Predicting the Outcome of an Election

Carlos Fernandez-Granda

www.cims.nyu.edu/~cfgranda

cSplash 2018

Acknowledgements

Support by NSF award DMS-1616340

Election

- Candidate  against candidate 
- New York: 8 million
- Goal: Estimate fraction of people who will vote for 

Experiment

- True fraction is 0.547
- We ask 1000 people at random
- Outcome:

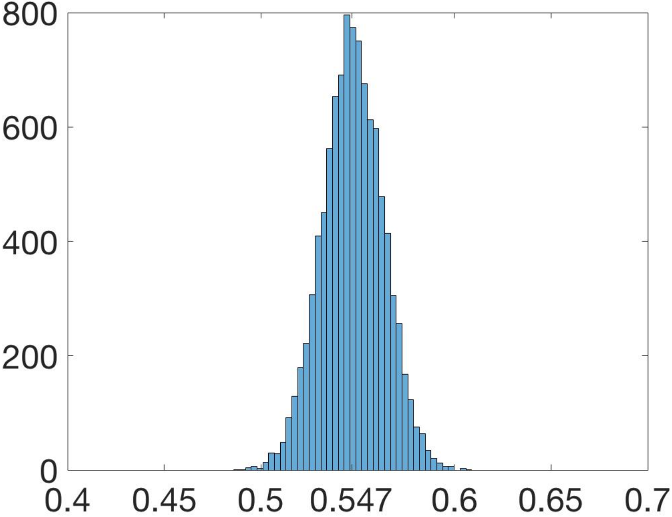
Experiment

- True fraction is 0.547
- We ask 1000 people at random
- Outcome: 545! (estimate: 0.545)

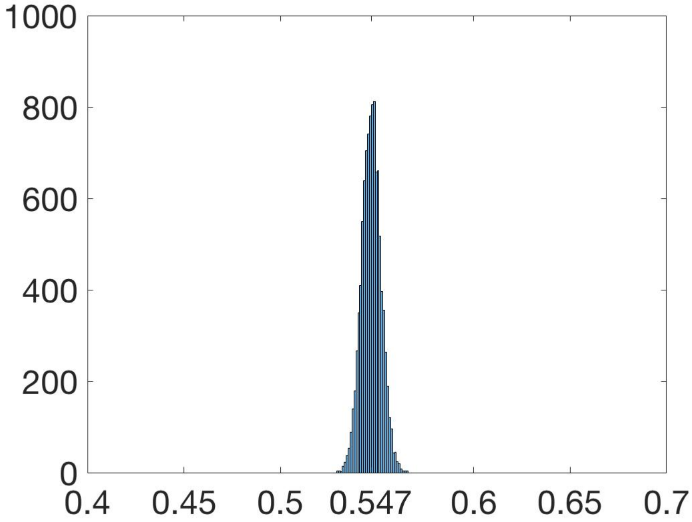
Experiment

- True fraction is 0.547
- We ask 1000 people at random
- Outcome: 545! (estimate: 0.545)
- Did we get lucky?

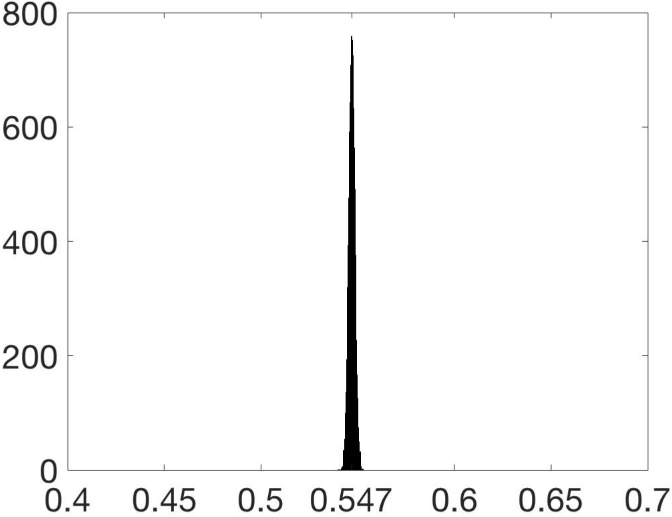
Poll of 1 000 people (repeated 10 000 times)



Poll of 10 000 people (repeated 10 000 times)



Poll of 100 000 people (repeated 10 000 times)



Question to think about

Does the total population (8 million) matter?

Interesting phenomenon

$$\frac{\#\ \text{voters in poll}}{\#\text{people in the poll}} \quad \text{is close to} \quad \frac{\#\ \text{voters}}{\#\ \text{people}}$$

Aim of the talk: Understand why this happens

Not so easy

 voters in poll changes every time: its value is uncertain

We need to reason probabilistically

We need mathematical tools to analyze uncertain quantities

Random variable

Mathematical objects that model uncertain quantities

A random variable X has a set of possible outcomes

Sampling X results in one of those outcomes

Probability

Maps outcomes to a number between 0 and 1

The probability of an outcome quantifies how likely it is

Intuitively

$$\mathrm{P}\left(\text{outcome } i\right) = \frac{\#\text{samples equal to outcome } i}{\#\text{ samples}}$$

when the number of samples is very large

Events

We can group outcomes in sets called events

An event occurs if we sample an outcome belonging to the event

$X \in \{0, 1\}$, $Y \leq 10$, $Z \geq 1.2$

The probability of an event quantifies how likely it is

Probability

Intuitively

$$\mathrm{P}\left(\text{event}\right) = \frac{\#\text{times event happens}}{\#\text{ samples}}$$

when the number of samples is very large

$$\mathrm{P}\left(X \in \{0, 3\}\right) \approx \frac{\#\text{ samples equal to 0 or 3}}{\#\text{ samples of } X}$$

Properties of probability

Probability is nonnegative, like mass or length

Properties of probability

If events can't happen simultaneously, we can add their probabilities

$$\mathrm{P}(X \in \{0, 4, 7\}) = \mathrm{P}(X = 0) + \mathrm{P}(X \in \{4, 7\})$$

Properties of probability

If events can't happen simultaneously, we can add their probabilities

$$\mathrm{P}(X \in \{0,4,7\}) = \mathrm{P}(X = 0) + \mathrm{P}(X \in \{4,7\})$$

Makes sense:

$$\mathrm{P}(X \in \{0,4,7\}) \approx \frac{\#\text{ samples equal to 0, 4 or 7}}{\#\text{ samples of } X}$$

