CMPU 100 · Programming with Data

Sampling and Distributions

Class 20

Where are we?

Roman coin featuring Pompey the Great and a ship

If you flip a fair coin 100 times, *what is the chance* of getting between 40 and 60 heads?

The intuition of probability*: If we repeat an event (flipping a coin) many times (say, 1 million) … what fraction of the times would we see the outcome we care about (heads)?*

Probability

$P(A)$: Chance that an event A will occur

Never	*Half*	*Always*
$P(A) = 0$	$P(A) = 0.5$	$P(A) = 1$
(or 0%)	(or 50%)	(or 100%)

$P(heads) = 1/2$

How to calculate an event's probability?

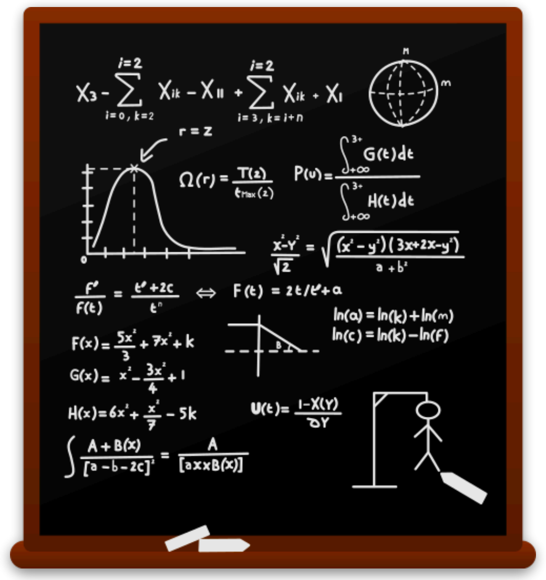
Computers (simulation)

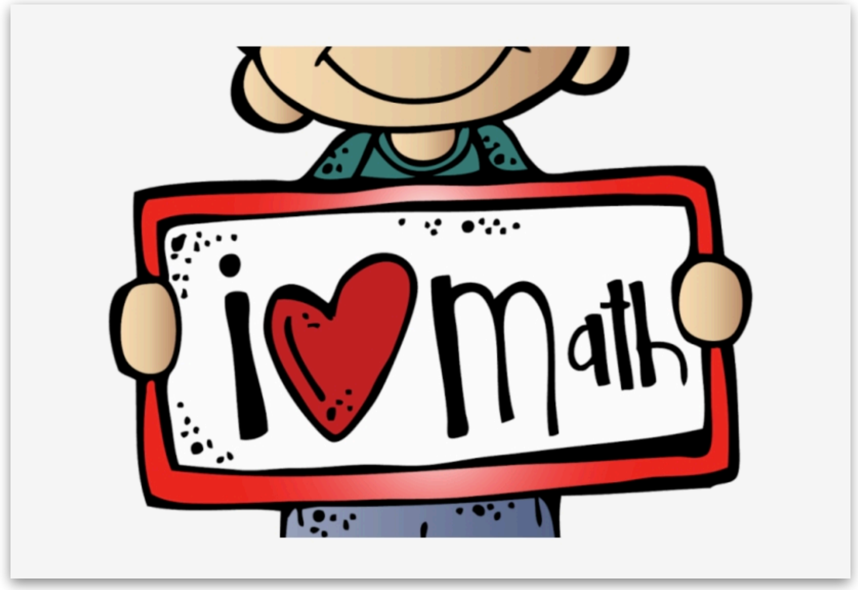
```
N = 1000000 #Roll the dice 1 million times
option_a = np.random.choice(dice, N) + np.random.c
option_b = 2 * np.random.choice(dice, N)

print("Option A Mean: ", np.mean(option_a))
print("Option B Mean: ", np.mean(option_b))

Option A Mean:  7.003198
Option B Mean:  6.99884

samples = Table().with_columns("Option A", option_
samples.hist("Option A", bins=np.arange(0,14))
samples.hist("Option B", bins=np.arange(0,14))
```

*Rooted in **algorithms***

- ✗ Approximate solutions
- ✓ Often convincing
- ✓ Non-trivial problems potentially captured cleanly with code

Math (analytical)





*Rooted in **rules (axioms)***

- ✓ Exact solutions
- ✓ Straightforward for simple problems
- ✗ Non-trivial problems potentially difficult to analyze/verify

How to calculate an event’s probability?

Computers (simulation)

```
N = 1000000 #Roll the dice 1 million times
option_a = np.random.choice(dice, N) + np.random.c
option_b = 2 * np.random.choice(dice, N)

print("Option A Mean: ", np.mean(option_a))
print("Option B Mean: ", np.mean(option_b))

Option A Mean:  7.003198
Option B Mean:  6.99884

samples = Table().with_columns("Option A", option_
samples.hist("Option A", bins=np.arange(0,14))
samples.hist("Option B", bins=np.arange(0,14))
```

Rooted in algorithms

- ✗ Approximate solutions
- ✓ Often convincing
- ✓ Non-trivial problems potentially captured cleanly with code

Repeat many times:

- Simulate one trial
- Record the outcome

Analyze outcomes for all trials

How to calculate an event’s probability?

Computers (simulation)

```
N = 1000000 #Roll the dice 1 million times
option_a = np.random.choice(dice, N) + np.random.c
option_b = 2 * np.random.choice(dice, N)

print("Option A Mean: ", np.mean(option_a))
print("Option B Mean: ", np.mean(option_b))

Option A Mean:  7.003198
Option B Mean:  6.99884

samples = Table().with_columns("Option A", option_
samples.hist("Option A", bins=np.arange(0,14))
samples.hist("Option B", bins=np.arange(0,14))
```

Rooted in algorithms

- ✗ Approximate solutions
- ✓ Often convincing
- ✓ Non-trivial problems potentially captured cleanly with code

Repeat many times:

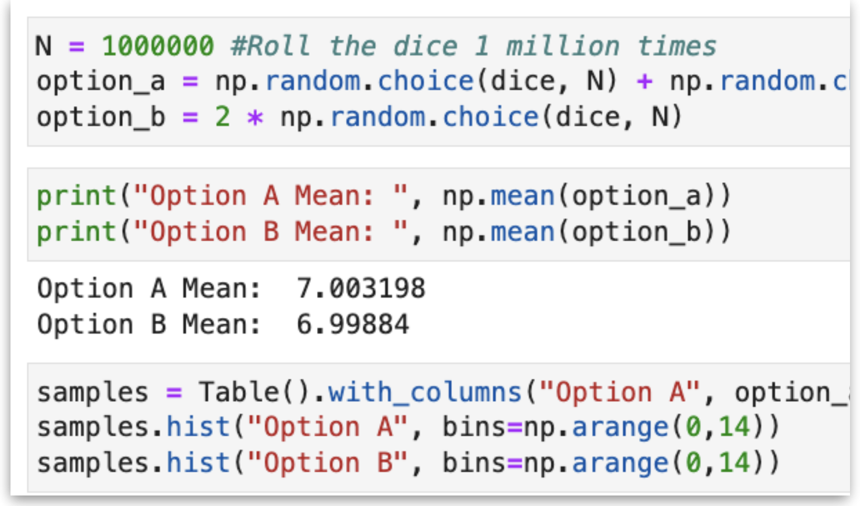
Flip 100 coins

Record the outcome

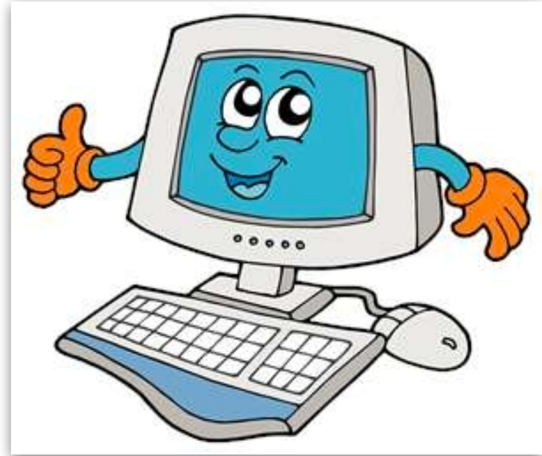
Analyze outcomes for all trials

How to calculate an event's probability?

Computers (simulation)



```
N = 1000000 #Roll the dice 1 million times
option_a = np.random.choice(dice, N) + np.random.c
option_b = 2 * np.random.choice(dice, N)

print("Option A Mean: ", np.mean(option_a))
print("Option B Mean: ", np.mean(option_b))

Option A Mean:  7.003198
Option B Mean:  6.99884

samples = Table().with_columns("Option A", option_
samples.hist("Option A", bins=np.arange(0,14))
samples.hist("Option B", bins=np.arange(0,14))
```



Rooted in algorithms

- ✗ Approximate solutions
- ✓ Often convincing
- ✓ Non-trivial problems potentially captured cleanly with code

Repeat many times:

Flip 100 coins

Record *the number of heads*

Analyze outcomes for all trials

How to calculate an event's probability?

Computers (simulation)

```
N = 1000000 #Roll the dice 1 million times
option_a = np.random.choice(dice, N) + np.random.c
option_b = 2 * np.random.choice(dice, N)

print("Option A Mean: ", np.mean(option_a))
print("Option B Mean: ", np.mean(option_b))

Option A Mean:  7.003198
Option B Mean:  6.99884

samples = Table().with_columns("Option A", option_
samples.hist("Option A", bins=np.arange(0,14))
samples.hist("Option B", bins=np.arange(0,14))
```

Rooted in algorithms

- ✗ Approximate solutions
- ✓ Often convincing
- ✓ Non-trivial problems potentially captured cleanly with code

Repeat many times:

Flip 100 coins

Record *the number of heads*

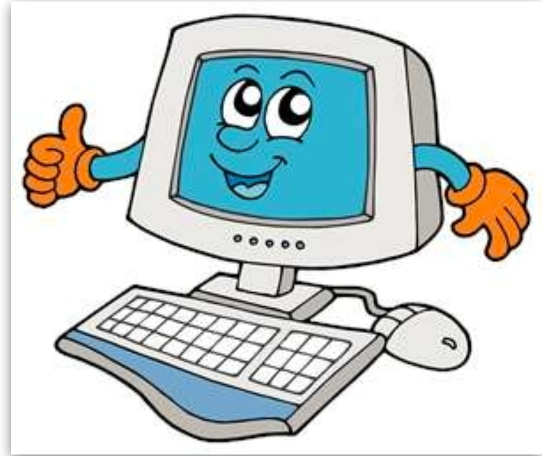
Analyze *what proportion of the times the number of heads was between 40 and 60*

Notebook: *Simulation*

How to calculate an event's probability?

Computers (simulation)

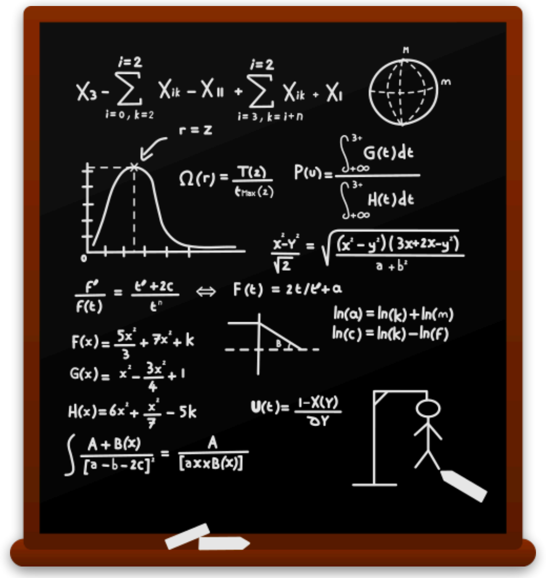
```
N = 1000000 #Roll the dice 1 million times
option_a = np.random.choice(dice, N) + np.random.c
option_b = 2 * np.random.choice(dice, N)

print("Option A Mean: ", np.mean(option_a))
print("Option B Mean: ", np.mean(option_b))

Option A Mean:  7.003198
Option B Mean:  6.99884

samples = Table().with_columns("Option A", option_
samples.hist("Option A", bins=np.arange(0,14))
samples.hist("Option B", bins=np.arange(0,14))
```

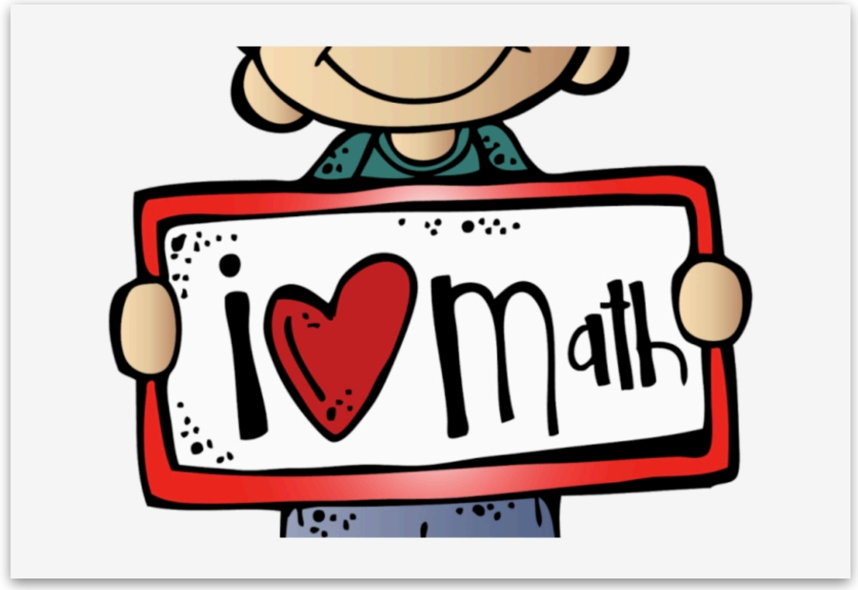


Rooted in ***algorithms***

- ✗ Approximate solutions
- ✓ Often convincing
- ✓ Non-trivial problems potentially captured cleanly with code

Math (analytical)





Rooted in ***rules (axioms)***

- ✓ Exact solutions
- ✓ Straightforward for simple problems
- ✗ Non-trivial problems potentially difficult to analyze/verify

Basic probability

If all outcomes are equally likely – for example, rolling a fair die or flipping a fair coin – then it’s easy to compute a probability by counting:

$$P(A) = \frac{\text{number of outcomes that make } A \text{ happen}}{\text{total number of outcomes}}$$

How likely are you to get an even number when rolling a die?

