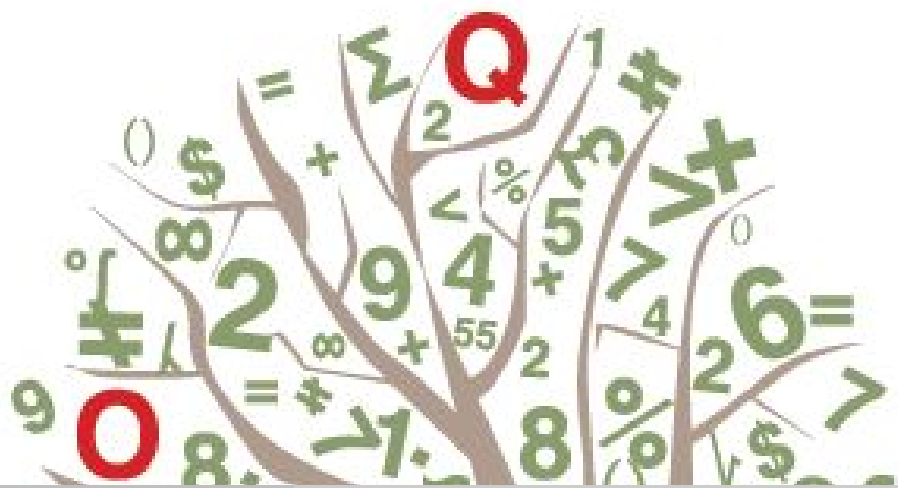
Q-Step: Week 7 Lecture

Statistical Inference

Spyros Kosmidis

Oxford

February 27, 2022



Roadmap

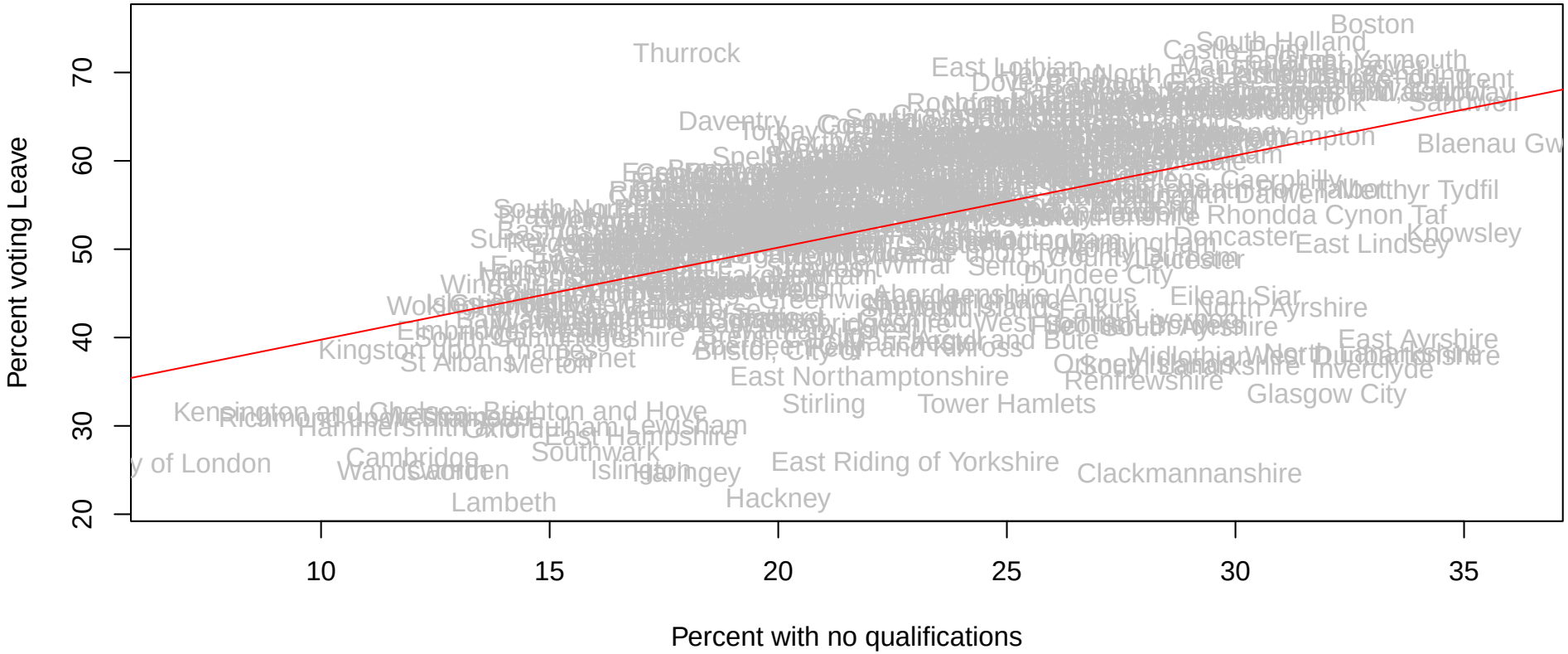
Previously

- Research Design
- Concepts and Measurement
- Descriptive statistics
- Case Selection
- Correlation & Regression

