Inference

POLI SCI 210

Introduction to Empirical Methods in Political Science

AI Prompts

- Central limit theorem
- Weak law of large numbers
- Confidence interval
- Standard deviation vs. standard error
- What is a p-value? How to interpret it?

Reminder: These are just suggestions, feel free to be creative with your prompts!

Last week

- Theory as explanation
- Elements of a good theory
- Data as the raw material from which we draw inferences
- Types of data (variables)

Why do we care so much about data?

This week: Inference

- We use data to conduct **inference**
- **Inference:** Using what we know to understand something we do not know
- ***Statistical* inference:** Using the *data* we have to understand something for which we do *not* have data
- Data alone does not help unless we **summarize** it
- **TUESDAY:** *Estimation* and *uncertainty*
- **THURSDAY:** Hypothesis testing

Running example

Position	Height (inches)
1	62.5
2	63.0
3	63.0
4	66.0
5	66.0
6	67.0
7	68.5
8	70.0
9	70.0
10	70.5
11	70.7

- How would you describe this group of people?
- Taller/shorter than most?
- How do we summarize this data?

Running example

Position	Height (inches)
1	62.5
2	63.0
3	63.0
4	66.0
5	66.0
6	67.0
7	68.5
8	70.0
9	70.0
10	70.5
11	70.7

Measures of central tendency

- **Mean:** $\frac{1}{n} \sum x_i$
- **Median:** Value in the middle (after ordering)
- **Mode:** Value most often repeated

Running example

Position	Height (inches)
1	62.5
2	63.0
3	63.0
4	66.0
5	66.0
6	67.0
7	68.5
8	70.0
9	70.0
10	70.5
11	70.7

Measures of central tendency

- **Mean =** 67
- **Median =** 67.0
- **Mode =** 63.0, 66.0, 70.0

Running example

Position	Height (inches)
1	62.5
2	63.0
3	63.0
4	66.0
5	66.0
6	67.0
7	68.5
8	70.0
9	70.0
10	70.5
11	70.7

Measures of central tendency

- **Mean =** 67
- **Median =** 67.0
- **Mode =** 63.0, 66.0, 70.0

Which is a better summary?

Running example

Position	Height (inches)
1	62.5
2	63.0
3	63.0
4	66.0
5	66.0
6	67.0
7	68.5
8	70.0
9	70.0
10	70.5
11	70.7

Measures of central tendency

- **Mean =** 67
- **Median =** 67.0
- **Mode =** 63.0, 66.0, 70.0

Which is a better summary?

Short answer: The **mean** because it has some nice statistical properties

Long answer: Take a step back

Question from the 2022 General Social Survey

In 2020, you remember that Joe Biden ran for President on the Democratic ticket against Donald Trump for the Republicans.

Long answer: Take a step back

Question from the 2022 General Social Survey

In 2020, you remember that Joe Biden ran for President on the Democratic ticket against Donald Trump for the Republicans. *Do you remember for sure whether or not you voted in that election?*

- Yes, I voted
- No, I did not vote
- I was not eligible to vote

Learn more about the GSS at gss.norc.org

Dataset

```
# A tibble: 3,876 × 1
     vote
     <chr>
   1 yes
   2 yes
   3 yes
   4 yes
   5 yes
   6 yes
   7 no
   8 yes
   9 yes
  10 yes
  11 yes
  12 yes
  13 yes
```

Dataset

vote	n
no	998
yes	2878

- Proportion yes = 0.74
- Pretend this is the entire population

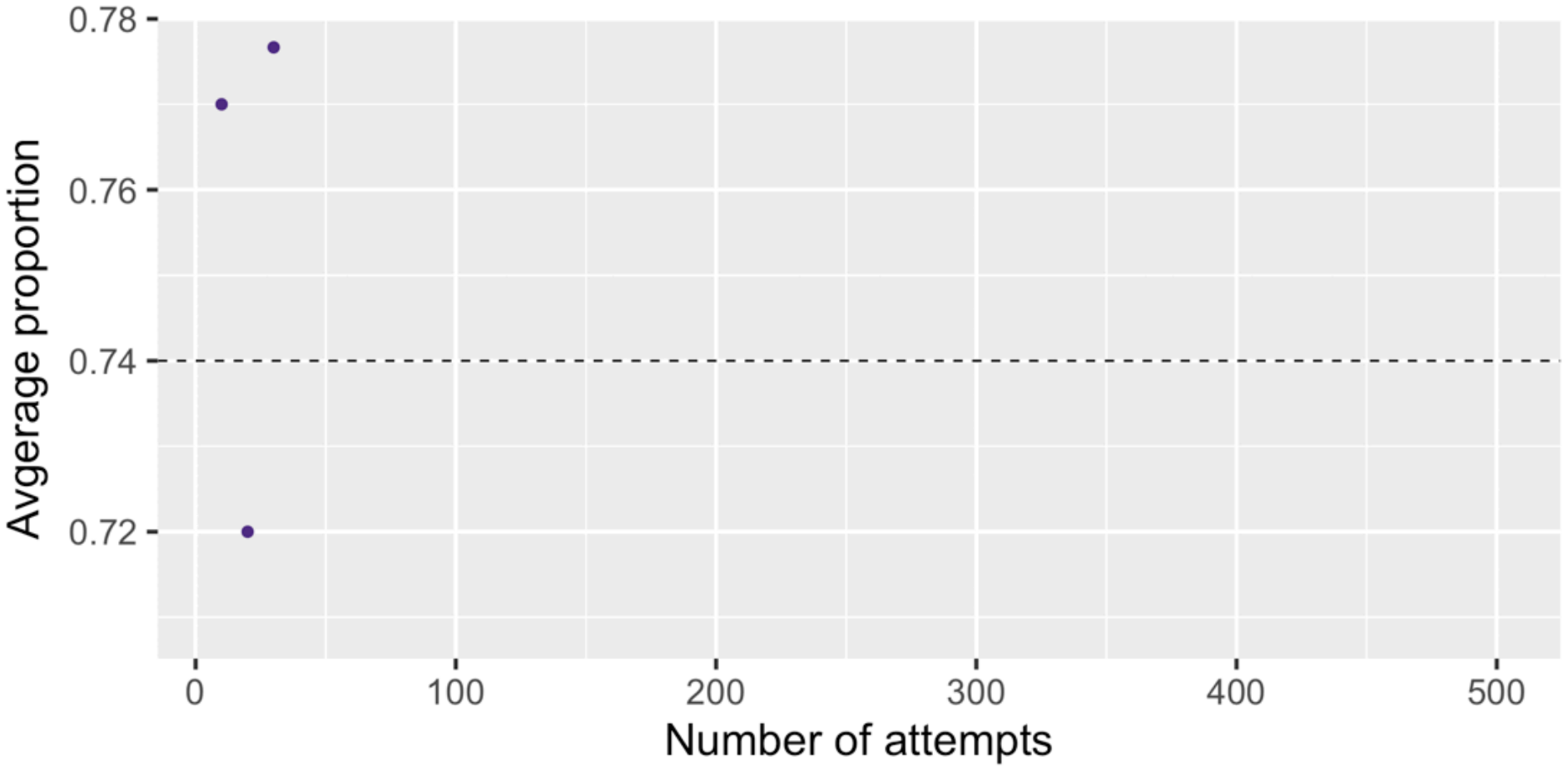
- But we cannot ask almost 4,000 people directly
- We only have time for 10
- So let's take a *random sample* and see what happens