Even = {⚁, ⚃, ⚅}

All = {⚀, ⚁, ⚂, ⚃, ⚄, ⚅}

In math, {…} denotes a set, and |…| denotes the number of elements.

|*Even*| / |*All*| = 3 / 6 = 50%

When the are two or more ways an event can happen, the ***addition rule*** says its probability is the sum of the probabilities for the different ways:

$$P(A) = P(\text{first way } A \text{ can happen}) + P(\text{second way } A \text{ can happen}) + \ldots$$

For example, the probability of rolling a total of six when rolling two dice is:

P(roll six) = *P*(⚄, ⚀) +
P(⚀, ⚄) +
P(⚃, ⚁) +
P(⚁, ⚃) +
P(⚂, ⚂)

The chance that event A will ***not*** occur:

$$P(\text{not } A) = 1 - P(A)$$

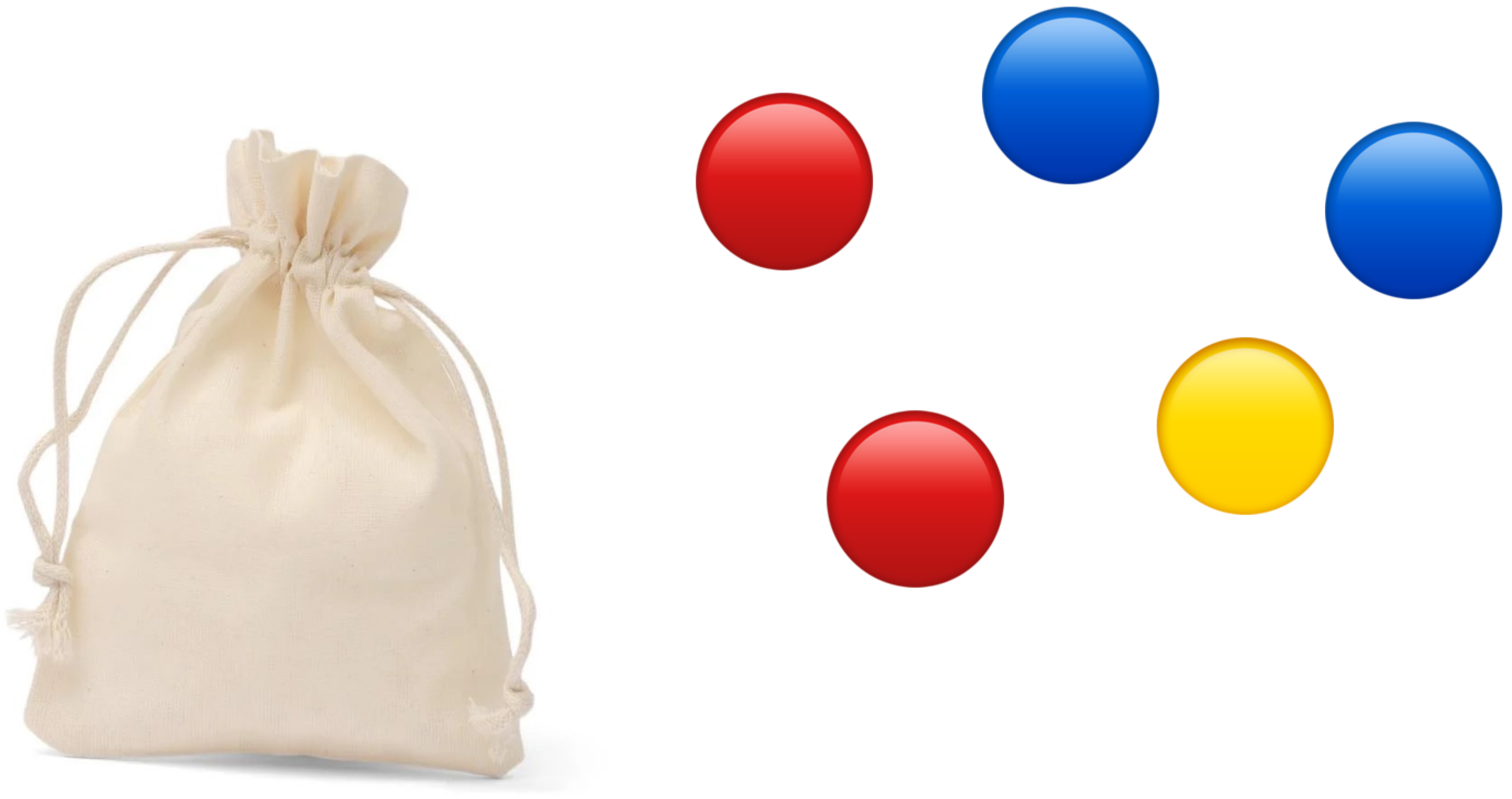
So, if *P*(🌧️) = 0.7 (that is, a 70% chance of rain)

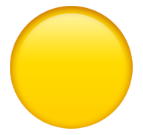
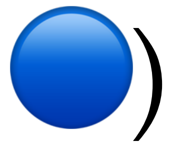
then *P*(not 🌧️) = 1 − 0.7
= 0.3 (that is, 30%)

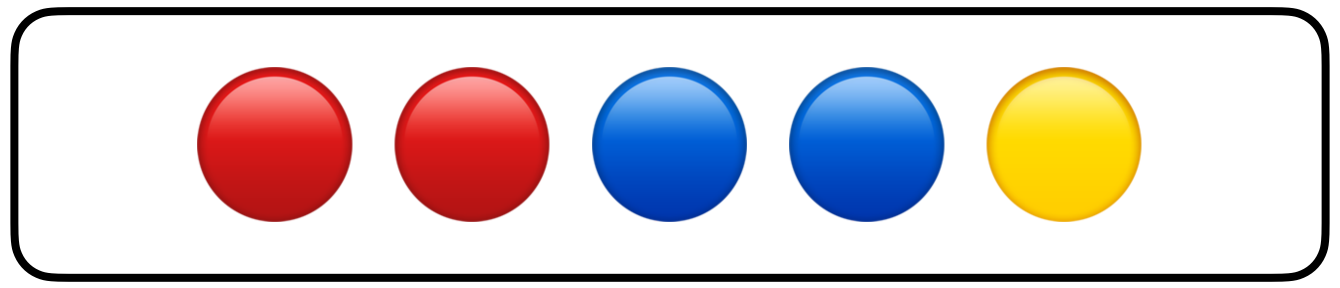
When you want to know the probability of events *A* and *B* ***both*** happening, you use the ***multiplication rule***:

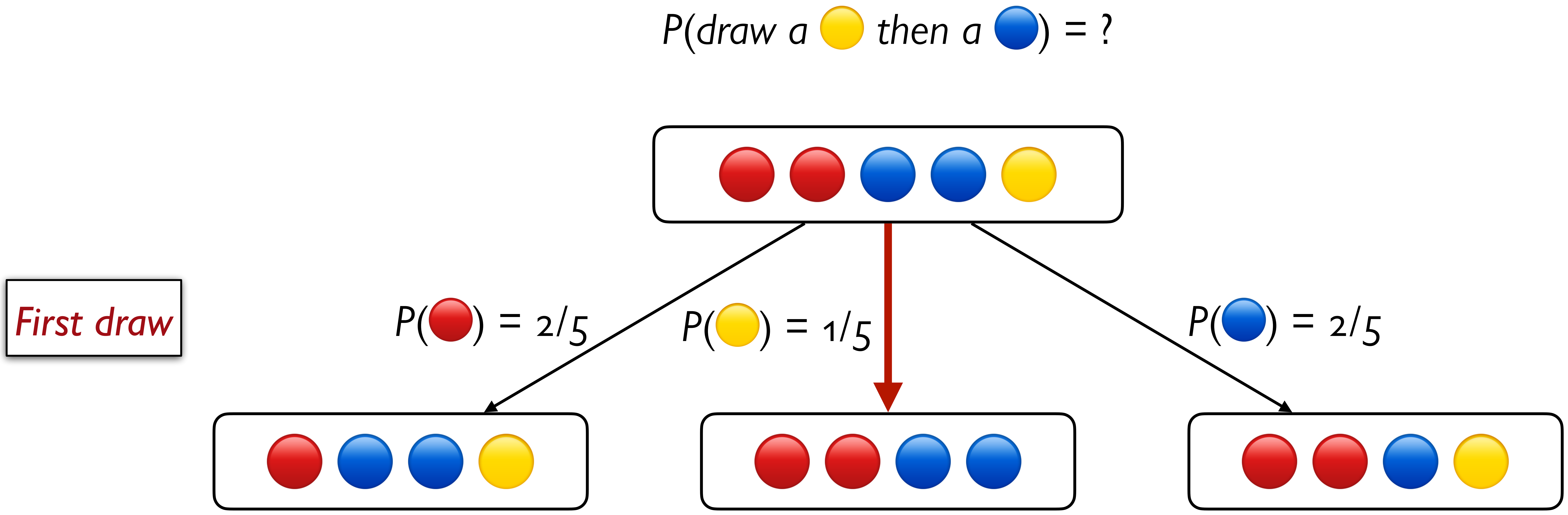
$$P(A \text{ and } B) = P(A) \cdot P(B \text{ given that } A \text{ happened})$$

For example, if you are taking colored marbles out of a bag, what is the probability of drawing a yellow marble *then* a blue marble?

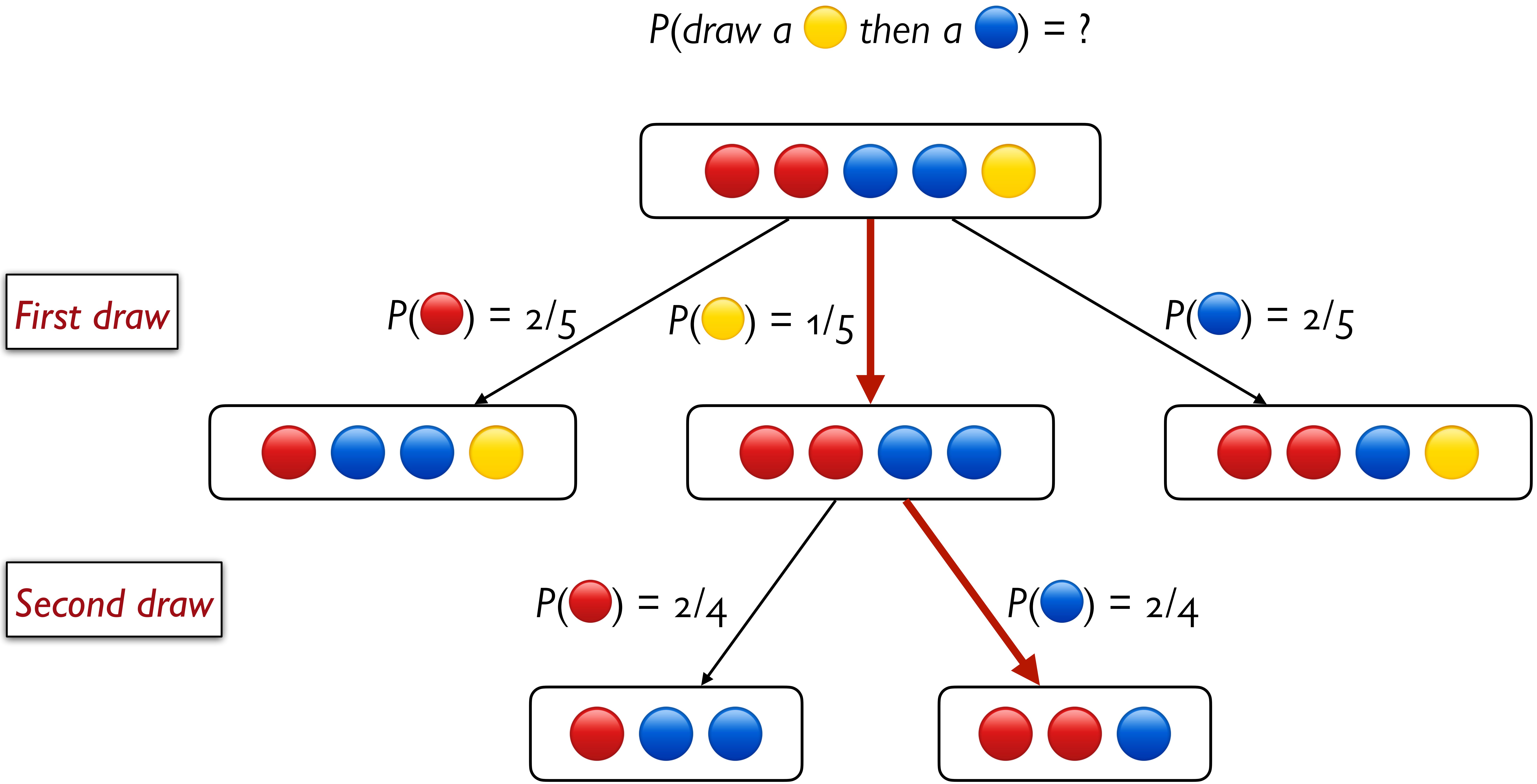


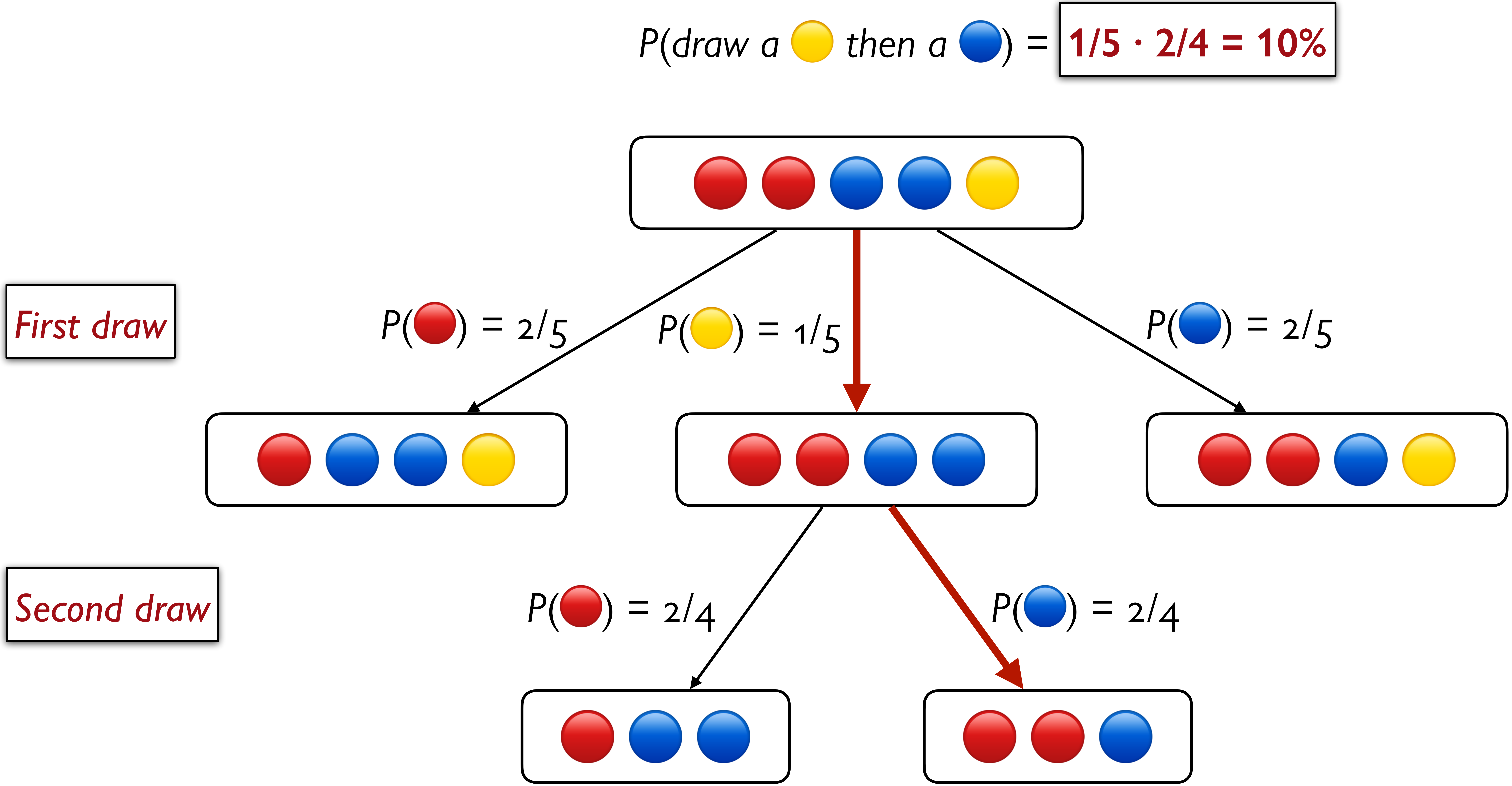
P(draw a  *then a* *) = ?*





P(draw a then a) = ?
First draw
P( ) = 2/5
P( ) = 1/5
P( ) = 2/5


P(draw a then a) = ?
First draw
P( ) = 2/5
P( ) = 1/5
P( ) = 2/5
Second draw
P( ) = 2/4
P( ) = 2/4



P(draw a then a) = 1/5 · 2/4 = 10%
First draw
P() = 2/5
P() = 1/5
P() = 2/5
Second draw
P() = 2/4
P() = 2/4

Practice

Probability the sum of two dice is 12?