Today

- Statistical Inference

Bivariate OLS



The OLS output in R

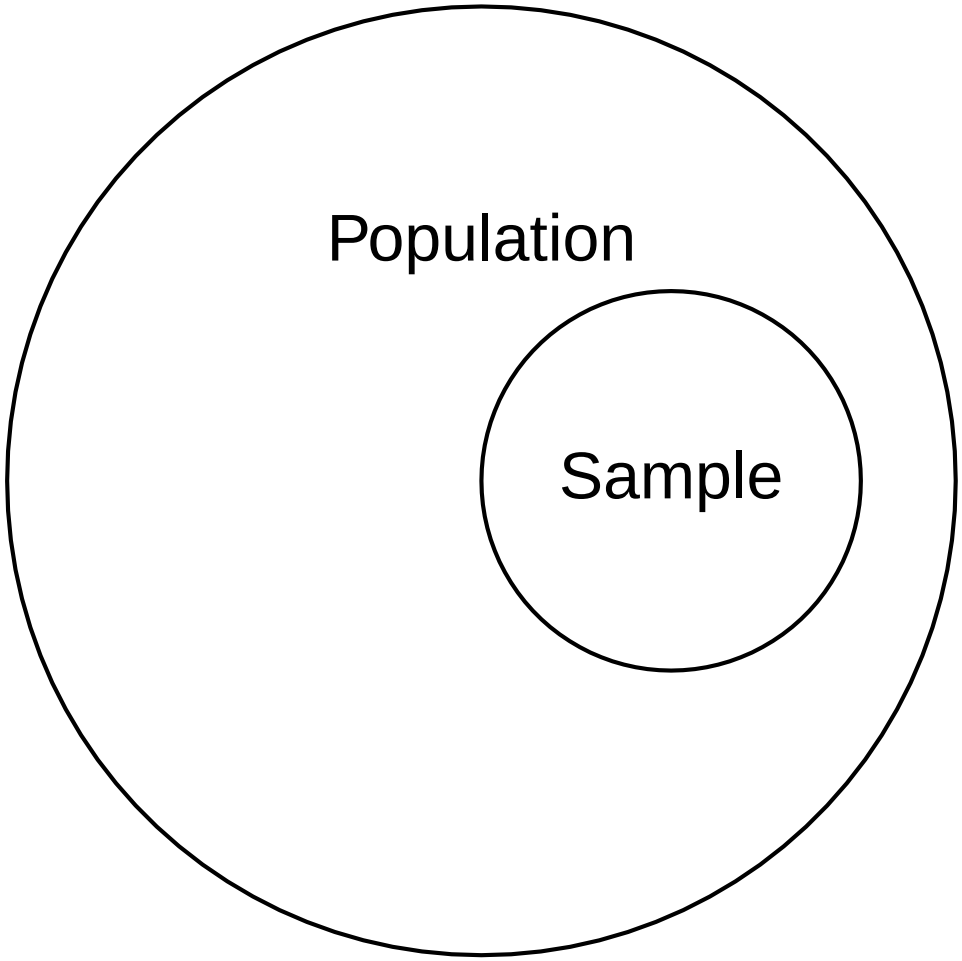
```
##
## Call:
## lm(formula = brexit$leave ~ brexit$noqual)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.855  -3.593   1.971   5.958  24.182
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   29.33773    2.08661   14.06   <2e-16 ***
## brexit$noqual  1.04234    0.08911   11.70   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.945 on 377 degrees of freedom
## Multiple R-squared:  0.2663, Adjusted R-squared:  0.2643
## F-statistic: 136.8 on 1 and 377 DF,  p-value: < 2.2e-16
```

Samples and Populations

- Population: The entire set of cases that our theory applies to
- Sample: Subset of cases that we analyse

Samples and Populations

- Population: The entire set of cases that our theory applies to
- Sample: Subset of cases that we analyse



Samples and Populations

- A key aim of social science is to use a sample to say something about a wider population
- This is what statistical inference is
- Samples are not perfect;

Samples and Populations

- A key aim of social science is to use a sample to say something about a wider population
- This is what statistical inference is
- Samples are not perfect;
  - When we draw a sample there is always a margin of error
  - When we estimate the slope of a regression, there is a standard error
- We make guesses about the population of interest and our confidence is summarized in the p-values/confidence intervals we get in our outputs
- We want to know whether the results we have are systematic or random!

Samples and Populations

- Imagine you are interested in drawing a sample from the population of GB.
- You know that the actual population parameter of interest is k.
- This k can be the average support a given policy
- The sample you draw allows you to estimate a sample mean $\bar{x}$.
- If there is no bias in your sample then it might well be that $k = \bar{x}$
- If there is sampling error then $k \neq \bar{x}$
- But statistical theory allows us to say that $k = \bar{x} + error$

Samples and Populations: A simulated Example

```
sample1=sample(x=c(0,1), replace=T, size=1000,
prob=c(0.48,0.52))
mean(sample1)
```

```
## [1] 0.518
```

Samples and Populations: A simulated Example

```
sample1=sample(x=c(0,1), replace=T, size=1000,
prob=c(0.48,0.52))
mean(sample1)
```

```
## [1] 0.518
```

```
sample2=sample(x=c(0,1), replace=T, size=1000,
prob=c(0.48,0.52))
mean(sample2)
```

```
## [1] 0.551
```

```
sample3=sample(x=c(0,1), replace=T, size=1000,
prob=c(0.48,0.52))
mean(sample3)
```

```
## [1] 0.539
```

```
sample4=sample(x=c(0,1), replace=T, size=1000,
prob=c(0.48,0.52))
mean(sample4)
```

```
## [1] 0.512
```

Samples and Populations: A simulated Example

```
sample1=sample(x=c(0,1), replace=T, size=1000,
prob=c(0.48,0.52))
mean(sample1)
```

```
## [1] 0.518
```

```
sample2=sample(x=c(0,1), replace=T, size=1000,
prob=c(0.48,0.52))
mean(sample2)
```

```
## [1] 0.551
```

```
sample3=sample(x=c(0,1), replace=T, size=1000,
prob=c(0.48,0.52))
mean(sample3)
```

