Regression and inference

Andy Eggers
Assoc. Professor
Department of Politics and International Relations

We want you to understand:

Dependent variable: Nobel Prizes awarded per capita (in log scale)

	(1)	(2)	(3)
Intercept	-1.629* (0.509)	-3.166* (0.511)	-2.982* (0.527)
Chocolate consumption per capita (log scale)	2.092* (0.298)	1.026* (0.326)	0.709 (0.415)
GDP/capita (thousands of USD)		0.105* (0.024)	0.106* (0.024)
NW Europe			0.549 (0.452)
R^2	0.70	0.85	0.86
N	34	34	34

Standard errors in parentheses. * Indicates p<0.05

- what a dependent variable is
- what an independent variable is
- what the coefficients mean (intercept, slopes)
- what the stars mean (i.e. what p<0.05 means)
- what the standard errors mean

Sample vs population (1)

Sample vs population (1)

Say you have the data from a **representative survey** of British voters in April 2016 asking how respondents plan to vote in the referendum and whether they went to university or not.

Sample vs population (1)

Say you have the data from a **representative survey** of British voters in April 2016 asking how respondents plan to vote in the referendum and whether they went to university or not.

Your research questions are:

Sample vs population (1)

Say you have the data from a **representative survey** of British voters in April 2016 asking how respondents plan to vote in the referendum and whether they went to university or not.

Your research questions are:

1. How much support was there for Brexit?

Sample vs population (1)

Say you have the data from a **representative survey** of British voters in April 2016 asking how respondents plan to vote in the referendum and whether they went to university or not.

Your research questions are:

1. How much support was there for Brexit?
2. How was support for Brexit related to education?

Sample vs population (1)

Say you have the data from a **representative survey** of British voters in April 2016 asking how respondents plan to vote in the referendum and whether they went to university or not.

Your research questions are:

1. How much support was there for Brexit?
2. How was support for Brexit related to education?

How would you answer these questions?

Sample vs population (1)

Say you have the data from a **representative survey** of British voters in April 2016 asking how respondents plan to vote in the referendum and whether they went to university or not.

Your research questions are:

1. How much support was there for Brexit?
2. How was support for Brexit related to education?

How would you answer these questions?

Is there any **uncertainty** in your answers?

Sample vs population (2)

Sample vs population (2)

Is there uncertainty? Depends on who you are asking about.

Sample vs population (2)

Is there uncertainty? Depends on who you are asking about.

1. How much support is there for Brexit *among respondents to this survey*?
2. How is support for Brexit related to education *among respondents to this survey*?

Sample vs population (2)

Is there uncertainty? Depends on who you are asking about.

1. How much support is there for Brexit *among respondents to this survey*?
2. How is support for Brexit related to education *among respondents to this survey*?

(About the sample)

Sample vs population (2)

Is there uncertainty? Depends on who you are asking about.

1. How much support is there for Brexit *among respondents to this survey*?
2. How is support for Brexit related to education *among respondents to this survey*?

(About the sample)

1. How much support is there for Brexit *among all voters*?
2. How is support for Brexit related to education *among all voters*?

Sample vs population (2)

Is there uncertainty? Depends on who you are asking about.

1. How much support is there for Brexit *among respondents to this survey*?
2. How is support for Brexit related to education *among respondents to this survey*?

(About the sample)

1. How much support is there for Brexit *among all voters*?
2. How is support for Brexit related to education *among all voters*?

(About the population)

Sample vs population (2)

Is there uncertainty? Depends on who you are asking about.

1. How much support is there for Brexit *among respondents to this survey*?
2. How is support for Brexit related to education *among respondents to this survey*?

(About the sample)

1. How much support is there for Brexit *among all voters*?
2. How is support for Brexit related to education *among all voters*?

(About the population)

No real uncertainty.

(Maybe about measurement.)

Sample vs population (2)

Is there uncertainty? Depends on who you are asking about.

1. How much support is there for Brexit *among respondents to this survey*?
2. How is support for Brexit related to education *among respondents to this survey*?

(About the sample)

No real uncertainty.
(Maybe about measurement.)

1. How much support is there for Brexit *among all voters*?
2. How is support for Brexit related to education *among all voters*?

(About the population)

Uncertainty due to **sampling variation**.

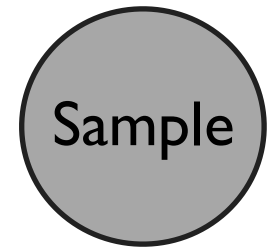
Sample vs population (3)

Sample vs population (3)

Generally, we have data from a **sample**

Sample vs population (3)

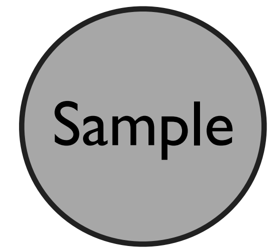
Generally, we have data from a **sample**



Sample vs population (3)

Generally, we have data from a **sample**

but we want to say something about a (larger) **population**.

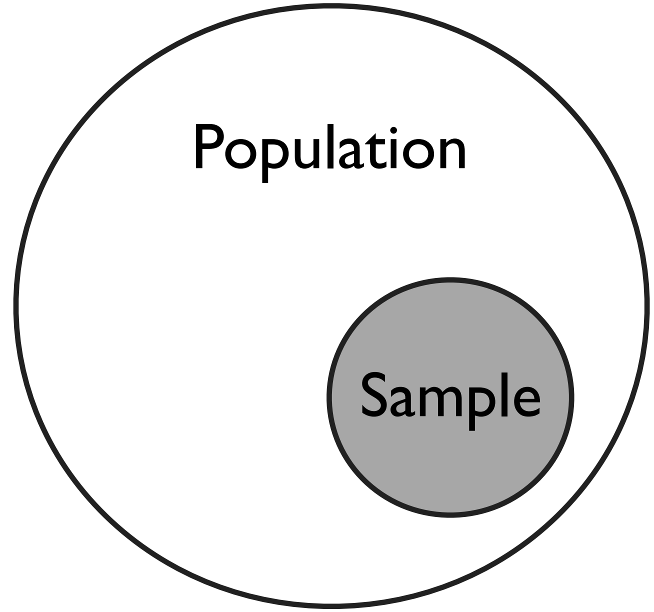


Sample vs population (3)

Generally, we have data from a **sample**

but we want to say something about a (larger) **population**.

This is **statistical inference**.

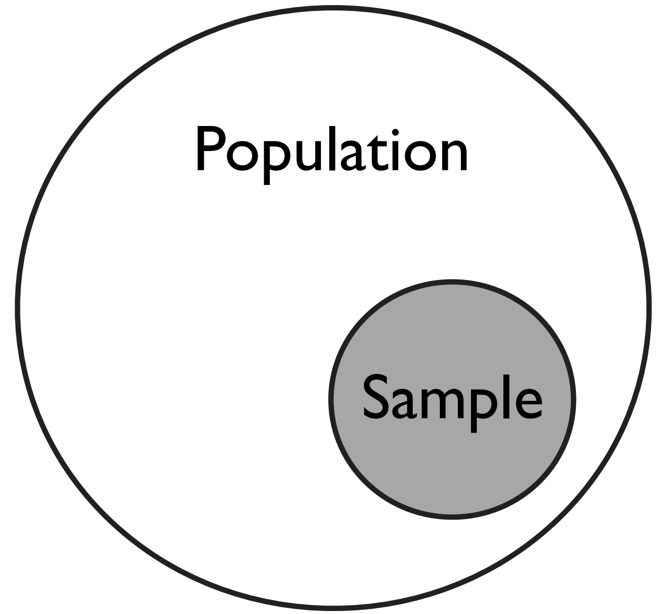


Sample vs population (3)

Generally, we have data from a **sample**

but we want to say something about a (larger) **population**.

This is **statistical inference**.



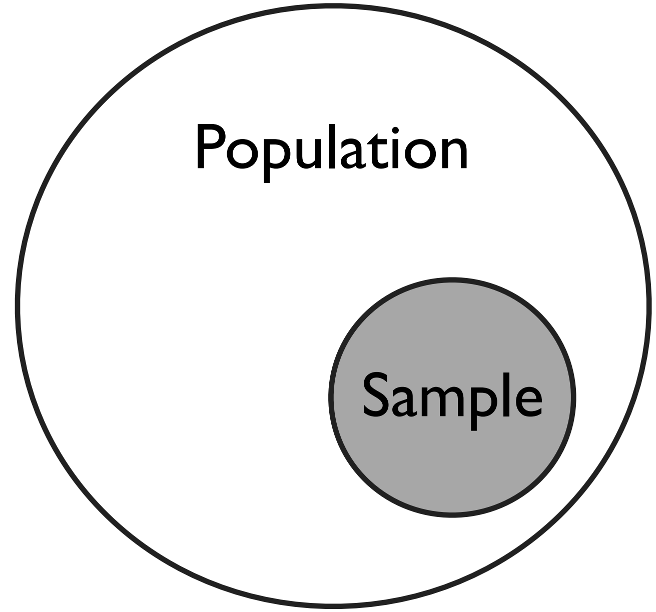
In hypothesis testing, we use data from a **sample** to assess conjectures about the **population**.

Sample vs population (3)

Generally, we have data from a **sample**

but we want to say something about a (larger) **population**.

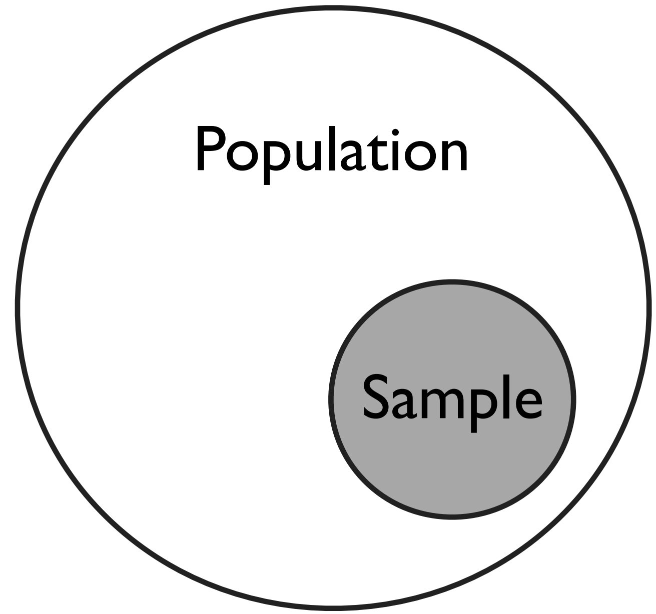
This is **statistical inference**.



In hypothesis testing, we use data from a **sample** to assess conjectures about the **population**.

Because the sample is not the population:

Sample vs population (3)



Generally, we have data from a **sample**

but we want to say something about a (larger) **population**.

