Data science for everyone

Prof. Jones-Rooy & Prof. Policastro

March 23, 2020

8.1: Statistics

ANNOUNCEMENTS

1. Prof. Policastro takes over the course on Monday, March 30
2. All grades & feedback are current up to the midterm and available on Gradescope
   - Please check and reach out to your TA with questions or issues (e.g., excused exceptions for late assignments)
   - We are emailing you feedback now, your TAs will explain. If you got a 0 for autograder you may have forgotten to submit on JupyterHub
   - Submission process has not changed
   - Releasing midterm grades after class
   - Regrade requests for everything through the midterm are due by **Friday** end of day EDT.
3. CAS has updated its Pass/Fail policy – see the email for official guidelines
   - This semester **only** if you earn a **Pass** in a DS course, it can count for the major/minor and future DS prerequisites
4. Per my email over break:
   - No formal Lab 5 = gift to you of automatic 2% full credit to your grade
   - HW 3 and HW 4 are combined = for added flexibility in our new situations
     - HW 3 was originally March 23-April 6 (2 weeks)
     - HW 4 was originally April 13-April 27 (2 weeks)
     - Now: HW 3/4 is April 6-April 27 (3 weeks)

My office hours are now by appointment

Your TA will let you know about office hours adjustments

Deadlines are still in EDT. Most are 8p EDT, so whatever that is in your local time!

PROJECT

- Deadline is still the same!
  - Due May 4th 8p
  - 6 weeks from today
- Recommended mini-deadlines
  - April 6: Have a dataset decided on
  - April 20: Be at least halfway through the questions
  - April 27: Be finished with all the coding and analysis (only writing, checking left)
  - May 4: Totally done and proofread by this day, upload well before 8p deadline
- We will point you to example datasets you can use, but you're also encouraged to explore and use your own (get permission from your TA in this case)
- No datasets we've used from class
  - But FYI Happiness 2020 is out, which you can use, but you can't research the same questions we've asked you to do in class

Reminder: Project assignment is on JupyterHub and Classes!

POSSIBLE DATA SOURCES

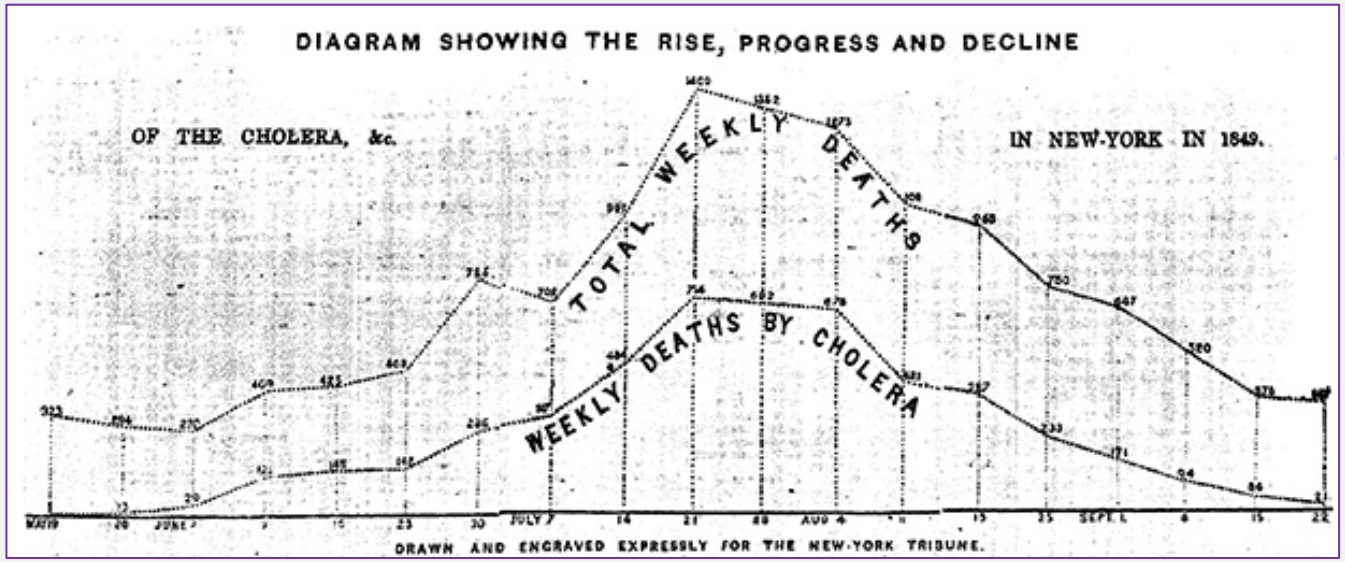
- FiveThirtyEight data from their articles: https://github.com/fivethirtyeight/data
  - Common topics: Sports, politics, education, movies & TV
- Buzzfeed news from their articles: https://github.com/BuzzFeedNews
  - By article: https://github.com/BuzzFeedNews/everything
  - Common topics: Politics, Twitter, tech, environment, violence, public health
- Open Case Studies: https://opencasestudies.github.io
  - Two projects: Health expenditure in the US, Relationship between fatal police shootings and firearm legislation in the US
- 19 Free Public Data Sets for Your Data Science Project: https://www.springboard.com/blog/free-public-data-sets-data-science-project/
  - Common topics: US Government (CDC, FBI, Census, BLS), international organizations (IMF, UNICEF), industry (Yelp, Airbnb, Walmart)
- Covid tracking project, Covid-19 Open Research Dataset
- Titanic dataset

JOHN SNOW & BROAD STREET PUMP

- London in the 1850s (but common in many cities)
- Waves of cholera, killing tens of thousands of people each
- John Snow: Doctor treating patients. Massive outbreak in summer 1854.

A chilling flashback to Lecture 1.2

WHAT SHOULD WE DO?!



NOTICE.

PREVENTIVES OF

CHOLERA!

Published by order of the Sanatory Committee, under the sanction of the Medical Counsel.

BE TEMPERATE IN EATING & DRINKING!

Avoid Raw Vegetables and Unripe Fruit !.

Abstain from COLD WATER, when heated, and above all from *Ardent Spirits*, and if habit have rendered them indispensable, take much less than usual.

SLEEP AND CLOTHE WARM !

DO NOT SLEEP OR SIT IN A DRAUGHT OF AIR.

Avoid getting Wet !

Attend immediately to all disorders of the Bowels.

TAKE NO MEDICINE WITHOUT ADVICE.

Medicine and Medical Advice can be had by the poor, at all hours of the day and night, by applying at the Station House in each Ward.

CALEB S. WOODHULL, *Mayor.*

JAMES KELLY, *Chairman of Sanatory Committee.*

OBSERVATIONS: WHAT DOES SNOW KNOW?

- Immediately ***deadly*** – you die within days of contracting
- Patterns of death:
  - Often people *within* one house would *all* die
  - But their *neighbors* weren't infected
- Symptoms:
  - Digestive problems

Scientists are going through this exercise right now with COVID-19

SNOW THINKS LIKE A SCIENTIST

A chilling flashback to Lecture 2.1

SCIENTIFIC METHOD

DATA

1. Observation
2. Question
3. Theory
4. Hypothesis

DATA

5. Test

6. Update theory

DATA

7. Repeat as desired

Snow's testable hypothesis

People who drink from the Broad Street Pump are more likely to get cholera than those who don't

Update theory

We cannot rule out the Broad Street Pump, but we can still go deeper to better understand **why**

Notice this language: We haven't "proven" anything; we have only failed to rule it out, or disprove

Repeat & extend

We still haven't established causality. Need more rigor!