Properties of probability

If events can't happen simultaneously, we can add their probabilities

$$\mathrm{P}(X \in \{0,4,7\}) = \mathrm{P}(X = 0) + \mathrm{P}(X \in \{4,7\})$$

Makes sense:

$$\mathrm{P}(X \in \{0,4,7\}) \approx \frac{\#\text{ samples equal to 0, 4 or 7}}{\#\text{ samples of } X}$$
$$= \frac{\#\text{ samples equal to 0 } + \#\text{ samples equal to 4 or 7}}{\#\text{ samples of } X}$$

Properties of probability

If events can't happen simultaneously, we can add their probabilities

$$\mathrm{P}\left(X \in \{0,4,7\}\right) = \mathrm{P}\left(X = 0\right) + \mathrm{P}\left(X \in \{4,7\}\right)$$

Makes sense:

$$\begin{aligned}
\mathrm{P}\left(X \in \{0,4,7\}\right) &\approx \frac{\#\text{ samples equal to 0, 4 or 7}}{\#\text{ samples of } X} \\
&= \frac{\#\text{ samples equal to 0 } + \#\text{ samples equal to 4 or 7}}{\#\text{ samples of } X} \\
&\approx \mathrm{P}\left(X = 0\right) + \mathrm{P}\left(X \in \{4,7\}\right)
\end{aligned}$$

Properties of probability

If events can't happen simultaneously, we can add their probabilities

$$\mathrm{P}(X \in \{0,4,7\}) = \mathrm{P}(X = 0) + \mathrm{P}(X \in \{4,7\})$$

Makes sense:

$$\begin{aligned}
\mathrm{P}(X \in \{0,4,7\}) &\approx \frac{\#\text{ samples equal to 0, 4 or 7}}{\#\text{ samples of } X} \\
&= \frac{\#\text{ samples equal to 0 } + \#\text{ samples equal to 4 or 7}}{\#\text{ samples of } X} \\
&\approx \mathrm{P}(X = 0) + \mathrm{P}(X \in \{4,7\})
\end{aligned}$$

Also like mass or length

Properties of probability

The probability of all events that can't happen simultaneously adds to one

$$\sum_{i=1}^{m} \mathrm{P}\left(X = o_i\right) = 1$$

where $\{o_1, \ldots, o_m\}$ are the possible outcomes of X

Not like mass or length!

Properties of probability

Makes sense:

$$\sum_{i=1}^{m} \mathrm{P}\left(X = o_i\right)$$

$$\approx \sum_{i=1}^{m} \frac{\#\text{ samples equal to } o_i}{\#\text{ samples of } X}$$

Properties of probability

Makes sense:

$$\sum_{i=1}^{m} \mathrm{P}\left(X = o_i\right)$$

$$\approx \sum_{i=1}^{m} \frac{\#\text{ samples equal to } o_i}{\#\text{ samples of } X}$$

$$= \frac{\#\text{ samples equal to } o_1 + \#\text{ samples equal to } o_2 + \cdots + \#\text{ samples equal to } o_m}{\#\text{ samples of } X}$$

Properties of probability

Makes sense:

$$
\begin{aligned}
&\sum_{i=1}^{m} \mathrm{P}\left(X = o_i\right) \\
&\approx \sum_{i=1}^{m} \frac{\#\text{ samples equal to } o_i}{\#\text{ samples of } X} \\
&= \frac{\#\text{ samples equal to } o_1 + \#\text{ samples equal to } o_2 + \cdots + \#\text{ samples equal to } o_m}{\#\text{ samples of } X} \\
&= \frac{\#\text{ samples of } X}{\#\text{ samples of } X} = 1
\end{aligned}
$$

Modeling a coin flip

When we flip a coin, it lands heads a fraction p of the time

- Possible outcomes?
- Probability of outcomes?

Modeling a coin flip

When we flip a coin, it lands heads a fraction p of the time

- Possible outcomes? 0 (tails) or 1 (heads)
- Probability of outcomes?

$$\mathrm{P}(X = 1) =$$

$$\mathrm{P}(X = 0) =$$

Modeling a coin flip

When we flip a coin, it lands heads a fraction p of the time

- Possible outcomes? 0 (tails) or 1 (heads)
- Probability of outcomes?

$$\mathrm{P}\left(X = 1\right) = p$$

$$\mathrm{P}\left(X = 0\right) =$$

Modeling a coin flip

When we flip a coin, it lands heads a fraction p of the time

- Possible outcomes? 0 (tails) or 1 (heads)
- Probability of outcomes?

$$\mathrm{P}\left(X = 1\right) = p$$

$$\mathrm{P}\left(X = 0\right) = 1 - p$$

Election

We poll a voter at random from a population of T people

Chosen voter is a random variable X

Possible outcomes? 1, 2, . . . , T

Assumption: We are equally likely to pick any voter

Probability that we pick a specific person?

Election

All possible outcomes must sum to one

$$\sum_{i=1}^{T} \mathrm{P}\left(X = i\right) = 1$$

and

$$\mathrm{P}\left(X = 1\right) = \mathrm{P}\left(X = 2\right) = \cdots = \mathrm{P}\left(X = T\right)$$

Election

All possible outcomes must sum to one

$$\sum_{i=1}^{T} \mathrm{P}\left(X = i\right) = 1$$

and

$$\mathrm{P}\left(X = 1\right) = \mathrm{P}\left(X = 2\right) = \cdots = \mathrm{P}\left(X = T\right) = \frac{1}{T}$$

Election

Pick a voter at random from T people from which $\#$ are voters

New random variable

$V = 1$ if voter is voter

otherwise $V = 0$

Probability that we pick a voter? $\mathrm{P}\left(V = 1\right)$

Election

Let's order the voters, first $\#$  are  voters

$\mathrm{P}\left(V = 1\right) = \mathrm{P}\left(X \in \{1, \ldots, \#$  $\}\right)$

Election

Let's order the voters, first # are voters

$$
\begin{aligned}
\mathrm{P}(V = 1) &= \mathrm{P}(X \in \{1, \ldots, \#\}) \\
&= \sum_{i=1}^{\#} \mathrm{P}(X = i)
\end{aligned}
$$

Election

Let's order the voters, first # are voters

$$\mathrm{P}\left(V=1\right)=\mathrm{P}\left(X\in\{1,\ldots,\#\}\right)$$
$$=\sum_{i=1}^{\#}\mathrm{P}\left(X=i\right)$$
$$=\frac{\#}{T}$$

Just like coin flip with $p =$

Election

Let's order the voters, first # are voters

$$
\begin{aligned}
\mathrm{P}(V = 1) &= \mathrm{P}(X \in \{1, \ldots, \#\}) \\
&= \sum_{i=1}^{\#} \mathrm{P}(X = i) \\
&= \frac{\#}{T}
\end{aligned}
$$

Just like coin flip with $p = \# / T$

Election

If $T = 8$ million and the fraction of  voters is 0.547

What is the probability that we choose a  voter?