This is **statistical inference**.

In hypothesis testing, we use data from a **sample** to assess conjectures about the **population**.

Because the sample is not the population:

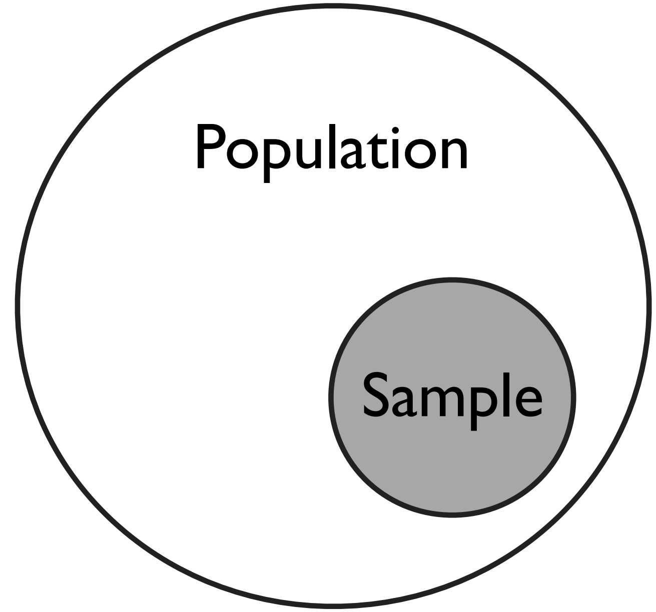
- polls have a **margin of error**

Sample vs population (3)

Generally, we have data from a **sample**

but we want to say something about a (larger) **population**.

This is **statistical inference**.



In hypothesis testing, we use data from a **sample** to assess conjectures about the **population**.

Because the sample is not the population:

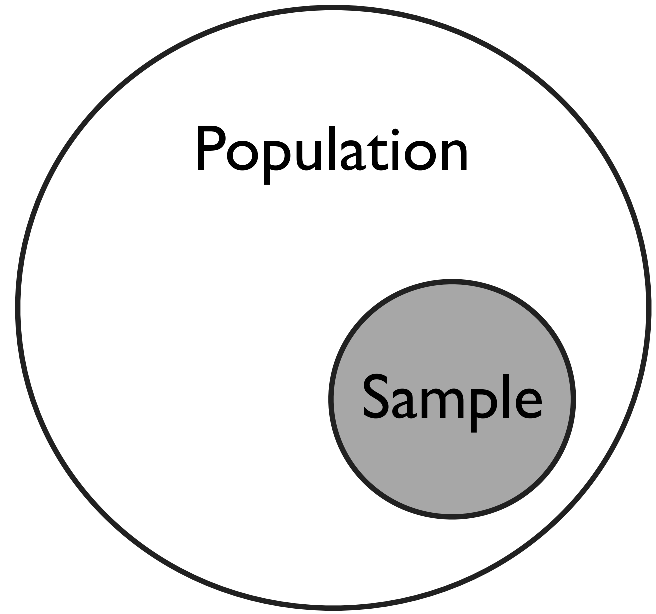
- polls have a **margin of error**
- regression coefficients have **standard errors**

Sample vs population (3)

Generally, we have data from a **sample**

but we want to say something about a (larger) **population**.

This is **statistical inference**.



In hypothesis testing, we use data from a **sample** to assess conjectures about the **population**.

Because the sample is not the population:

- polls have a **margin of error**
- regression coefficients have **standard errors**
- our conclusions in hypothesis testing are guesses, with confidence summarized by **p-values**

Thought experiment

Thought experiment

Suppose you actually **know** the proportion of people who supported leave on April 10, 2016, but you only survey a random sample.

Will the level of support in your sample be close to the true average support in the population on that date?

Thought experiment

Suppose you actually **know** the proportion of people who supported leave on April 10, 2016, but you only survey a random sample.

Will the level of support in your sample be close to the true average support in the population on that date?

If truly a random sample, there is no **bias**: you should expect to get the true value on average.

But the level of support in any given sample will differ from the true value due to **random error** (i.e. sampling variation).

Thought experiment

Suppose you actually **know** the proportion of people who supported leave on April 10, 2016, but you only survey a random sample.

Will the level of support in your sample be close to the true average support in the population on that date?

If truly a random sample, there is no **bias**: you should expect to get the true value on average.

But the level of support in any given sample will differ from the true value due to **random error** (i.e. sampling variation).

What would the magnitude of this **random error** depend on?

- size of sample (1,006 GB adults vs. 10,000,000)
- true level of support (what if 100% supported remaining in EU?)

Simulating the thought experiment in R

Simulating the thought experiment in R

Suppose **we know** that 52% of all voters supported **Leave. We want to know** how much the result of a poll might deviate from the true level of support.

Let's find out using R!

Simulating the thought experiment in R

Suppose **we know** that 52% of all voters supported **Leave. We want to know** how much the result of a poll might deviate from the true level of support.

Let's find out using R!

Using R, I can randomly draw 10 ones and zeros, where the probability of drawing a one is 0.52:

Simulating the thought experiment in R

Suppose **we know** that 52% of all voters supported **Leave. We want to know** how much the result of a poll might deviate from the true level of support.

Let's find out using R!

Using R, I can randomly draw 10 ones and zeros, where the probability of drawing a one is 0.52:

```
> sample(x = c(0,1), size = 10, replace = T, prob = c(.48, .52))
 [1] 1 1 0 0 1 1 0 0 0 0
```

Simulating the thought experiment in R

Suppose **we know** that 52% of all voters supported **Leave. We want to know** how much the result of a poll might deviate from the true level of support.

Let's find out using R!

Using R, I can randomly draw 10 ones and zeros, where the probability of drawing a one is 0.52:

```
> sample(x = c(0,1), size = 10, replace = T, prob = c(.48, .52))
 [1] 1 1 0 0 1 1 0 0 0 0
```

I can do it again:

```
> sample(x = c(0,1), size = 10, replace = T, prob = c(.48, .52))
 [1] 1 1 0 0 1 1 0 1 1 0
```

Simulating the thought experiment in R

Suppose **we know** that 52% of all voters supported **Leave. We want to know** how much the result of a poll might deviate from the true level of support.

Let's find out using R!

Using R, I can randomly draw 10 ones and zeros, where the probability of drawing a one is 0.52:

```
> sample(x = c(0,1), size = 10, replace = T, prob = c(.48, .52))
 [1] 1 1 0 0 1 1 0 0 0 0
```

I can do it again:

```
> sample(x = c(0,1), size = 10, replace = T, prob = c(.48, .52))
 [1] 1 1 0 0 1 1 0 1 1 0
```

I can increase the number of "respondents" to 1,006:

```
> sample(x = c(0,1), size = 1006, replace = T, prob = c(.48, .52))
   [1] 0 0 0 0 0 0 0 1 1 1 1 1 0 1 1 1 1 1 1 1 1 0 0 1 1 0 0 0 0 0 0 0
  [39] 0 0 0 1 1 1 1 0 0 0 1 0 0 1 0 1 1 1 0 0 0 1 1 0 0 0 0 0 0 0 0 0
  [77] 1 0 0 1 1 1 0 0 0 0 0 1 1 1 1 1 1 1 0 0 0 0 1 0 0 0 0 0 1 1 0 0
 [115] 1 0 1 1 0 1 0 0 1 0 0 0 0 1 1 0 0 0 0 1 1 1 1 0 1 0 1 0 1 0 0 1
 [153] 0 1 1 0 0 1 1 1 0 0 1 0 1 0 1 1 1 1 0 0 0 1 0 1 0 1 1 1 1 1 1 0
 [191] 1 0 0 1 1 1 1 0 1 0 1 0 0 0 1 0 0 1 0 0 0 1 1 1 0 0 0 0 0 1 0 1
```

Simulating the thought experiment (2)

Simulating the thought experiment (2)

I can store the sample and take the mean:

```
> samp = sample(x = c(0,1), size = 1006, replace = T, prob = c(.48, .52))
> mean(samp)
[1] 0.5318091
```

Simulating the thought experiment (2)

I can store the sample and take the mean:

```
> samp = sample(x = c(0,1), size = 1006, replace = T, prob = c(.48, .52))
> mean(samp)
[1] 0.5318091
```

I can do it again:

```
> samp = sample(x = c(0,1), size = 1006, replace = T, prob = c(.48, .52))
> mean(samp)
[1] 0.5119284
```

Simulating the thought experiment (2)

I can store the sample and take the mean:

```
> samp = sample(x = c(0,1), size = 1006, replace = T, prob = c(.48, .52))
> mean(samp)
[1] 0.5318091
```

I can do it again:

```
> samp = sample(x = c(0,1), size = 1006, replace = T, prob = c(.48, .52))
> mean(samp)
[1] 0.5119284
```

I can do it 10,000 times and look at the histogram of support:

Simulating the thought experiment (2)

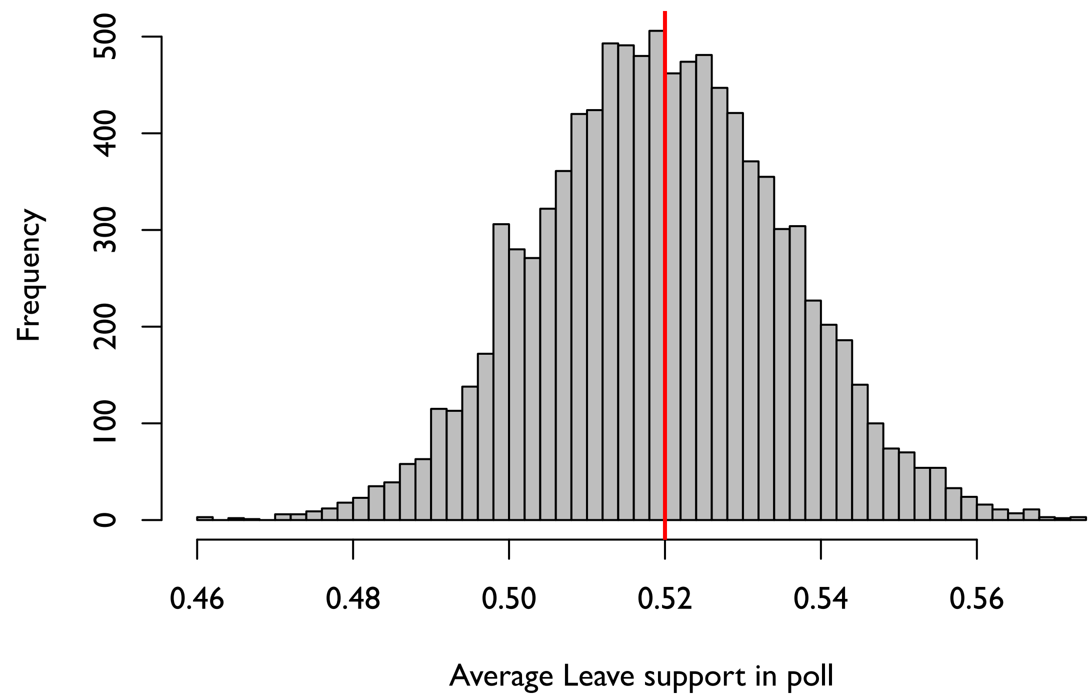
I can store the sample and take the mean:

```
> samp = sample(x = c(0,1), size = 1006, replace = T, prob = c(.48, .52))
> mean(samp)
[1] 0.5318091
```