COURSE JOURNEY

Thinking like a scientist

Programming fundamentals

Programming & data

Working with data

Tables & organizing data

Visualizations

Midterm

Statistics & prediction

Ethics

Next steps in DS

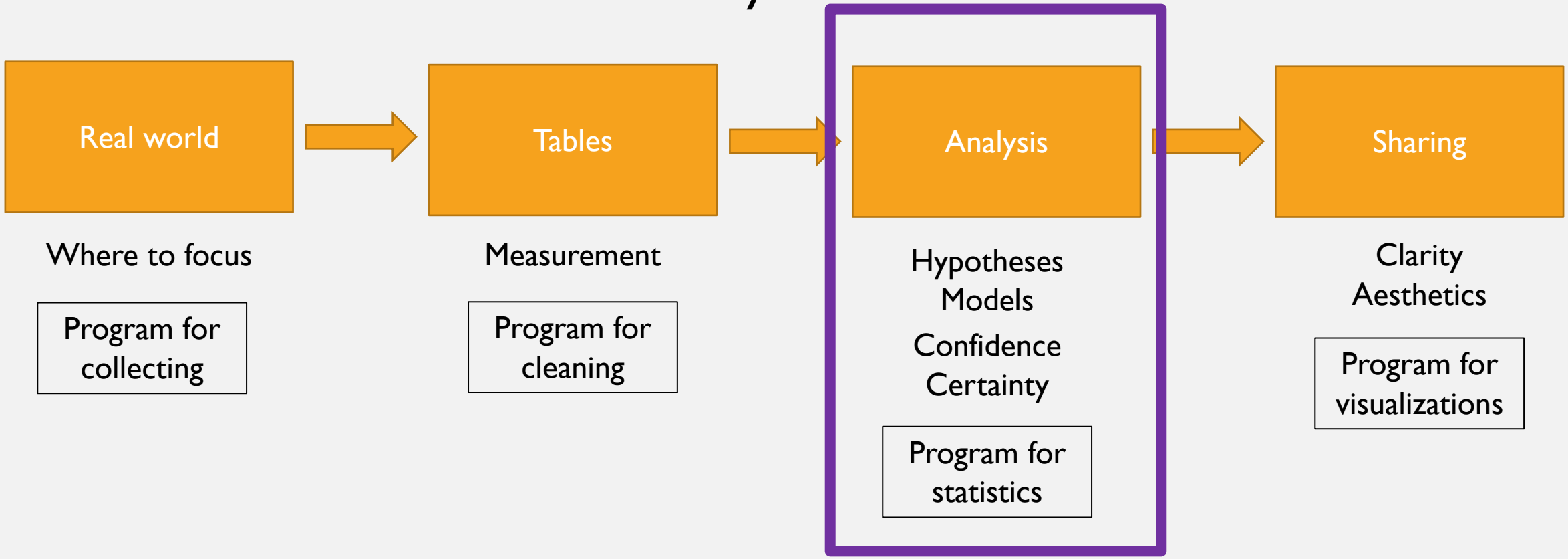
Final

Data science is a science

Data science is an art

TURNING THE WORLD INTO INSIGHTS

Fundamentally human exercise

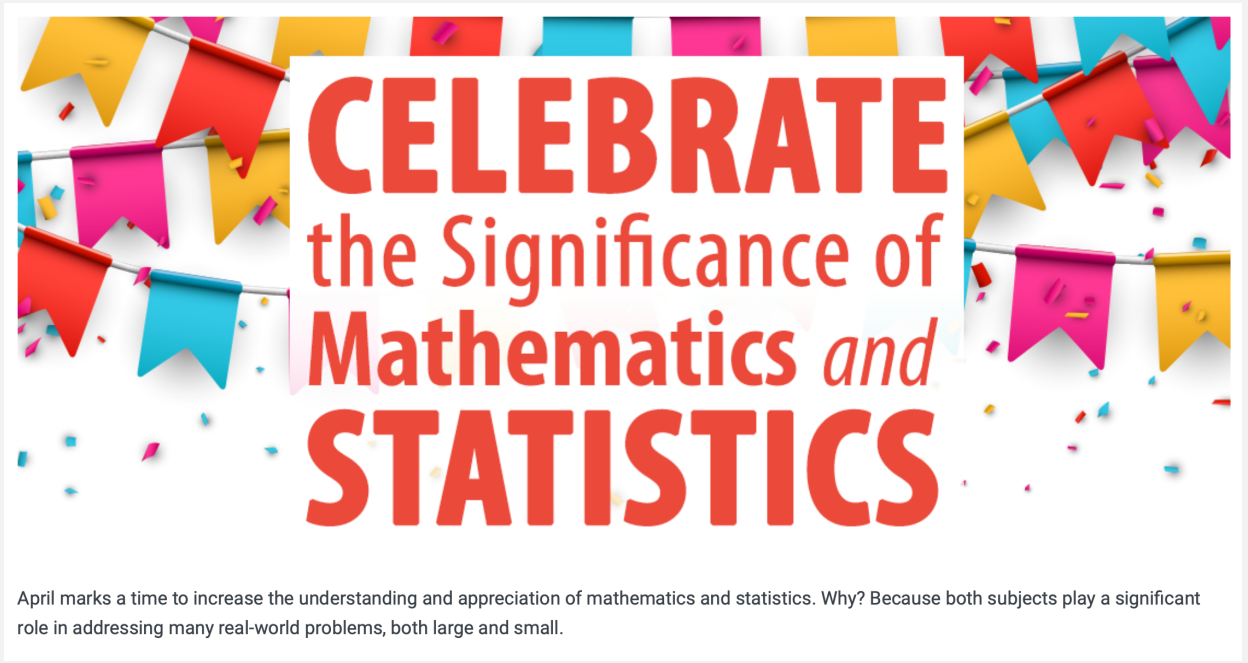


Outline

1. Statistics overview
2. Sample statistic vs. population parameter
3. Law of Large Numbers

GOOD NEWS!

- April is **Mathematics and Statistics Awareness Month!**
- Read more: American Mathematical Society & American Statistical Association



WHAT IS STATISTICS?

Jo Hardin, professor and chair in the department of mathematics and statistics at Pomona College, says, "Statistics is using data and knowledge about randomness to condense, communicate, and contextualize information and provide insight into the setting from which the data came."

Jeri Mulrow, ASA vice president, explains, "Statistics is the best area to be in because statistics are everywhere! They are all around us in our daily lives. It is important to be able to think critically about all of the data and information that surround us. Statistics and statistical thinking help us to make sense out of all of it."

Hadley Wickham, chief data scientist at R Studio says, "Statistics is an important tool in the data analysis/science toolbox. Statistics provides a coherent framework for thinking about random variation, and tools to partition data into signal and noise."

Source for quotes: American Statistical Association

OUR WORKING DEFINITION OF STATISTICS

Using data to generate insights we didn't already have, and to help us think about how confident or certain we should be about these discoveries

Three main components:

Making sense of lots of data

Imperfect snapshots

Subjective and normative decisions around whether and how to measure

Measurement

Biases

Errors

To learn something new

Hypothesis testing (H_0 v. H_a)

Coefficients

Confounding or missing variables

Correlation

Causation

Randomness

Descriptive statistics

And think about how confident we can be in that lesson

Confidence intervals

p-values and statistical significance

Probability

Uncertainty

Population vs. sample

Prediction

Outline

1. Statistics overview
2. Sample statistic vs. population parameter
3. Law of Large Numbers

You already know this!