Practice

Probability the sum of two dice is 12?

$P(\text{sum is } 12) = P(\text{roll ⚅}) \cdot P(\text{roll ⚅})$

Practice

Probability the sum of two dice is 12?

P(sum is 12) = P(roll ⚅) · P(roll ⚅)

= 1/6 · 1/6

Practice

Probability the sum of two dice is 12?

$$
\begin{aligned}
P(\text{sum is 12}) &= P(\text{roll ⚅}) \cdot P(\text{roll ⚅}) \\
&= 1/6 \cdot 1/6 \\
&= 1/36
\end{aligned}
$$

Practice





Probability of all 4s in five rolls?

Practice

Probability of all 4s in five rolls?

$$P(\text{five ⚃s}) = P(\text{roll ⚃})^5$$

Practice

Probability of all 4s in five rolls?

$$P(\text{five ⚃s}) = P(\text{roll ⚃})^5$$
$$= (1/6)^5$$

Practice

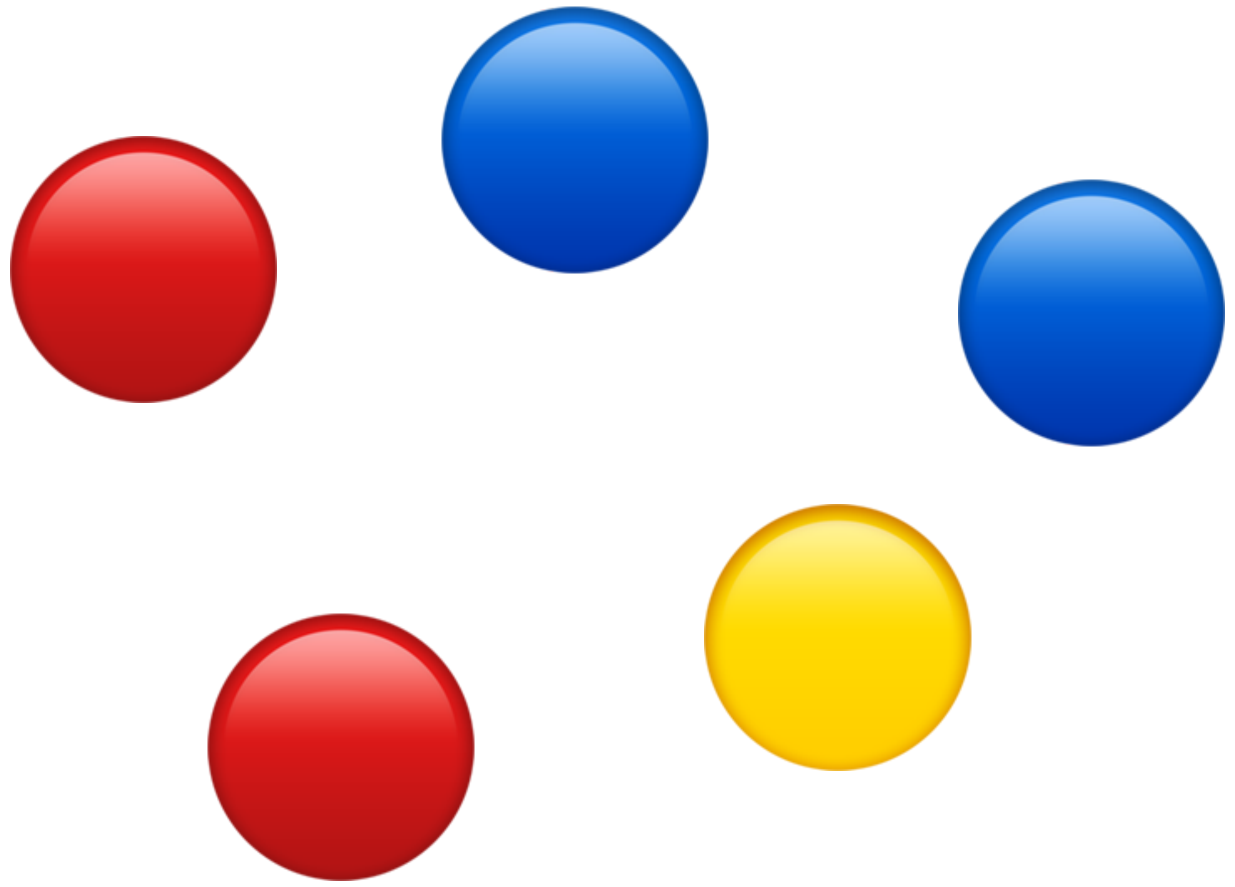
Probability of all 4s in five rolls?

$$P(\text{five ⚃s}) = P(\text{roll ⚃})^5$$

$$= (1/6)^5$$

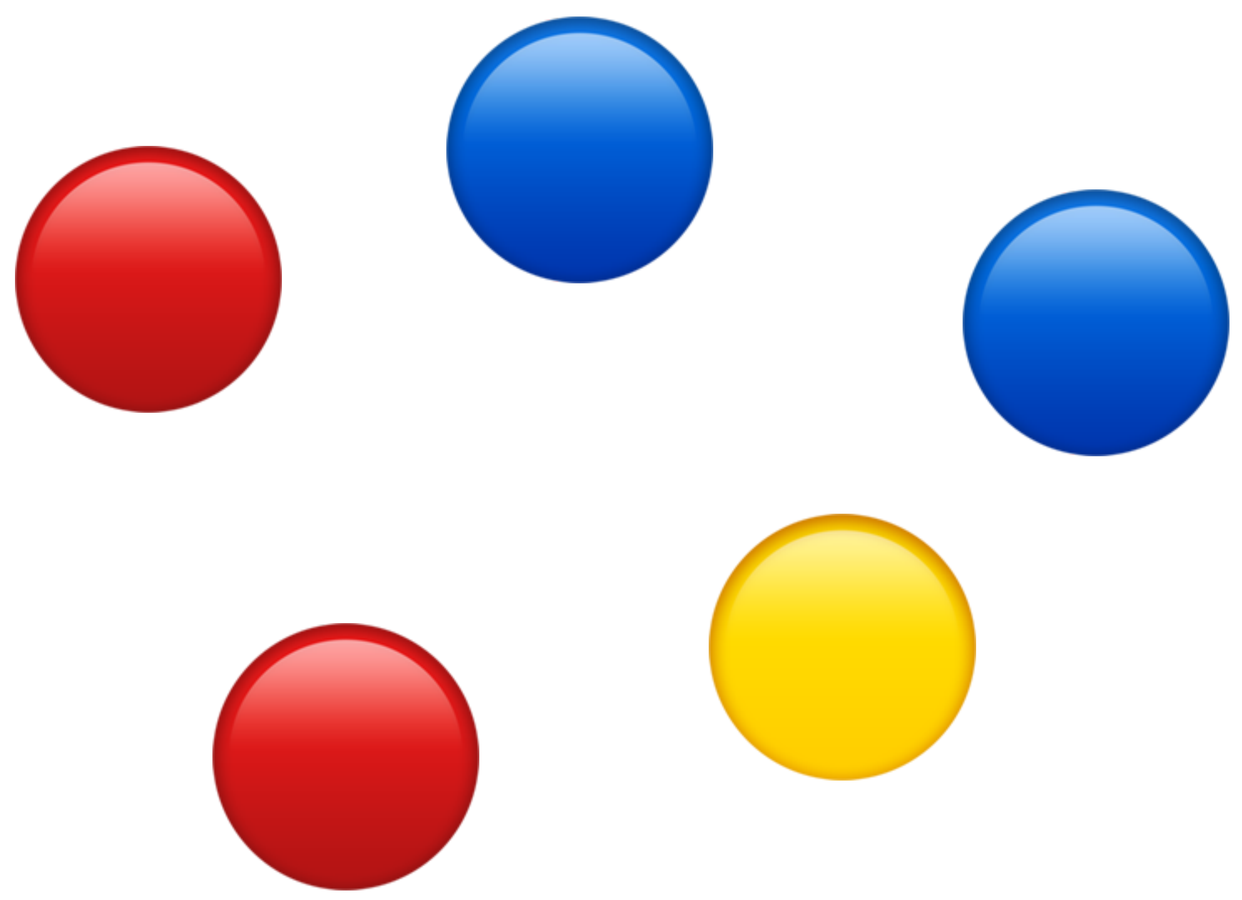
$$= 0.00013$$

Practice



Probability of drawing two of the same color?

Practice



Probability of drawing two of the same color?

P(two of same color) $= P($🔵 then 🔵$) + P($🔴 then 🔴$) + P($🟡 then 🟡$)$

$= 2/5 \cdot 1/4 + 2/5 \cdot 1/4 + 1/5 \cdot 0$

$= 2/20 + 2/20 + 0$

$= 1/5$

Probability and distributions

If you have a random quantity with various possible values, then its *probability* (*exact*) *distribution* associates

all the possible values of the quantity

with the probability of each of those values.

Sum of two dice rolls

If you have a random quantity with various possible values, then its *probability* (*exact*) *distribution* associates

all the possible values of the quantity

with the probability of each of those values.

Sum of two dice rolls

Random quantity

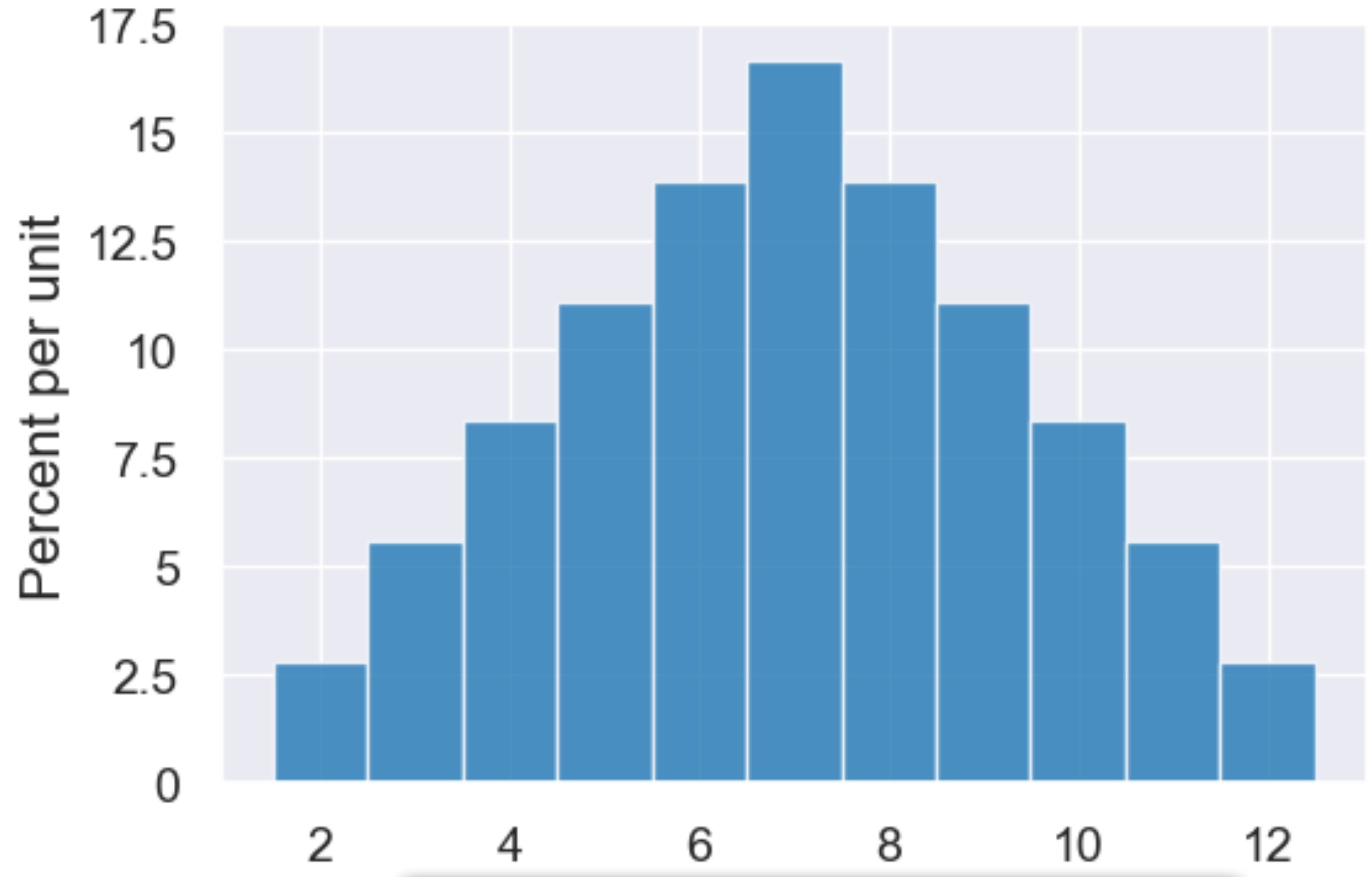
If you have a random quantity with various possible values, then its *probability* (*exact*) *distribution* associates

all the possible values of the quantity

with the probability of each of those values.

Sum of two dice rolls

Random quantity



Every possible outcome

If you can do the math, you can work out the probability distribution without ever simulating it:

Sum of two dice rolls

Employ probability rules

$P(\textit{sum is } 2) = P(⚀, ⚀)$

$= 1/6 \cdot 1/6$

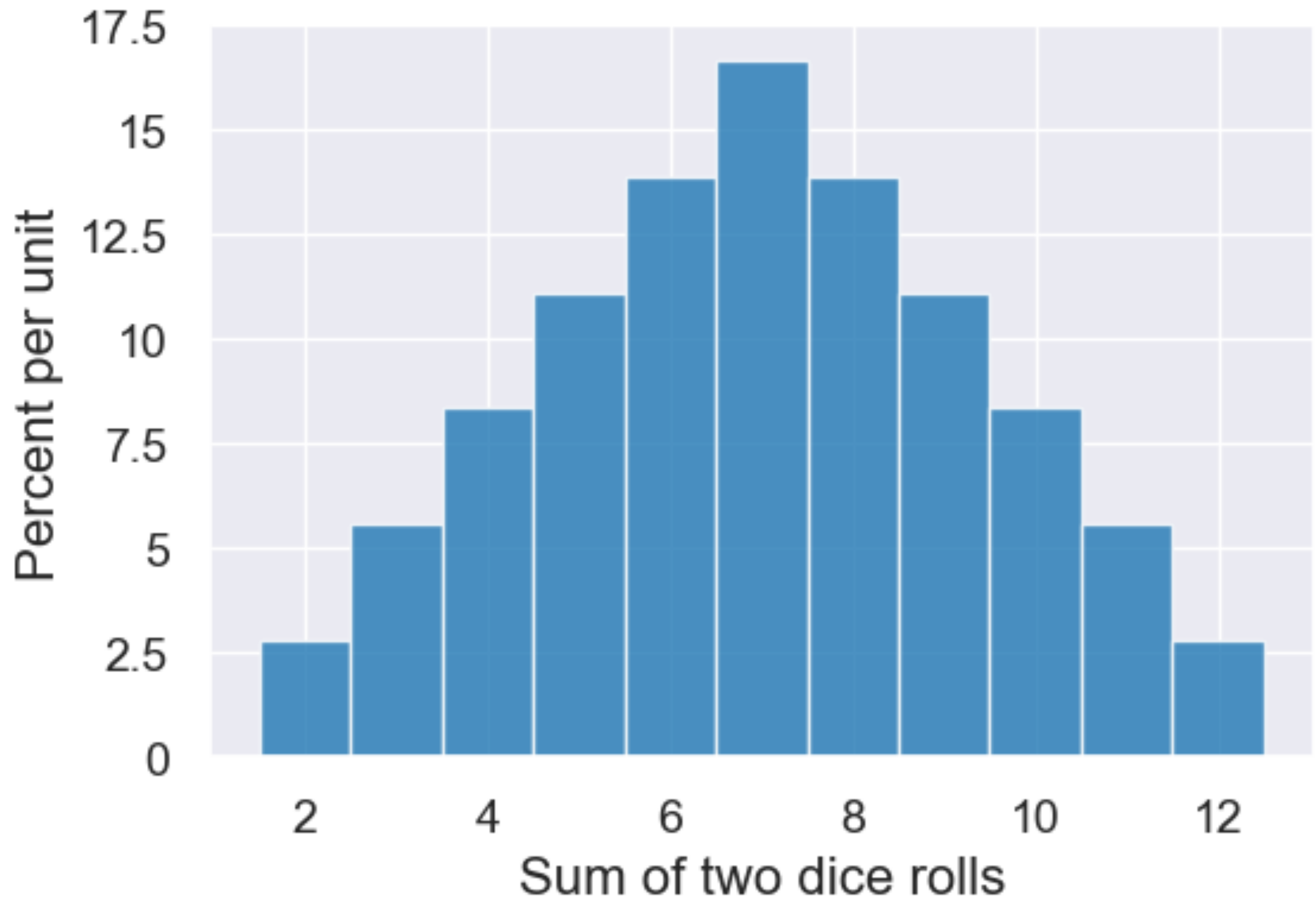
$P(\textit{sum is } 3) = P(⚀, ⚁) +$

$P(⚁, ⚀)$

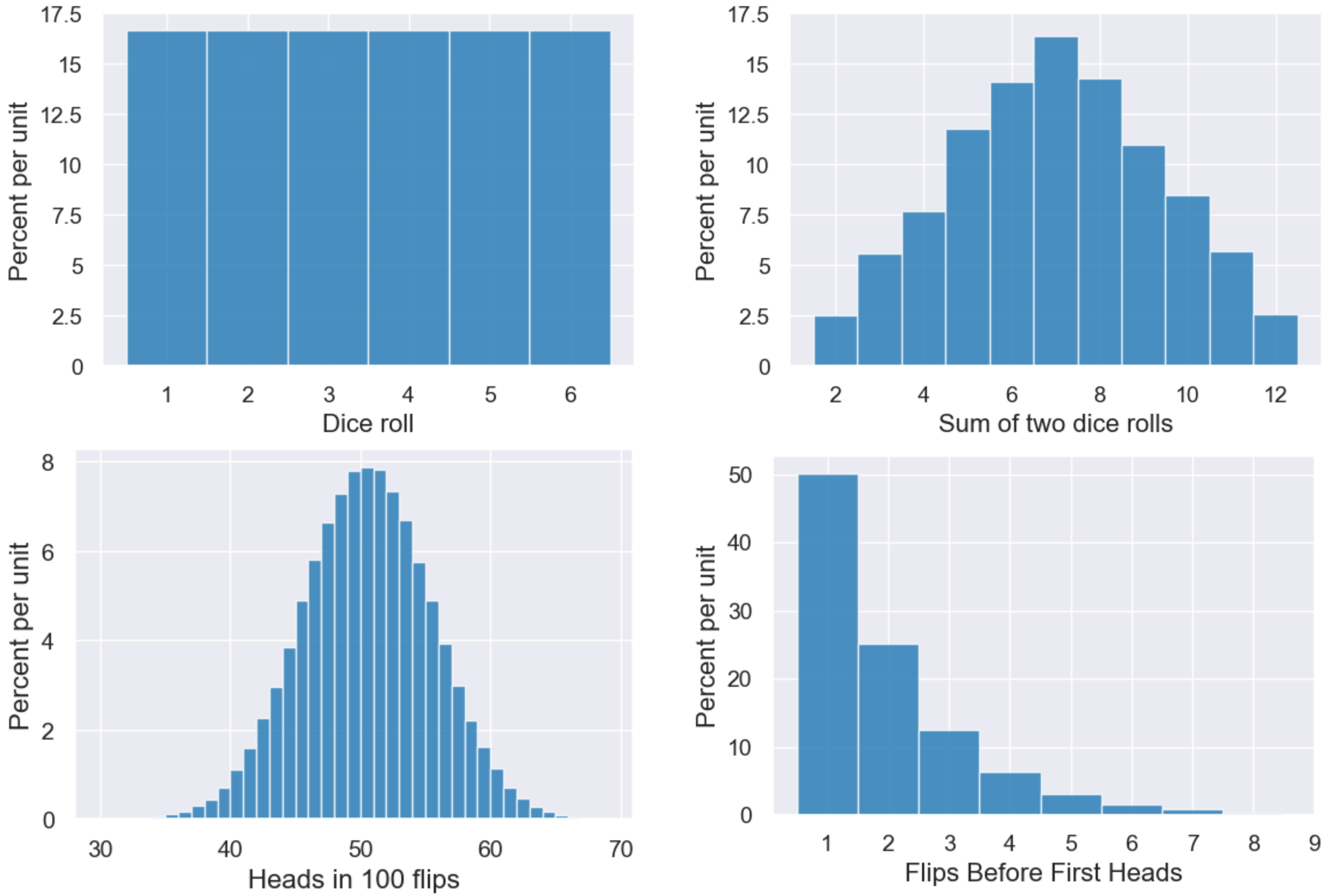
$= \ldots$

$\ldots$

$P(\textit{sum is } 12) = \ldots$

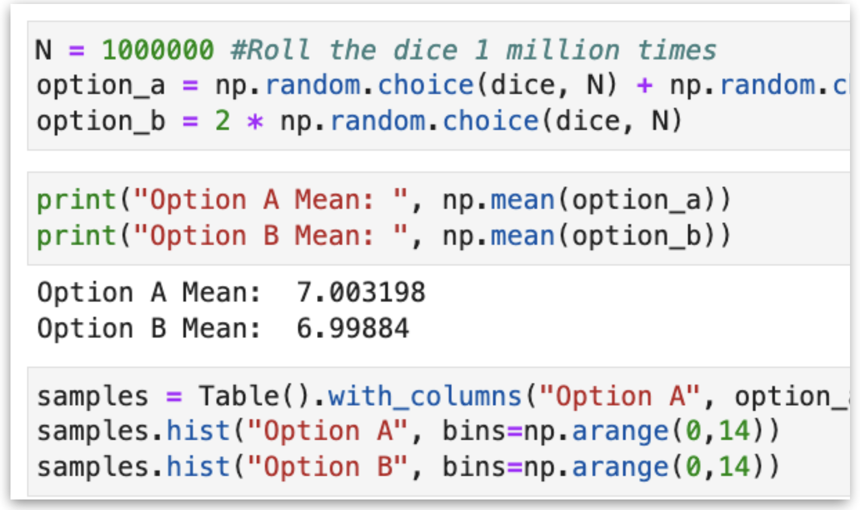


Probability distributions for other random quantities



How to calculate an event's probability?

Computers (simulation)

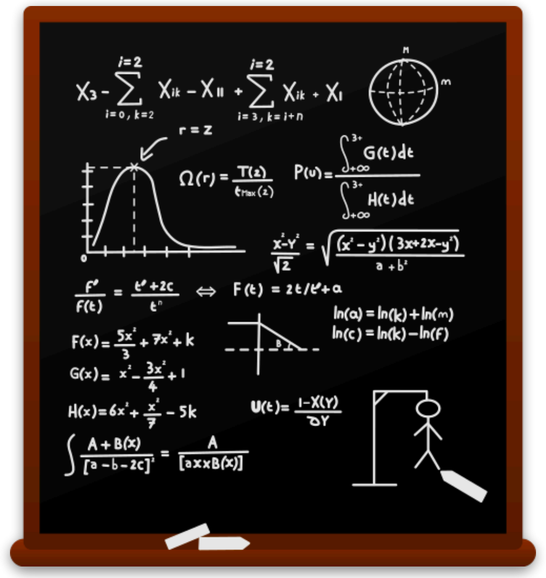


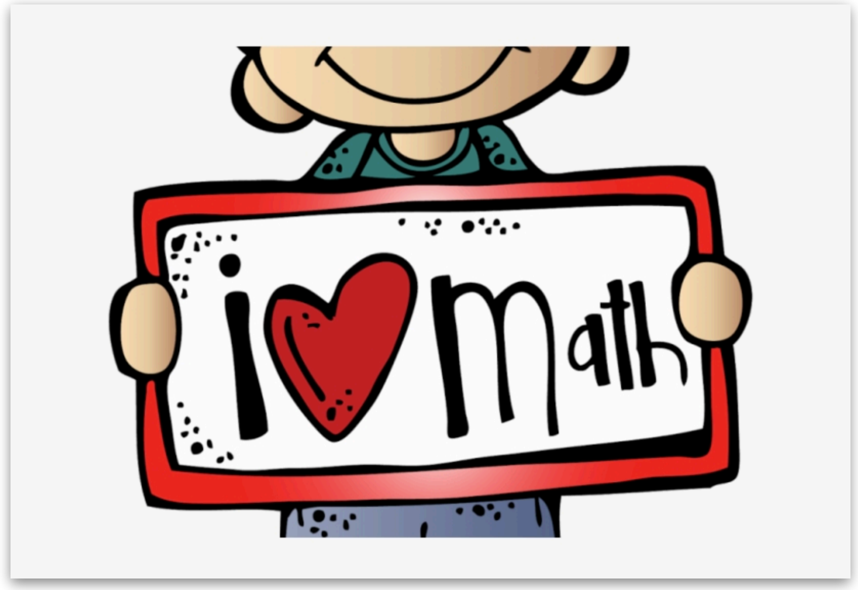
```
N = 1000000 #Roll the dice 1 million times
option_a = np.random.choice(dice, N) + np.random.c
option_b = 2 * np.random.choice(dice, N)

print("Option A Mean: ", np.mean(option_a))
print("Option B Mean: ", np.mean(option_b))

Option A Mean:  7.003198
Option B Mean:  6.99884

samples = Table().with_columns("Option A", option_
samples.hist("Option A", bins=np.arange(0,14))
samples.hist("Option B", bins=np.arange(0,14))
```

*Rooted in **algorithms***

- ✗ Approximate solutions
- ✓ Often convincing
- ✓ Non-trivial problems potentially captured cleanly with code

Math (analytical)





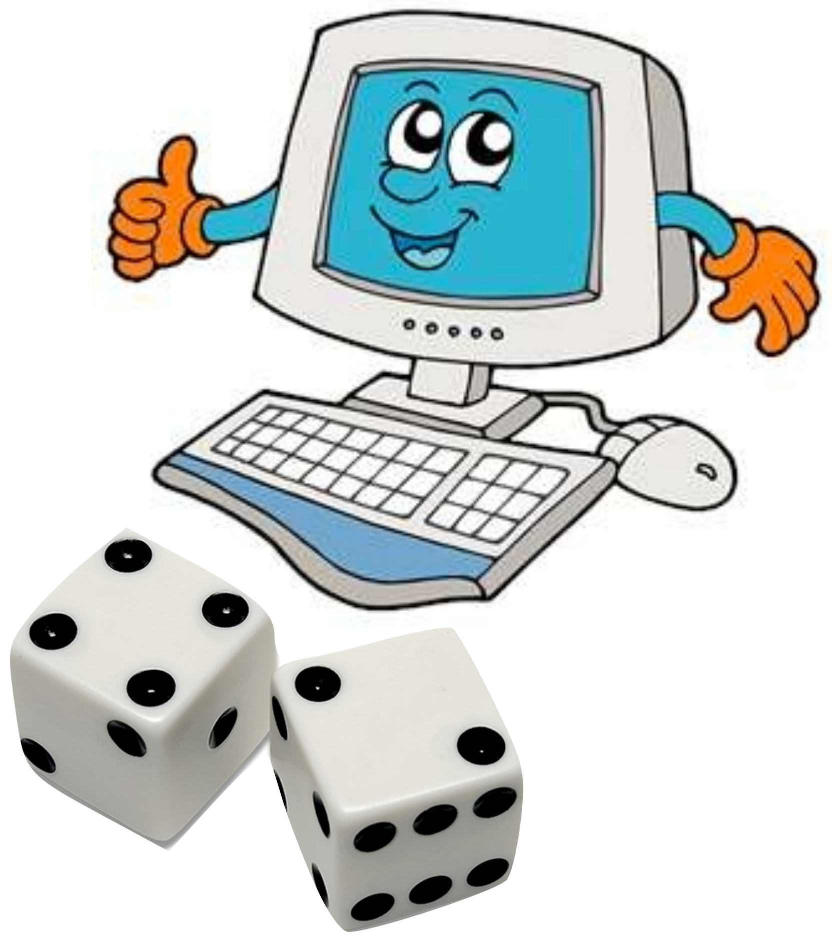
*Rooted in **rules (axioms)***

- ✓ Exact solutions
- ✓ Straightforward for simple problems
- ✗ Non-trivial problems potentially difficult to analyze/verify

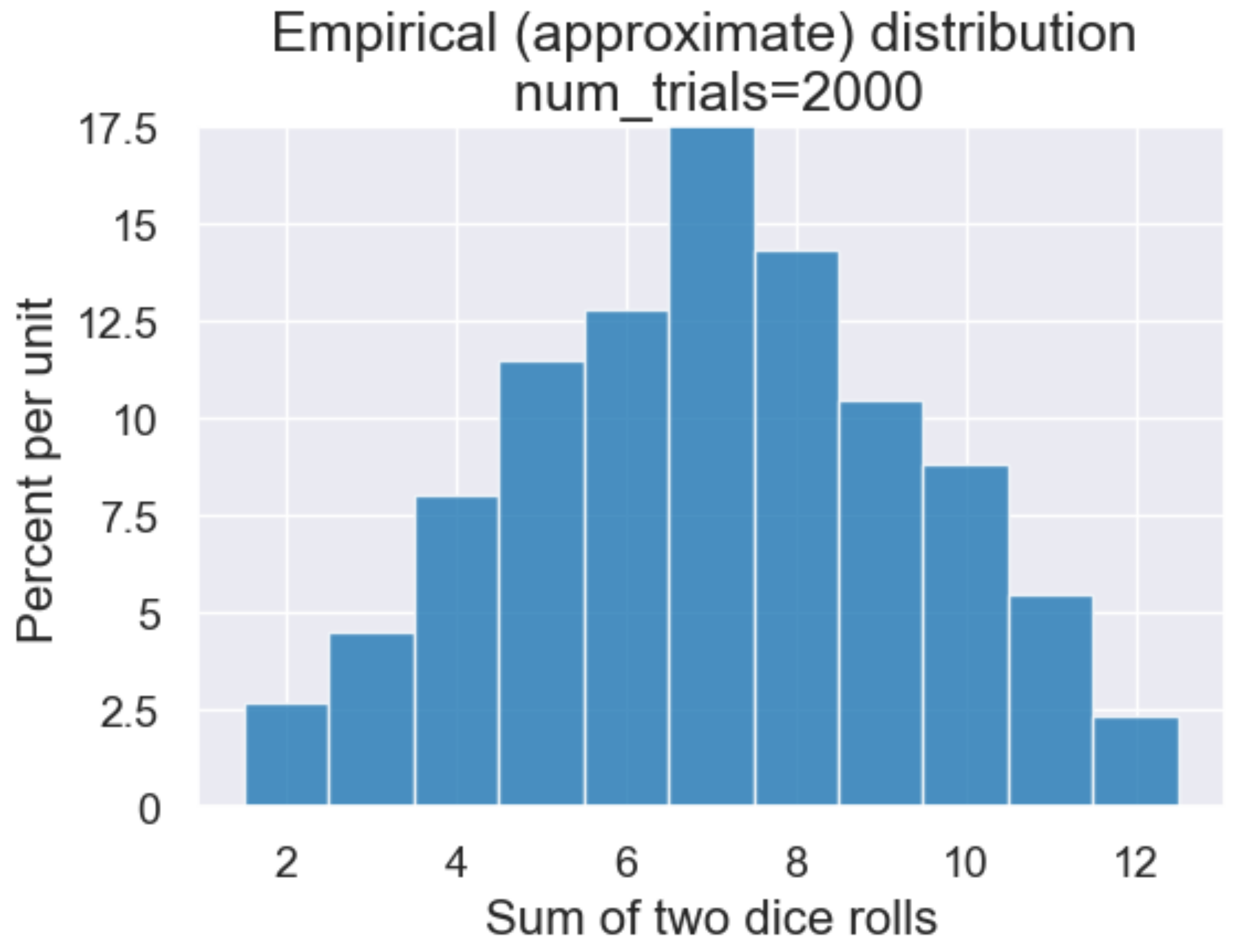
An *empirical* (*approximate*) *distribution* consists of observations, which can be from repetitions of an experiment. It associates

all the unique values you actually *observed* with

the proportion of times each value appeared.