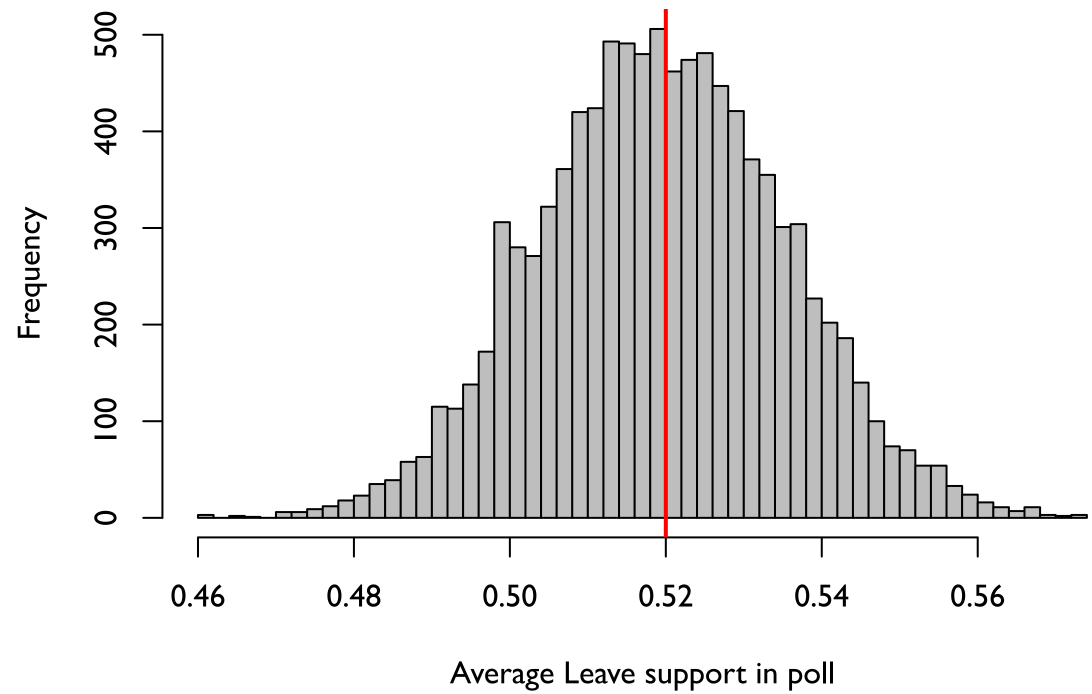
I can do it again:

```
> samp = sample(x = c(0,1), size = 1006, replace = T, prob = c(.48, .52))
> mean(samp)
[1] 0.5119284
```

I can do it 10,000 times and look at the histogram of support:

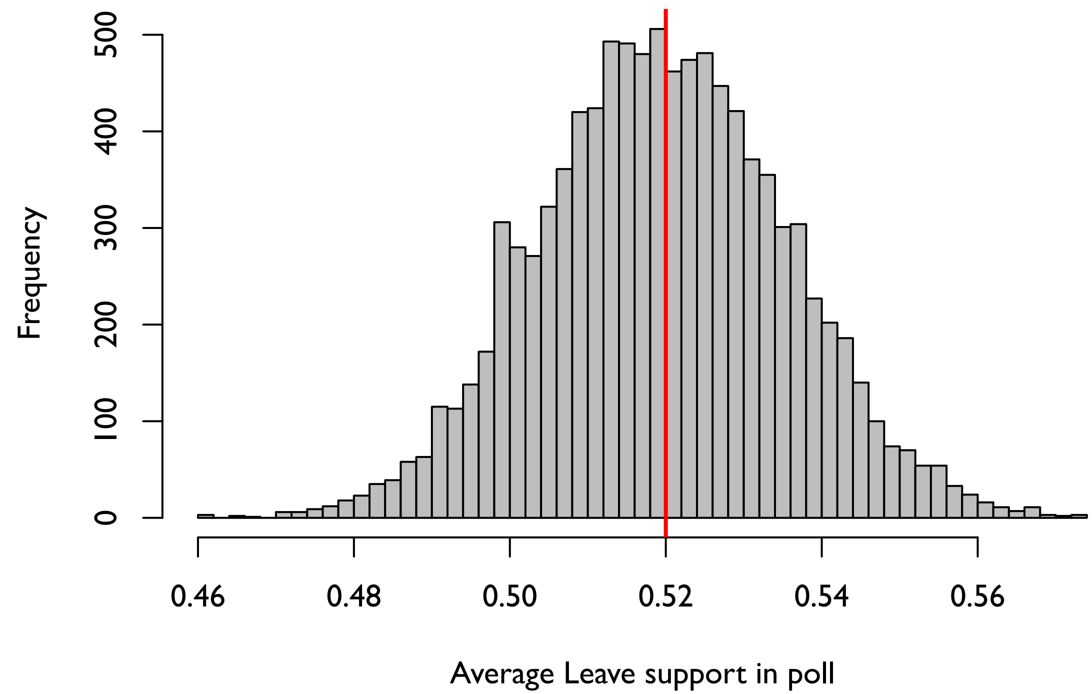


Simulating the thought experiment (3)



Simulating the thought experiment (3)

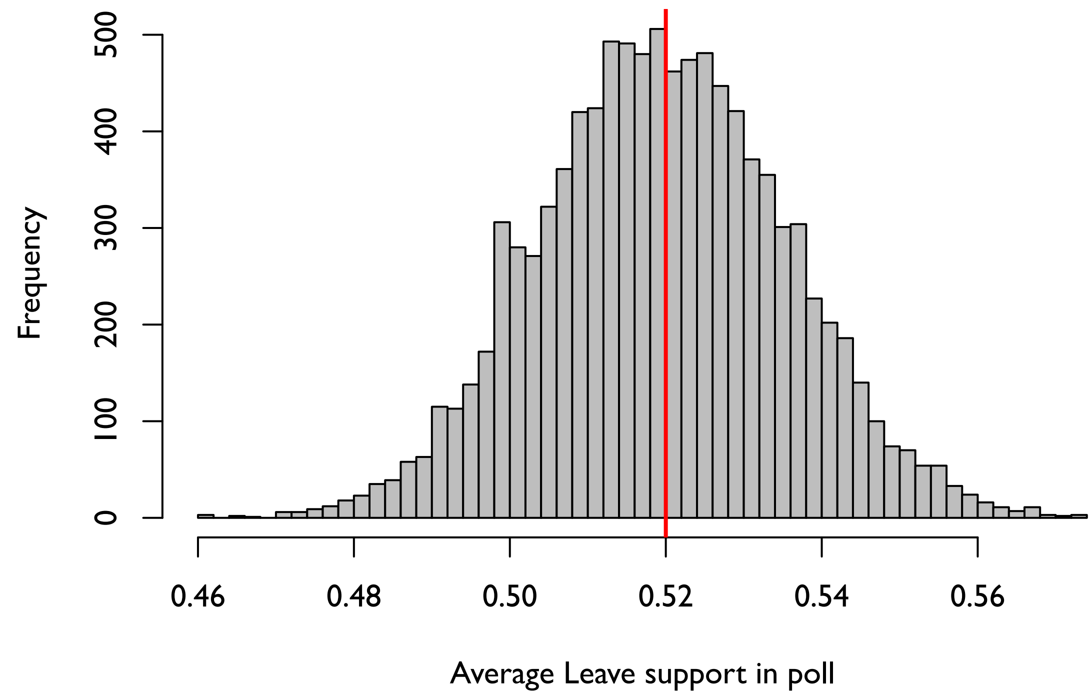
The results vary across our 10,000 "surveys" because of **sampling error**.



Simulating the thought experiment (3)

The results vary across our 10,000 "surveys" because of **sampling error**.

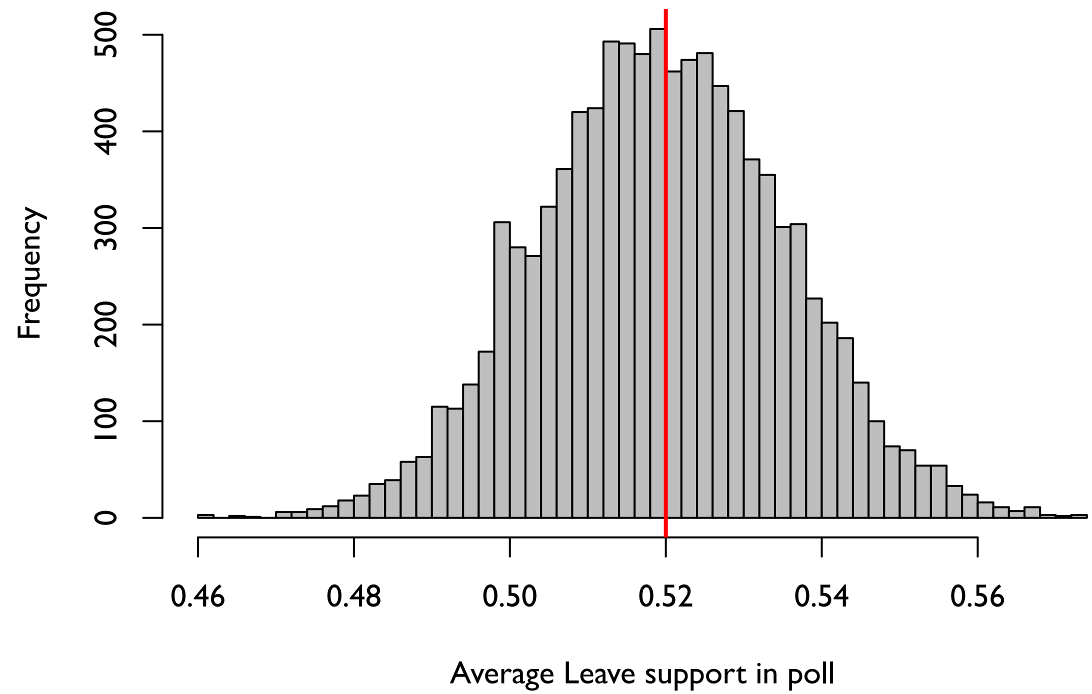
How much sampling error is there in our simulation?