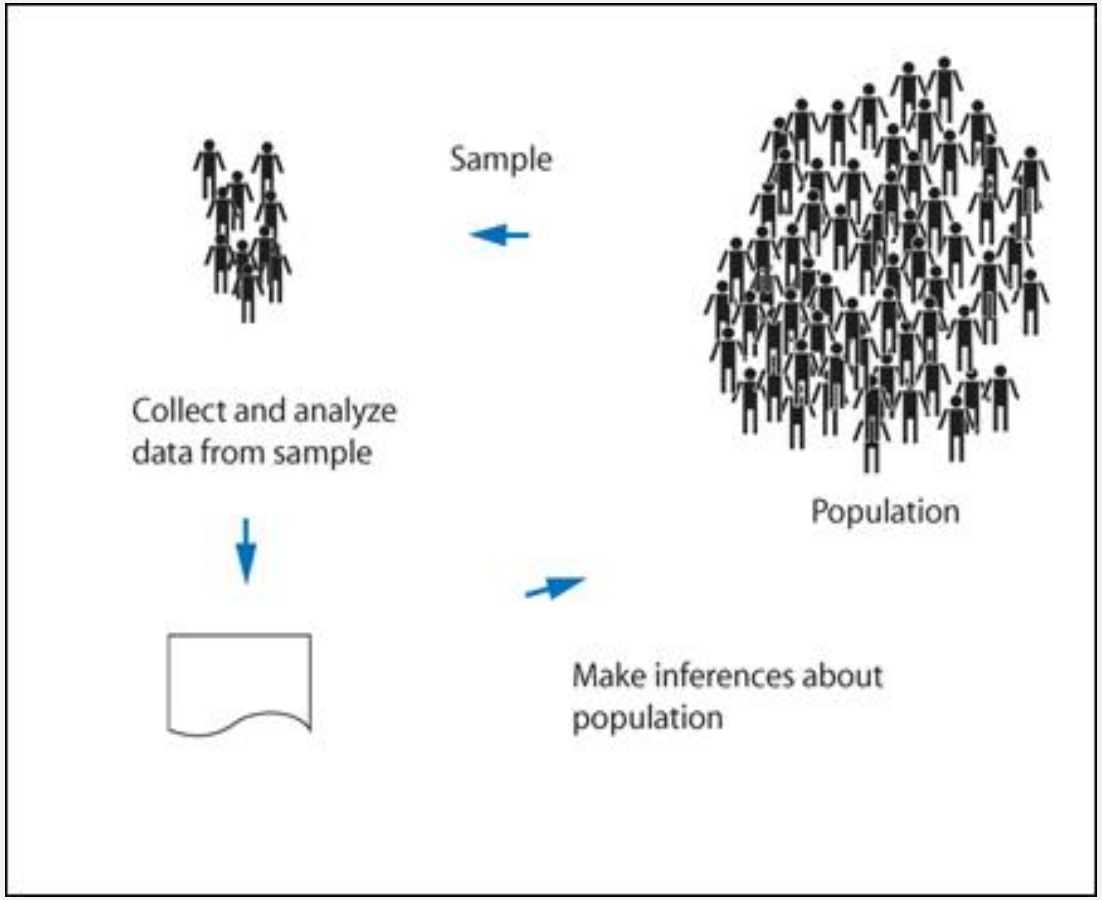
POPULATION

- The ***population*** is the universe of cases we want to describe. We could ask:
  - How will the United States (every voter) vote in the 2020 Presidential election?
  - What is the average income in China (every household or every person)?
- We call the characteristic we care about (vote intention, income) the ***population parameter***.
- Unfortunately, we can only rarely study the population directly

You already know this!

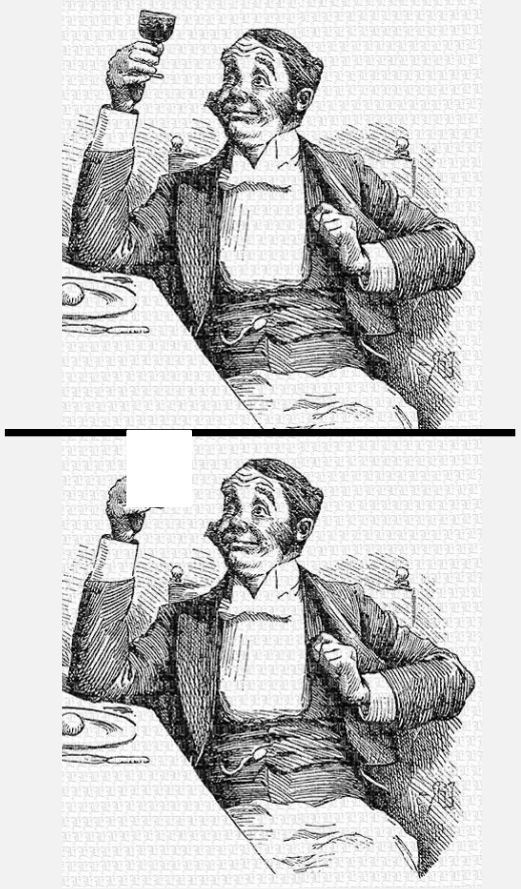
SAMPLES

- So, instead we will use ***samples*** from the population, and use those to ***estimate*** the parameters of the population.
- We'll be able to talk about our ***uncertainty*** around those estimates.
- Focus on the properties of ***large***, ***random*** samples.



You already know this!

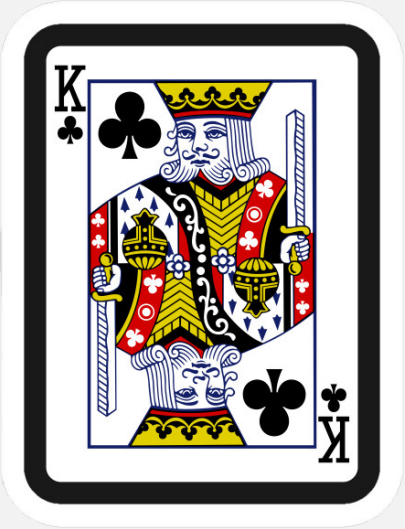
RANDOMNESS



- By ***random*** we mean units are chosen in a ***non-deterministic*** way, by ***chance***.
  - Or at least in a manner ***orthogonal*** to what we are trying to understand.
- We met this idea when talking about how we would like to assign people to ***treatment*** and ***control*** in a randomized experiment: we don't want them to self select!
- ***Computers*** are very good at helping us generate random numbers and random samples.

PROBABILITY SAMPLE

- ***Probability sample***: before we draw the sample, we can calculate the chance that any given unit will appear in it.
- In a ***simple random sample*** this chance is $1/n$, where n is the size of the sample
  - Example: if there are 365 balls (one for each day of the year), and I pick one, what is the probability it is Sept 3?
  - Or, if I pick a random card from a pack, what is the chance it is the King of Clubs?