Does this depend on T?

Multiple random variables

We can consider several random variables at the same time

Every time we sample, we sample all the random variables

Events can include any of the random variables

$\{X = 0 \text{ and } Y \leq 10\}$, $\{Z = 1.2 \text{ or } W \in \{10, 21\}\}$

Probability

The probability of the event still quantifies how likely it is

Same intuition

$$\mathrm{P}\left(\text{event}\right) = \frac{\#\text{times event happens}}{\#\text{ samples}}$$

when the number of samples is very large

$$\mathrm{P}\left(X = 0 \text{ and } Y \leq 10\right) \approx \frac{\#\text{ samples for which } X = 0 \text{ and } Y \leq 10}{\#\text{ samples of } (X, Y)}$$

Conditional probability

If we know an event $\mathcal{B}$ (for example $Y \leq 10$)

How likely is that another event $\mathcal{A}$ (for example $X = 0$) also happened?

$\mathrm{P}(\mathcal{B} \mid \mathcal{A})$, the conditional probability of $\mathcal{B}$ given $\mathcal{A}$

Conditional probability

Intuition

$$\mathrm{P}\left(\text{event } \mathcal{B} \,|\, \text{event } \mathcal{A}\right) = \frac{\#\text{samples for which } \mathcal{A} \text{ and } \mathcal{B} \text{ happen}}{\#\text{ samples for which } \mathcal{A} \text{ happens}}$$

when the number of samples is very large

$$\mathrm{P}\left(X = 0 \,|\, Y \leq 10\right) \approx \frac{\#\text{ samples for which } X = 0 \text{ and } Y \leq 10}{\#\text{ samples for which } Y \leq 10}$$

Chain rule

$$\mathrm{P}(\mathcal{A} \text{ and } \mathcal{B}) = \mathrm{P}(\mathcal{A})\,\mathrm{P}(\mathcal{B} \,|\, \mathcal{A})$$

Makes sense:

$$\mathrm{P}(\mathcal{A})\,\mathrm{P}(\mathcal{B} \,|\, \mathcal{A}) \approx \frac{\#\ \mathcal{A} \text{ happens}}{\#\ \text{samples}} \cdot \frac{\#\ \mathcal{A} \text{ and } \mathcal{B} \text{ happen}}{\#\ \mathcal{A} \text{ happens}}$$

Chain rule

$$\mathrm{P}(\mathcal{A} \text{ and } \mathcal{B}) = \mathrm{P}(\mathcal{A})\,\mathrm{P}(\mathcal{B} \mid \mathcal{A})$$

Makes sense:

$$\begin{aligned}
\mathrm{P}(\mathcal{A})\,\mathrm{P}(\mathcal{B} \mid \mathcal{A}) &\approx \frac{\#\ \mathcal{A} \text{ happens}}{\#\ \text{samples}} \cdot \frac{\#\ \mathcal{A} \text{ and } \mathcal{B} \text{ happen}}{\#\ \mathcal{A} \text{ happens}} \\
&= \frac{\#\ \mathcal{A} \text{ and } \mathcal{B} \text{ happen}}{\#\ \text{samples}} \\
&\approx \mathrm{P}(\mathcal{A} \text{ and } \mathcal{B})
\end{aligned}$$

Independence

If knowing that $\mathcal{A}$ happened does not affect how likely $\mathcal{B}$ is

$\mathcal{A}$ and $\mathcal{B}$ are independent

$$\mathrm{P}(\mathcal{B} \,|\, \mathcal{A}) = \mathrm{P}(\mathcal{B})$$

In that case

$$\mathrm{P}(\mathcal{A} \text{ and } \mathcal{B}) = \mathrm{P}(\mathcal{A})\,\mathrm{P}(\mathcal{B} \,|\, \mathcal{A}) = \mathrm{P}(\mathcal{A})\,\mathrm{P}(\mathcal{B})$$

Modeling coin flips

When we flip a coin, it lands heads a fraction p of the time

If we flip it n times, what is the probability that k flips are heads?

Assumptions:

- Probability of heads in a single flip is p
- Flips are independent

Modeling coin flips

We model the problem using random variables:

$V_i = 1$ if ith flip is heads, 0 if it isn't

$V_1, V_2, \ldots, V_n$ are independent

$S = V_1 + \cdots + V_n$, number of heads

Example

3 flips, probability of 2 heads?

$$\begin{aligned}\mathrm{P}(S = 2) = \mathrm{P}(\ &\{V_1 = 1, V_2 = 1, V_3 = 0\} \\ &\text{or } \{V_1 = 1, V_2 = 0, V_3 = 1\} \\ &\text{or } \{V_1 = 0, V_2 = 1, V_3 = 1\})\end{aligned}$$

Example

$$\begin{aligned}\mathrm{P}\left(V_1 = 1 \text{ and } V_2 = 1 \text{ and } V_3 = 0\right) &= \mathrm{P}\left(V_1 = 1\right)\mathrm{P}\left(V_2 = 1\right)\mathrm{P}\left(V_3 = 0\right) \\ &= p^2\left(1 - p\right)\end{aligned}$$

Example

$$\begin{aligned}\mathrm{P}(V_1 = 1 \text{ and } V_2 = 1 \text{ and } V_3 = 0) &= \mathrm{P}(V_1 = 1)\,\mathrm{P}(V_2 = 1)\,\mathrm{P}(V_3 = 0) \\ &= p^2(1-p)\end{aligned}$$

$\mathrm{P}(V_1 = 0 \text{ and } V_2 = 1 \text{ and } V_3 = 1)$?

Example

$$\begin{aligned}\mathrm{P}\left(V_1 = 1 \text{ and } V_2 = 1 \text{ and } V_3 = 0\right) &= \mathrm{P}\left(V_1 = 1\right)\mathrm{P}\left(V_2 = 1\right)\mathrm{P}\left(V_3 = 0\right) \\ &= p^2\left(1 - p\right)\end{aligned}$$

$\mathrm{P}\left(V_1 = 0 \text{ and } V_2 = 1 \text{ and } V_3 = 1\right)$? Does not depend on order!