Simulating the thought experiment (3)

The results vary across our 10,000 "surveys" because of **sampling error**.

How much sampling error is there in our simulation?

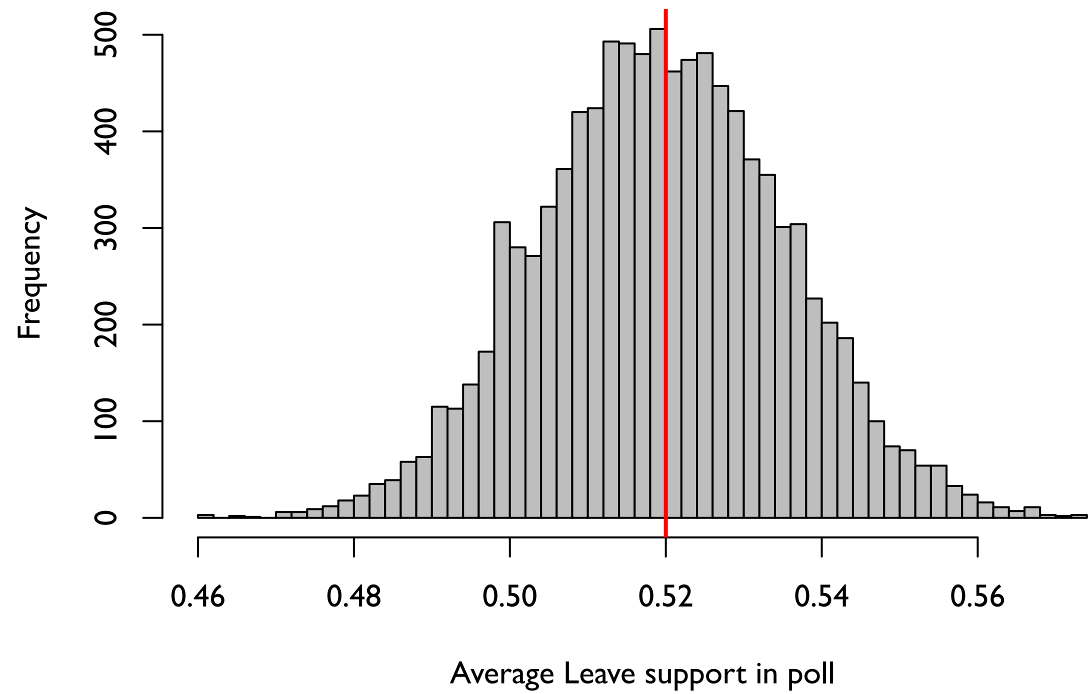


The standard deviation:

```
> sd(poll.results)
[1] 0.01572411
```

Simulating the thought experiment (3)

The results vary across our 10,000 "surveys" because of **sampling error.**



How much sampling error is there in our simulation?

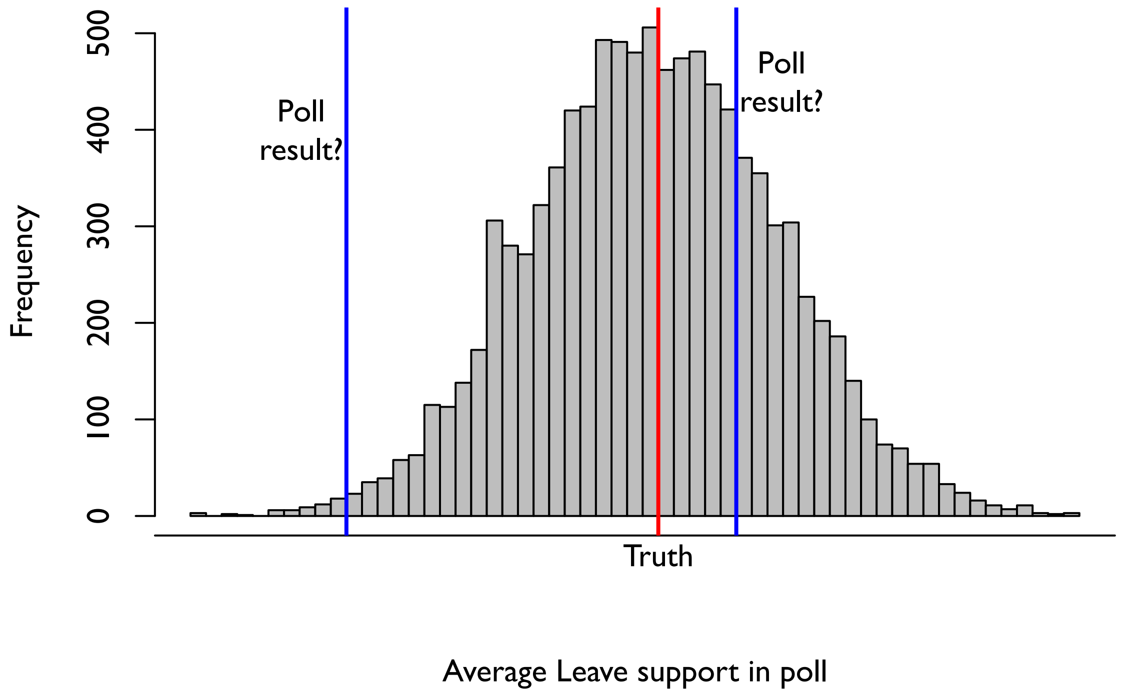
The standard deviation:

```
> sd(poll.results)
[1] 0.01572411
```

95% of the samples had a mean between 0.49 and 0.55:

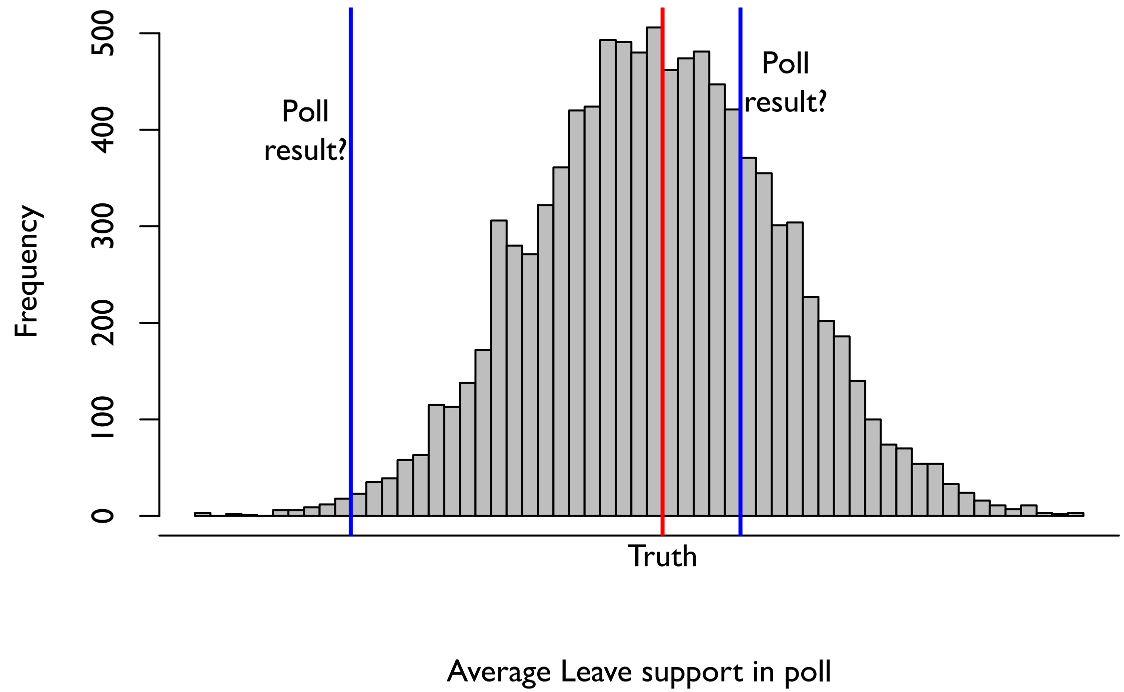
```
> quantile(poll.results, c(.025, .975))
     2.5%     97.5%
0.4890656 0.5497018
```

From thought experiment to margin of error



From thought experiment to margin of error

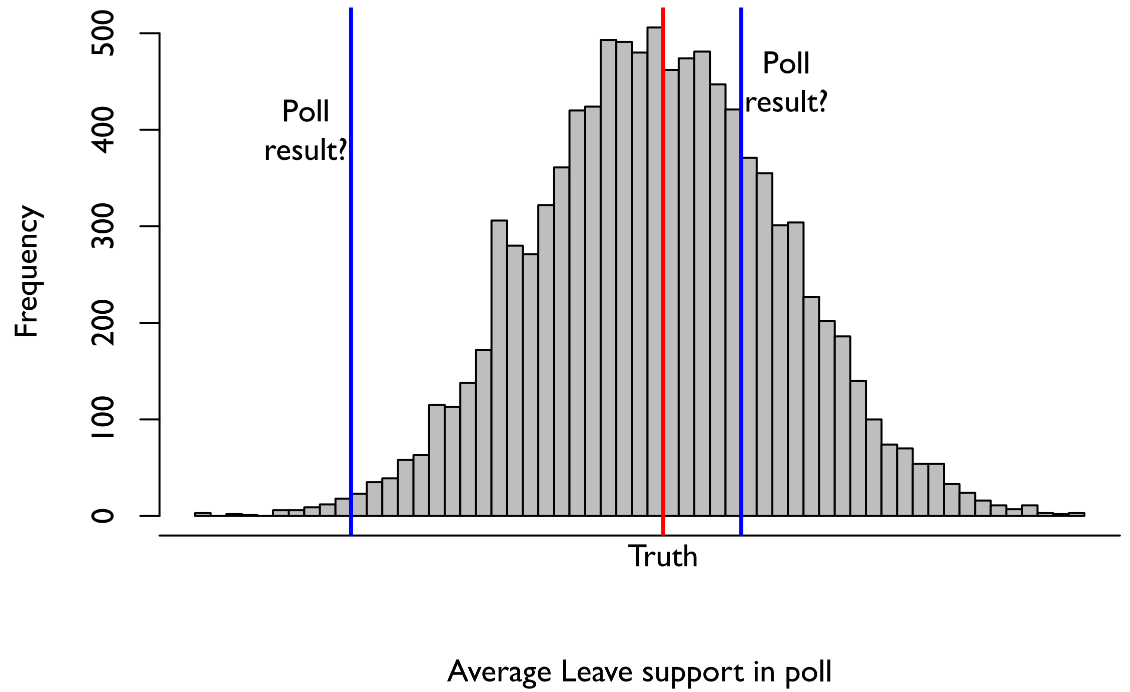
In a real survey, you don't know the answer; all you get is a **single number,** i.e. your poll result.