SAMPLING WITH AND WITHOUT REPLACEMENT

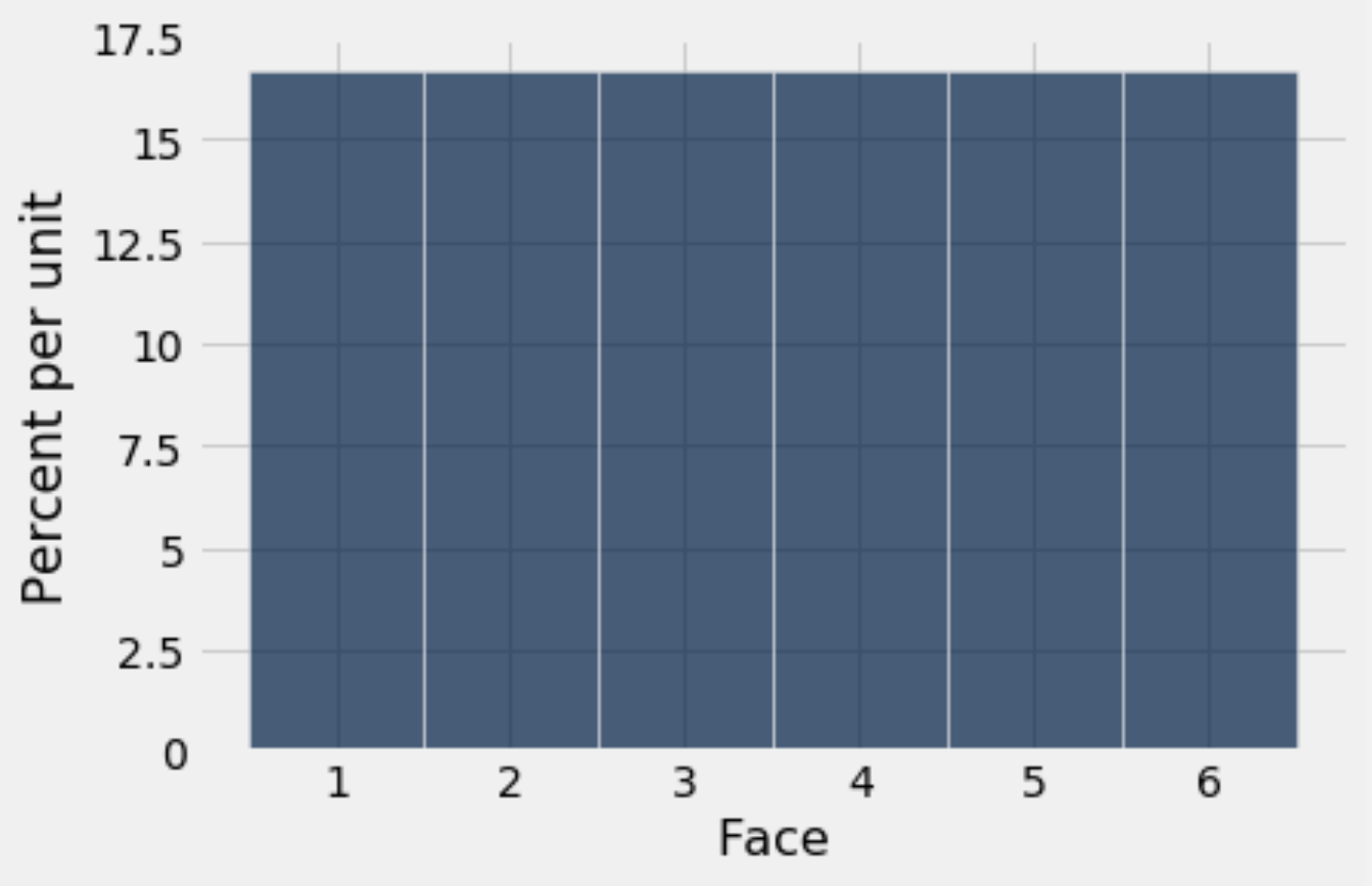
- If I pick a given date-ball from the draft lottery, it has probability 1/365. I could then…
  - replace it, which means I put it back. Now the chance of picking any given date-ball is 1/365 again. This is ***sampling with replacement***.
  - or not replace it. Now the chance of picking any remaining date-ball is 1/364 (why?). This is ***sampling without replacement***.

PROBABILITY DISTRIBUTION

- Suppose I have a fair 6-sided die.
- What ***percentage of the time*** do I expect it to come up 1?
  - And 2? And 3? And 4? And 5? And 6?
- We can ***plot*** all the values it can take (1,2,3,4,5,6) against the percentage of times it will come up that way.
- This is called a ***probability distribution***.
  - Remember, in this case it is ***discrete***, meaning the die can only take certain finite values (not continuous, e.g., no 3.3, 4.5, etc.)

PROBABILITY DISTRIBUTION

- 1/6 is about 17%
- Every face is as likely as any other, so we say it is a (discrete) ***uniform*** distribution



EMPIRICAL DISTRIBUTIONS

- We've seen how the die *should* behave, in theory.
- ***Empirical distributions*** are how the ***actual observed data*** stacks up in terms of all the values it could take
  - Imagine I rolled my die 10 times.
  - Would I expect exactly 16.6% of the rolls to be 1, 16.6% to be 2, 16.6% to be 3 etc.?
  - No (not actually possible in this case)
- Let's roll my die 10 times and draw a histogram of the distribution of the outcomes

SIMULATION

- Ah ha! Computers FTW!
- Let's have Python ***simulate*** a (fair) die
  - by drawing from numbers 1 through 6, ***with replacement*** each time
  - and doing that 10 times

SIMULATING ROLLING A DIE IN PYTHON

```
1  import pandas as pd
2  import numpy as np
3  import matplotlib.pyplot as plt
4  #data = pd.read_csv('polity4.csv')
```

```
1  dice = np.random.randint(1, 7, 10)
2  dice
```

```
array([2, 6, 3, 3, 6, 6, 1, 5, 1, 1])
```

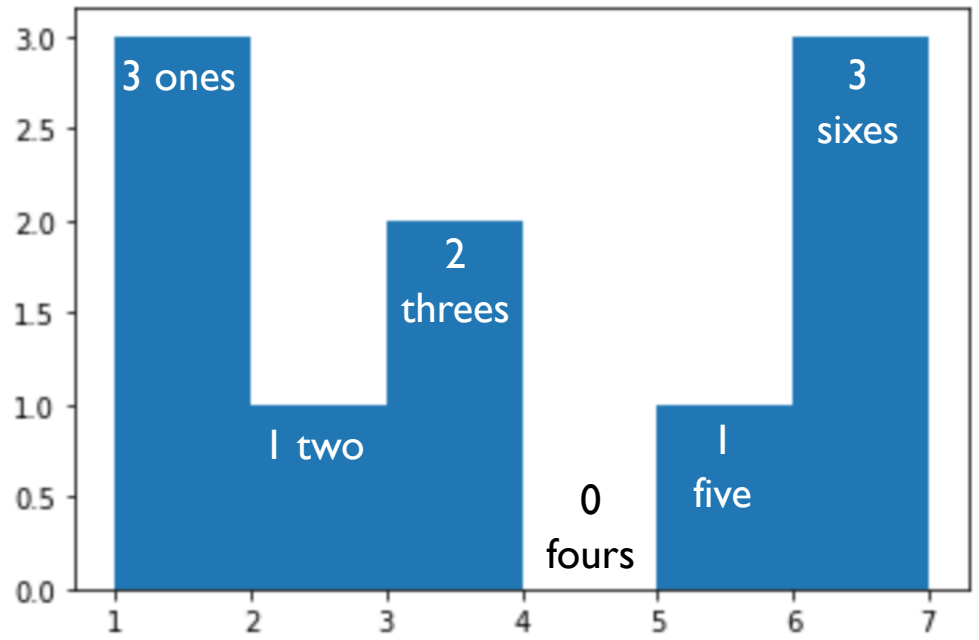
Same syntax we know from **range**

(lowest value, highest value (not inclusive), interval)

See also Lecture 8.1 Example Code on Classes and JupyterHub!