Example

$$\mathrm{P}(V_1 = 1 \text{ and } V_2 = 1 \text{ and } V_3 = 0) = \mathrm{P}(V_1 = 1)\,\mathrm{P}(V_2 = 1)\,\mathrm{P}(V_3 = 0)$$
$$= p^2(1-p)$$

$\mathrm{P}(V_1 = 0 \text{ and } V_2 = 1 \text{ and } V_3 = 1)$? Does not depend on order!

In general, k heads and $n-k$ tails in fixed order

$$\mathrm{P}(k \text{ heads and } n-k \text{ tails}) = p^k(1-p)^{n-k}$$

Example

3 flips, probability of 2 heads?

$$
\begin{aligned}
\mathrm{P}(S = 2) = \mathrm{P}(\ &\{V_1 = 1, V_2 = 1, V_3 = 0\} \\
&\text{or } \{V_1 = 1, V_2 = 0, V_3 = 1\} \\
&\text{or } \{V_1 = 0, V_2 = 1, V_3 = 1\}) \\
=& 3p^2(1 - p)
\end{aligned}
$$

Modeling coin flips

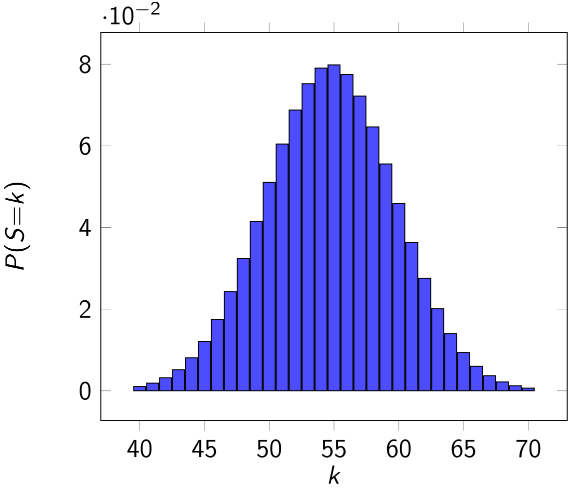
In general

$$
\begin{aligned}
\mathrm{P}(S=k) &= \#\text{ combinations of } k \text{ heads in } n \text{ flips } \cdot \mathrm{P}(k \text{ heads in fixed order}) \\
&= \binom{n}{k} p^k (1-p)^{n-k}
\end{aligned}
$$

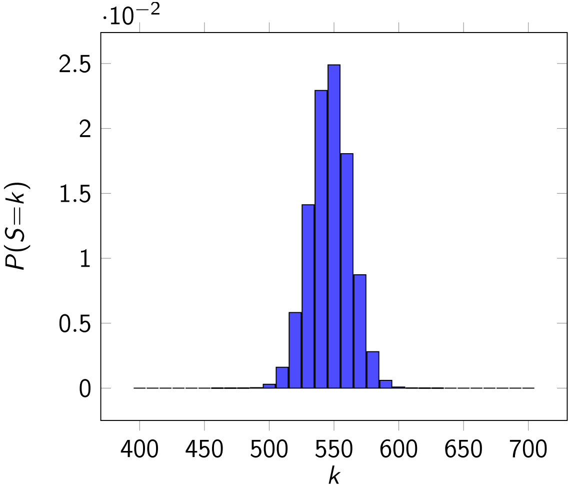
$$
\binom{n}{k} := \frac{n!}{k!\,(n-k)!}
$$

S is a binomial random variable with parameters n and p

Binomial random variable $n = 100$, $p = 0.547$



Binomial random variable $n = 1000$, $p = 0.547$



Election

Population of T people with #   voters

We poll n voters at random

How many are  voters? Random variable S

Assumption 1: We are equally likely to pick any voter every time we poll

Assumption 2: Picks are all independent

What kind of random variable is S?

Election

S is binomial with parameters n and $\#$  $/T$

But we are interested in fraction of  voters S/n !

$$\mathrm{P}\left(\frac{S}{n} = \frac{k}{n}\right) =$$

Election

S is binomial with parameters n and $\#$ $/T$

But we are interested in fraction of voters S/n !

$$\mathrm{P}\left(\frac{S}{n} = \frac{k}{n}\right) = \mathrm{P}\left(S = k\right)$$

Election

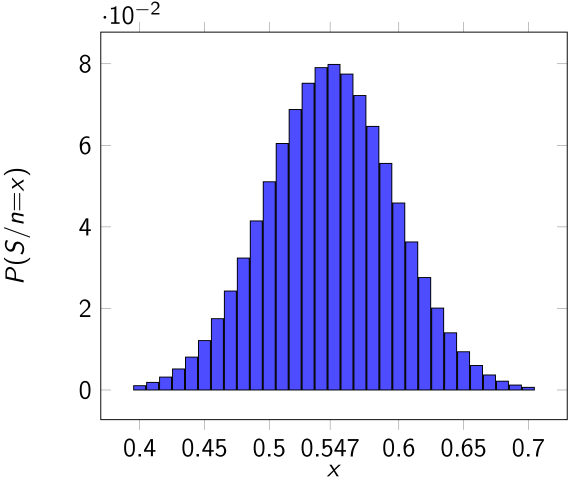
S is binomial with parameters n and # $/T$

But we are interested in fraction of  voters S/n !

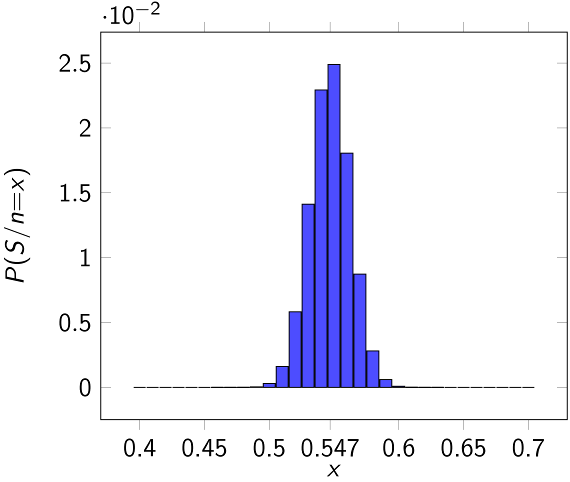
$$\mathrm{P}\left(\frac{S}{n} = \frac{k}{n}\right) = \mathrm{P}\left(S = k\right)$$
$$= \binom{n}{k} p^k \left(1 - p\right)^{n-k}$$

We can compute exact probability of poll outcome for any n!