From thought experiment to margin of error

In a real survey, you don't know the answer; all you get is a **single number,** i.e. your poll result.

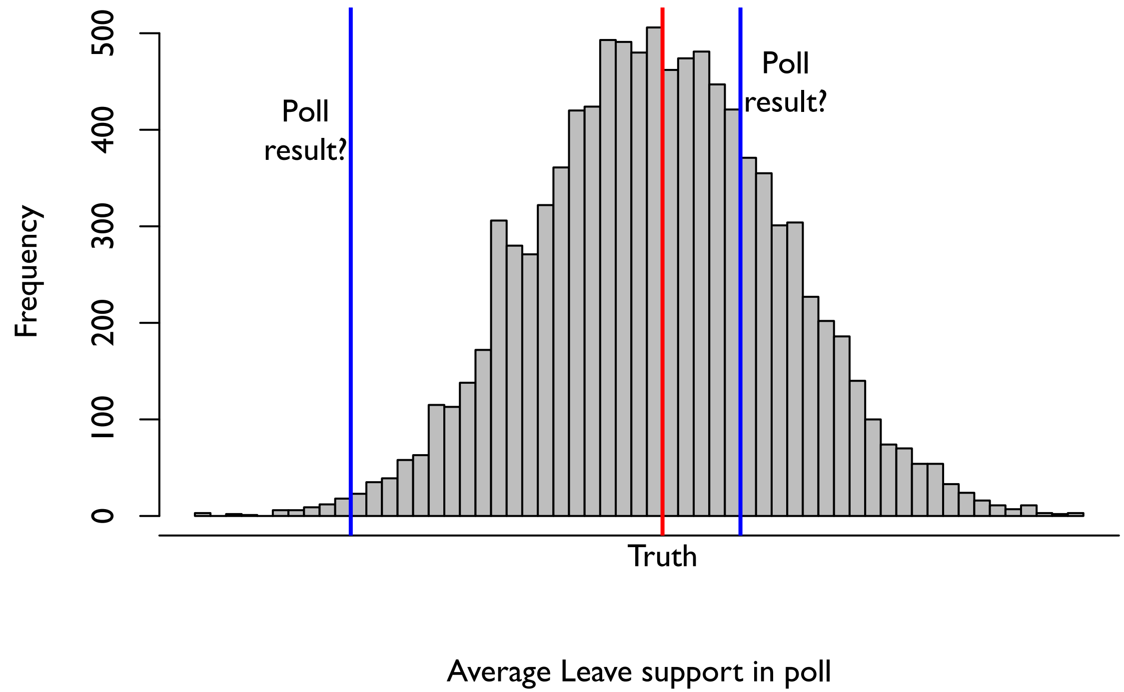
But the histogram from the thought experiment gives you a clue how close your number is to the "Truth".



From thought experiment to margin of error

In a real survey, you don't know the answer; all you get is a **single number,** i.e. your poll result.

But the histogram from the thought experiment gives you a clue how close your number is to the "Truth".

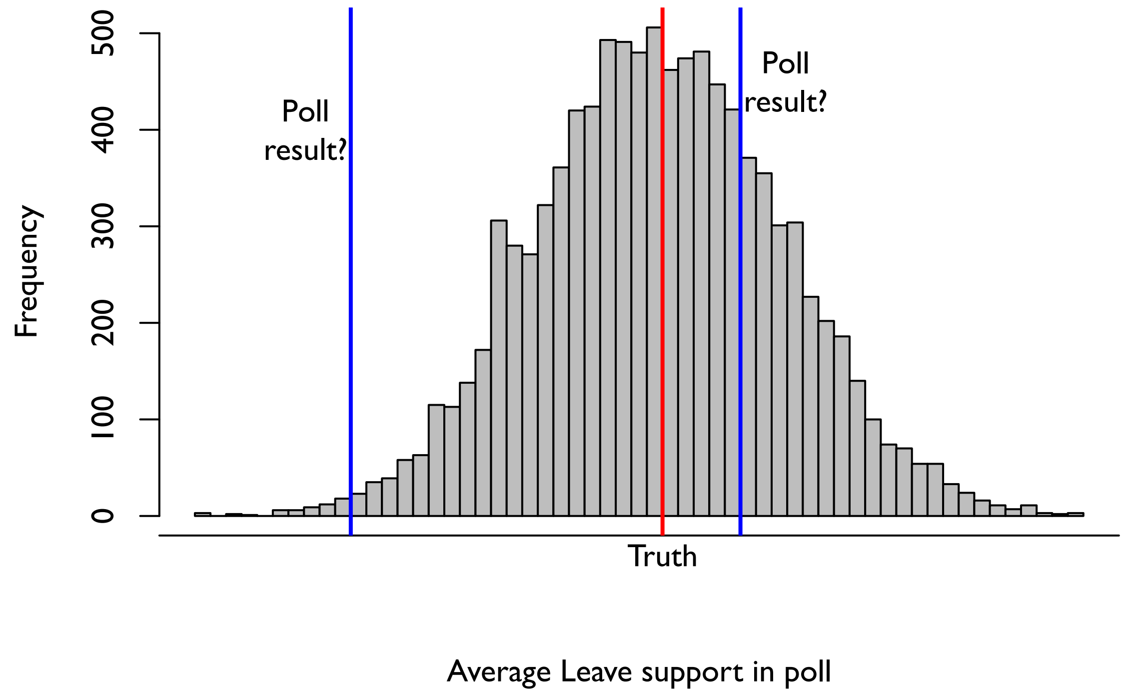


In our thought experiment (where we **know** the truth), 95% of the samples were within 0.031 of the truth.

From thought experiment to margin of error

In a real survey, you don't know the answer; all you get is a **single number,** i.e. your poll result.

But the histogram from the thought experiment gives you a clue how close your number is to the "Truth".



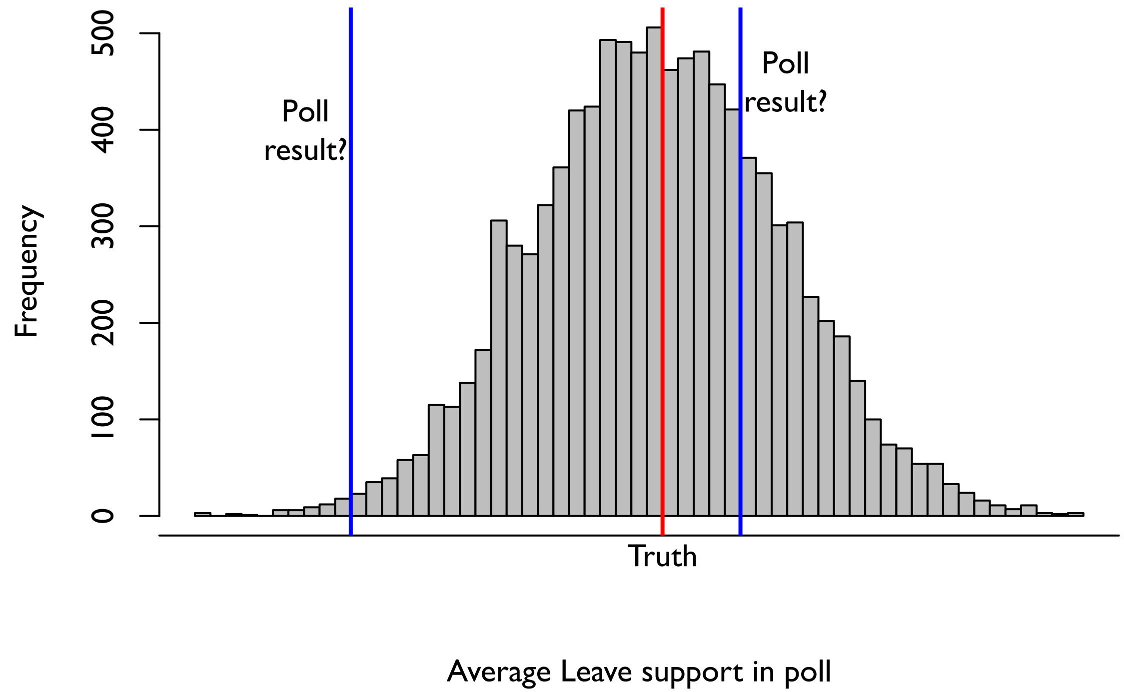
In our thought experiment (where we **know** the truth), 95% of the samples were within 0.031 of the truth.

In an actual survey (where we **don't know** the truth), we have 95% confidence that our estimate is within 0.031 of the truth.

From thought experiment to margin of error

In a real survey, you don't know the answer; all you get is a **single number,** i.e. your poll result.

But the histogram from the thought experiment gives you a clue how close your number is to the "Truth".



In our thought experiment (where we **know** the truth), 95% of the samples were within 0.031 of the truth.

In an actual survey (where we **don't know** the truth), we have 95% confidence that our estimate is within 0.031 of the truth.

Margin of error

From thought experiment to margin of error (2)

From thought experiment to margin of error (2)

So when we do a survey, we get:

From thought experiment to margin of error (2)

So when we do a survey, we get:

An estimate for Leave support (e.g. 49%)

From thought experiment to margin of error (2)

So when we do a survey, we get:

An estimate for Leave support (e.g. 49%)

(From the thought experiment:) An estimate of the standard deviation of poll results across samples: 0.0157 (called the **standard error** of the poll)

From thought experiment to margin of error (2)

So when we do a survey, we get:

An estimate for Leave support (e.g. 49%)

(From the thought experiment:) An estimate of the standard deviation of poll results across samples: 0.0157 (called the **standard error** of the poll)

(Combining the two:) A 95% confidence interval, which we expect to include the truth in 95% of samples: e.g. 49% ± 3.1% (3.1% is the **margin of error** of the poll)

Another way to get the margin of error (1)

Another way to get the margin of error (1)

Another way to get the margin
of error from a single sample:

Another way to get the margin of error (1)

Another way to get the margin of error from a single sample:

The **central limit theorem** says that the proportion of support in samples of size n will follow a Normal distribution centered on the truth with approximate standard deviation:

Another way to get the margin of error (1)

Another way to get the margin of error from a single sample:

The **central limit theorem** says that the proportion of support in samples of size n will follow a Normal distribution centered on the truth with approximate standard deviation:

$$\sqrt{\frac{\text{Variance of sample}}{n}}$$

Another way to get the margin of error (1)

Another way to get the margin of error from a single sample:

The **central limit theorem** says that the proportion of support in samples of size n will follow a Normal distribution centered on the truth with approximate standard deviation:

$$\sqrt{\frac{\text{Variance of sample}}{n}}$$

Our sample of e.g. 546 "Leaves" and 460 "Remains" has a variance of .248.