PLOTTING A DIE SIMULATION IN PYTHON

```
1  plt.hist(dice, bins=[1, 2, 3, 4, 5, 6, 7])
2  plt.show()
```



We know how to do this!

Recall our numbers for **dice**

```
array([2, 6, 3, 3, 6, 6, 1, 5, 1, 1])
```

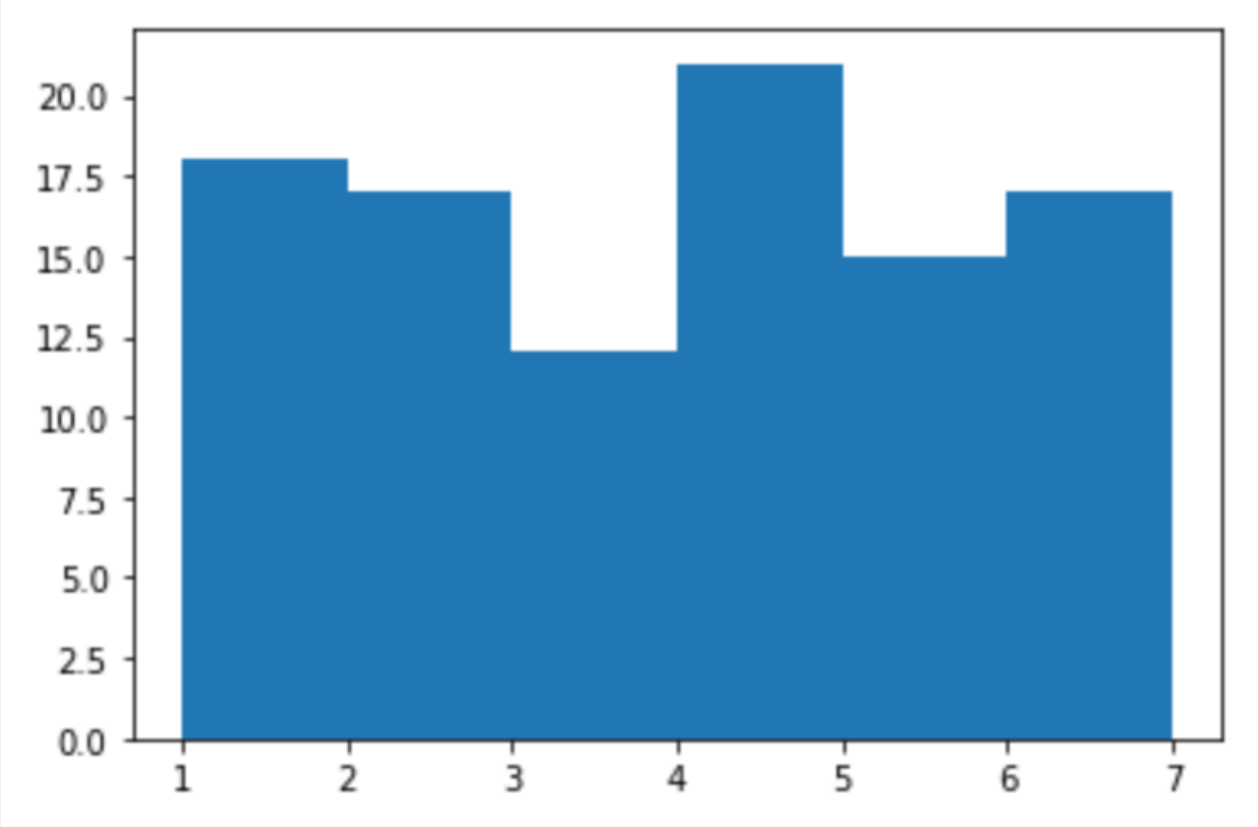
SIMULATION

- Let's do it ***100 times***

```
1  dice = np.random.randint(1, 7, 100)
2  dice
```

```
array([1, 5, 1, 6, 4, 4, 4, 1, 6, 1, 1, 4, 3, 4, 4, 1, 2, 4, 1, 5, 3, 1,
       2, 4, 4, 5, 3, 4, 3, 1, 2, 1, 5, 3, 6, 6, 3, 6, 5, 1, 1, 2, 2, 6,
       6, 5, 5, 5, 1, 5, 5, 6, 3, 1, 6, 6, 5, 4, 3, 6, 4, 5, 6, 6, 4, 1,
       6, 2, 2, 2, 4, 4, 4, 2, 1, 3, 5, 2, 4, 4, 5, 3, 3, 4, 3, 5, 4, 6,
       2, 2, 2, 6, 2, 2, 1, 4, 2, 6, 1, 2])
```

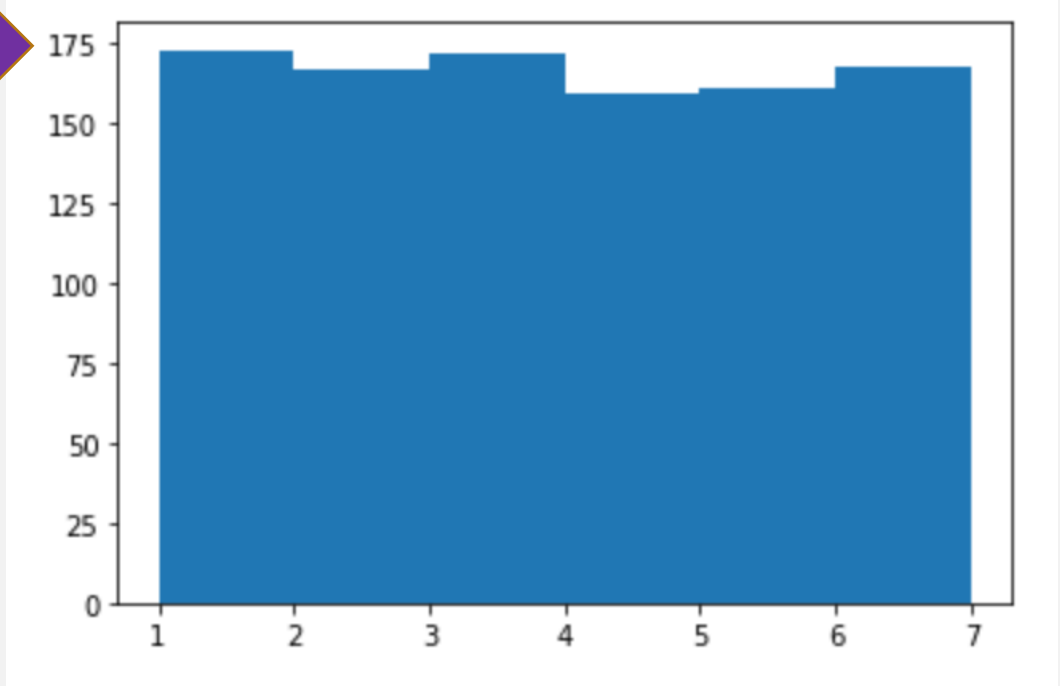
```
1  plt.hist(dice, bins=[1, 2, 3, 4, 5, 6, 7])
2  plt.show()
```