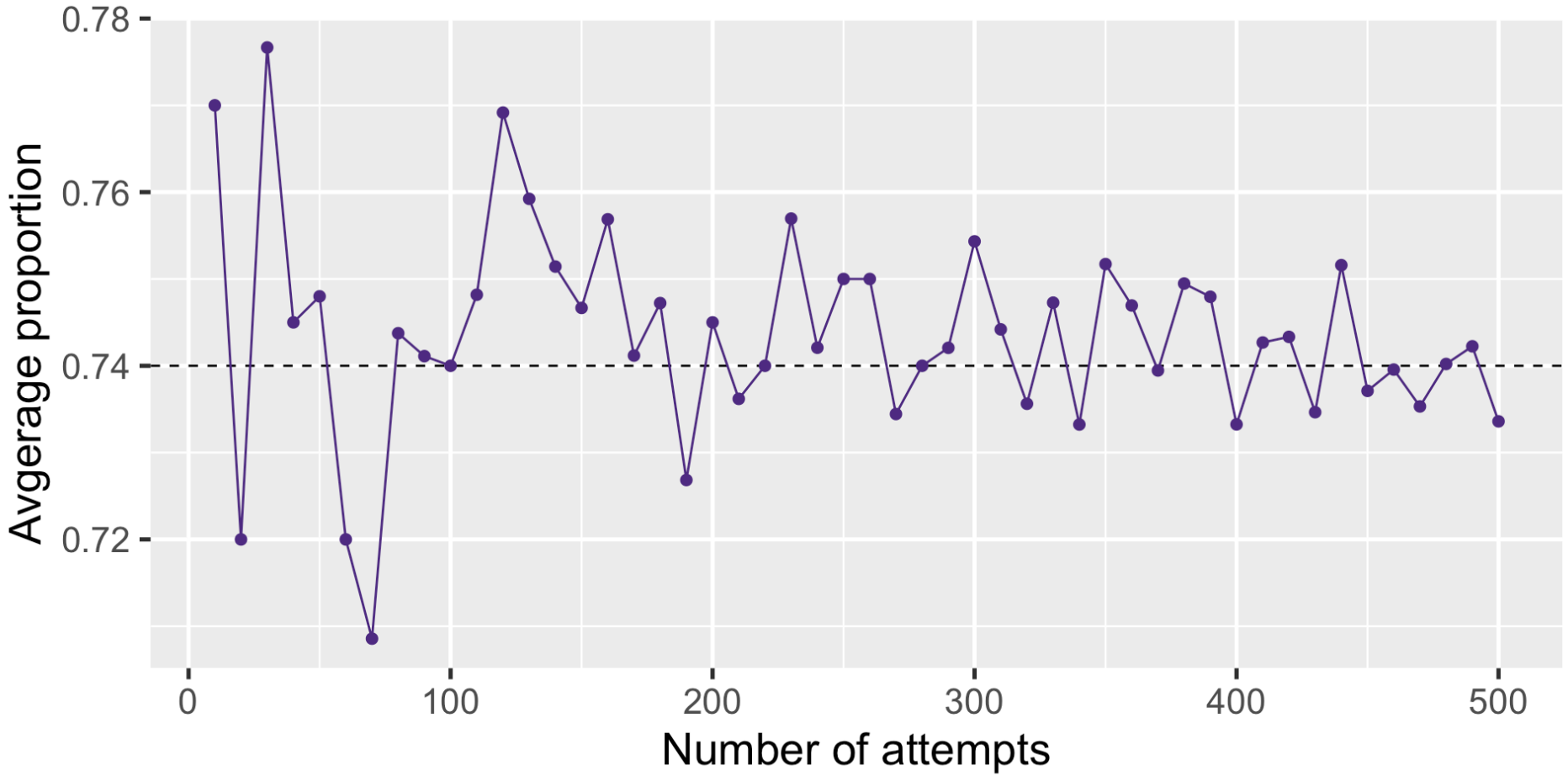
Asking 10 people

- Attempt 1: 0.8
- Attempt 2: 0.9
- Attempt 3: 0.9
- Attempt 4: 0.7
- Attempt 5: 0.7

Average: 0.78

What if we tried more and more attempts?


0.78
0.76
0.74
0.72
Avgerage proportion
0
100
200
300
400
500
Number of attempts



Weak law of large numbers: Sample mean *approximates* the true population mean over many repeated measurements

So what?