Another way to get the margin of error (1)

Another way to get the margin of error from a single sample:

The **central limit theorem** says that the proportion of support in samples of size n will follow a Normal distribution centered on the truth with approximate standard deviation:

$$\sqrt{\frac{\text{Variance of sample}}{n}}$$

Our sample of e.g. 546 "Leaves" and 460 "Remains" has a variance of .248.

So the estimated standard deviation (**standard error**) of our estimate is:

$$\sqrt{\frac{.248}{1006}} = 0.0157$$

Another way to get the margin of error (1)

Another way to get the margin of error from a single sample:

The **central limit theorem** says that the proportion of support in samples of size n will follow a Normal distribution centered on the truth with approximate standard deviation:

$$\sqrt{\frac{\text{Variance of sample}}{n}}$$

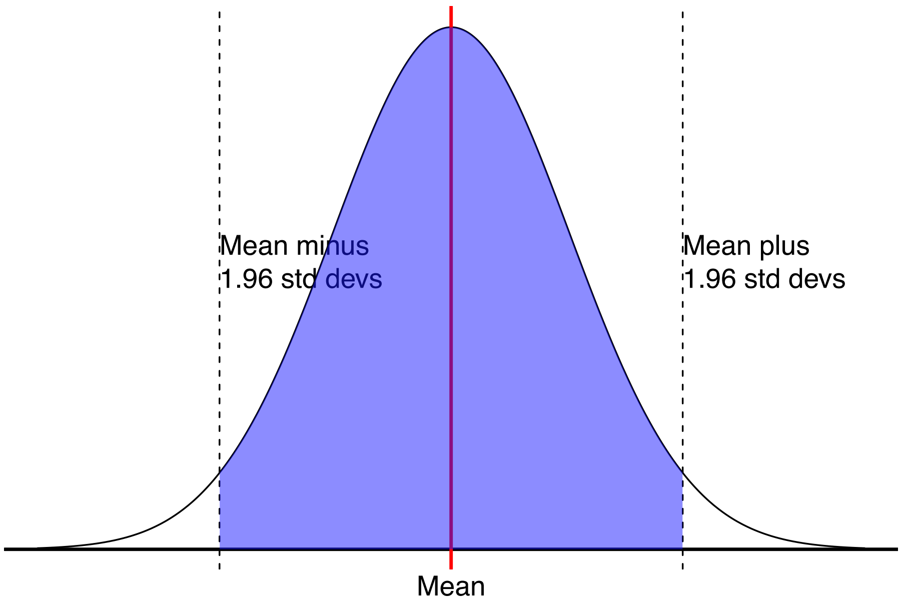
Our sample of e.g. 546 "Leaves" and 460 "Remains" has a variance of .248.

So the estimated standard deviation (**standard error**) of our estimate is:

$$\sqrt{\frac{.248}{1006}} = 0.0157$$

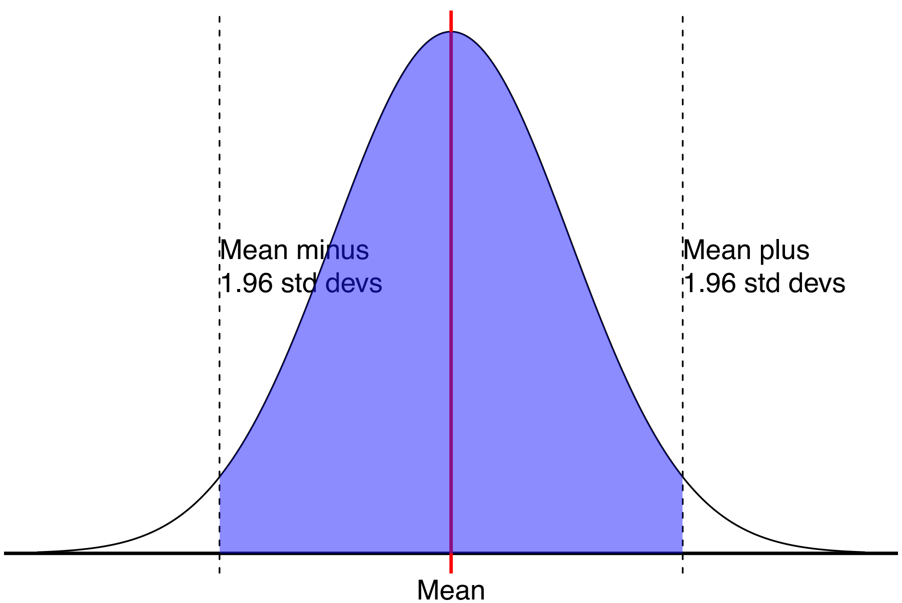
Compare: the standard deviation of our simulations was 0.0157

Another way to get the margin of error (2)



Another way to get the margin of error (2)

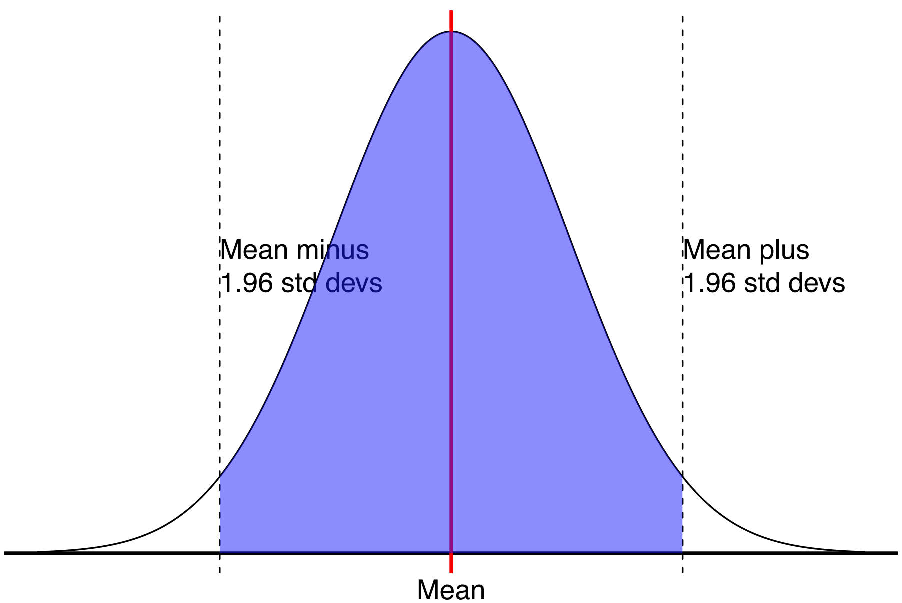
In a Normal distribution, about 95% of the draws are within 1.96 standard deviations of the mean.



Another way to get the margin of error (2)

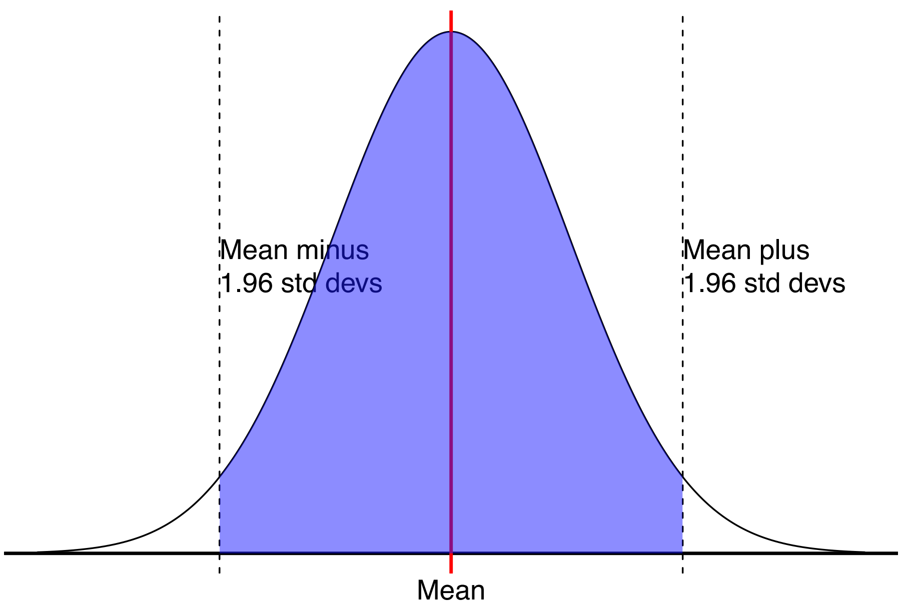
In a Normal distribution, about 95% of the draws are within 1.96 standard deviations of the mean.

This indicates that in 95% of surveys we run, our answer should be within 1.96 standard deviations of the truth.



Another way to get the margin of error (2)

In a Normal distribution, about 95% of the draws are within 1.96 standard deviations of the mean.

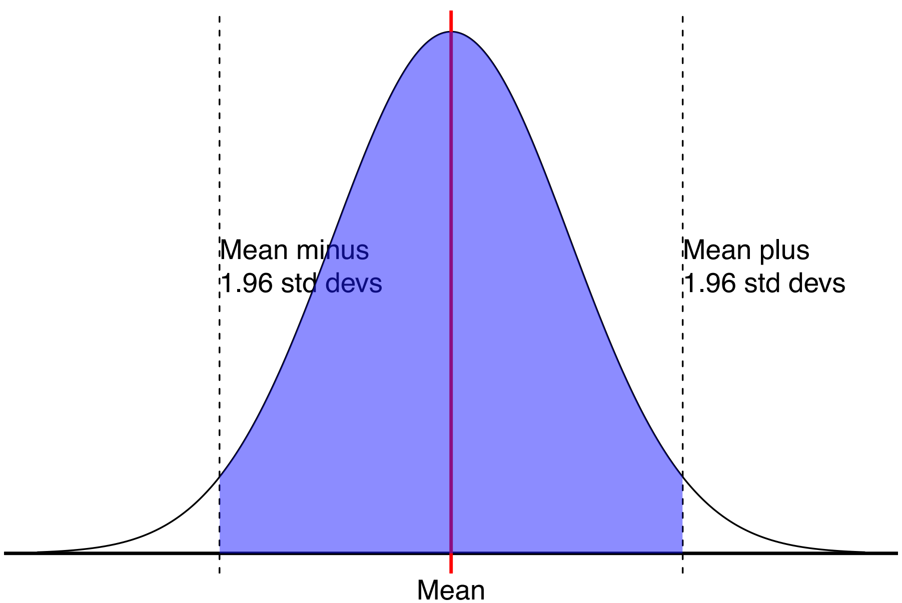


This indicates that in 95% of surveys we run, our answer should be within 1.96 standard deviations of the truth.

Given estimated standard deviation (standard error) of 0.0157, we have a **margin of error** (1.96 times standard error) of .031.

Another way to get the margin of error (2)

In a Normal distribution, about 95% of the draws are within 1.96 standard deviations of the mean.