SIMULATION

- Let's do it ***1000 times***

Hang on this y-axis isn't so clear

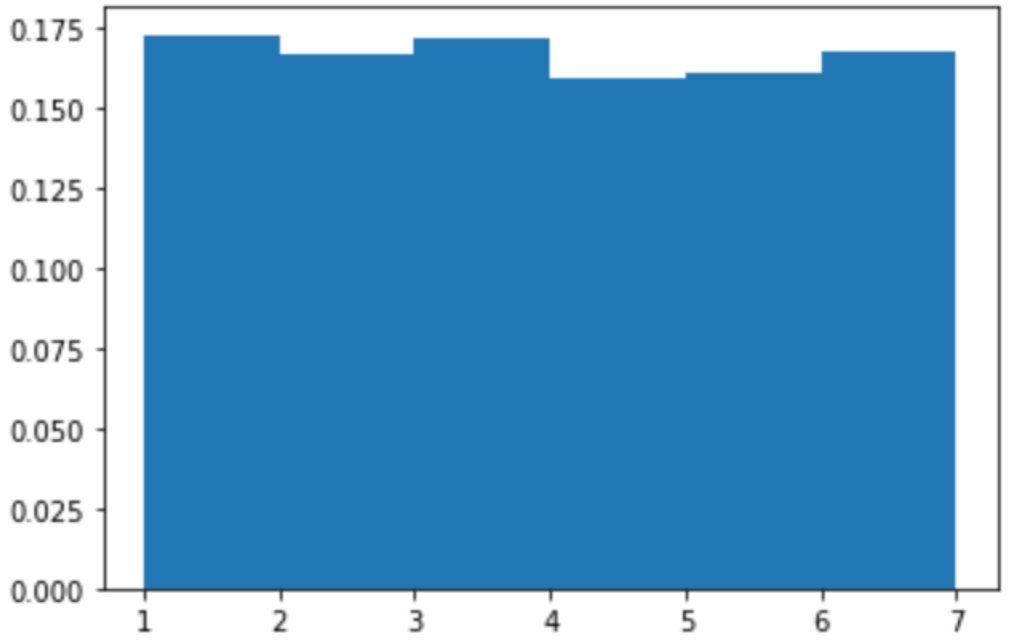



```
1  dice = np.random.randint(1, 7, 1000)
2  dice
```

- As we increase size of the sample (n), what do you notice is happening to the histogram?

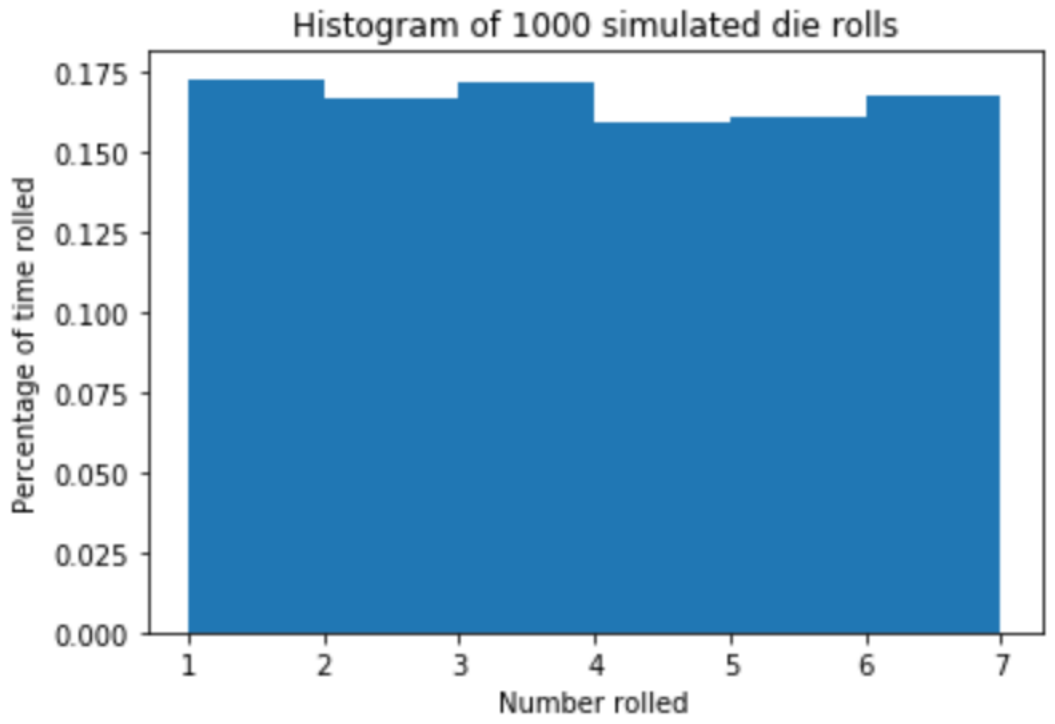
SIMULATION WITH PERCENTAGES RATHER THAN FREQUENCIES

```
1  plt.hist(dice, bins=[1, 2, 3, 4, 5, 6, 7], density=True)
2  plt.show()
```



OK, THAT, BUT MAKE IT LOOK NICER

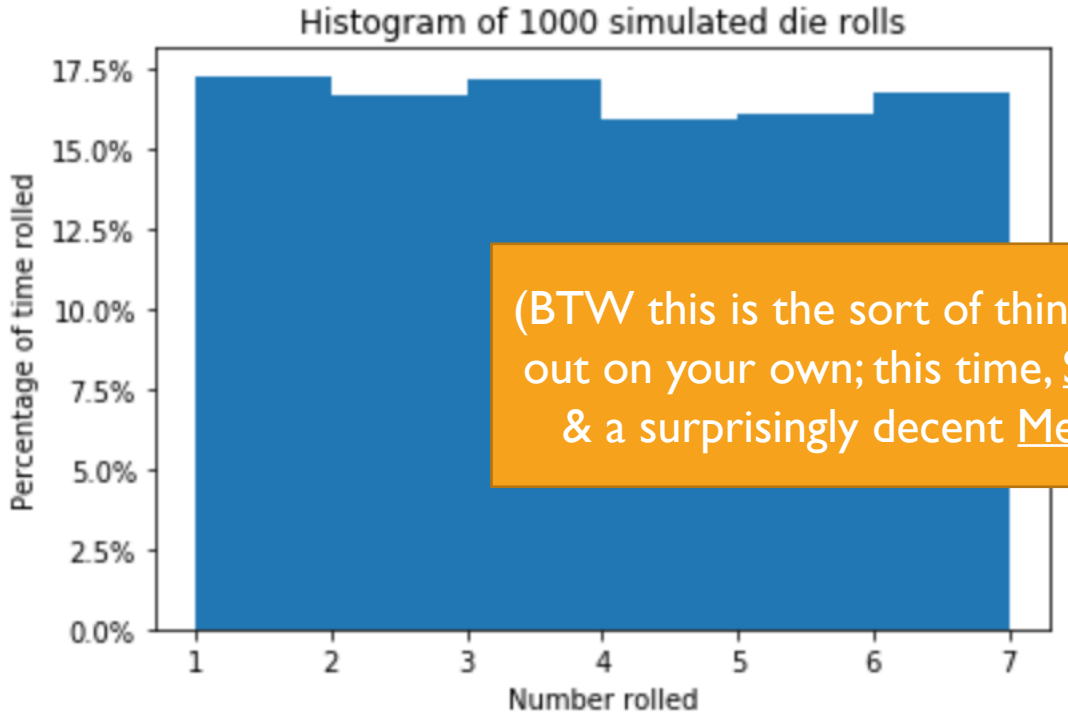
```
1  plt.hist(dice, bins=[1, 2, 3, 4, 5, 6, 7], density=True)
2  plt.xlabel('Number rolled')
3  plt.ylabel('Percentage of time rolled')
4  plt.title('Histogram of 1000 simulated die rolls')
5  #plt.legend
6  plt.show()
```