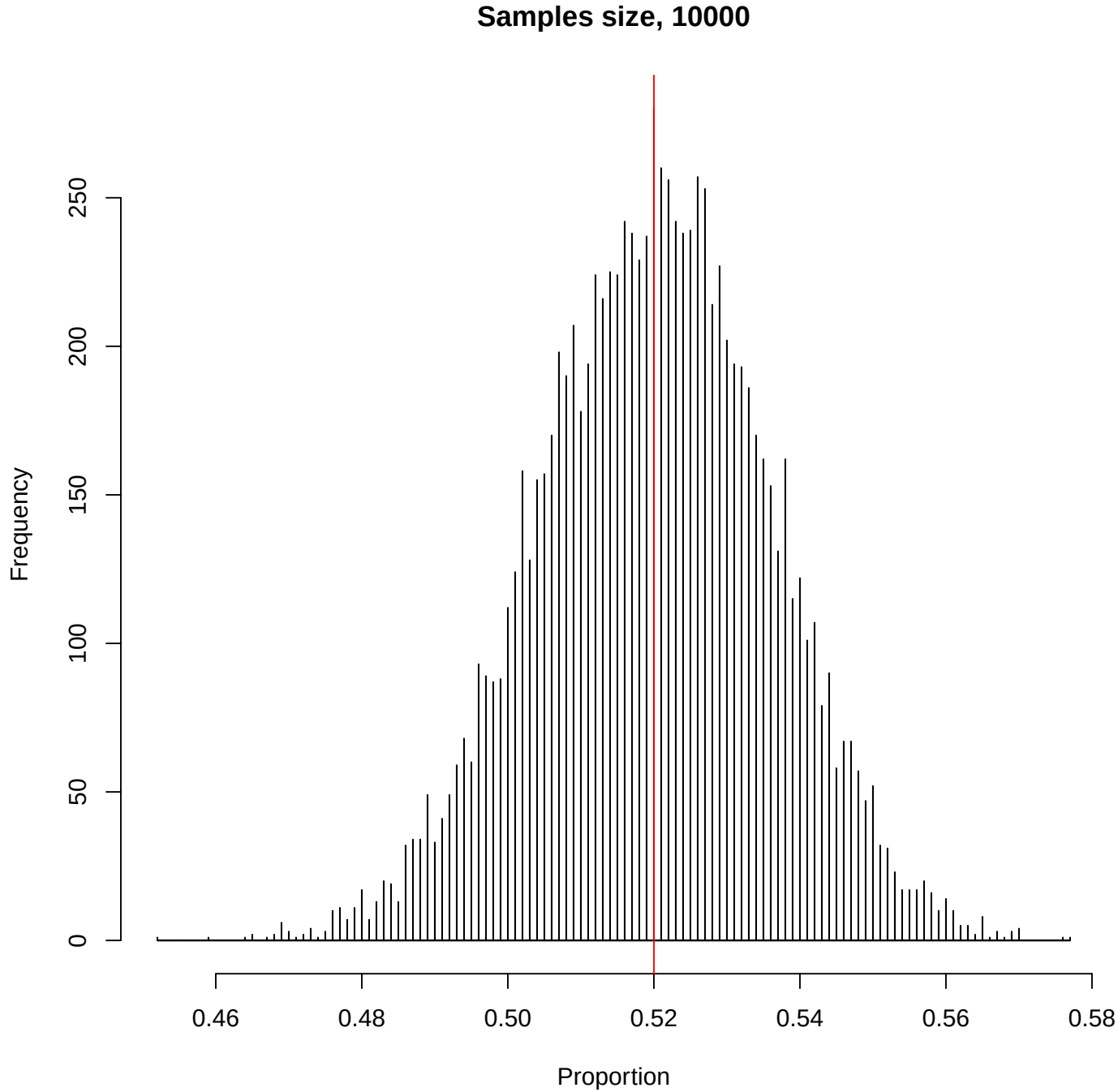
```
## [1] 0.539
```

```
sample4=sample(x=c(0,1), replace=T, size=1000,
prob=c(0.48,0.52))
mean(sample4)
```

```
## [1] 0.512
```

```
sample10000=sample(x=c(0,1), replace=T, size=1000,
prob=c(0.48,0.52))
mean(sample10000)
```

```
## [1] 0.511
```


Samples size, 10000
Frequency
0
50
100
150
200
250
0.46
0.48
0.50
0.52
0.54
0.56
0.58
Proportion

The error

- How can we get the error in a single sample?
- Answer: The Central Limit Theorem allows us to claim that in samples of a given size n, our variable will follow a normal distribution and the standard deviation will be given by

$$\sqrt{\frac{Variance}{n}}$$

The error

- How can we get the error in a single sample?
- Answer: The Central Limit Theorem allows us to claim that in samples of a given size n, our variable will follow a normal distribution and the standard deviation will be given by

$$\sqrt{\frac{Variance}{n}}$$

```
sqrt(var(sample1)/1000)
```

```
## [1] 0.01580905
```

```
sqrt(var(sample2)/1000)
```

```
## [1] 0.01573679
```

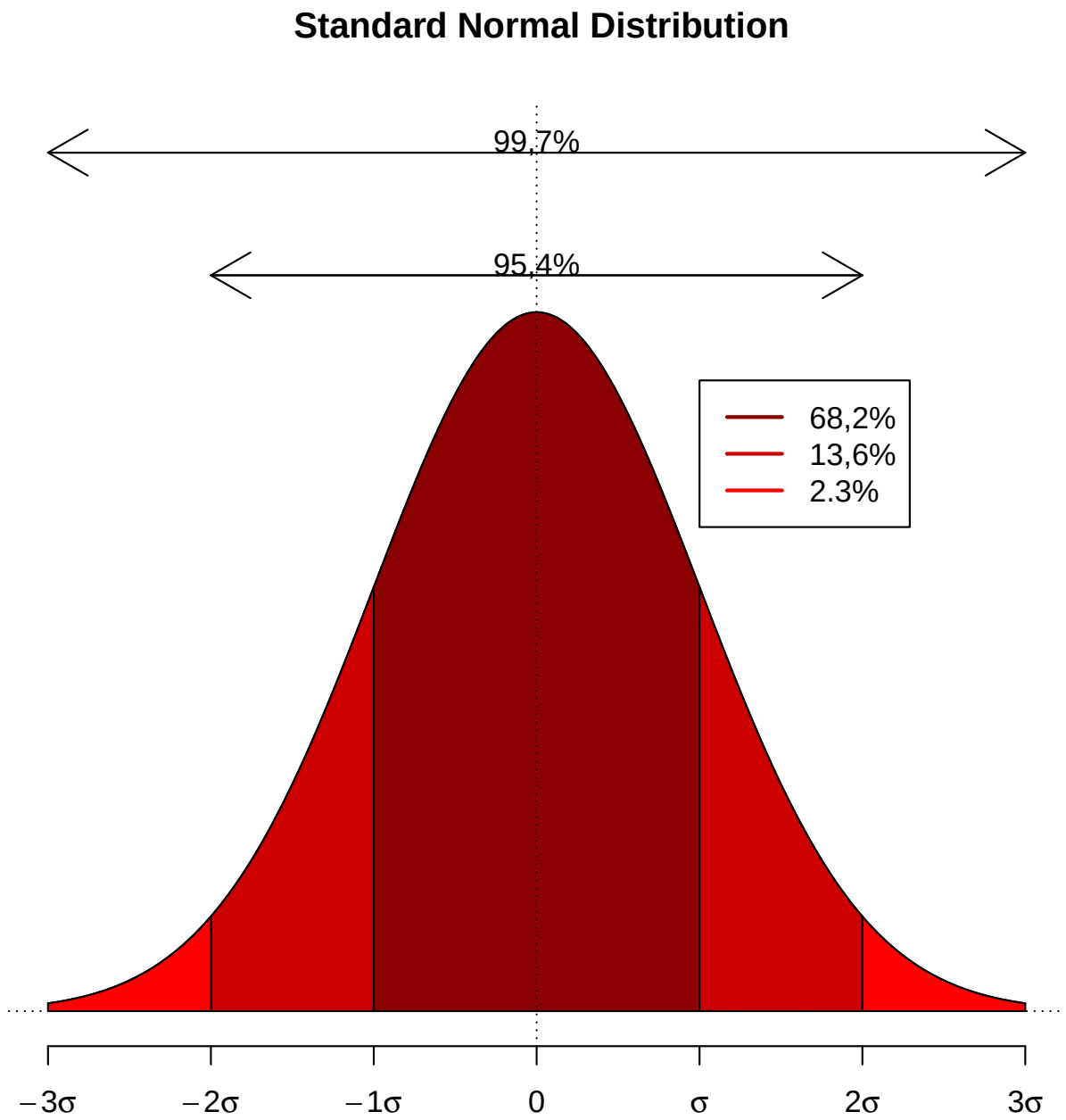
```
sqrt(var(sample10000)/1000)
```

```
## [1] 0.01581547
```