This indicates that in 95% of surveys we run, our answer should be within 1.96 standard deviations of the truth.

Given estimated standard deviation (standard error) of 0.0157, we have a **margin of error** (1.96 times standard error) of .031.

Compare: our simulations implied a margin of error of 0.031.

Quick recap: survey part

Quick recap: survey part

- **Standard error of poll**: an estimate of how much the estimate might vary due to random error (**sampling error**)

Quick recap: survey part

- **Standard error of poll**: an estimate of how much the estimate might vary due to random error (**sampling error**)
- In 95% of polls, the true value should be within **margin of error** (≈ 2 × standard error) of the estimate (assuming no bias)

Quick recap: survey part

- **Standard error of poll**: an estimate of how much the estimate might vary due to random error (**sampling error**)
- In 95% of polls, the true value should be within **margin of error** ($\approx$ 2 × standard error) of the estimate (assuming no bias)
- Two ways we got the margin of error:

Quick recap: survey part

- **Standard error of poll**: an estimate of how much the estimate might vary due to random error (**sampling error**)
- In 95% of polls, the true value should be within **margin of error** (≈ 2 × standard error) of the estimate (assuming no bias)
- Two ways we got the margin of error:
  - **Simulation** in R of 10,000 random samples of size 1,006 given a known level of support for "Leave"

Quick recap: survey part

- **Standard error of poll**: an estimate of how much the estimate might vary due to random error (**sampling error**)
- In 95% of polls, the true value should be within **margin of error** (≈ 2 × standard error) of the estimate (assuming no bias)
- Two ways we got the margin of error:
  - **Simulation** in R of 10,000 random samples of size 1,006 given a known level of support for "Leave"
  - **Central limit theorem:** approximation to a normal distribution

Thought experiment (2)

Thought experiment (2)

Suppose we know that

Thought experiment (2)

Suppose we know that

Support for Leave

57.9% of those who did not attend university

41.5% of those who did attend university

Thought experiment (2)

Suppose we know that

Support for Leave

57.9% of those who did not attend university

41.5% of those who did attend university

Thus if we ran the regression

$$\text{LeaveSupport} = \beta_0 + \beta_1 \text{ AttendedUniversity}$$

in the **full population data**, the coefficients will be:

Thought experiment (2)

Suppose we know that

Support for Leave

57.9% of those who did not attend university

41.5% of those who did attend university

Thus if we ran the regression

$$\text{LeaveSupport} = \beta_0 + \beta_1 \text{ AttendedUniversity}$$

in the **full population data**, the coefficients will be:

$$\beta_0 = .579, \beta_1 = -.164$$

Thought experiment (2)

Suppose we know that

Support for Leave

57.9% of those who did not attend university

41.5% of those who did attend university

Thus if we ran the regression

$$\text{LeaveSupport} = \beta_0 + \beta_1 \text{ AttendedUniversity}$$

in the **full population data**, the coefficients will be:

$$\beta_0 = .579, \beta_1 = -.164$$

But what if we draw a random sample and run this regression in our sample? How far off might the coefficients be?

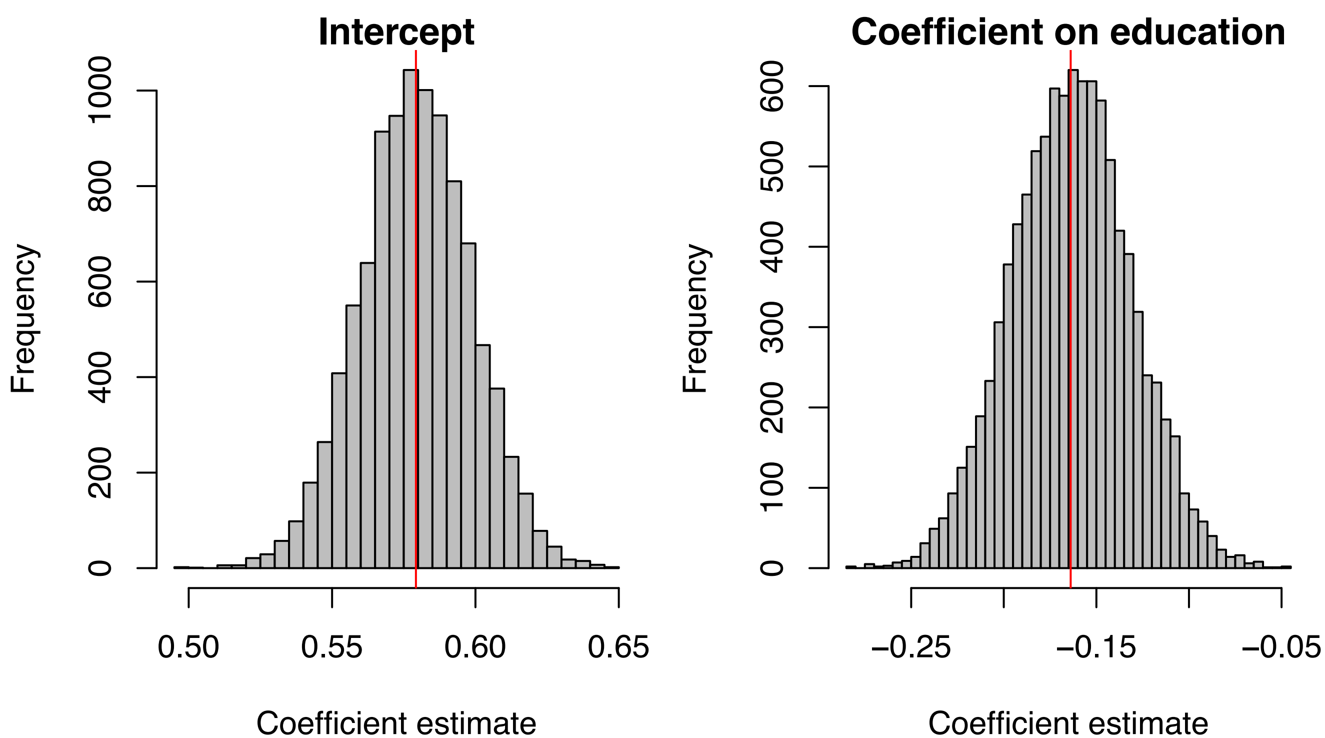
Thought experiment (2.1)

Thought experiment (2.1)

Across 10,000 simulated samples of size 1,006, the histograms for the two coefficients look like:

Thought experiment (2.1)

Across 10,000 simulated samples of size 1,006, the histograms for the two coefficients look like:



Thought experiment (2.2)

Thought experiment (2.2)

We can calculate the standard deviation of the coefficients across simulations:

Thought experiment (2.2)

We can calculate the standard deviation of the coefficients across simulations:

```
> apply(coef.mat, 2, sd)
 Intercept      Slope
0.01943883 0.03216419
```

Thought experiment (2.2)

We can calculate the standard deviation of the coefficients across simulations:

```
> apply(coef.mat, 2, sd)
 Intercept      Slope
0.01943883 0.03216419
```

I stored the estimates in a matrix called `coef.mat`. This command says "calculate the standard deviation of the columns".

Thought experiment (2.2)

We can calculate the standard deviation of the coefficients across simulations:

```
> apply(coef.mat, 2, sd)
 Intercept      Slope
0.01943883 0.03216419
```

I stored the estimates in a matrix called `coef.mat`. This command says "calculate the standard deviation of the columns".

Again, we call these **standard errors**.

Thought experiment (2.2)

We can calculate the standard deviation of the coefficients across simulations:

```
> apply(coef.mat, 2, sd)
 Intercept      Slope
0.01943883 0.03216419
```

I stored the estimates in a matrix called `coef.mat`. This command says "calculate the standard deviation of the columns".

Again, we call these **standard errors**.

As with the simple polling case, we can also use some statistical theory to estimate the standard errors given a sample (i.e. without doing a simulation).

Standard errors in regression output

Standard errors in regression output

Output from a regression for one sample:

Standard errors in regression output

Output from a regression for one sample:

```
> summary(lm(support.leave[indices.to.sample] ~ attended.uni[indices.to.sample]))

Call:
lm(formula = support.leave[indices.to.sample] ~ attended.uni[indices.to.sample])

Residuals:
    Min      1Q  Median      3Q     Max 
-0.5987 -0.5987  0.4013  0.4013  0.5550 

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)    
(Intercept)                      0.59874    0.01960  30.544  < 2e-16 ***
attended.uni[indices.to.sample] -0.15370    0.03219  -4.774 2.07e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4932 on 1004 degrees of freedom
Multiple R-squared:  0.0222,Adjusted R-squared:  0.02123 
F-statistic: 22.79 on 1 and 1004 DF,  p-value: 2.071e-06
```

Standard errors in regression output

Output from a regression for one sample:

```
> summary(lm(support.leave[indices.to.sample] ~ attended.uni[indices.to.sample]))

Call:
lm(formula = support.leave[indices.to.sample] ~ attended.uni[indices.to.sample])

Residuals:
    Min      1Q  Median      3Q     Max 
-0.5987 -0.5987  0.4013  0.4013  0.5550 

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)    
(Intercept)                      0.59874    0.01960  30.544  < 2e-16 ***
attended.uni[indices.to.sample] -0.15370    0.03219  -4.774 2.07e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4932 on 1004 degrees of freedom
Multiple R-squared:  0.0222,Adjusted R-squared:  0.02123 
F-statistic: 22.79 on 1 and 1004 DF,  p-value: 2.071e-06
```

Very close to our estimates of the standard errors from simulation.

The most important use of regression standard errors: hypothesis testing

The most important use of regression standard errors: hypothesis testing

Suppose the coefficient on *AttendedUniversity* in your sample is -0.154, as in this regression output.

The most important use of regression standard errors: hypothesis testing

Suppose the coefficient on *AttendedUniversity* in your sample is -0.154, as in this regression output.

We want to know: have you proven conclusively that university attendance is related to Brexit support *in the population*, or might it just be a fluke *in your sample*?

The most important use of regression standard errors: hypothesis testing

Suppose the coefficient on *AttendedUniversity* in your sample is -0.154, as in this regression output.