AND NOW MAKE THE Y-AXIS MORE READABLE

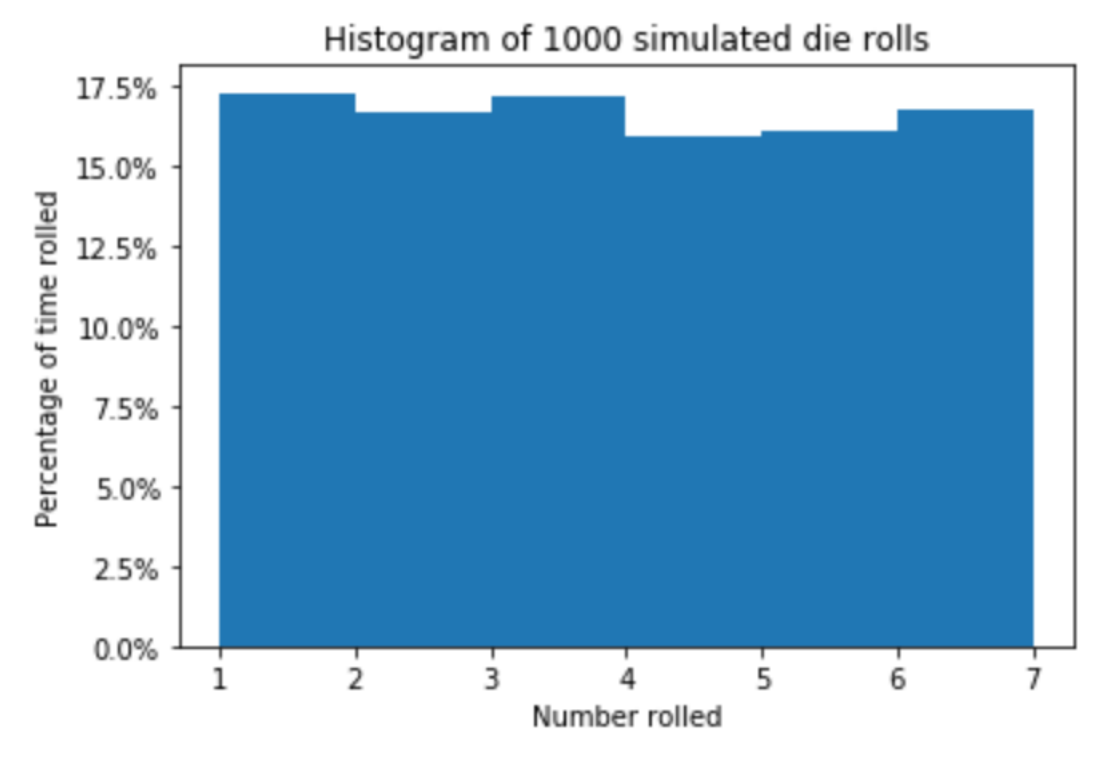

```
from matplotlib.ticker import PercentFormatter
```

```
plt.hist(dice, bins=[1, 2, 3, 4, 5, 6, 7], density=True)
plt.xlabel('Number rolled')
plt.ylabel('Percentage of time rolled')
plt.title('Histogram of 1000 simulated die rolls')
plt.gca().yaxis.set_major_formatter(PercentFormatter(1))
plt.show()
```



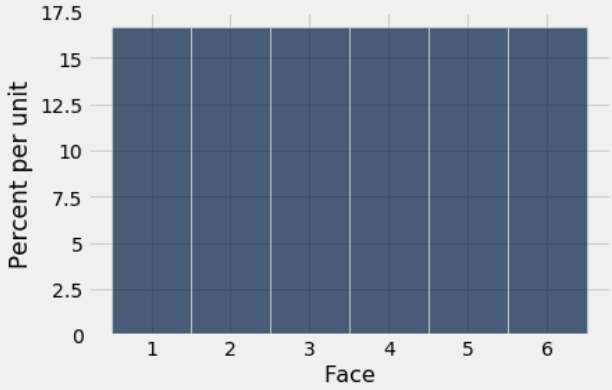
(BTW this is the sort of thing you can figure out on your own; this time, Stack Overflow & a surprisingly decent Medium article)

BACK TO THE MAIN QUESTION



As we increase size of the sample (*n*), what do you notice is happening to this histogram of the ***empirical distribution***?

It's looking more and more like the ***probability distribution!***



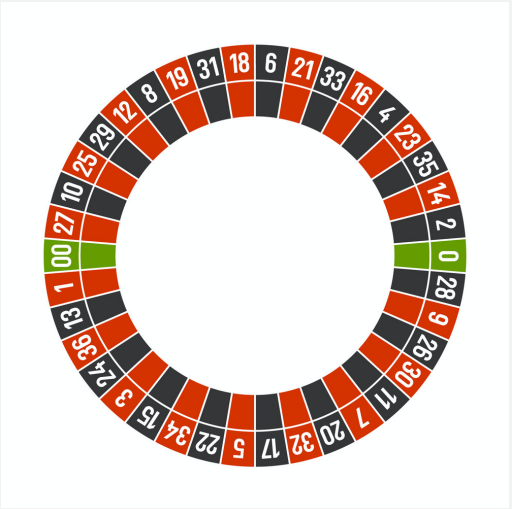
Outline

1. Statistics overview
2. Sample statistic vs. population parameter
3. Law of Large Numbers

LAW OF LARGE NUMBERS

- If we repeat an experiment many, many times, the proportion of times we see a given outcome (e.g., a 3 or a 6) ***empirically*** will converge to the ***theoretical*** value we would expect (1/6, or 16.6%)

- This is an implication of the ***Law of Large Numbers*** (sometimes called the "Law of Averages")