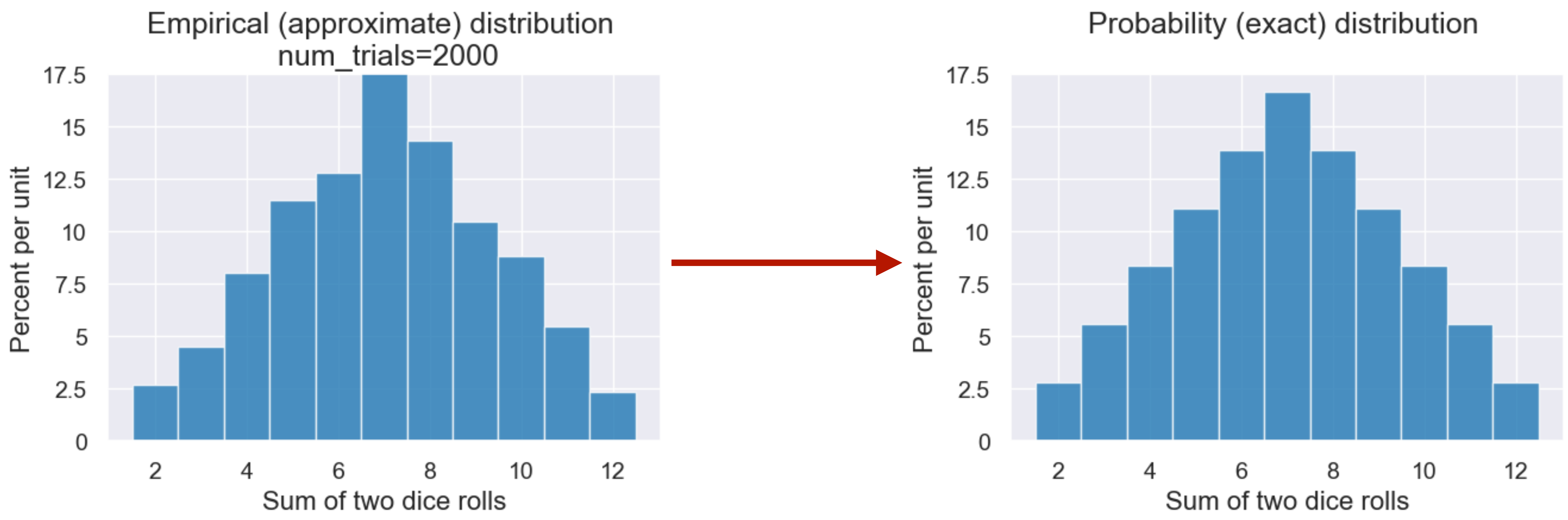
Outcomes: 9, 6, 8, 9, 5, 8, 8, 8, 8,
6, 9, 3, 3, 6, 9, 10, 6, 10, 7, 6, …



Notebook:
Empirical distribution: Sum of two dice

The *law of averages* (or *law of large numbers*):

If a chance experiment is repeated many times, independently and under the same conditions, then the proportion of times that an event occurs *gets closer* to the theoretical probability of the event.



Empirical distribution of a sample

If the sample size is large,

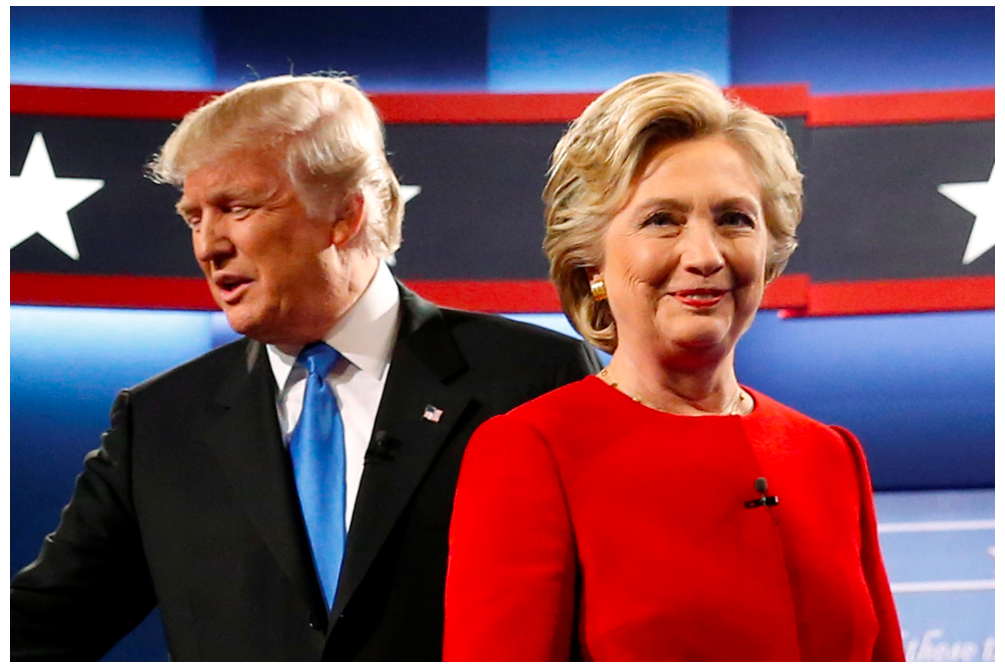
then the empirical distribution of a uniform random sample resembles the distribution of the population,

with high probability.

Real-world distributions and sampling

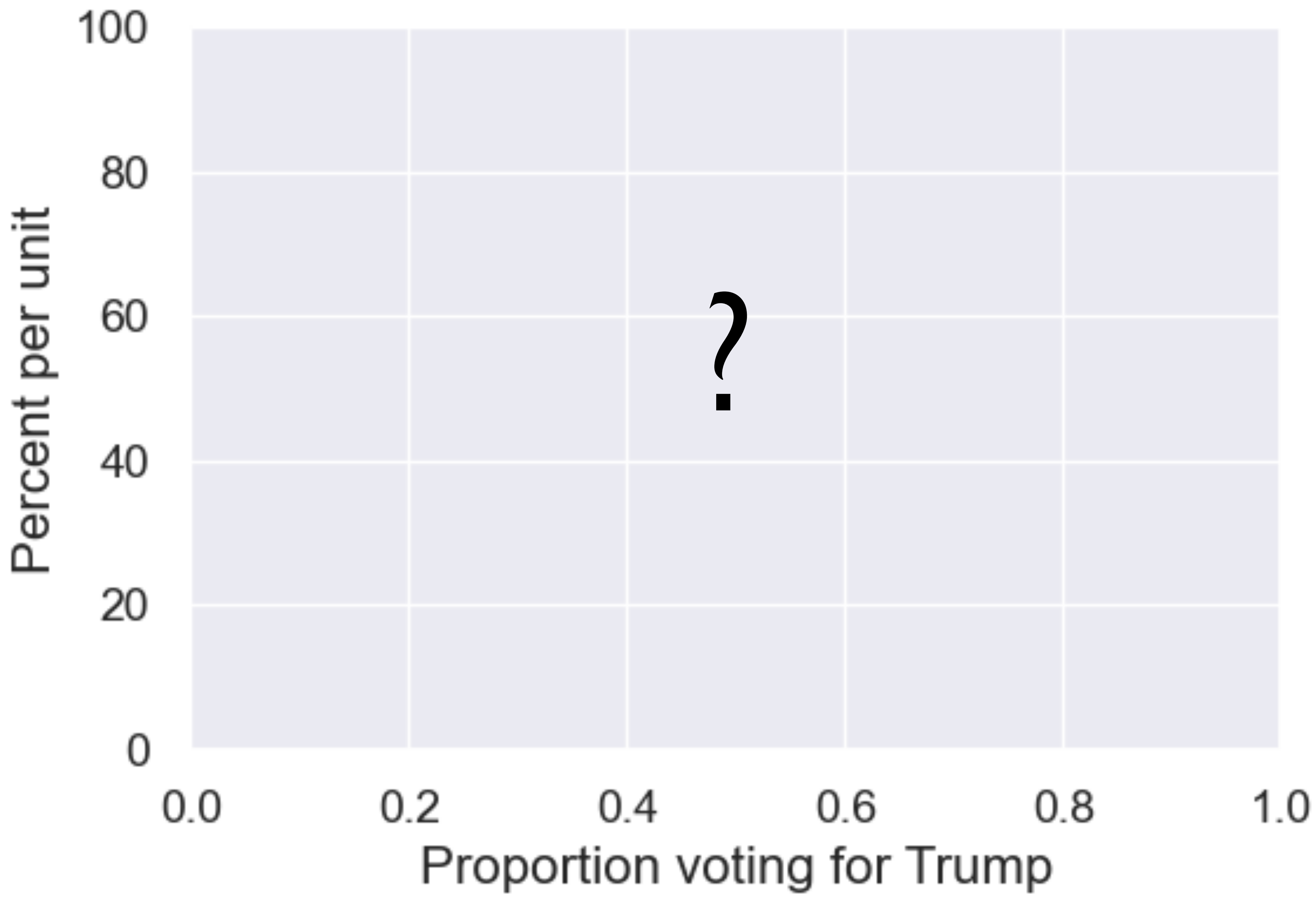
We could only simulate rolling two dice and taking their sum because we knew the true likelihood of each outcome for rolling a single die:

P(⚀) = 1/6
P(⚁) = 1/6
P(⚂) = 1/6
P(⚃) = 1/6
P(⚄) = 1/6
P(⚅) = 1/6

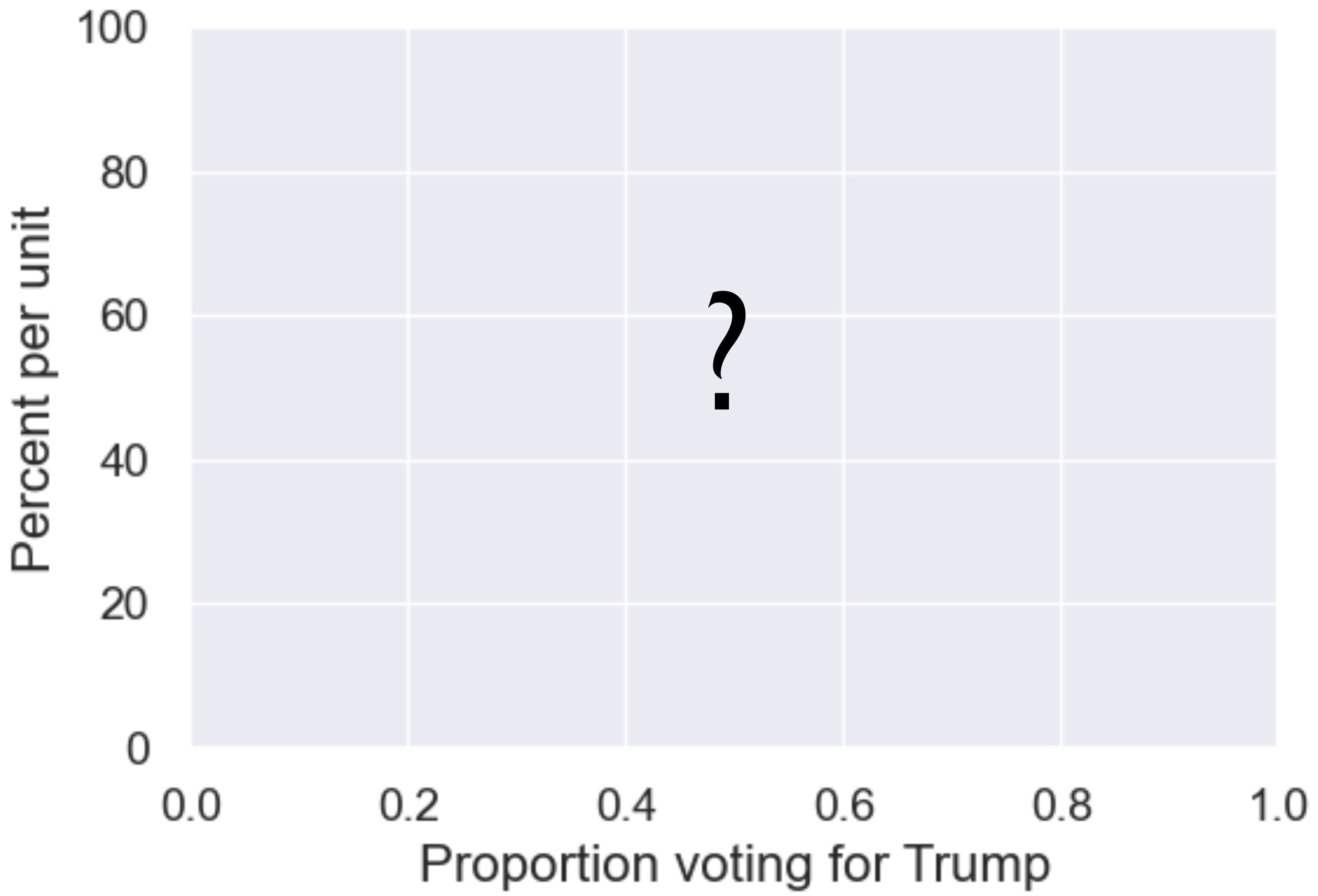


Circa 2016

What is the probability distribution for a candidate's chance of getting some percent of votes in an upcoming election?