If we have time to ask 10 people many times…

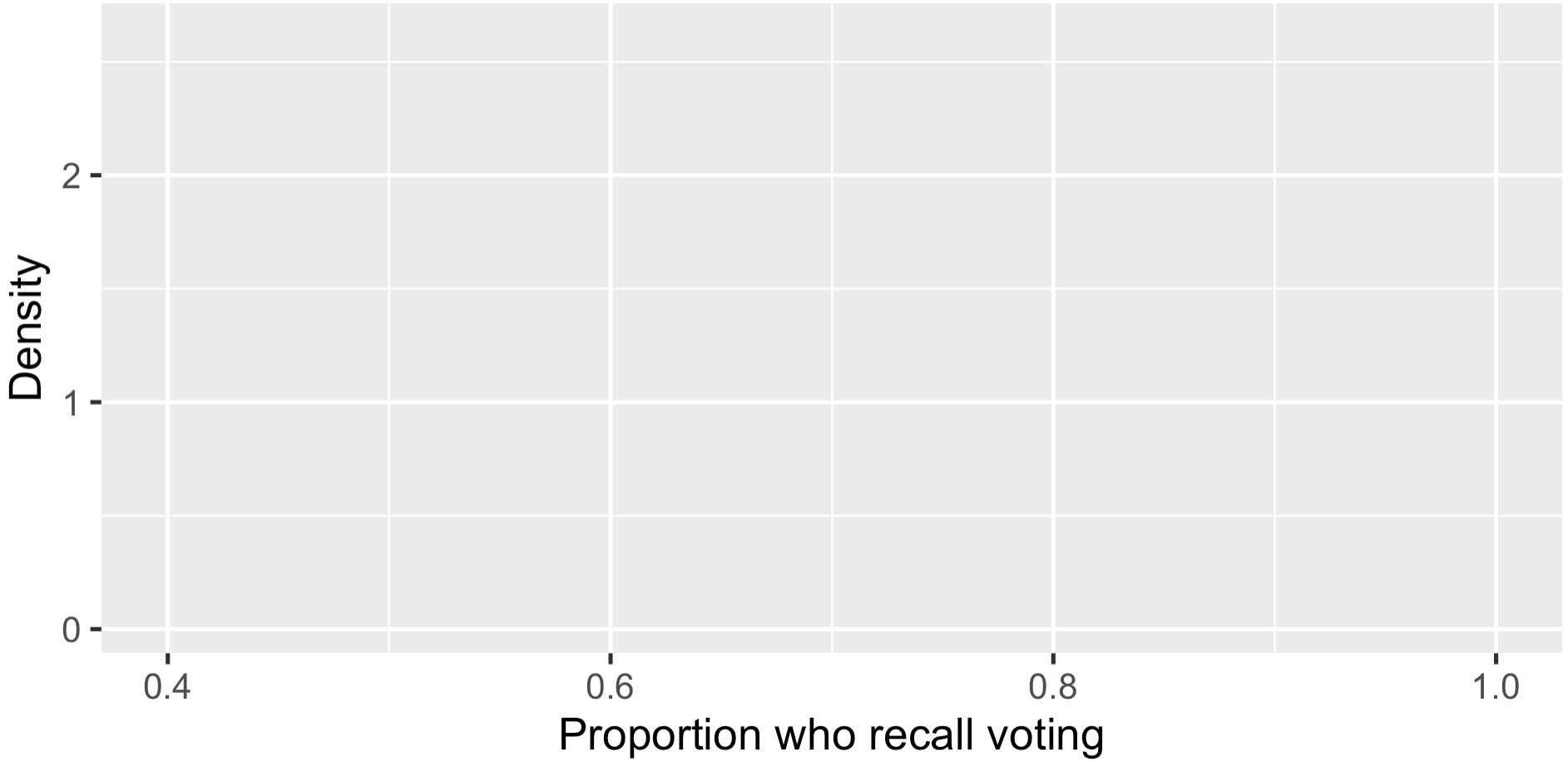
Maybe we also have time to ask many people one time?

How many people would be enough?

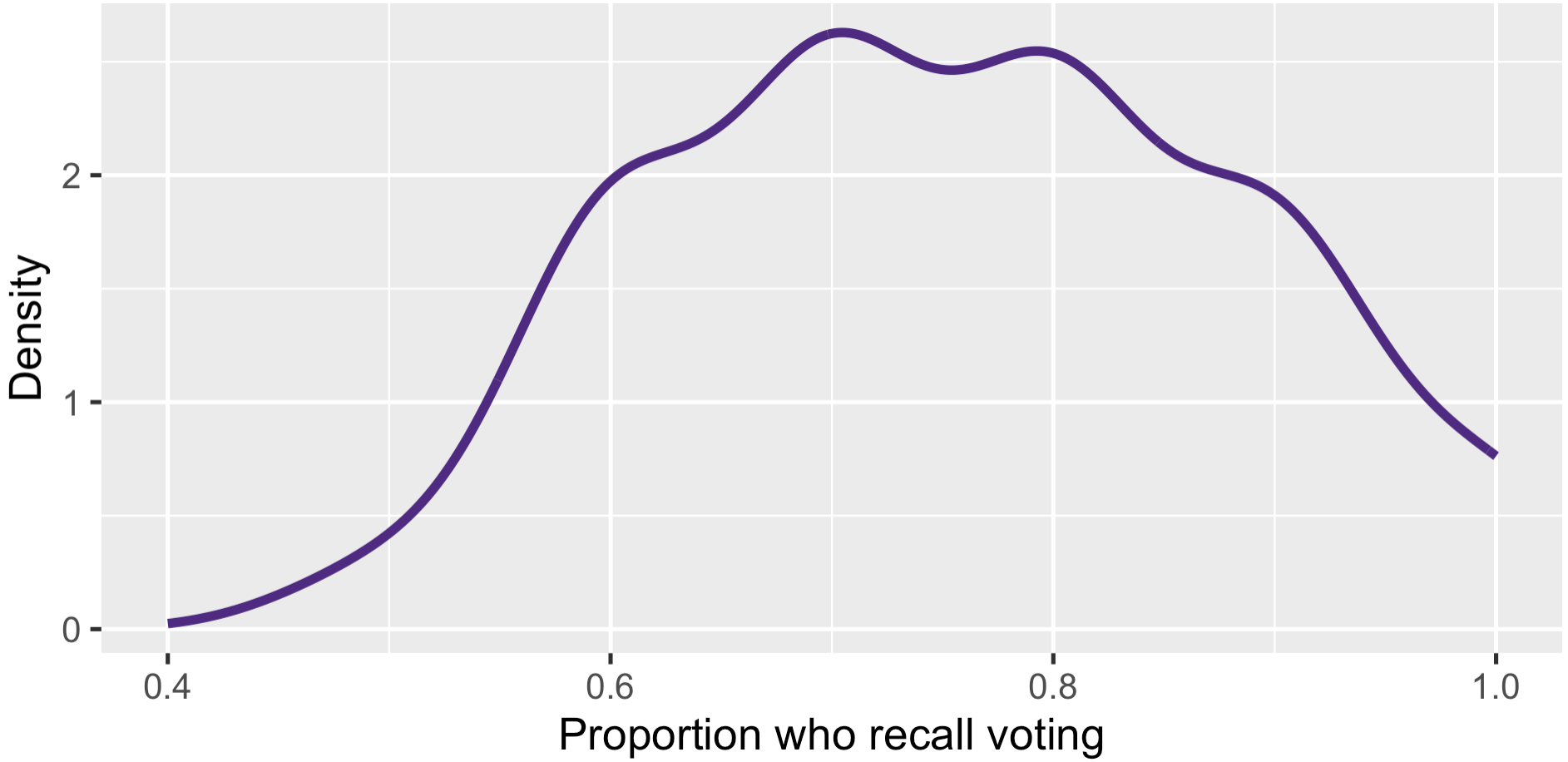
Another exercise:

- Pick a sample size
- Repeat the study with a new sample many times
- Calculate proportion who recall voting every time