Fraction of 🥷 voters in poll, $n = 100$, $\#$ 🥷$/T = 0.547$



Binomial random variable $n = 1000$, # 🥷/ $T = 0.547$



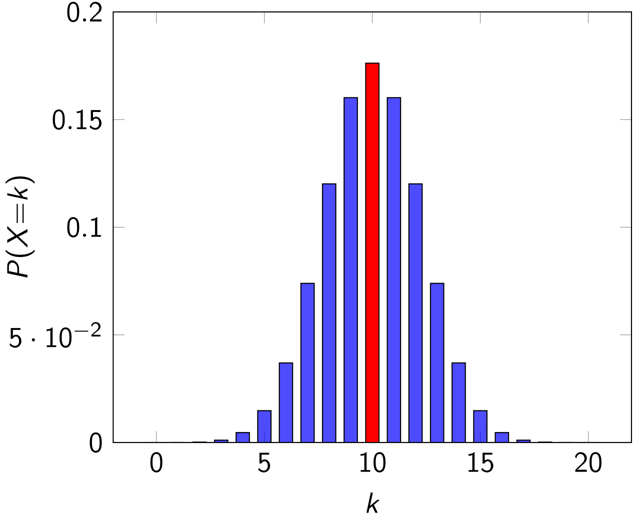
Mean

Average of a random variable

$$\begin{aligned} \mathrm{E}(X) &= \sum_{i=1} o_i \mathrm{P}(X = o_i) \\ &= o_1 \mathrm{P}(X = o_1) + o_2 \mathrm{P}(X = o_2) + \cdots + o_m \mathrm{P}(X = o_m) \end{aligned}$$

where $\{o_1, \ldots, o_m\}$ are the possible outcomes of X

The mean is the center of mass of the probabilities



Election

Population of T people with #   voters

We poll n voters at random

$S = \#$ of  voters in poll

What is the average estimate S/n?

Modeling a coin flip

When we flip a coin, it lands heads a fraction p of the time

Random variable X equal to 0 (tails) or 1 (heads)

Mean

$$\mathrm{E}(X) =$$

Modeling a coin flip

When we flip a coin, it lands heads a fraction p of the time

Random variable X equal to 0 (tails) or 1 (heads)

Mean

$$\mathrm{E}(X) = 0 \cdot \mathrm{P}(X = 0) + 1 \cdot \mathrm{P}(X = 1)$$

Modeling a coin flip

When we flip a coin, it lands heads a fraction p of the time

Random variable X equal to 0 (tails) or 1 (heads)

Mean

$$\begin{aligned} \mathrm{E}(X) &= 0 \cdot \mathrm{P}(X = 0) + 1 \cdot \mathrm{P}(X = 1) \\ &= p \end{aligned}$$

Mean of sum

The sum of the averages is the average of the sums

$$\mathrm{E}(X + Y) =$$

$$= \mathrm{E}(X) + \mathrm{E}(Y)$$

Mean of sum

The sum of the averages is the average of the sums

$$
\begin{aligned}
\mathrm{E}(X+Y) &= \sum_{i=1}^{m}\sum_{j=1}^{m}\left(o_i + o'_j\right)\mathrm{P}\left(X = o_i, Y = o'_j\right) \\
&= \sum_{i=1}^{m}\sum_{j=1}^{m} o_i \mathrm{P}\left(X = o_i, Y = o'_j\right) + \sum_{i=1}^{m}\sum_{j=1}^{m} o'_j \mathrm{P}\left(X = o_i, Y = o'_j\right) \\
&= \sum_{i=1}^{m} o_i \sum_{j=1}^{m}\mathrm{P}\left(X = o_i, Y = o'_j\right) + \sum_{j=1}^{m} o'_j \sum_{i=1}^{m}\mathrm{P}\left(X = o_i, Y = o'_j\right) \\
&= \sum_{i=1}^{m} o_i \mathrm{P}\left(X = o_i\right) + \sum_{j=1}^{m} o'_j \mathrm{P}\left(Y = o'_j\right) \\
&= \mathrm{E}(X) + \mathrm{E}(Y)
\end{aligned}
$$

Election

S is a sum of n coin flips with $p = \#$  $/ T$

$$S = \sum_{i=1}^{n} V_i$$

Election

S is a sum of n coin flips with $p = \#$  $/\, T$

$$S = \sum_{i=1}^{n} V_i$$

$$\mathrm{E}\left(S\right) = \mathrm{E}\left(\sum_{i=1}^{n} V_i\right)$$

Election

S is a sum of n coin flips with $p = \#$  $/ T$

$$S = \sum_{i=1}^{n} V_i$$

$$\begin{aligned} \mathrm{E}(S) &= \mathrm{E}\left(\sum_{i=1}^{n} V_i\right) \\ &= \sum_{i=1}^{n} \mathrm{E}(V_i) \end{aligned}$$

Election

S is a sum of n coin flips with $p = \#$  $/ T$

$$S = \sum_{i=1}^{n} V_i$$

$$\mathrm{E}(S) = \mathrm{E}\left(\sum_{i=1}^{n} V_i\right)$$

$$= \sum_{i=1}^{n} \mathrm{E}(V_i)$$

$$= \frac{n \, \# \, \text{(icon)}}{T}$$

What about $\frac{S}{n}$?

Expectation of scaled random variable

The average of a scaled random variable is just the scaled average

$$\mathrm{E}(c\,X) =$$

Expectation of scaled random variable

The average of a scaled random variable is just the scaled average

$$\mathrm{E}\left(c\,X\right) = \sum_{i=1}^{m} c\,o_i \mathrm{P}\left(X = o_i\right)$$

Expectation of scaled random variable

The average of a scaled random variable is just the scaled average

$$\begin{aligned} \mathrm{E}(c\,X) &= \sum_{i=1}^{m} c\,o_i \mathrm{P}(X = o_i) \\ &= c \sum_{i=1}^{m} o_i \mathrm{P}(X = o_i) \end{aligned}$$

Expectation of scaled random variable

The average of a scaled random variable is just the scaled average

$$\begin{aligned}
\mathrm{E}\left(c\,X\right) &= \sum_{i=1}^{m} c\,o_i \mathrm{P}\left(X = o_i\right) \\
&= c \sum_{i=1}^{m} o_i \mathrm{P}\left(X = o_i\right) \\
&= c\,\mathrm{E}\left(X\right)
\end{aligned}$$

Election

Population of T people with #   voters

We poll n voters at random

S = # of  voters

What is the average estimate S/n?

