using the simulation is dead trivial!!

# But All Dates Are NOT Equally Likely



Source: National Vital Statistics System natality data, as provided by Google BigQuery. Graph by Chris Mulligan (chmullig.com)

# How Does this Affect Probabilities?

---

- Do you expect a big change?
- Again, adjusting analytic model is a pain
- Adjusting simulation model easy

# Excerpt of Data

---

month,day,births

1,1,160369

1,2,169896

1,3,180036

1,4,182854

1,5,184145

1,6,186726

1,7,188277

1,8,185186

1,9,181511

1,10,183668

# Approximating Using a Simulation

```
def getBdays(toPlot = False):
    inFile = open('Births.csv')
    inFile.readline() #discard first line
    numBirths = [int(line.split(',')[2][:-1])\
                  for line in inFile]
    possibleDates = []
    for i in range(len(numBirths)):
        possibleDates += [i]*(numBirths[i])
    if toPlot:
        d = {i+1:numBirths[i] for i in range(len(numBirths))}
        vals = [d[k] for k in d.keys()]
        plt.plot(vals, 'bo')
        plt.xlim(-10, plt.xlim()[1])
        plt.xlabel('Day of Year')
        plt.ylabel('Number of Births')
        mean = 'Mean = ' + str(int(sum(vals)/len(vals)))
        std = 'Std = ' + str(int(numpy.std(vals)))
        plt.title('Frequency of Birthdates\n' +
                  mean + ', ' + std)
    return possibleDates
```

Create a very long list

# Approximating Using a Simulation

---

```
possibleDates = getBdays(False)
numShared = 3
for numPeople in [8, 16, 32, 64, 128, 256]:
    print('For', numPeople, 'probability of',
          numShared, 'shared birthdays:')
    print('  Estimated for actual distribution:',
          birthdayProb(numPeople, numShared, 10000,
                        possibleDates))
    print('  Estimated for uniform distribution:',
          birthdayProb(numPeople, numShared, 10000,
                        range(366)))
```

# Results

---

For 16 probability of 3 shared birthdays:  
Estimated for actual distribution: 0.0038  
Estimated for uniform distribution: 0.0026

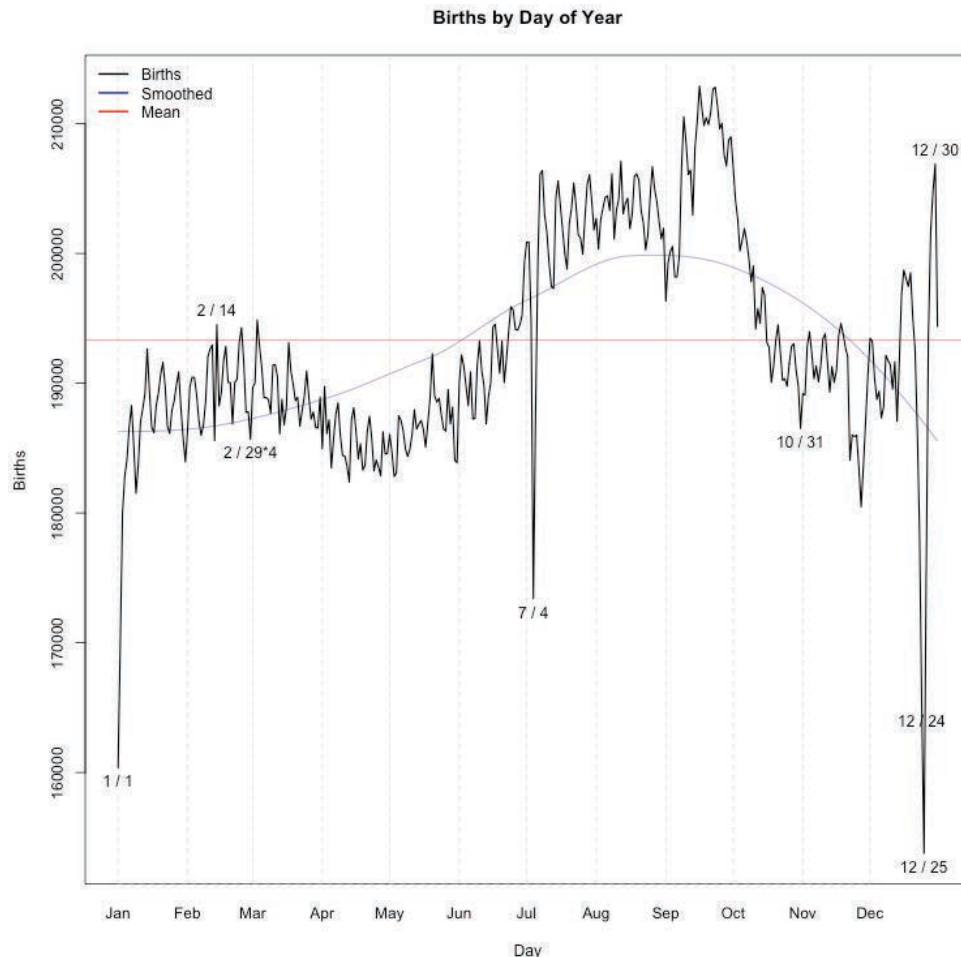
For 32 probability of 3 shared birthdays:  
Estimated for actual distribution: 0.035  
Estimated for uniform distribution: 0.0358