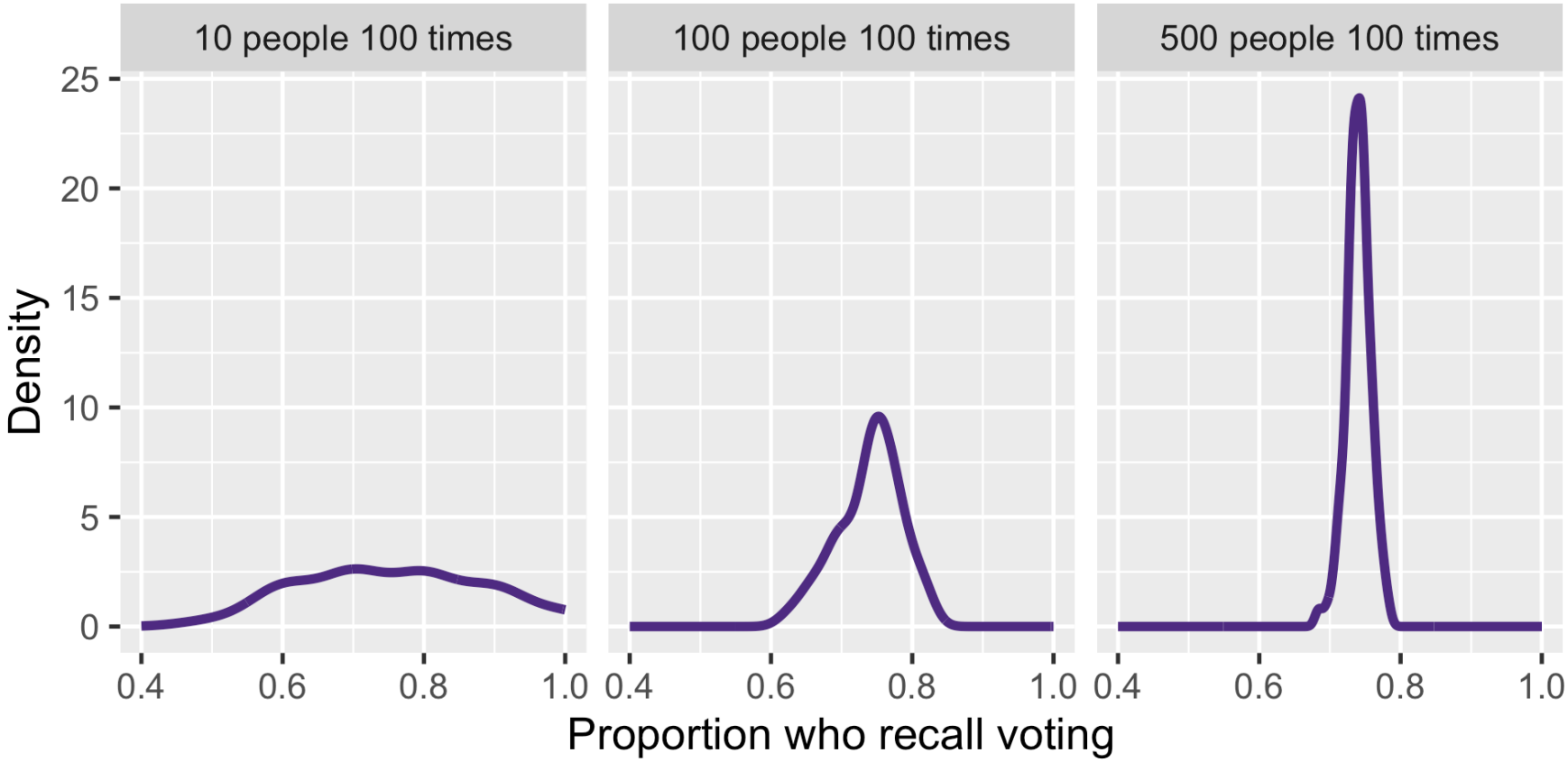
100 samples of 10 people each



100 samples of 10 people each



What about more people?



Central limit theorem: Mean converges to a normal distribution as the sample size increases

Statistical properties

The mean is *usually* a good summary of the data because it has good **finite sample** and **asymptotic** properties

Finite sample properties

- Over many repetitions with a fixed sample size
- **Example:** Weak law of large numbers

Asymptotic properties

- As sample size grows to infinity
- **Example:** Central limit theorem (CLT)

This is also true for fancier mean-like things that we will learn about this term

The mean is a good summary

The mean is a good summary

Ok, but how good?

The mean is a good summary

Ok, but how good?

Are we cooking or are we cooked?

The mean is a good summary

Ok, but how good?

Are we cooking or are we cooked?

We want to *quantify* our **confidence** or **uncertainty**

Back to heights

Position	Height
1	62.5
2	63.0
3	63.0
4	66.0
5	66.0
6	67.0
7	68.5
8	70.0
9	70.0
10	70.5
11	70.7

- The mean was 67
- The *range* is [62.5, 70.7]
- Is that informative?

Some extremes of height. Angus McAskill (1825-1863) and Charles Sherwood Stratton (1838-1883). McAskill was 7 feet 9 inches tall. Stratton, also known as Tom Thumb, was 2 feet 6 inches in height.

Maybe remove extremes?

Position	Height
1	62.5
2	63.0
3	63.0
4	66.0
5	66.0
6	67.0
7	68.5
8	70.0
9	70.0
10	70.5
11	70.7