Central Limit Theorem (in a nutshell)

- If you repeat an experiment a large number of times, your summary statistic (e.g., mean value, count of votes) will be dispersed around the expected "true" value.
- Sometimes you will be below the expected value, sometimes above. But more often you will be closer to the true value, and less often you'll be very far off.
- The more you repeat the experiment, the closer you will get to the true value if you average all your results from all the experiments (see also "law of large numbers").
- In fact, if you repeat the experiment many times, your results will be distributed in the form of a Bell curve (normal distribution) around the true expected value.

The Bell and its properties



Implications of the Error

- I guess you have already heard of the idea that we want to be 95% confident about our results
- If 1.96 standard deviations (or standard errors) above and below the mean contain 95% of the data then we can claim that our population parameter k is our sample estimate $\bar{X}$ plus 1.96 errors above and below the sample mean.
- This leads to claiming that the upper limit of confidence is

$$CI95\%_{upper} = \bar{X} + 1.96SE$$

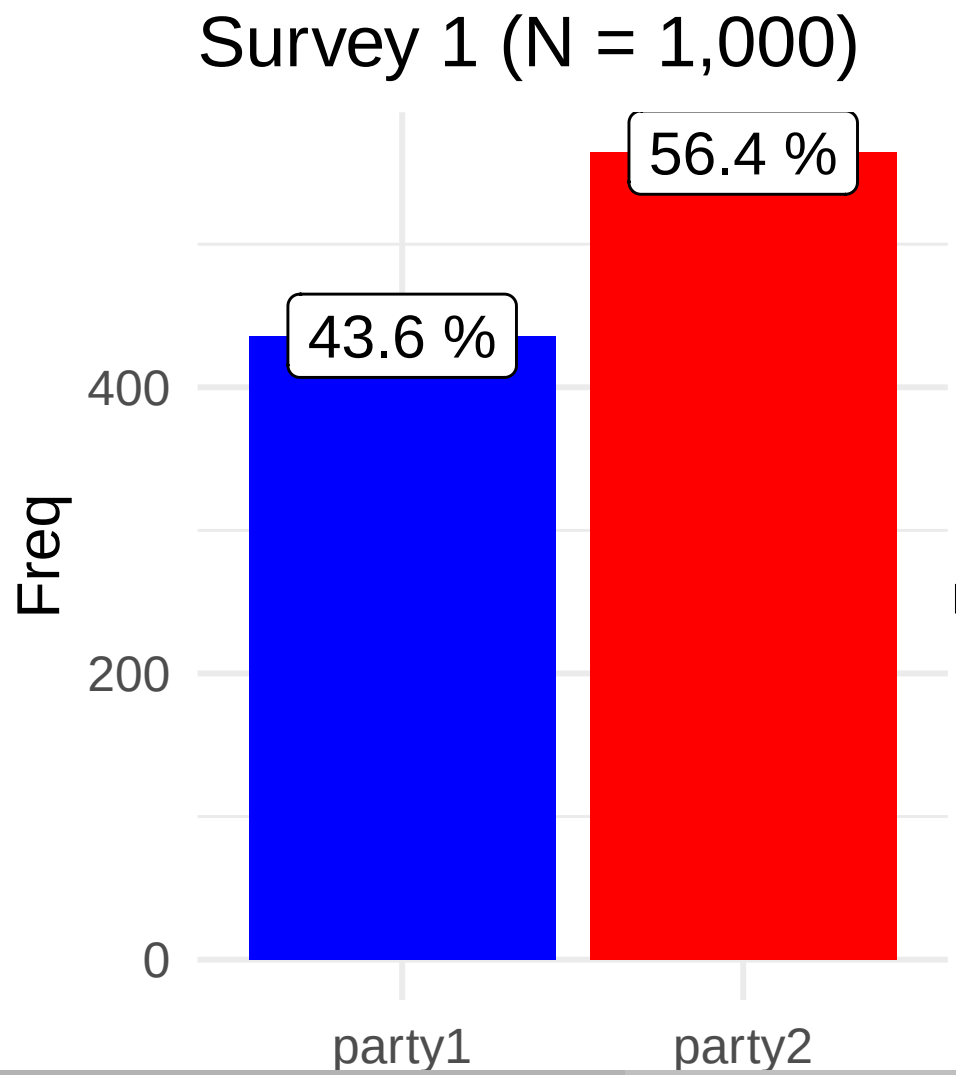
- The lower bound is given by

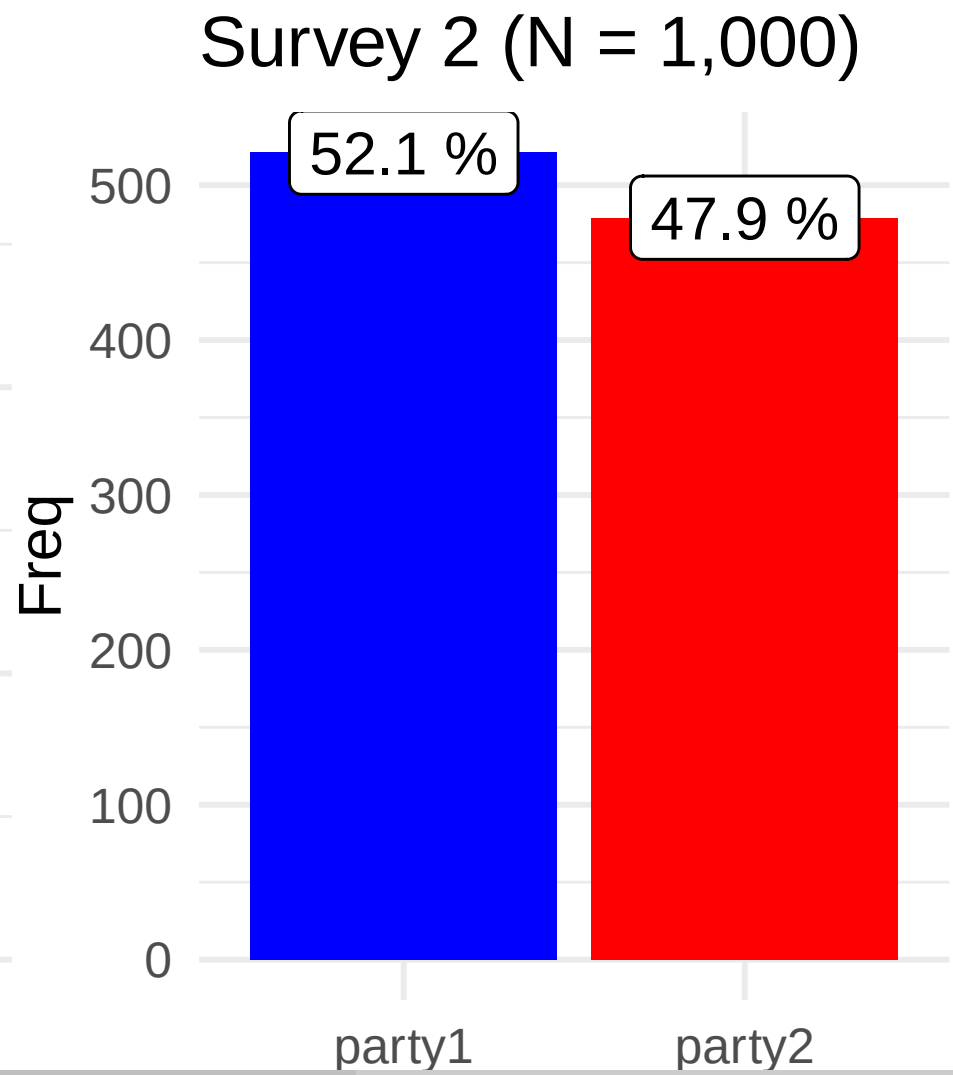
$$CI95\%_{lower} = \bar{X} - 1.96SE$$

- In other words, we want to be confident that our true parameter will be included in 19 out of 20 confidence intervals

A Second Illustration

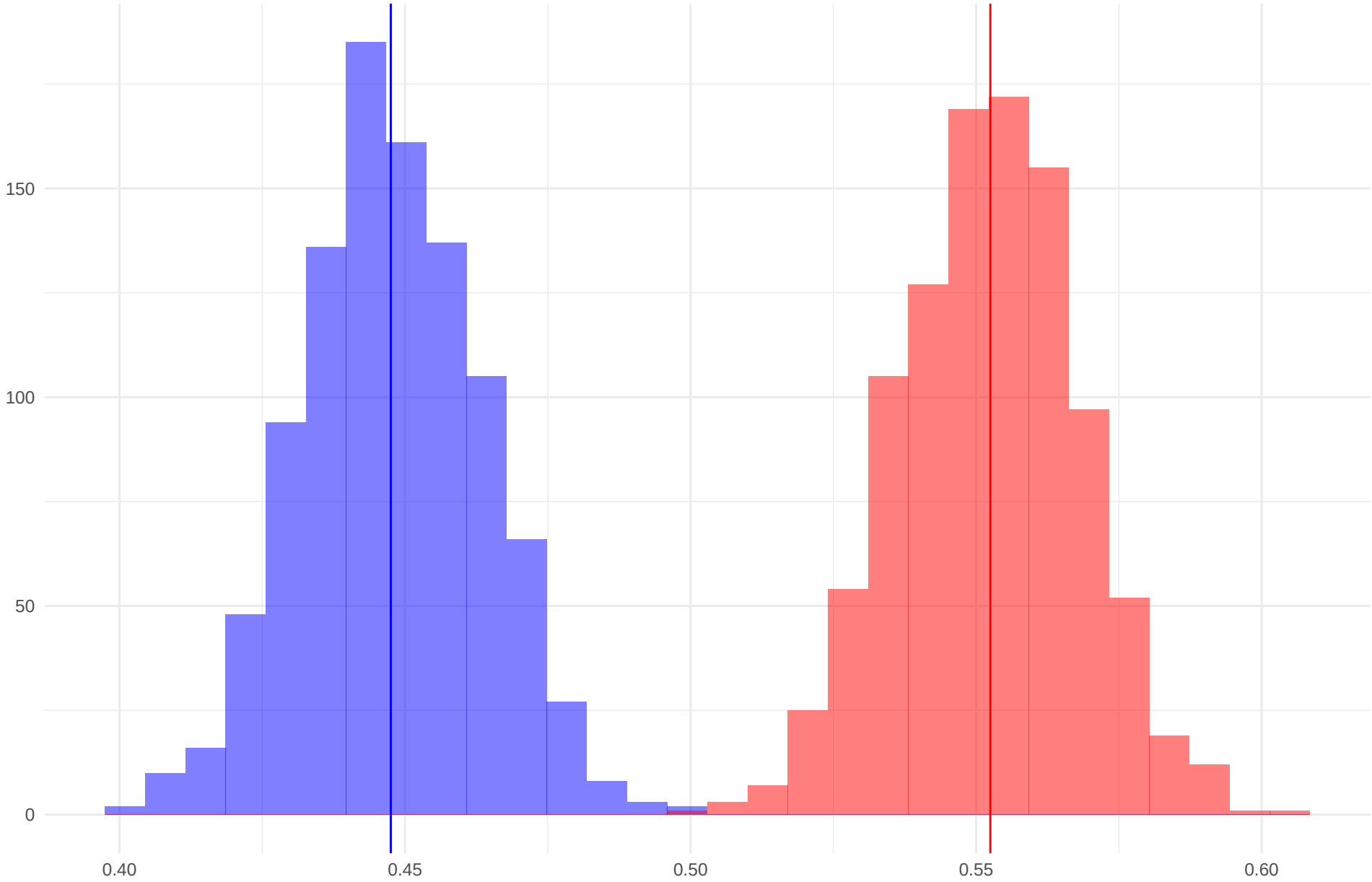
- Imagine an election where the true result is 0.44 for Party1 and 0.56 for Party2.
- On the day of the election we also conducted two separate surveys that gave us the following results.