For 64 probability of 3 shared birthdays:  
Estimated for actual distribution: 0.2432  
Estimated for uniform distribution: 0.2533

For 128 probability of 3 shared birthdays:  
Estimated for actual distribution: 0.8839  
Estimated for uniform distribution: 0.8822

For 256 probability of 3 shared birthdays:  
Estimated for actual distribution: 1.0  
Estimated for uniform distribution: 1.0

# Looking at Results

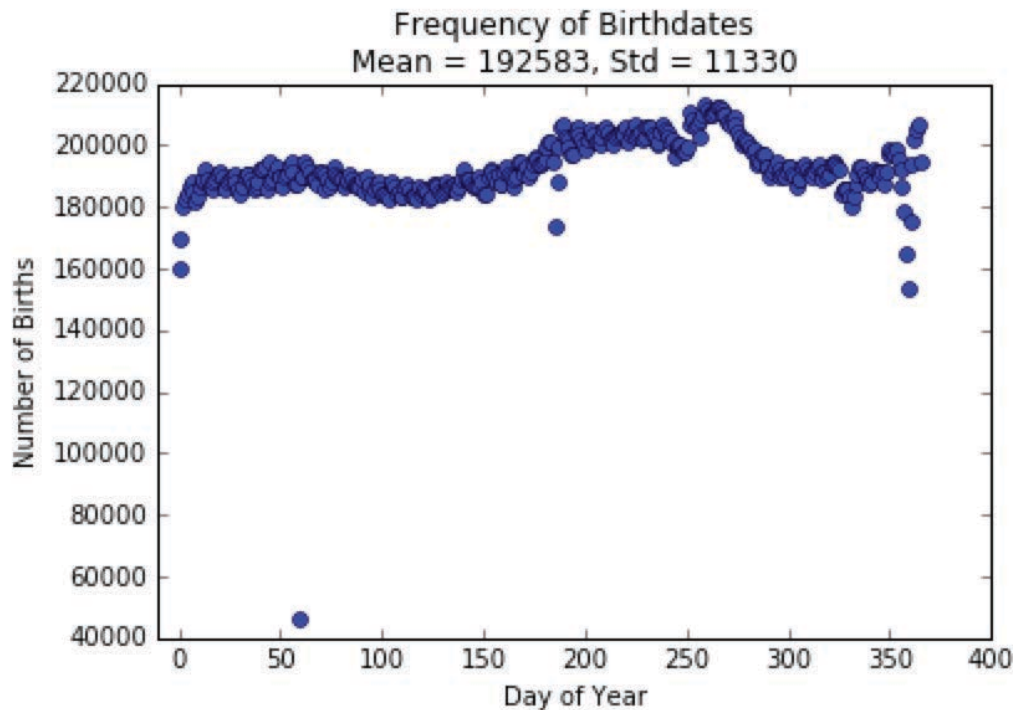


Source: National Vital Statistics System natality data, as provided by Google BigQuery. Graph by Chris Mulligan (chmullig.com)

Given apparent variance  
Shouldn't difference  
between actual and  
uniform distribution  
be larger?  
Or at least in  
consistent direction?

Numbers on y-axis  
hard to read  
Note y axis scaling

# A Better Plot



- Effect small
  - Frequency is pretty similar with a few exceptions
- Not a large enough sample to see effect consistently
- Much more on this later in the term

# 2M Trials

---

For 16 probability of 3 shared birthdays:  
Estimated for actual distribution: 0.00416  
Estimated for uniform distribution: 0.0040215

For 32 probability of 3 shared birthdays:  
Estimated for actual distribution: 0.0348315  
Estimated for uniform distribution: 0.0341425

For 64 probability of 3 shared birthdays:  
Estimated for actual distribution: 0.246033  
Estimated for uniform distribution: 0.243804

For 128 probability of 3 shared birthdays:  
Estimated for actual distribution: 0.8808115  
Estimated for uniform distribution: 0.879002

For 256 probability of 3 shared birthdays:  
Estimated for actual distribution: 1.0  
Estimated for uniform distribution: 0.9999995

# Summary

---

- As far as we can tell, the world is stochastic
- Therefore models need to estimate probabilities
- Analytic models feasible for reasonably simple situations
- For more complex situations, simulations often better
- Simulate stochastic process by:
  - Defining an event (e.g., rolling a die  $N$  times, flipping a coin  $N$  times, randomly picking birthdays for  $N$  people),
  - Running some number of trials,
  - Estimating probability of observing particular event (e.g., rolling 11111, finding three people with shared birthday)
- Much much **much** more on simulation models to come