$\mathrm{E}(S) = $ (n # ) / T

$$\mathrm{E}(S/n) =$$

Election

Population of T people with $\#$ voters

We poll n voters at random

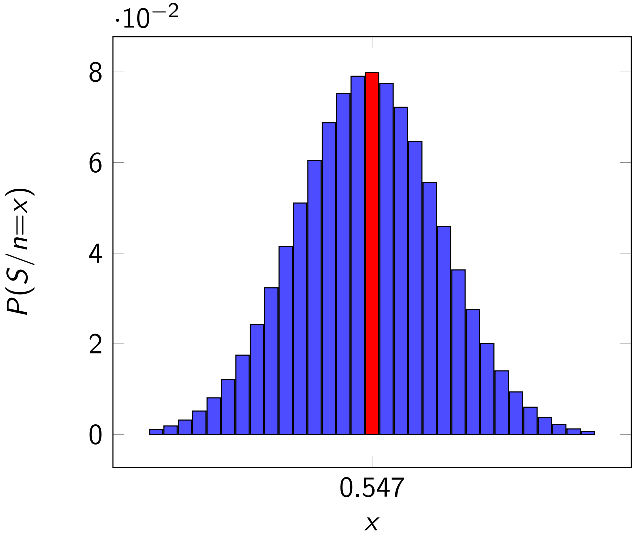
$S = \#$ of voters

What is the average estimate S/n?

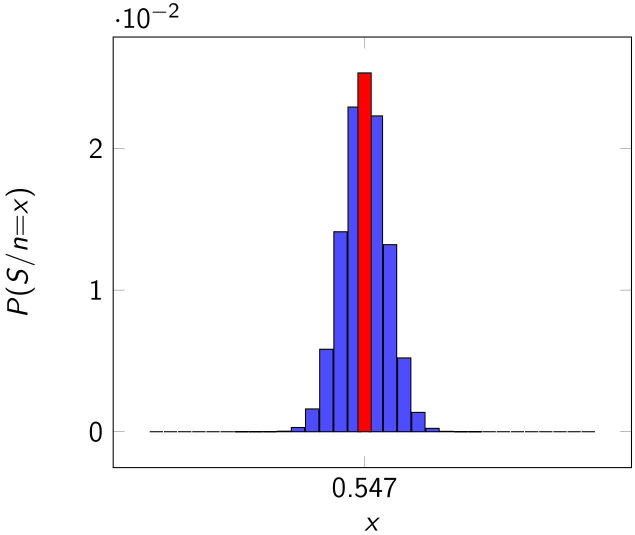
$$\mathrm{E}\left(S\right) = \frac{n\,\#}{T}$$

$$\mathrm{E}\left(S/n\right) = \frac{\#}{T} \quad (!)$$

Fraction of voters in poll, $n = 100$, # $/T = 0.547$



Fraction of 🥷 voters in poll $n = 1000$, # 🥷$/T = 0.547$



Variance

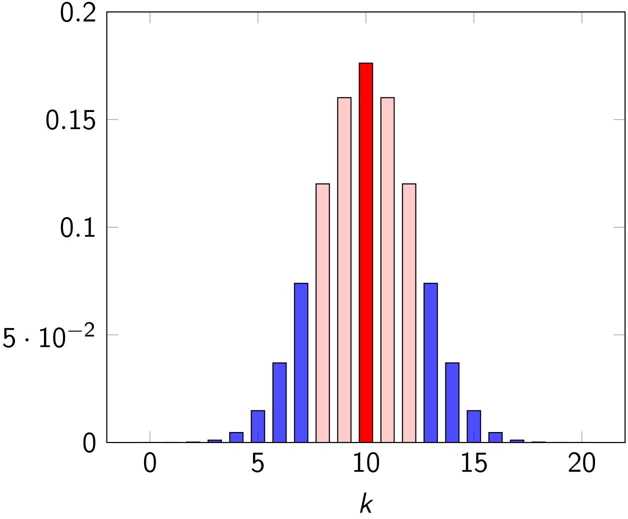
Square deviation from $D := (X - \mathrm{E}(X))^2$

The variance is the mean of the square deviation

$$\mathrm{Var}(X) = \mathrm{E}(Y)$$

Standard deviation $\sigma_X := \sqrt{\mathrm{Var}(X)}$

Standard deviation ≈ average variation around mean



Election

Population of T people with $\#$ voters

We poll n voters at random

On average

$$\mathrm{E}(S/n) = \frac{\#}{T}$$

Standard deviation $\sigma_{S/n}$ quantifies average deviation from truth

Key question: Does $\sigma_{S/n}$ decrease as n grows?

Coin flip

$$\mathrm{Var}\left(X\right) = \mathrm{E}\left(\left(X - \mathrm{E}\left(X\right)\right)^2\right)$$

Coin flip

$$\begin{aligned}\mathrm{Var}(X) &= \mathrm{E}\left((X-\mathrm{E}(X))^2\right) \\ &= \sum \text{possible values of } (X-\mathrm{E}(X))^2 \cdot \text{corresponding probability}\end{aligned}$$

Coin flip

$$
\begin{aligned}
\mathrm{Var}(X) &= \mathrm{E}\left((X - \mathrm{E}(X))^2\right) \\
&= \sum \text{possible values of } (X - \mathrm{E}(X))^2 \cdot \text{corresponding probability} \\
&= (1-p)^2\, p + p^2 (1-p)
\end{aligned}
$$

Coin flip

$$
\begin{aligned}
\operatorname{Var}(X) &= \operatorname{E}\left((X - \operatorname{E}(X))^2\right) \\
&= \sum \text{possible values of } (X - \operatorname{E}(X))^2 \cdot \text{corresponding probability} \\
&= (1-p)^2\, p + p^2\,(1-p) \\
&= (1-p)\, p
\end{aligned}
$$

Variance of a sum

$$\mathrm{Var}\,(X + Y) =$$

$$= \mathrm{Var}\,(X) + \mathrm{Var}\,(Y) + 2\mathrm{E}\,(XY) - 2\mathrm{E}\,(X)\,\mathrm{E}\,(Y)$$

Variance of a sum

$$
\begin{aligned}
\mathrm{Var}\left(X+Y\right) &= \mathrm{E}\left(\left(X+Y-\mathrm{E}\left(X+Y\right)\right)^2\right) \\
&= \mathrm{E}\left(\left(X-\mathrm{E}\left(X\right)\right)^2\right) + \mathrm{E}\left(\left(Y-\mathrm{E}\left(Y\right)\right)^2\right) \\
&\quad + 2\mathrm{E}\left(\left(X-\mathrm{E}\left(X\right)\right)\left(Y-\mathrm{E}\left(Y\right)\right)\right) \\
&= \mathrm{E}\left(\left(X-\mathrm{E}\left(X\right)\right)^2\right) + \mathrm{E}\left(\left(Y-\mathrm{E}\left(Y\right)\right)^2\right) \\
&\quad + 2\mathrm{E}\left(XY\right) - 2\mathrm{E}\left(X\right)\mathrm{E}\left(Y\right) \\
&= \mathrm{Var}\left(X\right) + \mathrm{Var}\left(Y\right) + 2\mathrm{E}\left(XY\right) - 2\mathrm{E}\left(X\right)\mathrm{E}\left(Y\right)
\end{aligned}
$$