100
80
60
40
20
0
Percent per unit
?
0.0
0.2
0.4
0.6
0.8
1.0
Proportion voting for Trump



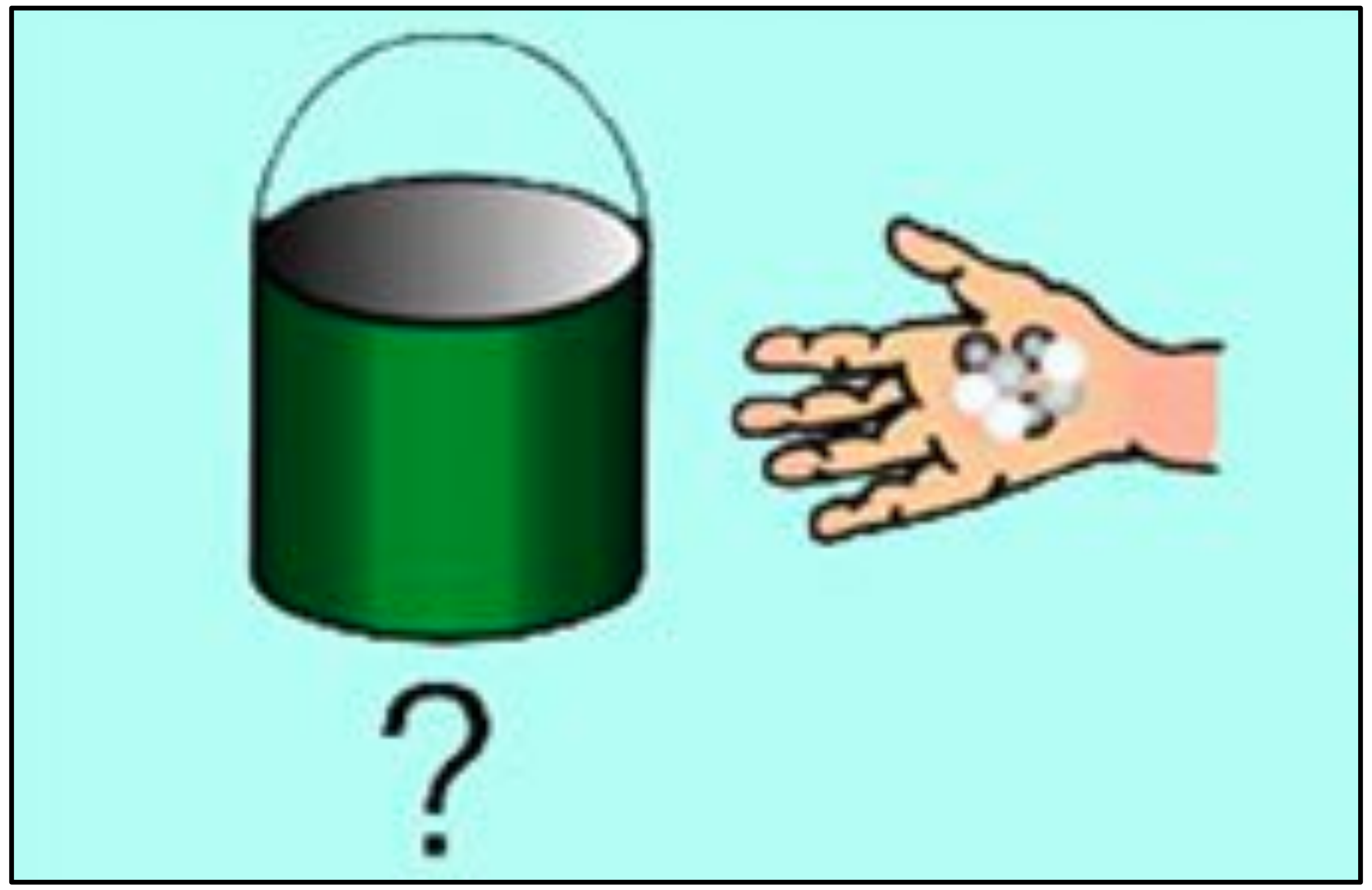
We don't know the true probability of whether each person will vote for a given candidate; we can't compute this distribution the way we did before!

Probability

Given the information in the pail, what is in your hand?

Probability

Given the information in the pail, what is in your hand?



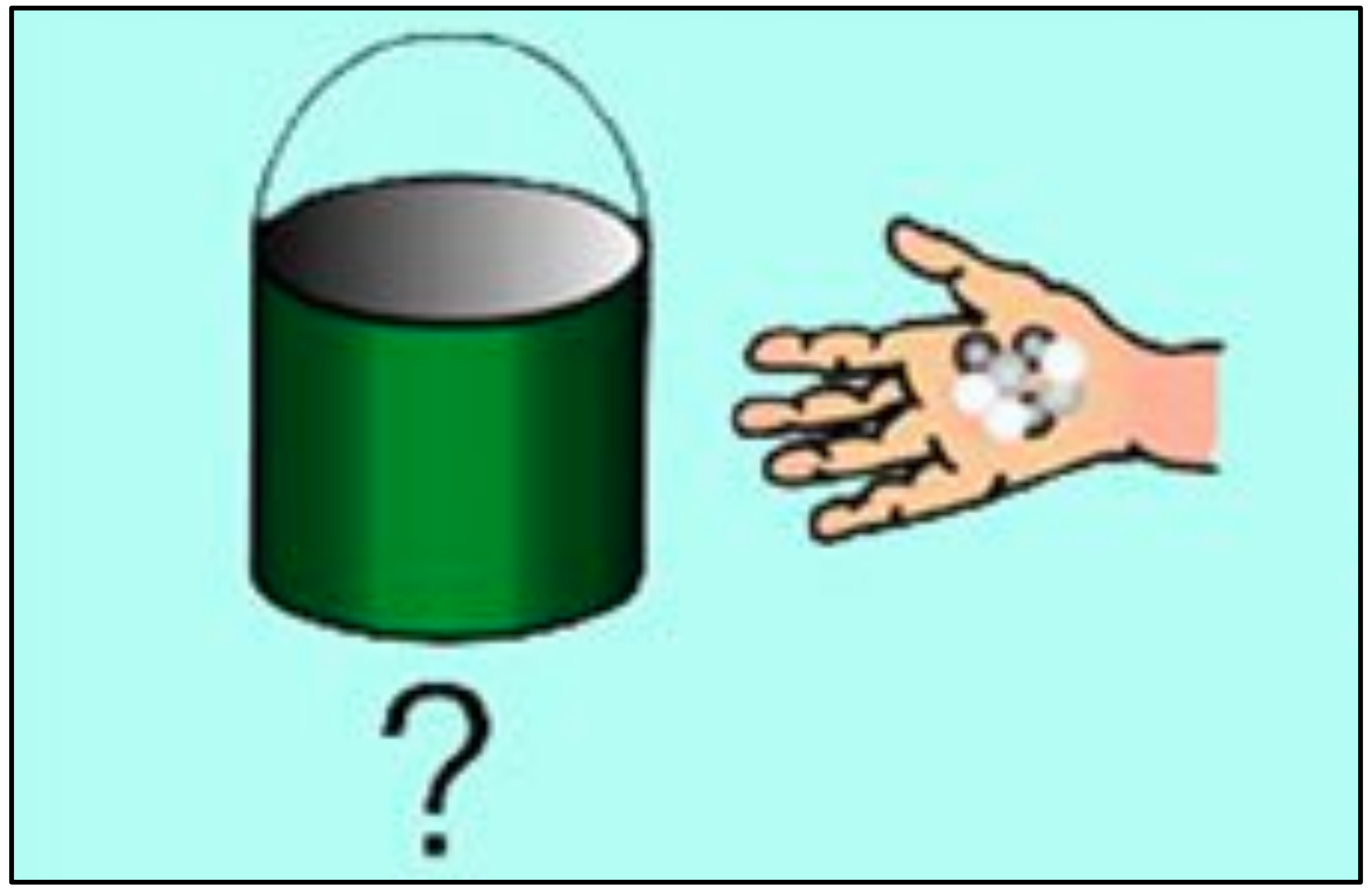
Statistics

Given the information in your hand, what is in the pail?

A *population* is a set of all elements from which a subset called a *sample* will be drawn.

How do we select our sample?

How do we draw meaningful conclusions using a sample?



Statistics

Given the information in your hand, what is in the pail?

Selecting a sample

Not all samples involve chance!

Here's an example of *deterministic sampling*:

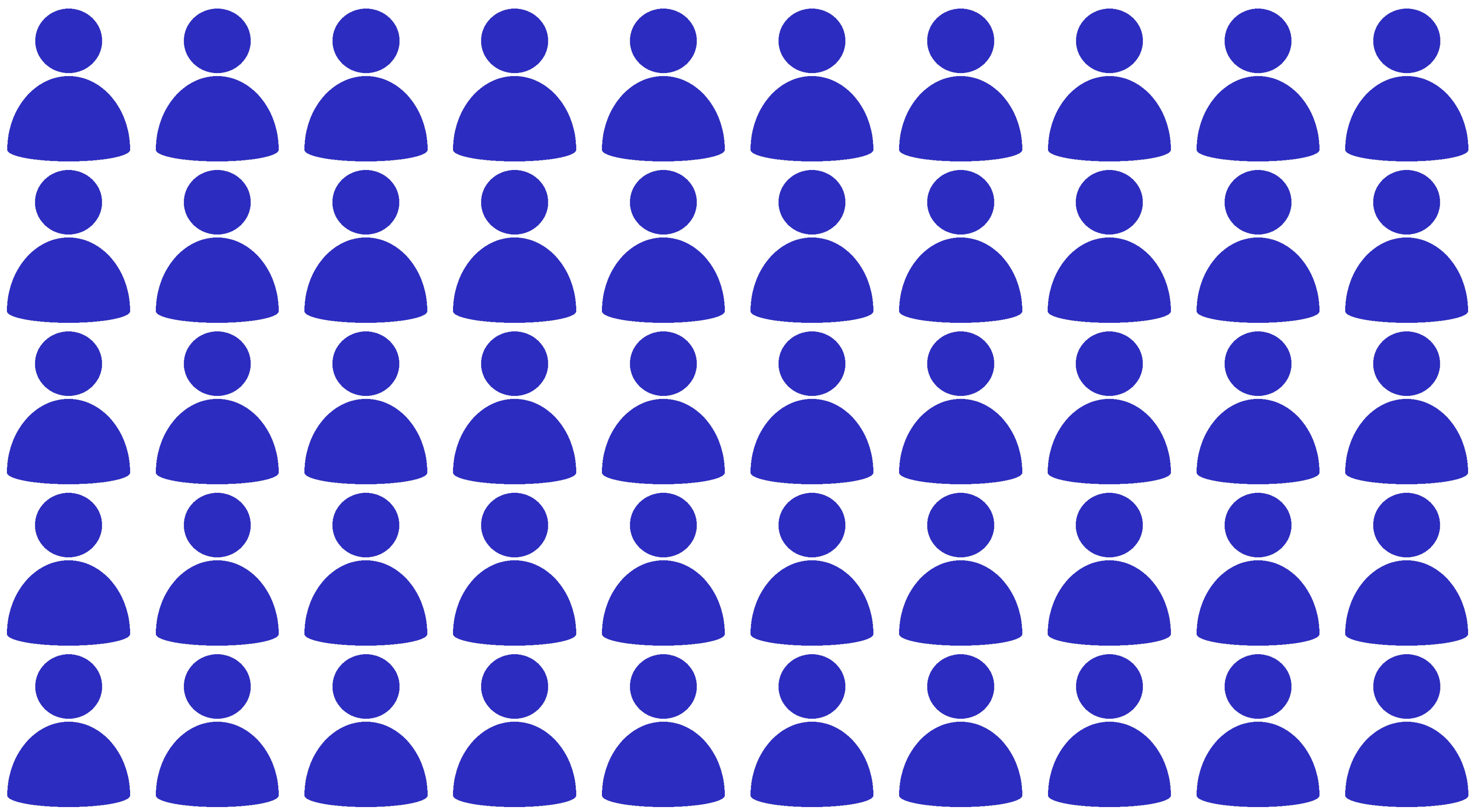
Pollster: Call every person with the first name "Bob" and ask him who he's voting for.

Not all samples involve chance!

Here's an example of *deterministic sampling*:

Pollster: Call every person with the first name "Bob" and ask him who he's voting for.

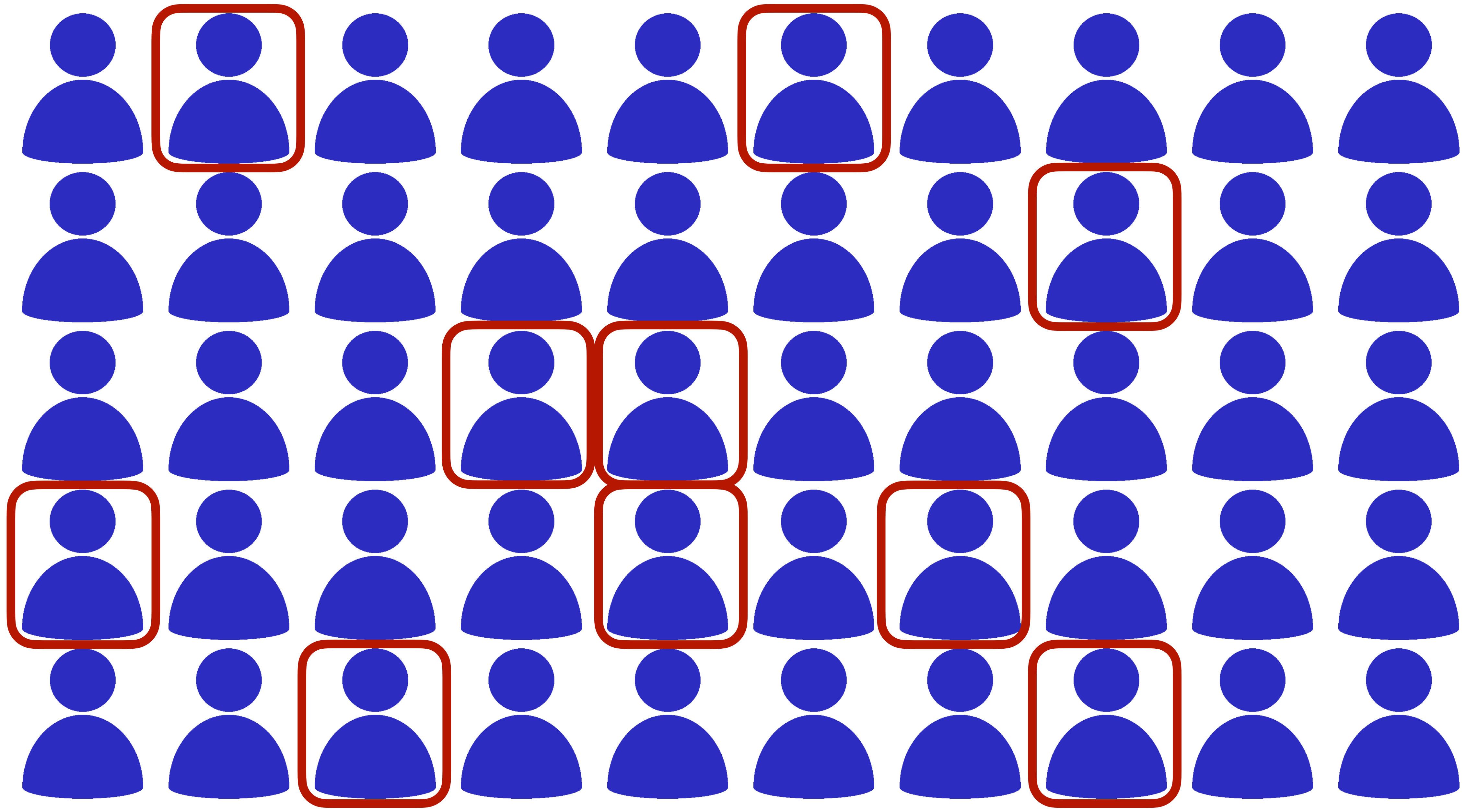
What's wrong with this?