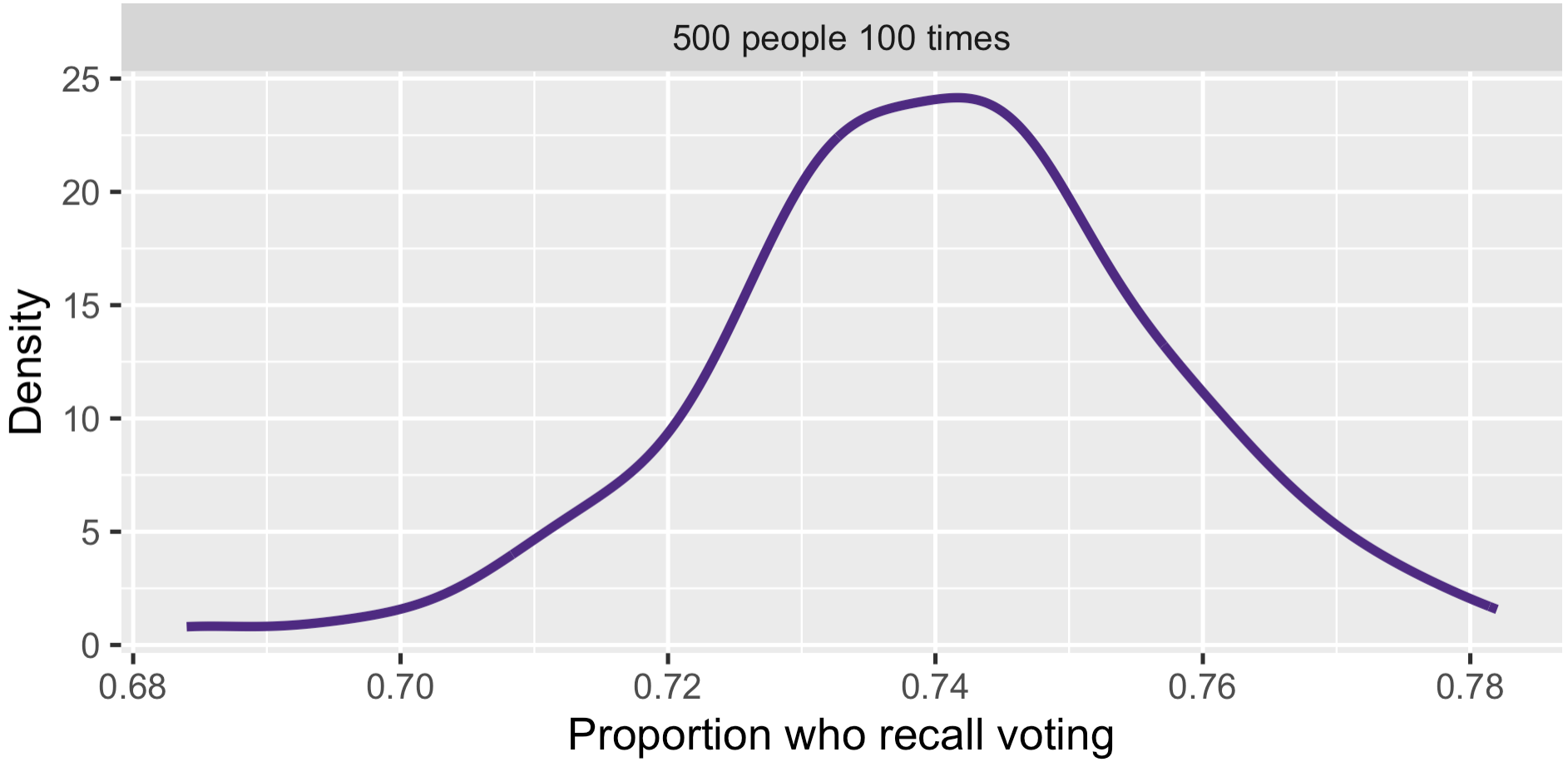
- Original range is [62.5, 70.7]
- Taking out extremes gives [63.0, 70.5]
- We can't keep doing this with a large sample!
- Is there a more principled way?

Use percentiles

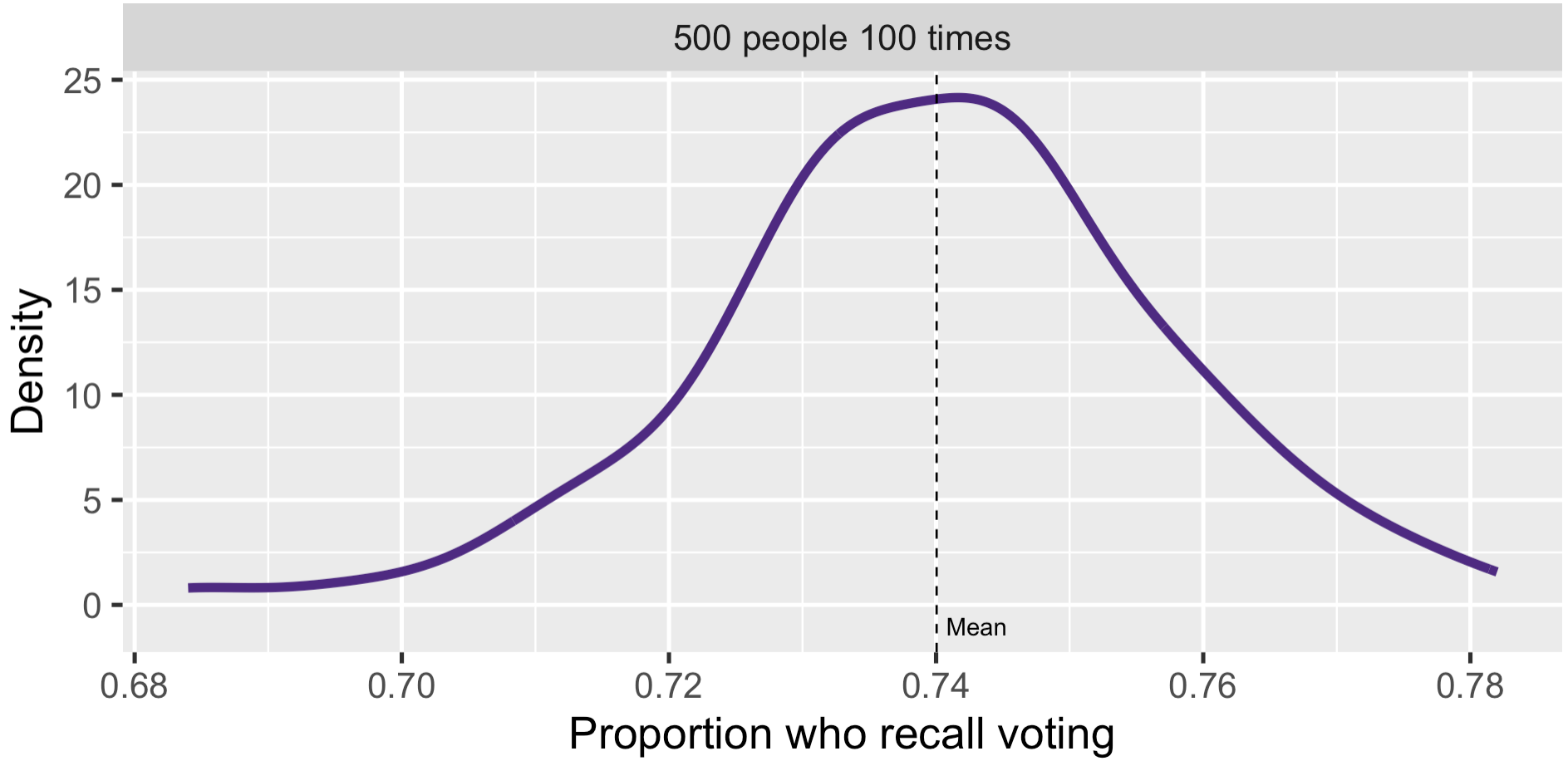
Position	Height	Percentile
1	62.5	0
2	63.0	10
3	63.0	20
4	66.0	30
5	66.0	40
6	67.0	50
7	68.5	60
8	70.0	70
9	70.0	80
10	70.5	90
11	70.7	100

- Taking away 10% on each side gives [63.0, 70.5]
- 90% of the observations lie in this range
- For the sample mean, this has a special name