What if we had 1000 surveys of 1000 respondents?

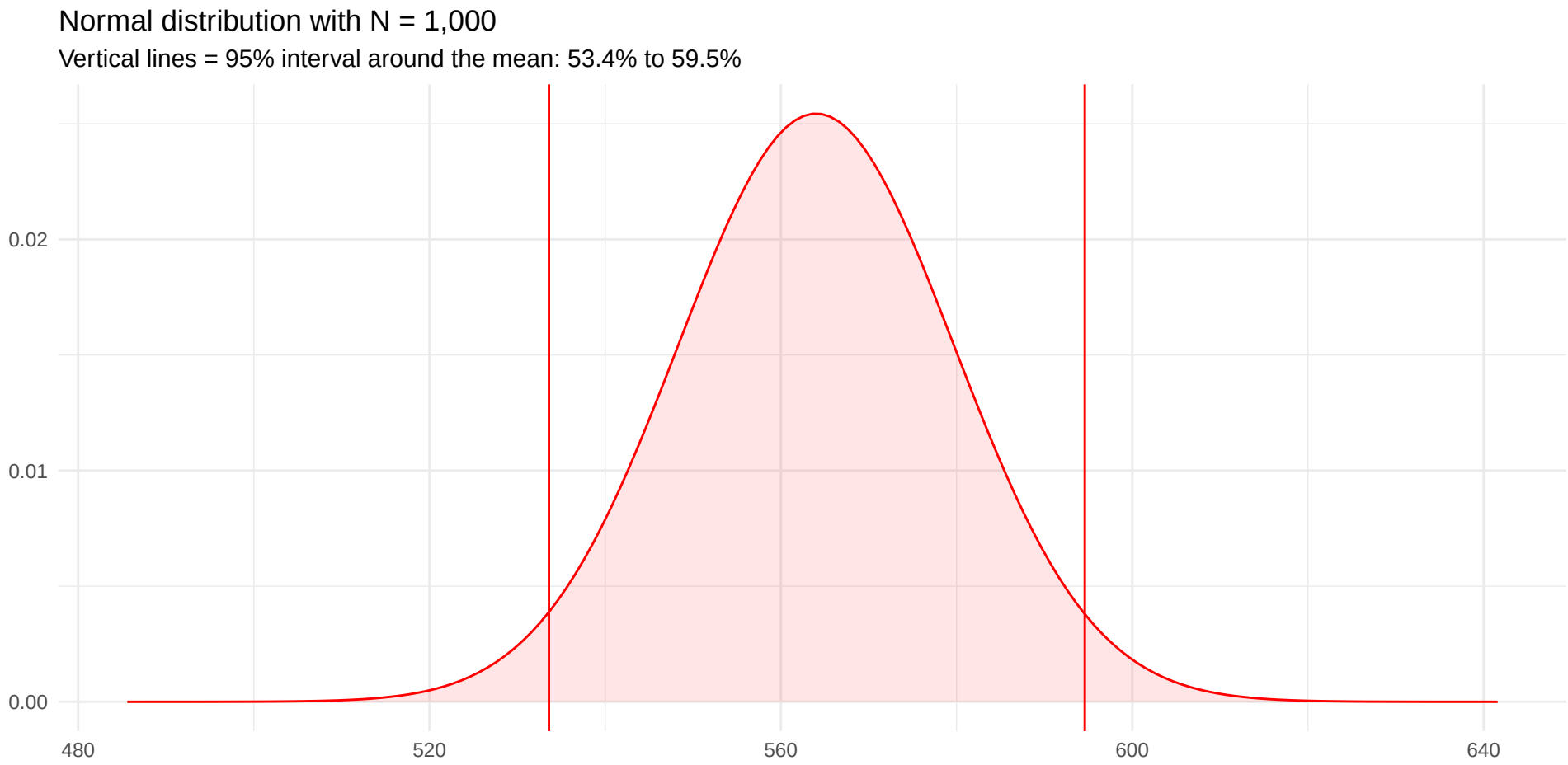
Histograms of simulated values for Party2 (red) and Party1 (blue) percentages



A first view of the confidence interval

Let's assume we are polling 1000 respondents and we find the 56% of voters prefer Party1