Variance of a sum

If X and Y are independent, then

$$\mathrm{E}(XY) =$$

Variance of a sum

If X and Y are independent, then

$$\mathrm{E}(XY) = \sum_{i=1}^{m} \sum_{j=1}^{m} o_i o'_j \mathrm{P}\left(X = o_i, Y = o'_j\right)$$

Variance of $X + Y$ if $Y := -X$

Variance of a sum

If X and Y are independent, then

$$\mathrm{E}(XY) = \sum_{i=1}^{m} \sum_{j=1}^{m} o_i o_j' \mathrm{P}\left(X = o_i, Y = o_j'\right)$$
$$= \sum_{i=1}^{m} \sum_{j=1}^{m} o_i o_j' \mathrm{P}(X = o_i)\, \mathrm{P}\left(Y = o_j'\right)$$

Variance of $X + Y$ if $Y := -X$

Not true if not independent!

Variance of a sum

If X and Y are independent, then

$$\begin{aligned} \mathrm{E}(XY) &= \sum_{i=1}^{m} \sum_{j=1}^{m} o_i o'_j \mathrm{P}\left(X = o_i, Y = o'_j\right) \\ &= \sum_{i=1}^{m} \sum_{j=1}^{m} o_i o'_j \mathrm{P}(X = o_i)\, \mathrm{P}\left(Y = o'_j\right) \\ &= \mathrm{E}(X)\, \mathrm{E}(Y) \end{aligned}$$

Variance of $X + Y$ if $Y := -X$

Not true if not independent!

Election

S is a sum of n coin flips with $p = \#$  $/\, T$

$$S = \sum_{i=1}^{n} V_i$$

$$\mathrm{Var}\left(S\right) =$$

Election

S is a sum of n coin flips with $p = \#$  $/ T$

$$S = \sum_{i=1}^{n} V_i$$

$$\mathrm{Var}\left(S\right) = \mathrm{Var}\left(\sum_{i=1}^{n} V_i\right)$$

Election

S is a sum of n coin flips with $p = \#$  $/T$

$$S = \sum_{i=1}^{n} V_i$$

$$\begin{aligned}\mathrm{Var}(S) &= \mathrm{Var}\left(\sum_{i=1}^{n} V_i\right) \\ &= \sum_{i=1}^{n} \mathrm{Var}(V_i)\end{aligned}$$

Election

S is a sum of n coin flips with $p = \#$  $/T$

$$S = \sum_{i=1}^{n} V_i$$

$$\begin{aligned}
\mathrm{Var}\left(S\right) &= \mathrm{Var}\left(\sum_{i=1}^{n} V_i\right) \\
&= \sum_{i=1}^{n} \mathrm{Var}\left(V_i\right) \\
&= np\left(1 - p\right)
\end{aligned}$$

Election

S is a sum of n coin flips with $p = \#$  $/ T$

$$S = \sum_{i=1}^{n} V_i$$

$$\begin{aligned}
\mathrm{Var}\left(S\right) &= \mathrm{Var}\left(\sum_{i=1}^{n} V_i\right) \\
&= \sum_{i=1}^{n} \mathrm{Var}\left(V_i\right) \\
&= np\left(1 - p\right)
\end{aligned}$$

What about S/n ?

Variance

$$\mathrm{Var}\left(c\,X\right) = \mathrm{E}\left(\left(c\,X - \mathrm{E}\left(c\,X\right)\right)^2\right)$$

Variance

$$\begin{aligned}\mathrm{Var}\left(c\,X\right) &= \mathrm{E}\left(\left(c\,X - \mathrm{E}\left(c\,X\right)\right)^2\right) \\ &= \mathrm{E}\left(c^2\left(X - \mathrm{E}\left(X\right)\right)^2\right)\end{aligned}$$

Variance

$$\begin{aligned}
\mathrm{Var}(c\,X) &= \mathrm{E}\left((c\,X - \mathrm{E}(c\,X))^2\right) \\
&= \mathrm{E}\left(c^2\,(X - \mathrm{E}(X))^2\right) \\
&= c^2\mathrm{E}\left((X - \mathrm{E}(X))^2\right)
\end{aligned}$$

Variance

$$\begin{aligned}
\operatorname{Var}(c\,X) &= \operatorname{E}\left((c\,X - \operatorname{E}(c\,X))^2\right) \\
&= \operatorname{E}\left(c^2\,(X - \operatorname{E}(X))^2\right) \\
&= c^2\operatorname{E}\left((X - \operatorname{E}(X))^2\right) \\
&= c^2\,\operatorname{Var}(X)
\end{aligned}$$

Election

S is a sum of n coin flips with $p = \#$  $/T$

$$S = \sum_{i=1}^{n} V_i$$

$$\mathrm{Var}\left(\frac{S}{n}\right)$$

Election

S is a sum of n coin flips with $p = \#$  $/ T$

$$S = \sum_{i=1}^{n} V_i$$

$$\mathrm{Var}\left(\frac{S}{n}\right) = \frac{1}{n^2}\mathrm{Var}\left(S\right)$$

Election

S is a sum of n coin flips with $p = \#$  $/T$

$$S = \sum_{i=1}^{n} V_i$$

$$\mathrm{Var}\left(\frac{S}{n}\right) = \frac{1}{n^2}\mathrm{Var}\left(S\right)$$

$$= \frac{np(1-p)}{n^2}$$

Election

S is a sum of n coin flips with $p = \#$  $/T$

$$S = \sum_{i=1}^{n} V_i$$

$$\mathrm{Var}\left(\frac{S}{n}\right) = \frac{1}{n^2}\mathrm{Var}\left(S\right)$$