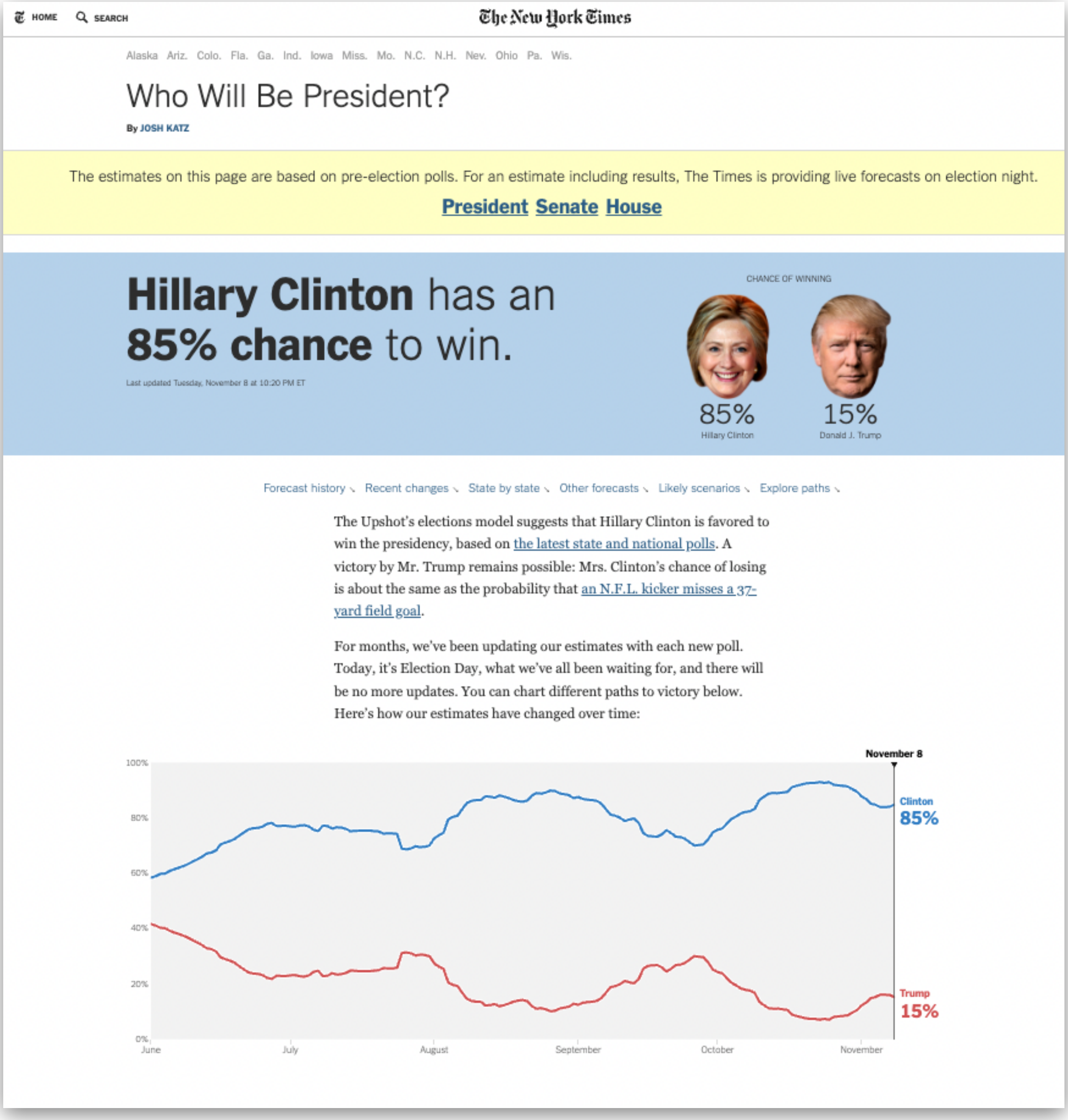
In ***random sampling***, each member of the population has some probability of being picked.

Example:

Random-digit dialing



In ***random sampling***, each member of the population has some probability of being picked.



The New York Times

HOME SEARCH

Alaska Ariz. Colo. Fla. Ga. Ind. Iowa Miss. Mo. N.C. N.H. Nev. Ohio Pa. Wis.

Who Will Be President?

By JOSH KATZ

The estimates on this page are based on pre-election polls. For an estimate including results, The Times is providing live forecasts on election night.

President Senate House

Hillary Clinton has an **85% chance** to win.

Last updated Tuesday, November 8 at 10:20 PM ET

CHANCE OF WINNING

85% Hillary Clinton

15% Donald J. Trump

Forecast history Recent changes State by state Other forecasts Likely scenarios Explore paths

The Upshot's elections model suggests that Hillary Clinton is favored to win the presidency, based on the latest state and national polls. A victory by Mr. Trump remains possible: Mrs. Clinton's chance of losing is about the same as the probability that an N.F.L. kicker misses a 37-yard field goal.

For months, we've been updating our estimates with each new poll. Today, it's Election Day, what we've all been waiting for, and there will be no more updates. You can chart different paths to victory below. Here's how our estimates have changed over time:

"All the News That's Fit to Print"

The New York Times

Late Edition

Today, cloudy, showers midday, high 56. Tonight, stray evening showers, clouds breaking late, colder, low 40. Tomorrow, sunshine, high 56. Weather map appears on Page B11.

VOL. CLXVI . . . No. 57,411 ++ © 2016 The New York Times Company NEW YORK, WEDNESDAY, NOVEMBER 9, 2016 $2.50

TRUMP TRIUMPHS

OUTSIDER MOGUL CAPTURES THE PRESIDENCY, STUNNING CLINTON IN BATTLEGROUND STATES

WORKING CLASS SPEAKS

Blue-Collar Whites Give Stinging Rebuke to Democratic Party

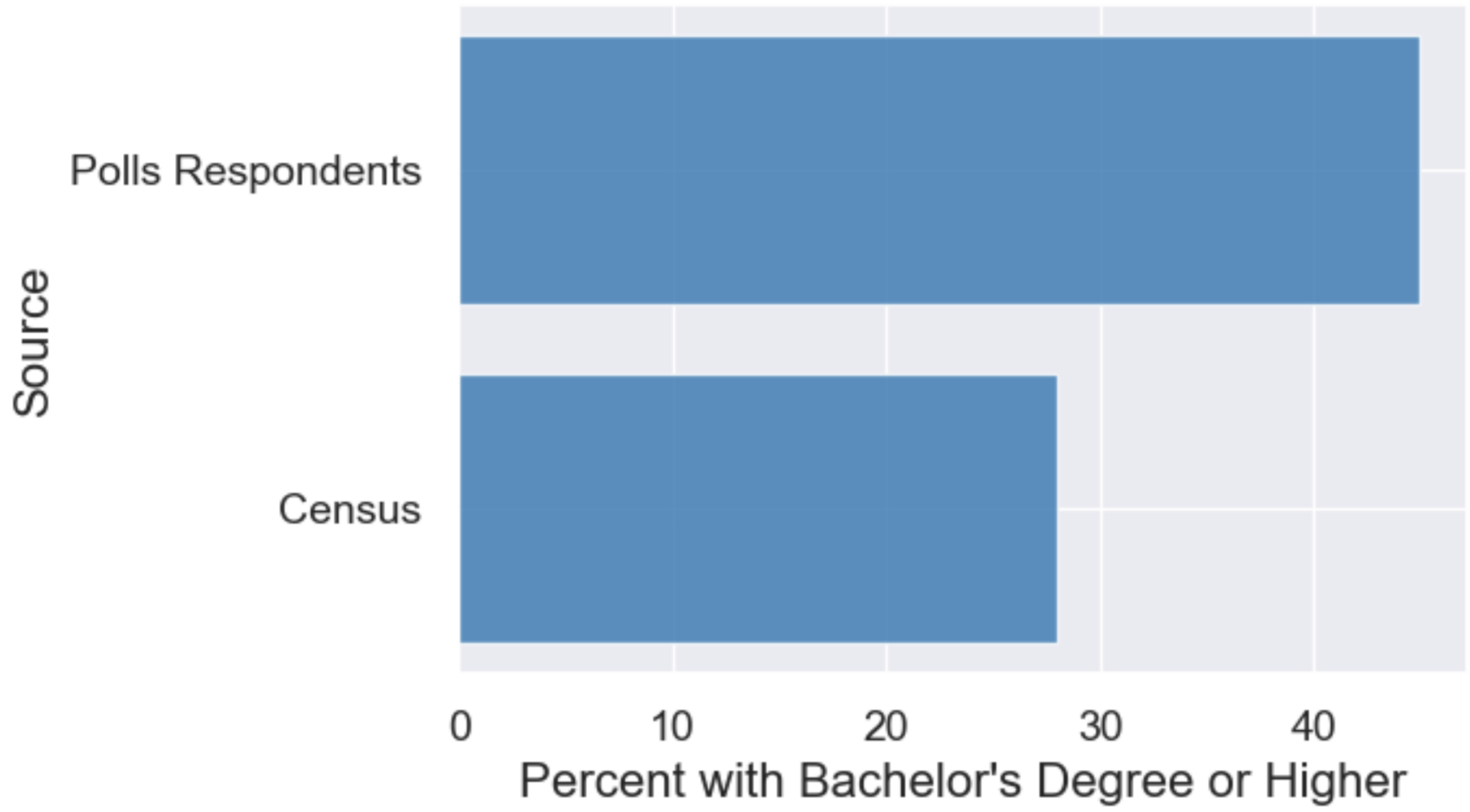
A Blue-Collar Town in Decline And in Despair Turns to Trump

Around the World, Uncertainty And Fear That 'All Bets Are Off'

Clarion of White Populist Rage Who Vowed 'I Am Your Voice'

New York Times,

Tuesday November 8, 2016

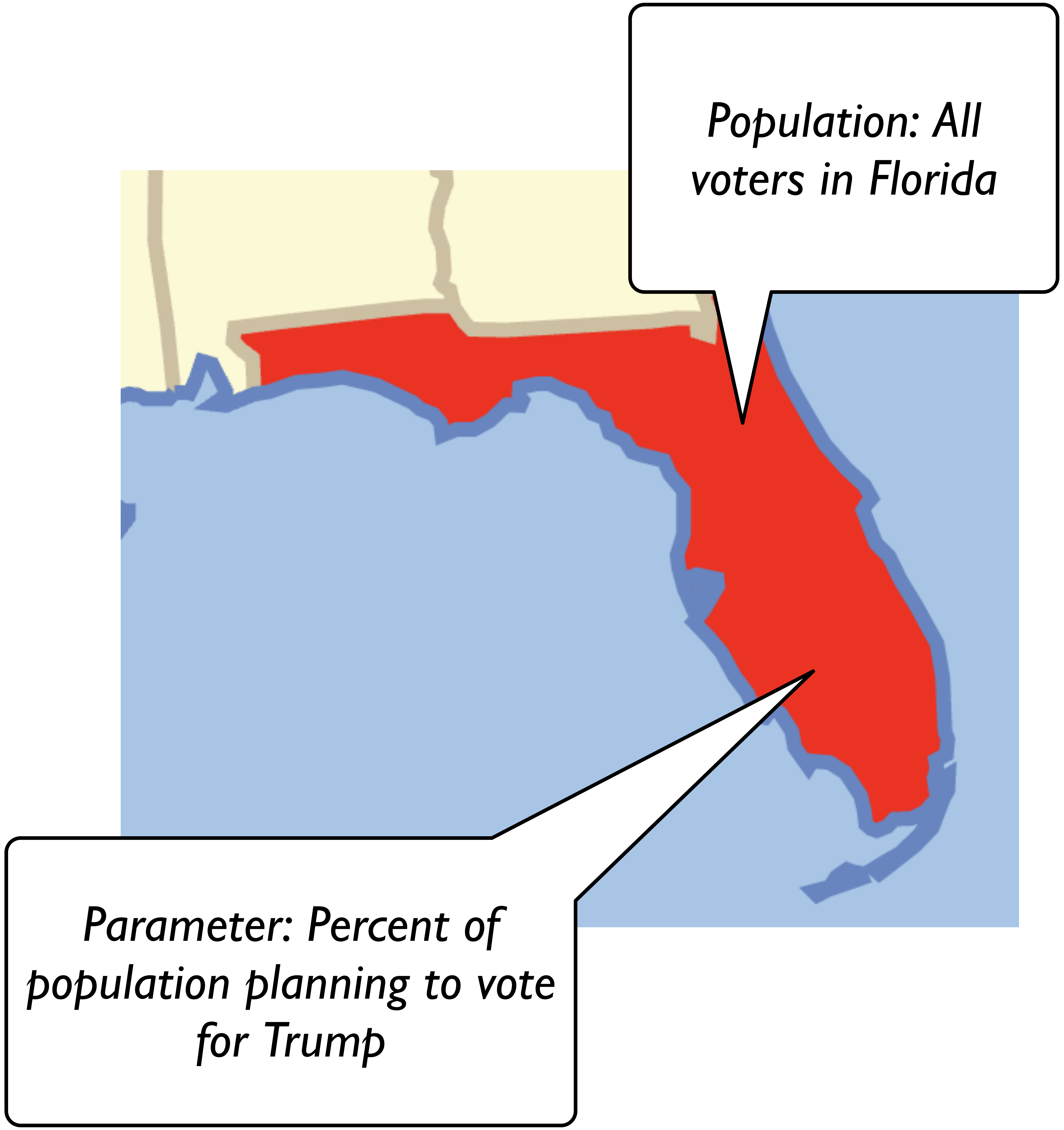


After the 2016 election, analysis showed individuals with higher education were *overrepresented* in polling samples.

A statistic

Terminology

Parameter: A fixed number associated with a *population*

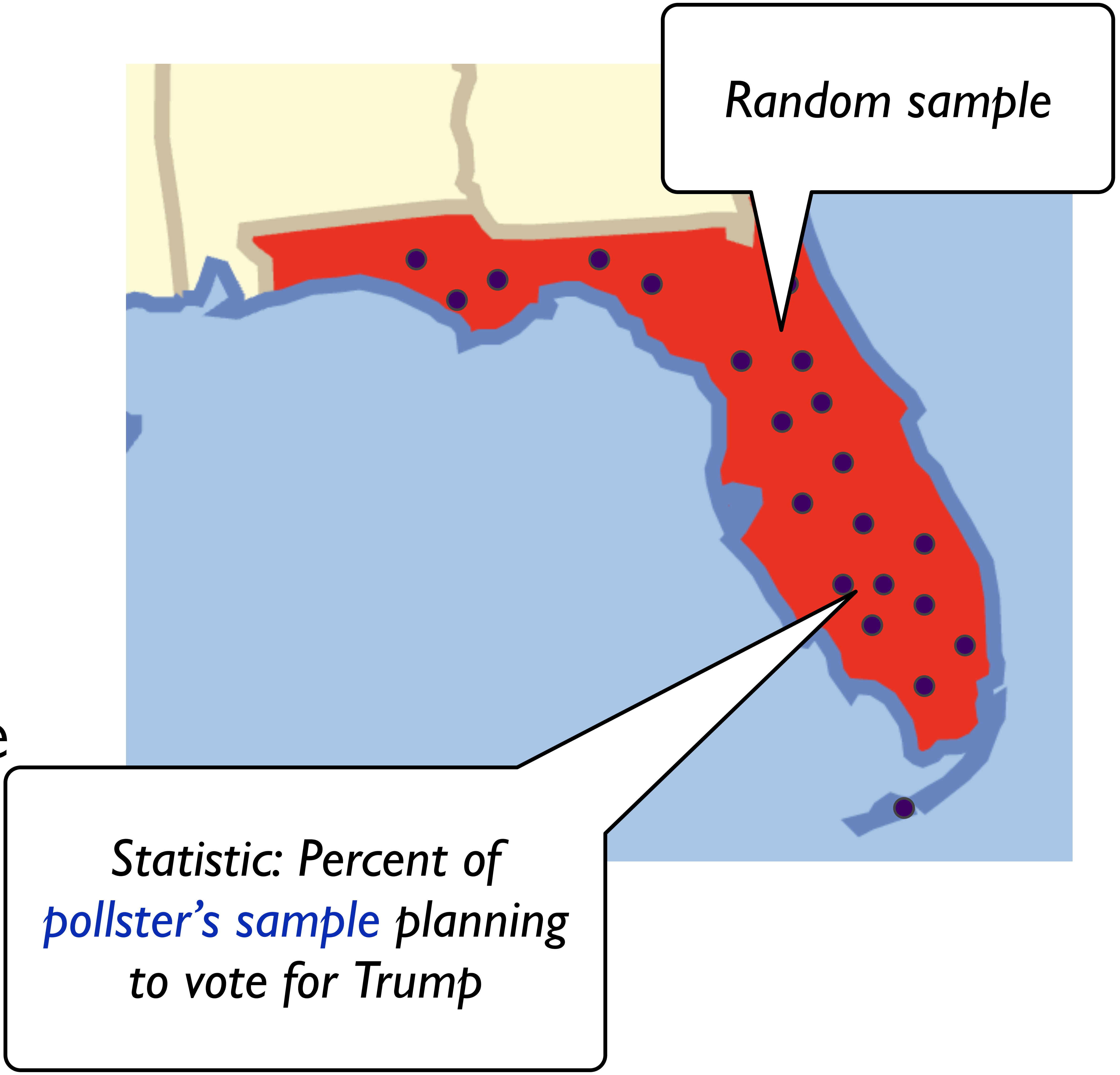


Terminology

Parameter: A fixed number associated with a *population*

Statistic: Any number computed using the data in a sample, e.g., the mean or median.