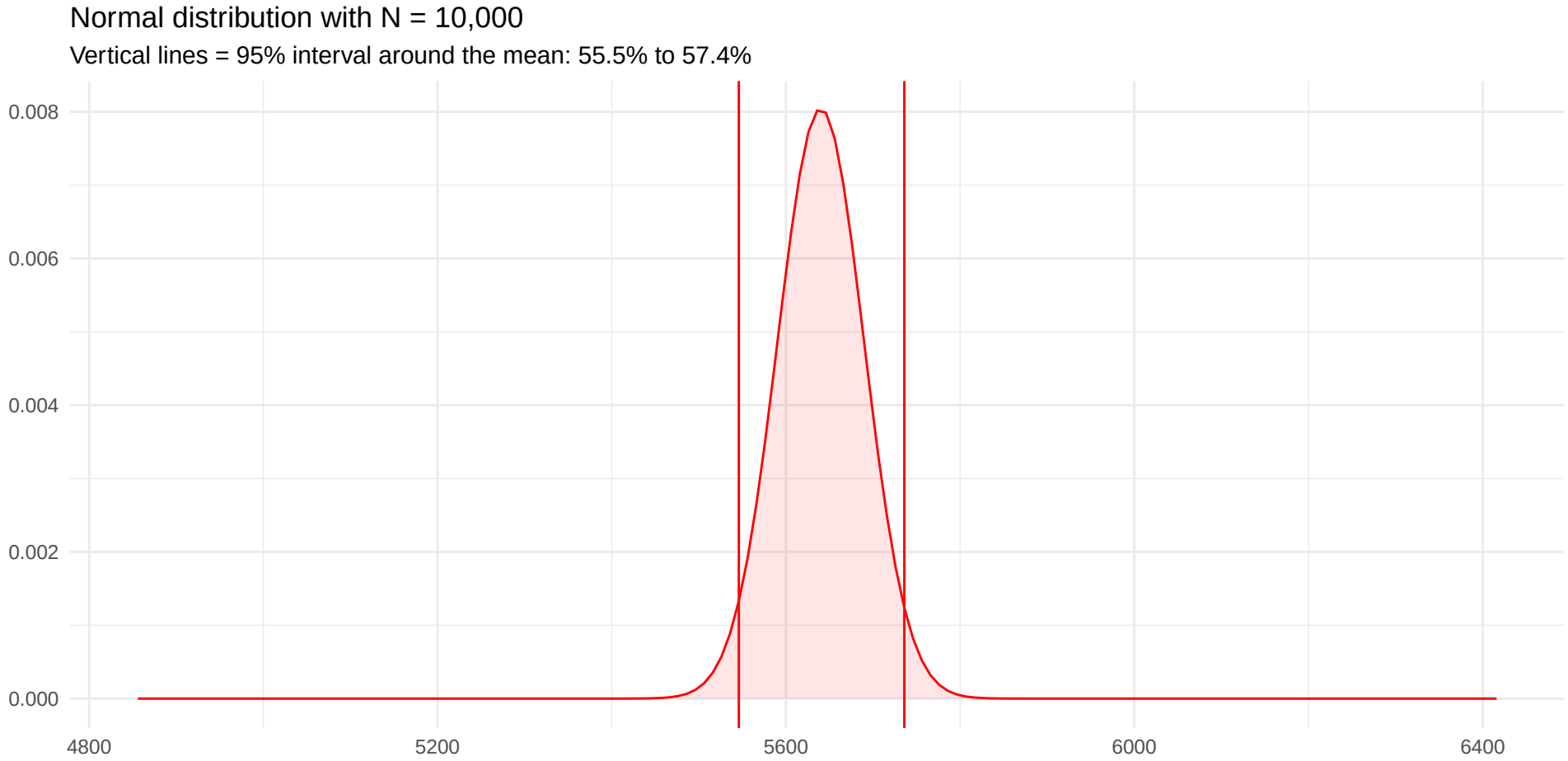
```
## [1] 0.44751 0.55249
```



A first view of the confidence interval

- What is this showing us?
- If the share of Party 1 votes was 56.4%, then repeating a survey many times with N = 1,000 voters, 95% of our results would range between 53.4% and 59.5%
- This is what we sometimes call the margin of error!
- With more N we can make it smaller:

Sample size and confidence



p-values?

- The previous figures show ranges of of values around the sample mean.
- If the point of interest is the 50% margin, then the question we ask ourselves is, "Do we cross that value when we also consider the 'error'?"
- An alternative way to ask the same questions is by saying, "How likely is it that what we calculated in our sample is far from the 'true' value?"
- We call this likelihood -or better this probability- as the p-value.
- What is the probability that the population parameter can be as big/extreme as our data (sample statistic)?

p-values

Null hypothesis: Party 1 votes = Party 2 votes

vs. the result of our survey

What about OLS coefficients?

- Coefficients behave in exactly the same way.
- They are sample parameters that describe a relationship at the level of the population
- These estimates come with uncertainty (SE) and we use the uncertainty to draw statistical inferences
- The p-values or the confidence interval allow us to say whether a given estimate is statistically different from a given benchmark. In our case 0!
- This is why we often call this the Null hypothesis
- We are asking; can we reject the Null hypothesis ($b = 0$)?
- If the p-value is smaller than 0.05 we choose the alternative hypothesis
- What does this mean in practice?

What about OLS coefficients?

- Coefficients behave in exactly the same way.
- They are sample parameters that describe a relationship at the level of the population
- These estimates come with uncertainty (SE) and we use the uncertainty to draw statistical inferences
- The p-values or the confidence interval allow us to say whether a given estimate is statistically different from a given benchmark. In our case 0!
- This is why we often call this the Null hypothesis
- We are asking; can we reject the Null hypothesis ($b = 0$)?
- If the p-value is smaller than 0.05 we choose the alternative hypothesis
- What does this mean in practice?Our results are statistically significant!
- Where is this coming from?

Hypothesis tests

- In common regression output you get three important pieces of information;

Hypothesis tests

- In common regression output you get three important pieces of information; the coefs,

Hypothesis tests

- In common regression output you get three important pieces of information; the coefs, the standard error,

Hypothesis tests

- In common regression output you get three important pieces of information; the coefs, the standard error, the t-ratio,

Hypothesis tests

- In common regression output you get three important pieces of information; the coefs, the standard error, the t-ratio, and a p-value

Hypothesis tests

- In common regression output you get three important pieces of information; the coefs, the standard error, the t-ratio, and a p-value
- These are the basic ingredients for a hypothesis test
- Is the coefficient we get a random fluke in the data (because of say sampling error)?
- Or is it systematic?
- A p-value tells how likely it is that one would get a coefficient that far from 0 in the sample, if the true coefficient were in fact 0?
- Using confidence intervals;

Hypothesis tests

- In common regression output you get three important pieces of information; the coefs, the standard error, the t-ratio, and a p-value
- These are the basic ingredients for a hypothesis test
- Is the coefficient we get a random fluke in the data (because of say sampling error)?
- Or is it systematic?
- A p-value tells how likely it is that one would get a coefficient that far from 0 in the sample, if the true coefficient were in fact 0?
- Using confidence intervals;Is 0 included in our calculated 95% CI?