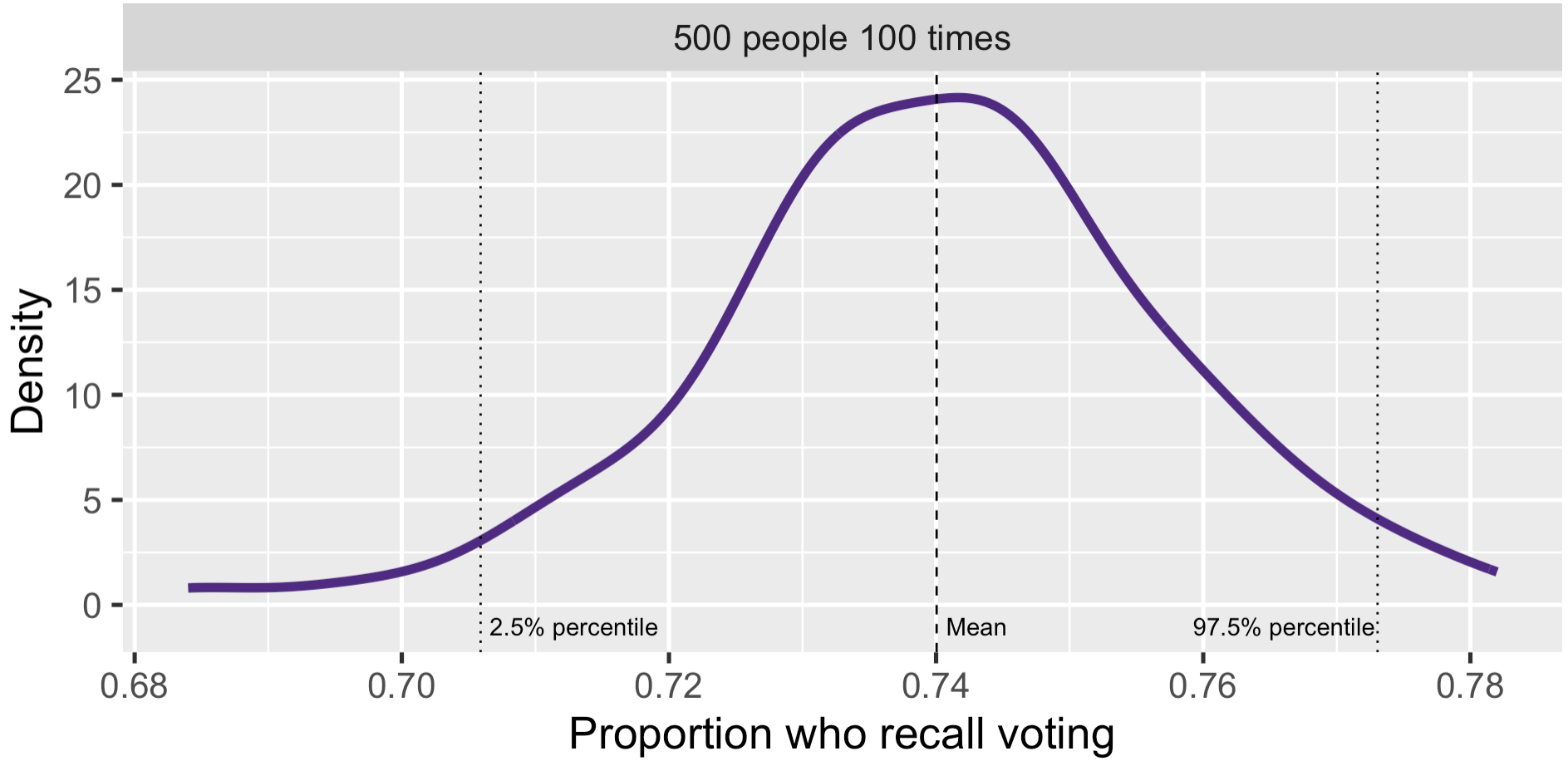
Confidence interval



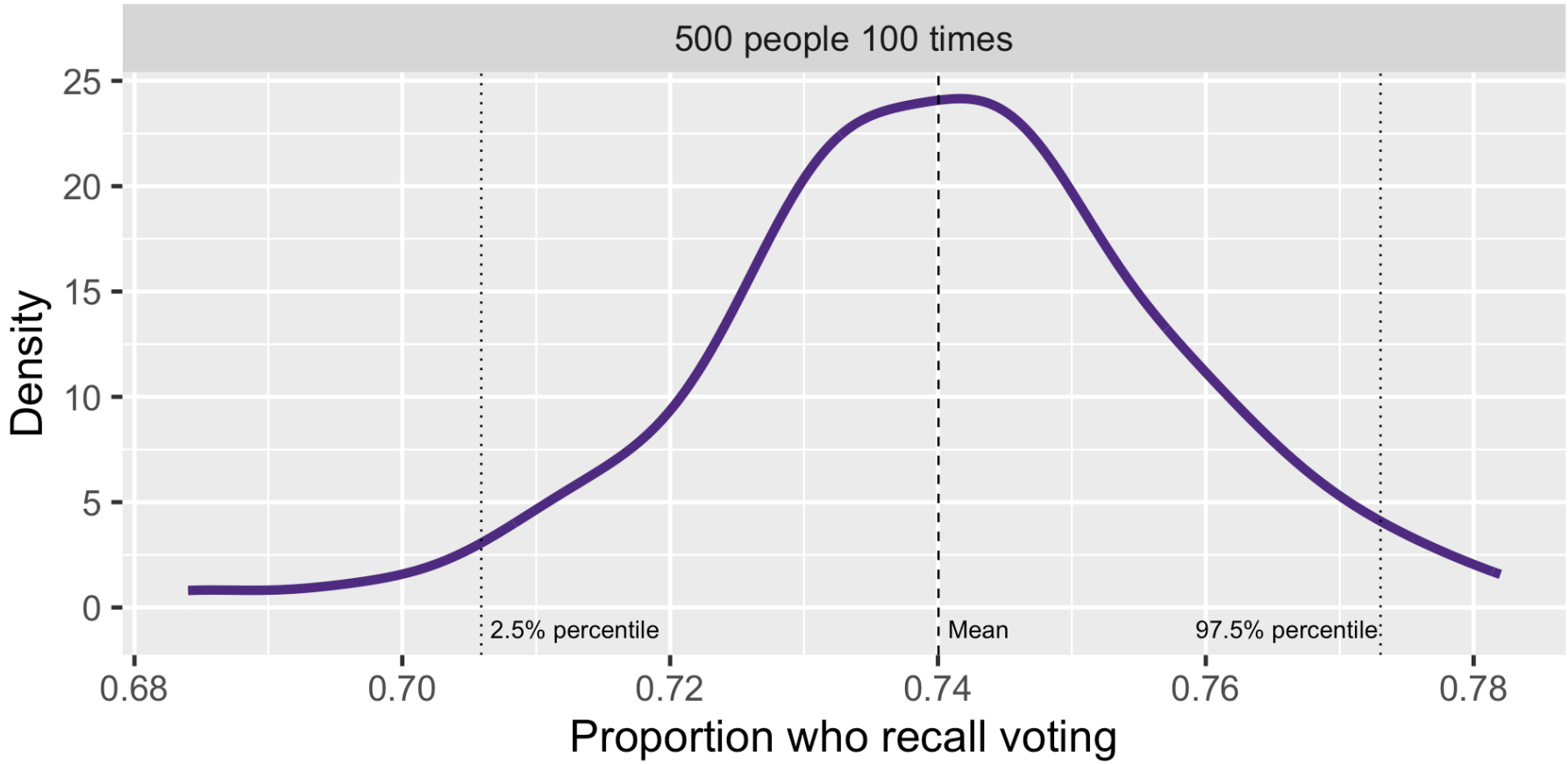
Confidence interval



Confidence interval



Confidence interval



Mean = 0.74. **95% confidence interval =** [0.71,0.77]

More formally

Confidence interval: Range of values in which X% of the estimates would fall

More formally

Confidence interval: Range of values in which X% of the estimates would fall **over many repetitions with a fixed sample size**

Informally: We are X% confident that our estimate falls in this range

How to calculate:

1. Non-parametric approximation (percentile method)
2. Analytic derivation (shortcut via aymptotic properties)