$$= \frac{np(1-p)}{n^2}$$

$$= \frac{p\left(1-p\right)}{n}$$

$$\sigma_{S/n} = \sqrt{\frac{p\left(1-p\right)}{n}}$$

Election

S is a sum of n coin flips with $p = \#$  $/T$

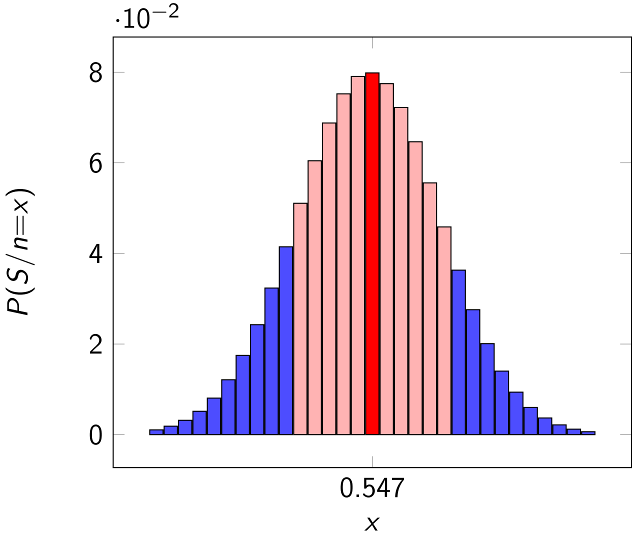
$$S = \sum_{i=1}^{n} V_i$$

$$\mathrm{Var}\left(\frac{S}{n}\right) = \frac{1}{n^2}\mathrm{Var}\left(S\right) = \frac{np(1-p)}{n^2} = \frac{p\left(1-p\right)}{n}$$

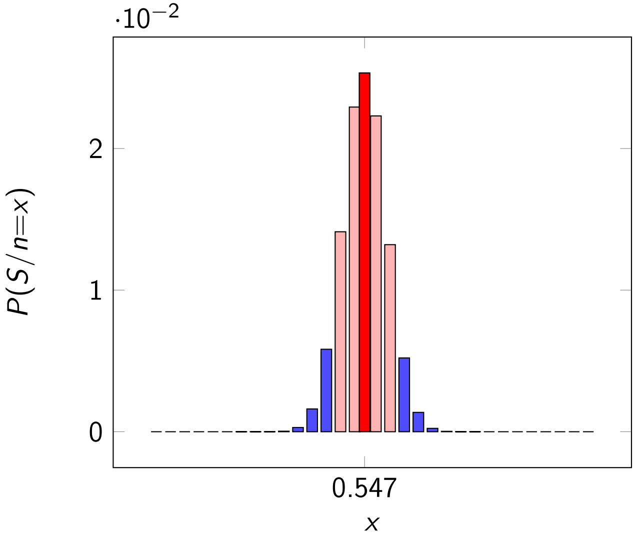
$$\sigma_{S/n} = \sqrt{\frac{p\left(1-p\right)}{n}}$$

Decreases as n grows!

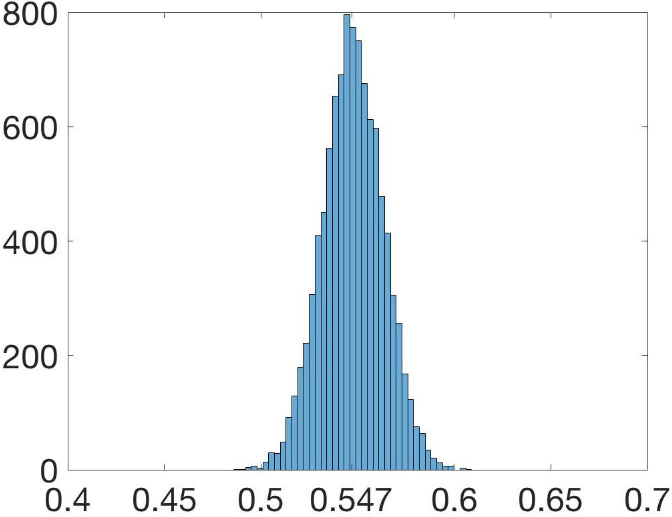
Fraction of voters in poll, $n = 100$, # $/T = 0.547$



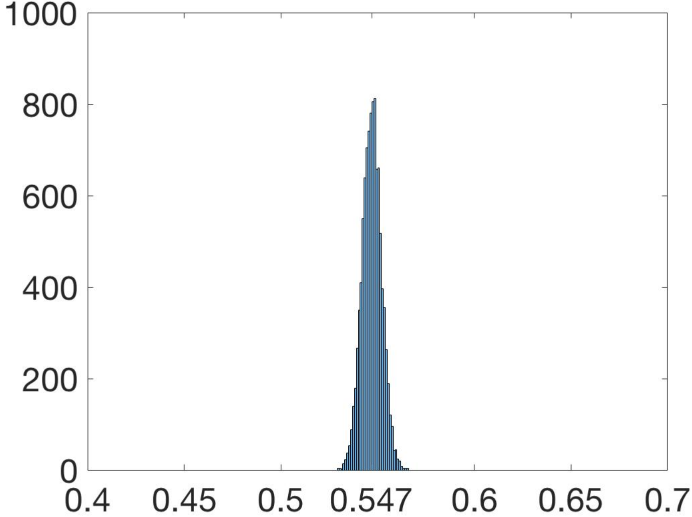
Fraction of voters in poll, $n = 1000$, # / $T = 0.547$



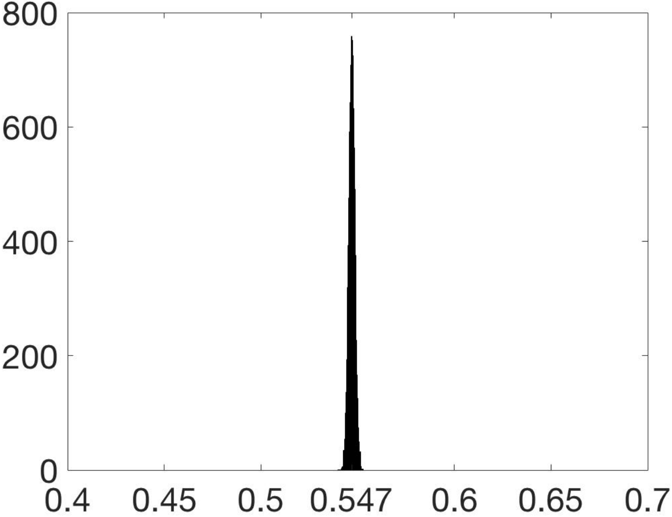
Poll of 1 000 people (repeated 10 000 times)



Poll of 10 000 people (repeated 10 000 times)



Poll of 100 000 people (repeated 10 000 times)



Some elections are difficult to predict

- It's very hard to sample a population uniformly
- Also hard to make samples independent
- Often not interested in just popular vote (e.g. presidential election)