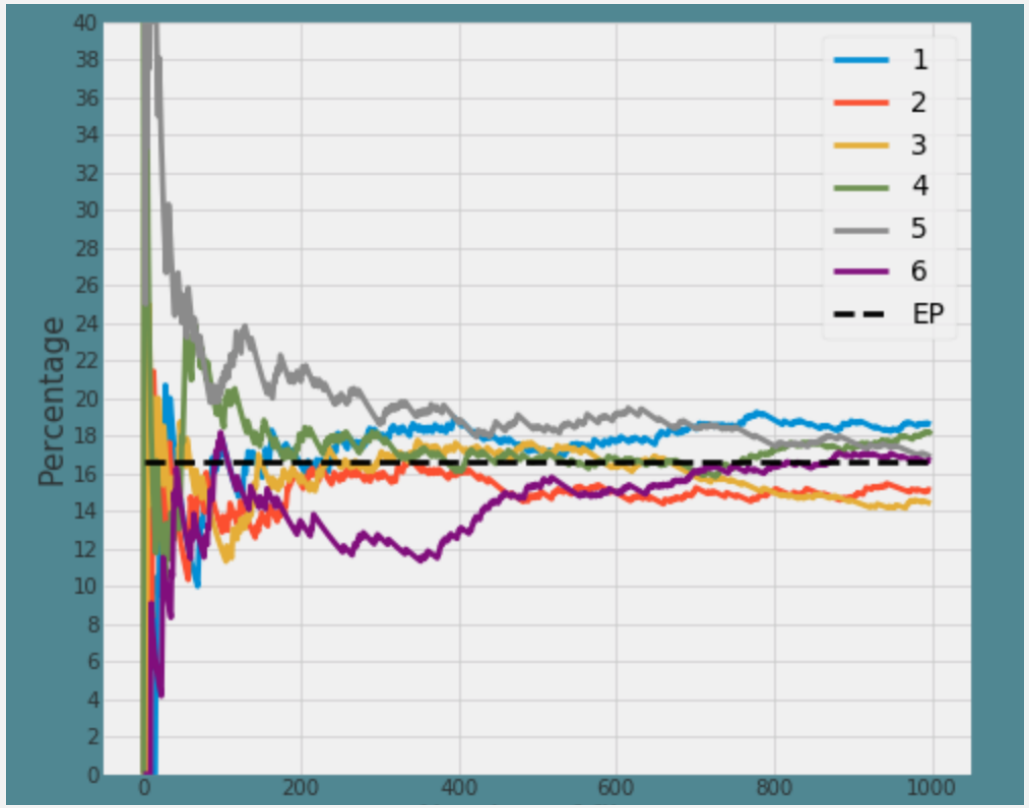


"Expected value" of $1 bet on roulette is -$0.0526.

LAW OF LARGE NUMBERS

"When the same random process is repeated a large number of times, the relative frequency of the possible outcomes will be approximately equal to their respective probabilities."

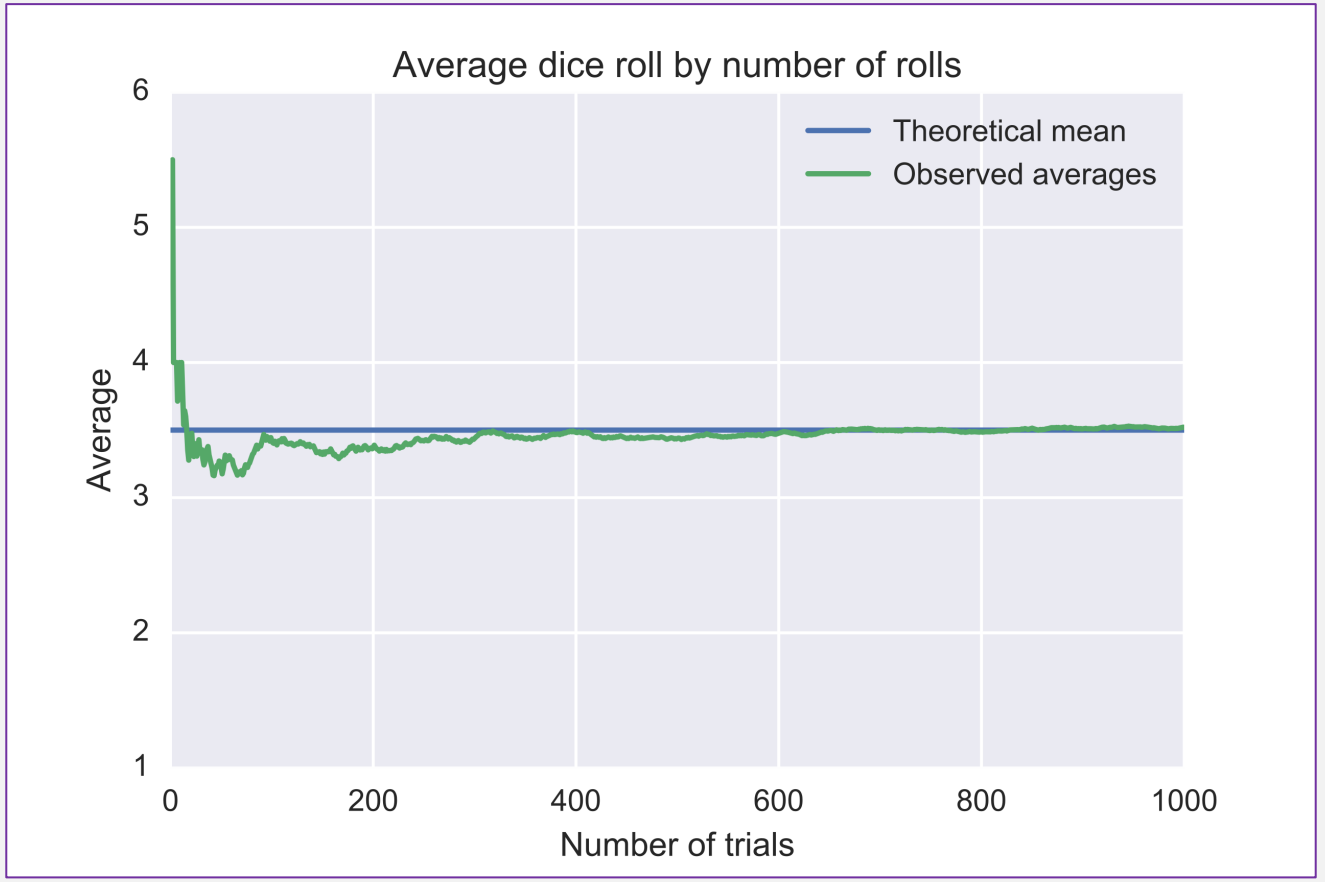
Optional further reading here!



Émile Borel

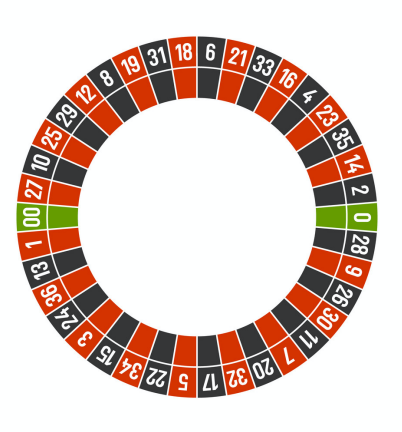
LAW OF LARGE NUMBERS

The average of the ***empirical*** results obtained from a large number of independent trials or experiments should be close to the ***theoretical***, or ***expected***, value, and will tend to become closer as more trials are performed



WHY THE LAW OF LARGE NUMBERS IS IMPORTANT

- Well, statistics doesn't really work without it
- As a probabilistic process is repeated a large number of times, the relative frequencies of its possible outcomes will get closer and closer to their respective probabilities
  - Example: Flipping a coin many times results in approximately 50% heads and 50% tails
  - The more times we flip a coin, the more the closer the results get to 50-50
- This basically means we have a bridge between probability theory and the real world
- Let's go back to the Roulette example – the expected value of a $1 bet is -0.05. If you bet, you might make money, or you might lose more than that, but if you play long enough, your results will approach -0.05 for every $1 bet.



"Expected value" of $1 bet on roulette is -$0.0526.

HOW LARGE IS LARGE ENOUGH FOR THE LAW OF LARGE NUMBERS?

- The LLN guarantees that the empirical relative frequency of an outcome will be ***approaching*** (getting closer to), the ***expected relative frequency*** as the number of trials goes up
- So the answer to this question depends on ***how close do you need to be*** to the expected relative frequency?
- The answer depends on your own tolerance for uncertainty/risk/error, which may depend on what you're trying to understand
  - We may not care if our dice rolling or coin flipping is 5% off
  - But we may care if, say, our vaccine is 5% off
- Another factor is ***variance***. It takes dice more trials to converge compared to coins because there are more possible values they can take on
- We will be discussing thresholds of uncertainty/errors and variance a lot going forward!

Outline

1. Statistics overview
2. Sample statistic vs. population parameter
3. Law of Large Numbers