The 95% is an arbitrary yet useful convention in the social sciences

Analytic derivation

The percentile method would require us to *actually* repeat the study many times or *simulate*

The **CLT** lets us use *measures of dispersion* as shortcuts

$$\text{Sample mean: } \overline{X} = \frac{1}{n} \sum x_i$$

$$\text{Sample variance: } V[X] = \frac{1}{(n-1)} \sum (x_i - \overline{X})^2$$

Analytic derivation

The percentile method would require us to *actually* repeat the study many times or *simulate*

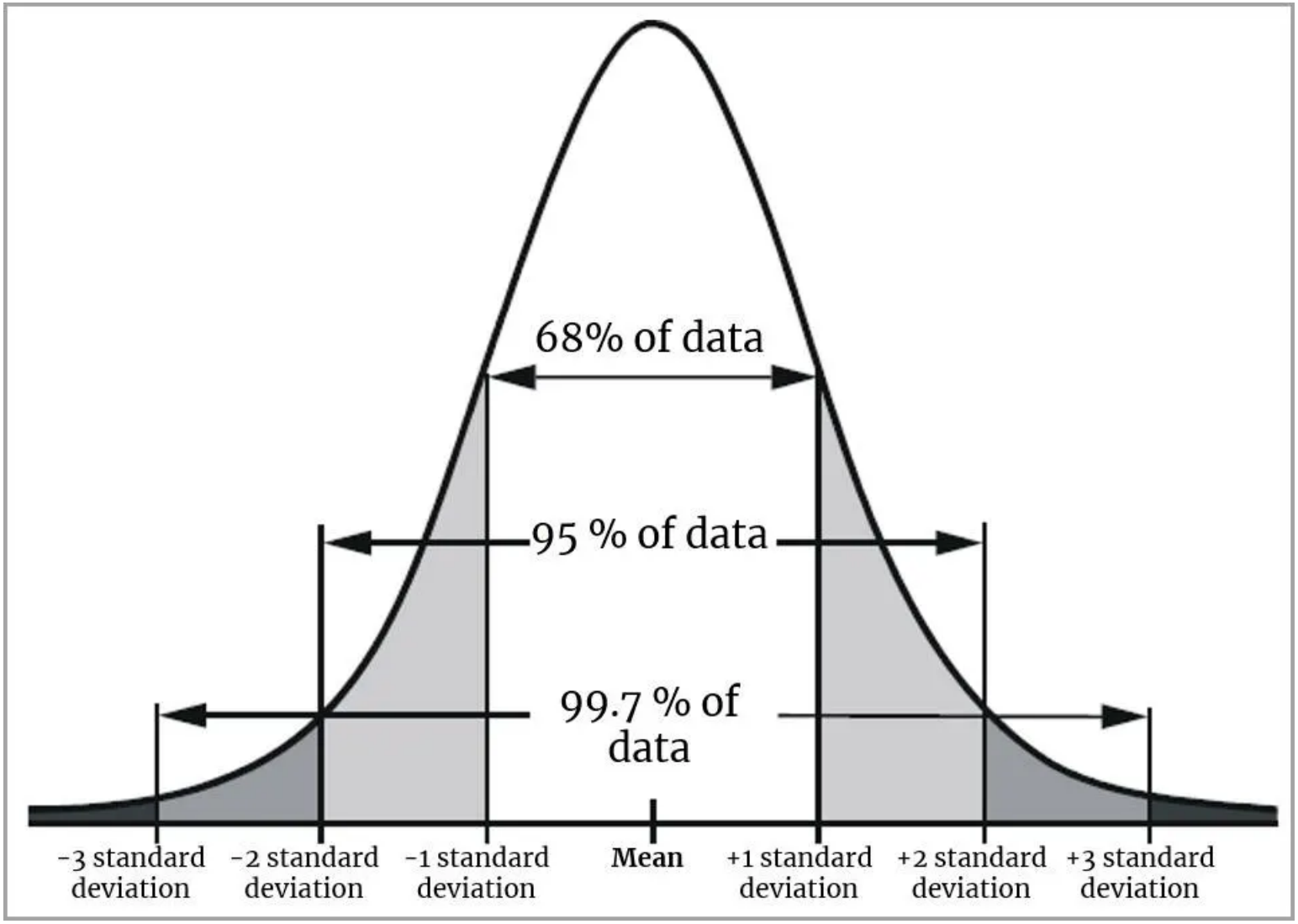
The **CLT** lets us use *measures of dispersion* as shortcuts