Back to OLS

```
##
## Call:
## lm(formula = brexit$leave ~ brexit$noqual)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.855  -3.593   1.971   5.958  24.182
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   29.33773    2.08661   14.06   <2e-16 ***
## brexit$noqual  1.04234    0.08911   11.70   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.945 on 377 degrees of freedom
## Multiple R-squared:  0.2663, Adjusted R-squared:  0.2643
## F-statistic: 136.8 on 1 and 377 DF,  p-value: < 2.2e-16
```

Back to OLS

```
##
## Call:
## lm(formula = brexit$leave ~ brexit$noqual)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.855  -3.593   1.971   5.958  24.182
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   29.33773    2.08661   14.06   <2e-16 ***
## brexit$noqual  1.04234    0.08911   11.70   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.945 on 377 degrees of freedom
## Multiple R-squared:  0.2663, Adjusted R-squared:  0.2643
## F-statistic: 136.8 on 1 and 377 DF,  p-value: < 2.2e-16
```

- Now the confidence intervals

Back to OLS

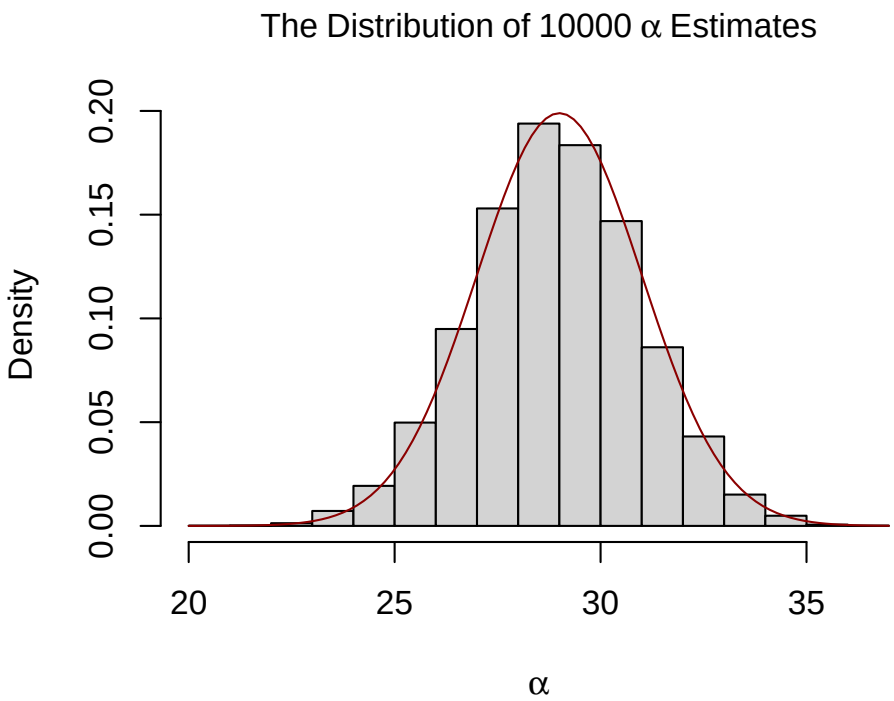
```
##
## Call:
## lm(formula = brexit$leave ~ brexit$noqual)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.855  -3.593   1.971   5.958  24.182
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   29.33773    2.08661   14.06   <2e-16 ***
## brexit$noqual  1.04234    0.08911   11.70   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.945 on 377 degrees of freedom
## Multiple R-squared:  0.2663, Adjusted R-squared:  0.2643
## F-statistic: 136.8 on 1 and 377 DF,  p-value: < 2.2e-16
```

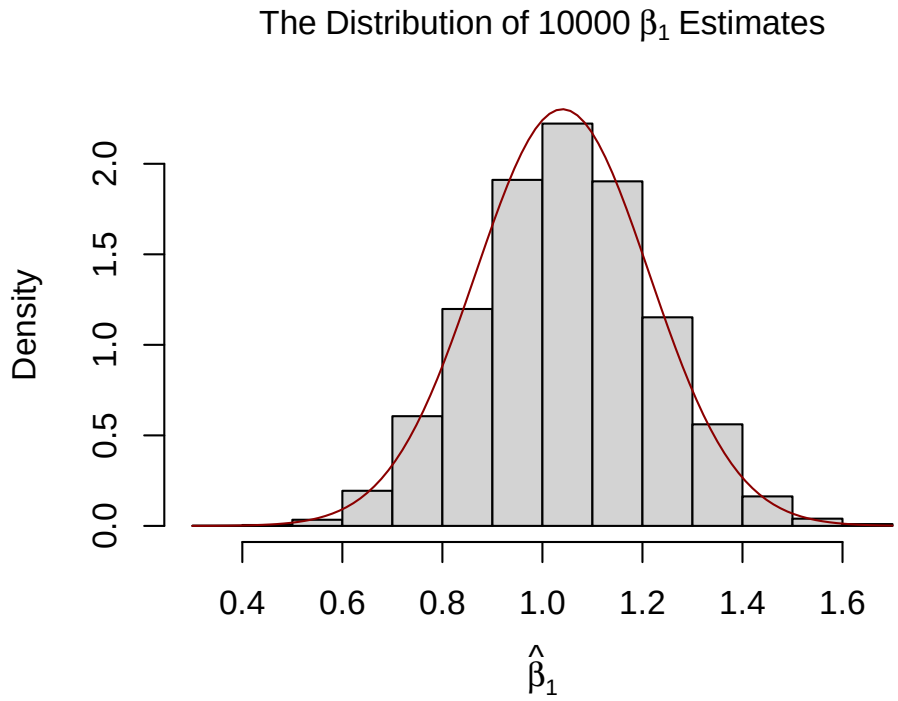
- Now the confidence intervals

```
confint.lm(noqual)
```

```
##                    2.5 %    97.5 %
## (Intercept)   25.2348677 33.440587
## brexit$noqual  0.8671266  1.217552
```

Sampling distribution of the coefficients





Hypothesis tests

- Our estimated sample statistic (i.e. the coef), is compared to a distribution of coefficients that given the sample characteristics it centered 0 (i.e. no effect).
- How likely is it that our estimated coefficient is included in that distribution?
- This is an alternative way of thinking about the p-value.
- In general, however, I would be happy if you can make the very basic evaluations of your output
- If you are able to say if an estimate is statistically significant, then you are in good shape
- Your second year Q-Step will include a more comprehensive account of inference!

Thank you!