Statistical inference: Estimate the value of the parameter with statistics

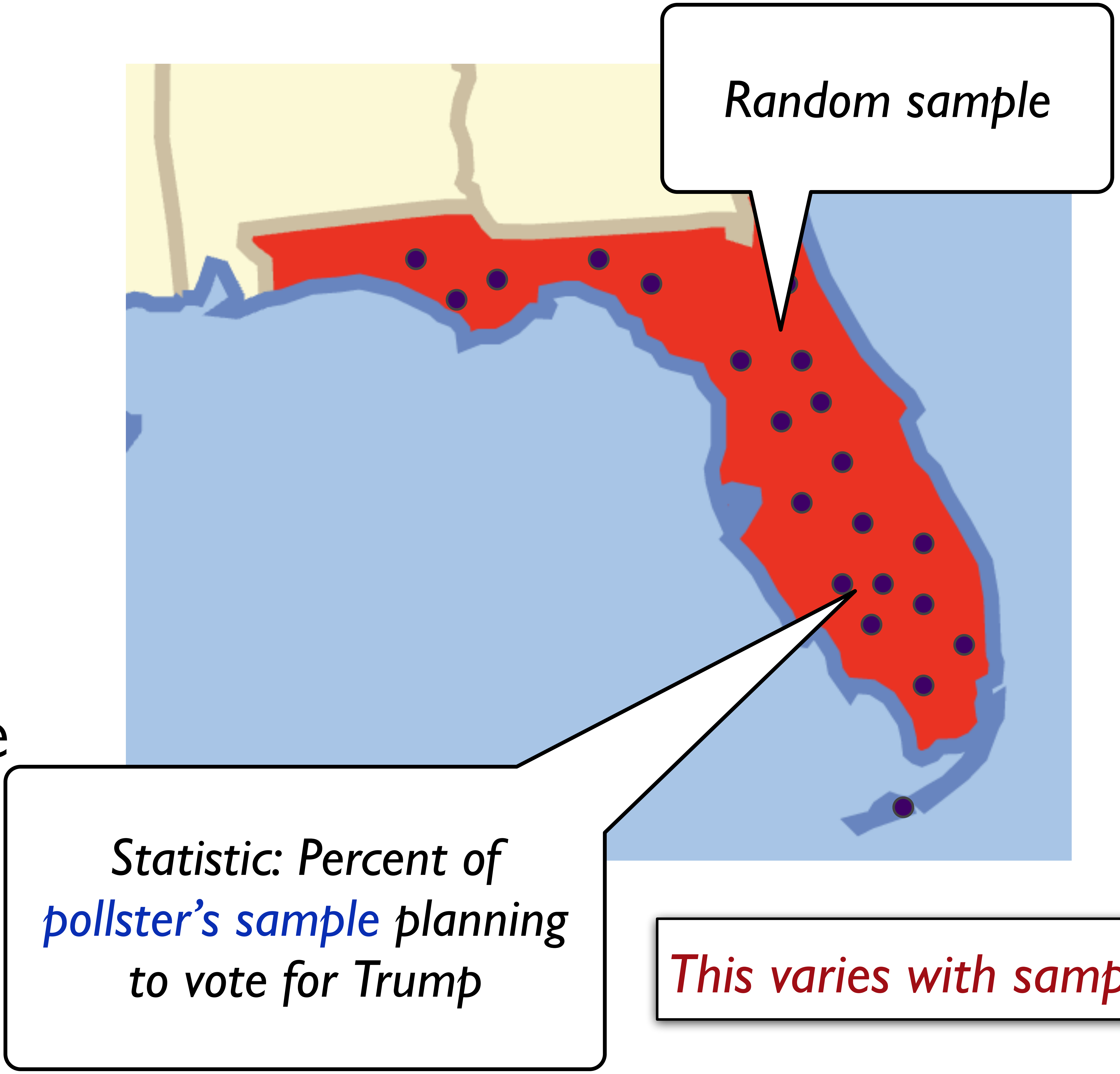


Terminology

Parameter: A fixed number associated with a *population*

Statistic: Any number computed using the data in a sample, e.g., the mean or median.

Statistical inference: Estimate the value of the parameter with statistics



This varies with sample!

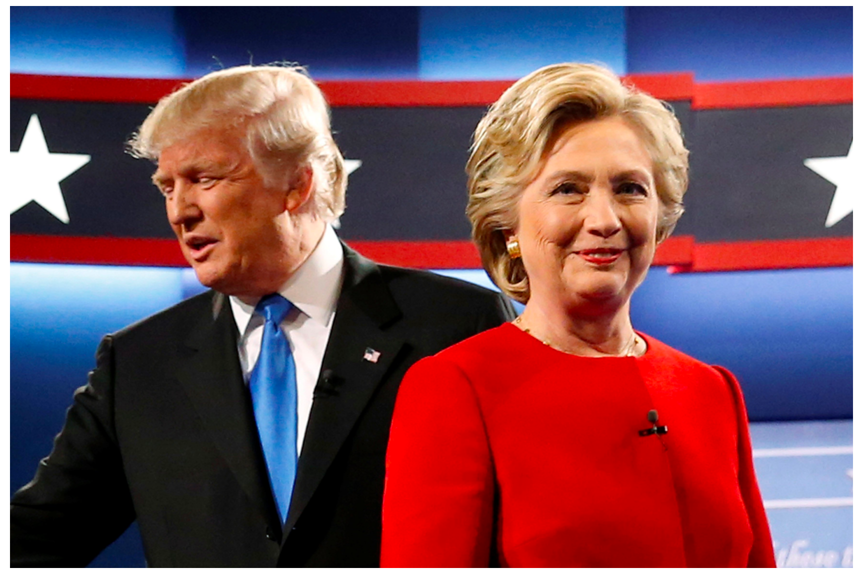
Notebook: *Random sampling: Florida votes in 2016*

Notebook: *General sampling function*

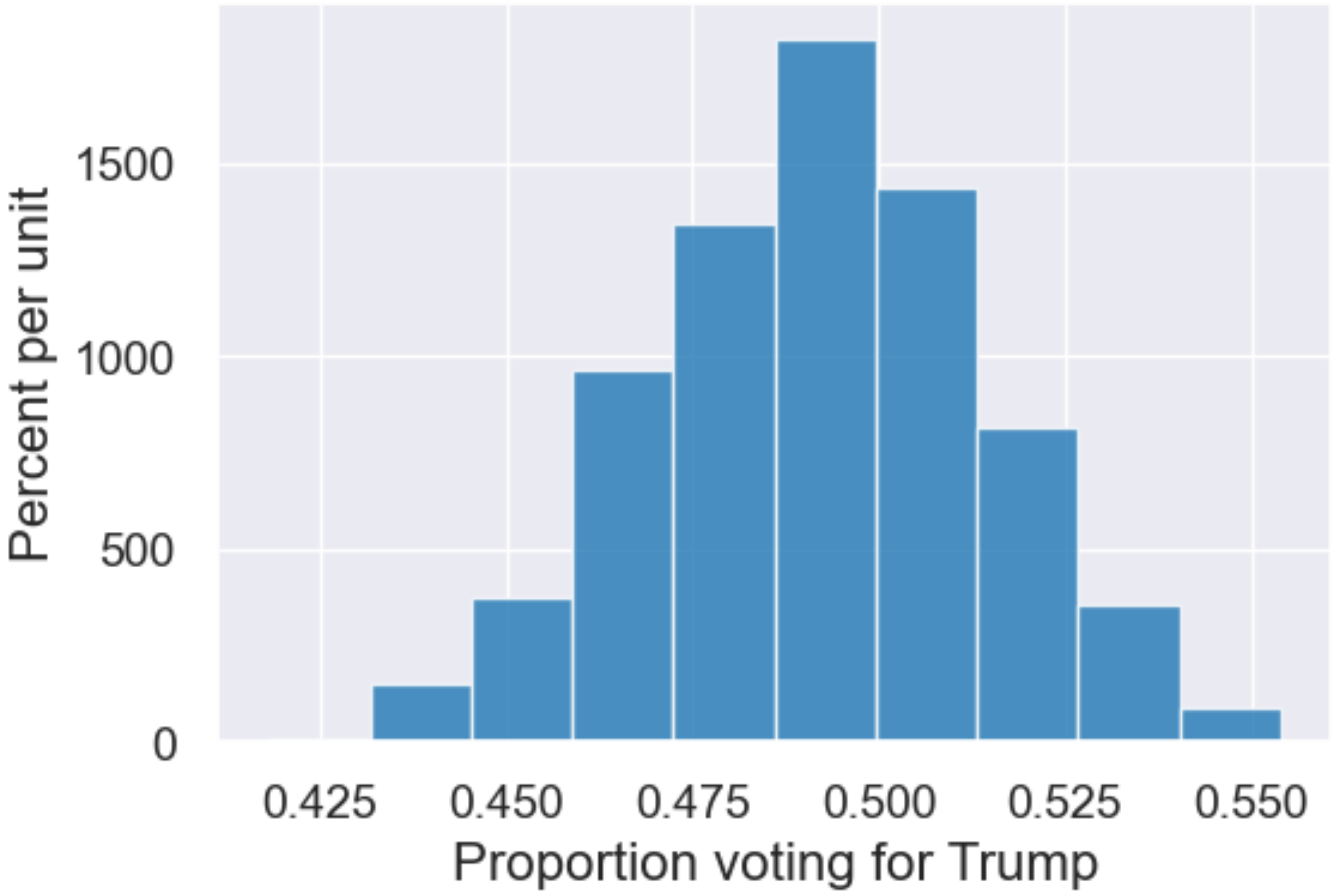
The *sampling distribution* or probability distribution of the statistic consists of all possible values of the statistic and their corresponding probabilities.

This can be hard to calculate!

- Either you need to do the math or you need to generate all possible samples and calculate the statistic based on each sample.



Empirical distribution of a statistic



1. Observe the statistic from repetitions of a (sampling) experiment or simulation.

2. Create a distribution of statistics (i.e., a histogram)

The *empirical distribution of the statistic* is

based on simulated values of the statistic and

consists of

all the observed values of the statistic and

the proportion of times each value appeared.

The empirical distribution is a good approximation to the probability distribution of the statistic – if the number of repetitions in the simulation is large!

A fundamental consideration in using any statistic based on a *random sample* is that

the sample could have come out differently, and

therefore the statistic could have come out differently too!

projects.fivethirtyeight.com/2024-election-forecast/

538 **President** Senate House Latest articles ▼ abc NEWS

538 uses polling, economic and demographic data to explore likely election outcomes.

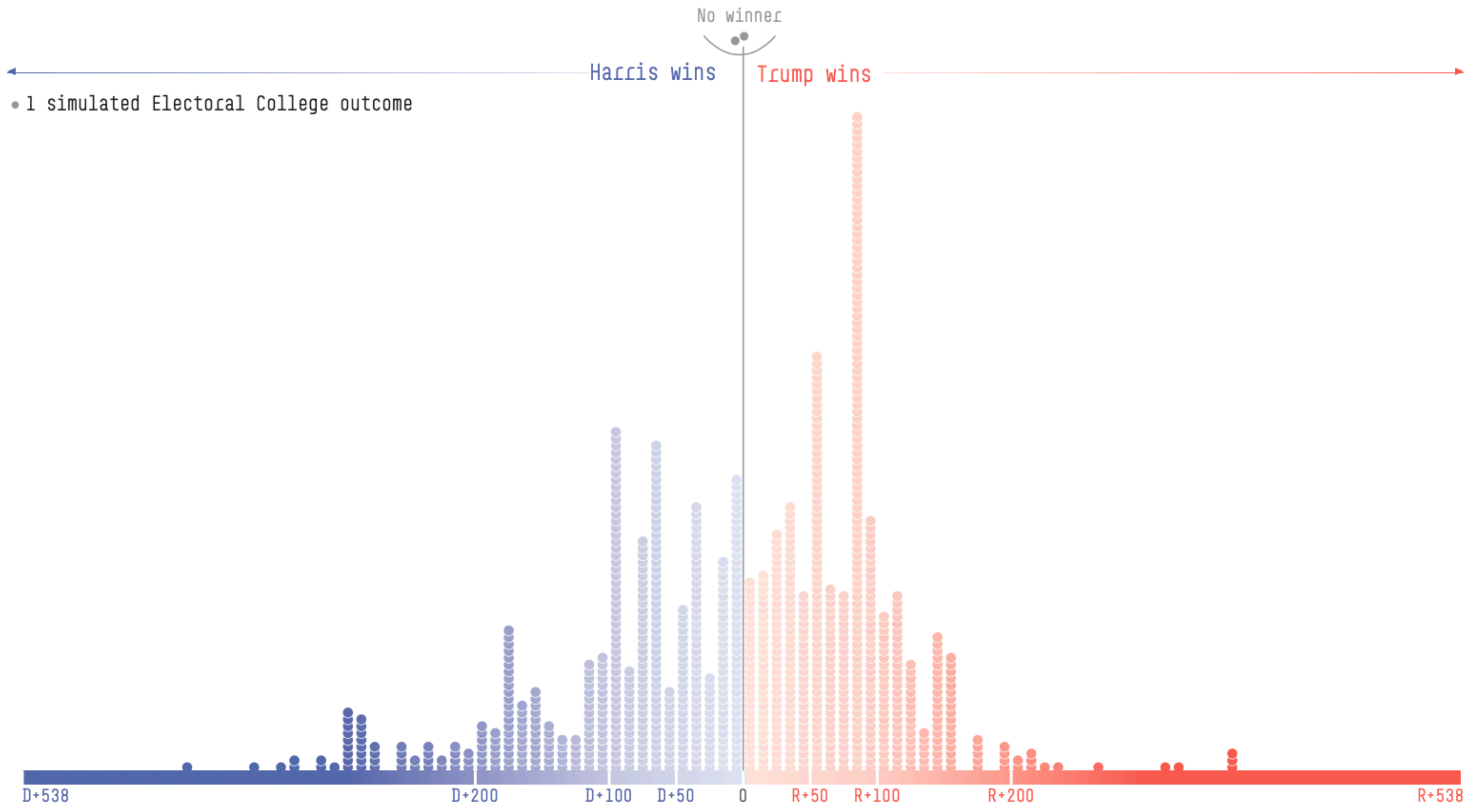
Trump wins 53 times out of 100

in our simulations of the 2024 presidential election.

Harris wins 46 times out of 100.

There is a less than 1-in-100 chance of no Electoral College winner.

Trump	534
Harris	464
No winner	2
	1,000 simulations



ELECTORAL VOTE MARGIN

National overview

GO TO A RANDOM STATE

Electoral College probability | **Latest polls** | Latest updates | Forecast over time | State probabilities | Closest races | Path to 270 | Electoral vs. popular vote | Polls vs. fundamentals | Possible scenarios

Acknowledgments

This class incorporates material from:

Stephen Freund and Katie Keith, Williams College

Data 8, University of California, Berkeley (CC BY-NC-SA)