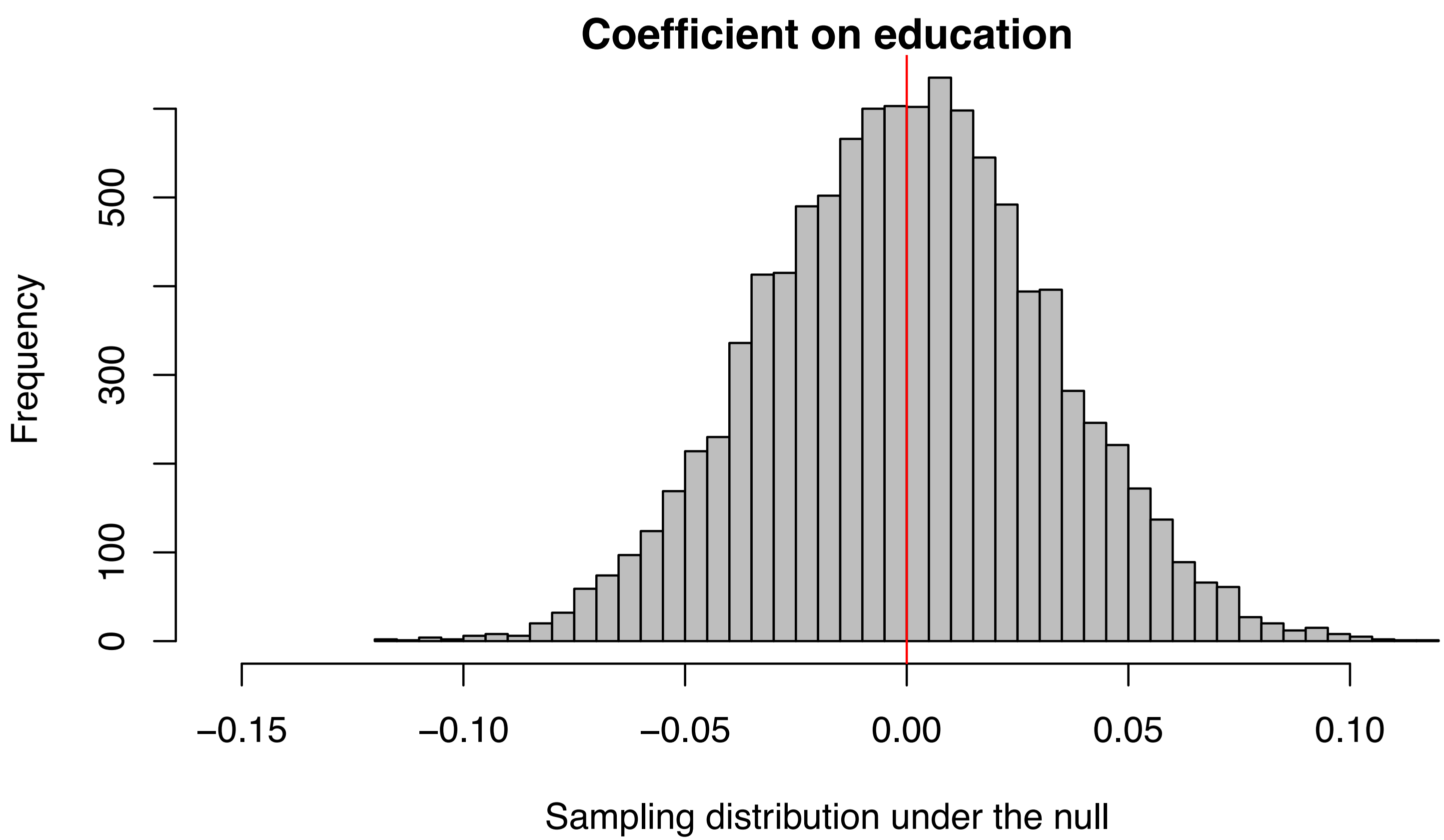
We want to know: have you proven conclusively that university attendance is related to Brexit support *in the population*, or might it just be a fluke *in your sample*?

Put differently: How likely is it that you would get a coefficient that far from 0 in your sample if the true coefficient were in fact 0?

Hypothesis testing for regression coefficients (2)

Hypothesis testing for regression coefficients (2)

The sampling distribution of the coefficient on AttendedUni, **if the true coefficient were 0**, would be something like



Hypothesis testing for regression coefficients (3)

Hypothesis testing for regression coefficients (3)

So what's the probability of getting a sample that yields a coefficient as far from zero as your estimate? (p-value)

Hypothesis testing for regression coefficients (3)

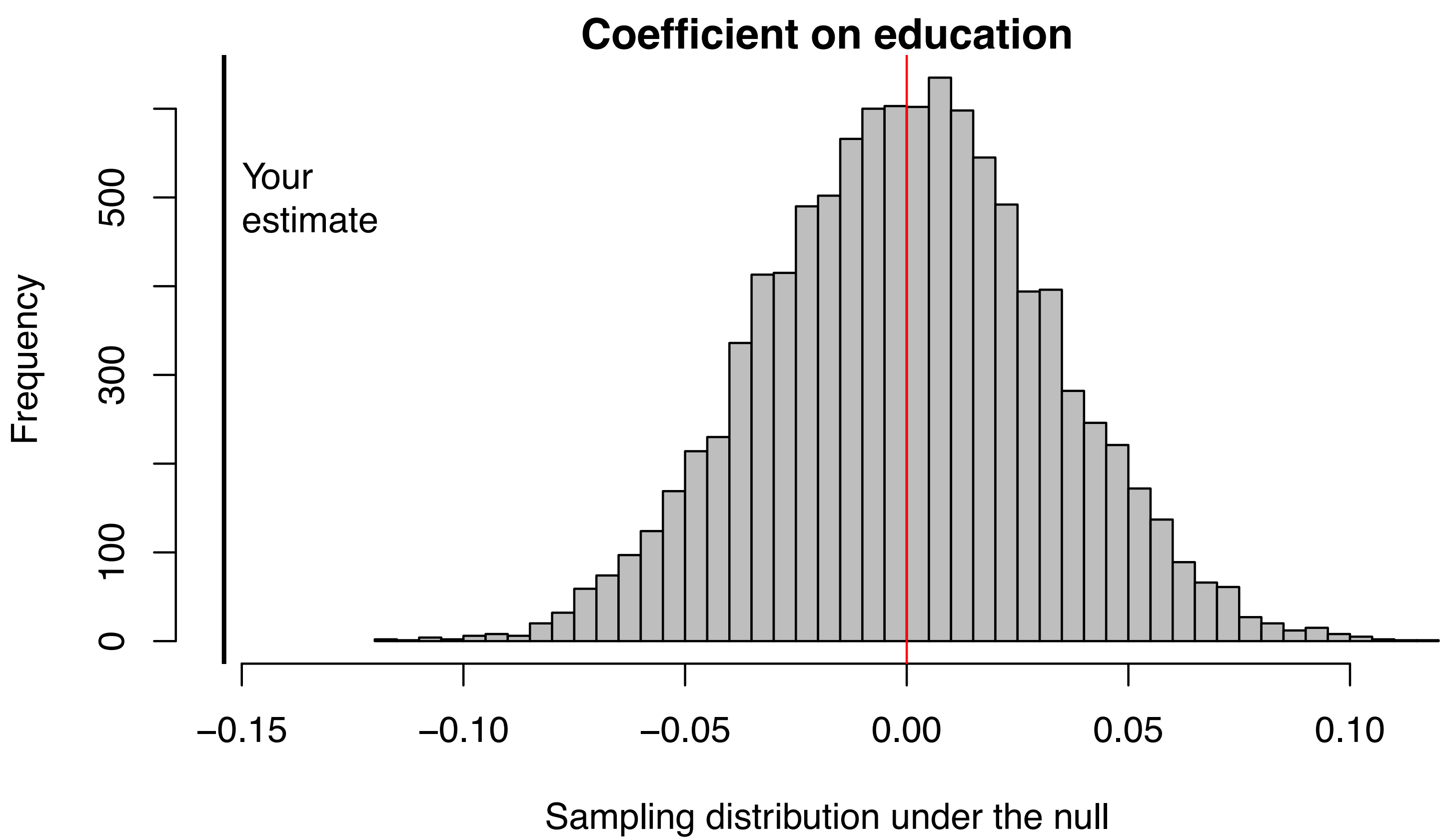
So what's the probability of getting a sample that yields a coefficient as far from zero as your estimate? (p-value)

Basically zero in this case!

Hypothesis testing for regression coefficients (3)

So what's the probability of getting a sample that yields a coefficient as far from zero as your estimate? (p-value)

Basically zero in this case!



Hypothesis testing in regression output

Hypothesis testing in regression output

Output from a regression for one sample:

Hypothesis testing in regression output

Output from a regression for one sample:

```
> summary(lm(support.leave[indices.to.sample] ~ attended.uni[indices.to.sample]))

Call:
lm(formula = support.leave[indices.to.sample] ~ attended.uni[indices.to.sample])

Residuals:
    Min      1Q  Median      3Q     Max 
-0.5987 -0.5987  0.4013  0.4013  0.5550 

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)    
(Intercept)                      0.59874    0.01960  30.544  < 2e-16 ***
attended.uni[indices.to.sample] -0.15370    0.03219  -4.774 2.07e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4932 on 1004 degrees of freedom
Multiple R-squared:  0.0222,Adjusted R-squared:  0.02123 
F-statistic: 22.79 on 1 and 1004 DF,  p-value: 2.071e-06
```

Hypothesis testing in regression output

Output from a regression for one sample:

```
> summary(lm(support.leave[indices.to.sample] ~ attended.uni[indices.to.sample]))

Call:
lm(formula = support.leave[indices.to.sample] ~ attended.uni[indices.to.sample])

Residuals:
    Min      1Q  Median      3Q     Max 
-0.5987 -0.5987  0.4013  0.4013  0.5550 

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)    
(Intercept)                      0.59874    0.01960  30.544  < 2e-16 ***
attended.uni[indices.to.sample] -0.15370    0.03219  -4.774 2.07e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4932 on 1004 degrees of freedom
Multiple R-squared:  0.0222,Adjusted R-squared:  0.02123 
F-statistic: 22.79 on 1 and 1004 DF,  p-value: 2.071e-06
```

Basically zero.

Now you should understand:

Dependent variable: Nobel Prizes awarded per capita (in log scale)

	(1)	(2)	(3)
Intercept	-1.629* (0.509)	-3.166* (0.511)	-2.982* (0.527)
Chocolate consumption per capita (log scale)	2.092* (0.298)	1.026* (0.326)	0.709 (0.415)
GDP/capita (thousands of USD)		0.105* (0.024)	0.106* (0.024)
NW Europe			0.549 (0.452)
R^2	0.70	0.85	0.86
N	34	34	34

Standard errors in parentheses. * Indicates p<0.05

- what a dependent variable is
- what an independent variable is
- what the coefficients mean (intercept, slopes)
- what the stars mean (i.e. what $p<0.05$ means)
- what the standard errors mean

To consider

To consider

In the thought experiments above, there are standard errors because of **sampling**. But what about when the sample **is** the population? (e.g. Lijphart's analysis of all countries that have been democratic since 1988)

To consider

In the thought experiments above, there are standard errors because of **sampling**. But what about when the sample **is** the population? (e.g. Lijphart's analysis of all countries that have been democratic since 1988)

The broader view is that history offers one sample, but if "re-run" it might have produced another. Less philosophically satisfying!