$$\text{Variance of sample mean: } V[\overline{X}] = \frac{V[X]}{n}$$

$$\text{Standard error: } SE[\overline{X}] = \sqrt{V[\overline{X}]}$$

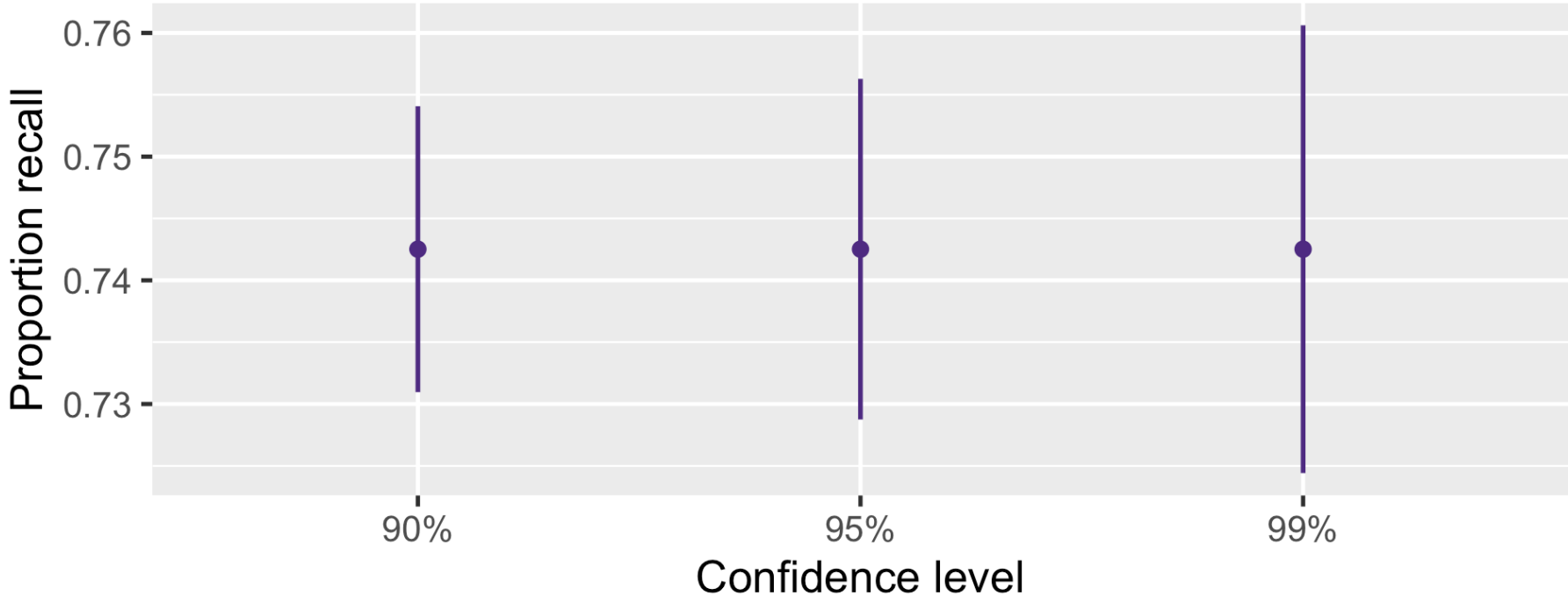
The **standard error** is a measure of dispersion for the **sample mean**

Calculate confidence intervals



Back to vote recall

level	estimate	conf.low	conf.high
90%	0.743	0.731	0.754
95%	0.743	0.729	0.756
99%	0.743	0.724	0.761



Summary

- We summarize data to conduct statistical inference
- The mean is *usually* good because of its statistical properties
- We use confidence intervals to convey uncertainty around the sample mean
- 95% confidence intervals are the convention in the social sciences

Inference

POLI SCI 210

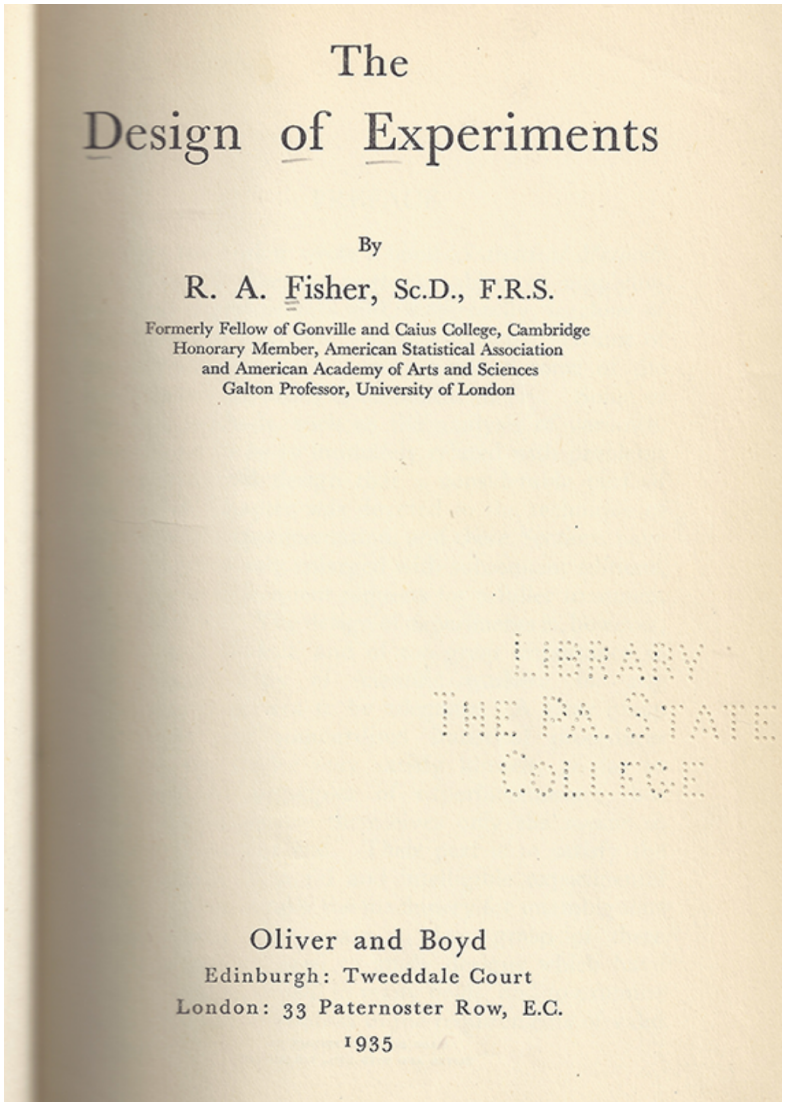
Introduction to Empirical Methods in Political Science

Last time

- We summarize data to conduct statistical inference
- The mean is *usually* a good summary
- We compute confidence intervals to *quantify uncertainty* around the mean

TODAY: *Hypothesis testing* as an alternative approach

Example: Fisher (1935) Chapter 2



The
Design of Experiments

By

R. A. Fisher, Sc.D., F.R.S.

Formerly Fellow of Gonville and Caius College, Cambridge
Honorary Member, American Statistical Association
and American Academy of Arts and Sciences
Galton Professor, University of London

Oliver and Boyd
Edinburgh: Tweeddale Court
London: 33 Paternoster Row, E.C.
1935

The lady tasting tea

A lady declares that by tasting a cup of tea made with milk she can discriminate whether the milk or the tea infusion was first added to the cup

How do you **evaluate** this statement?

An experiment

- Grab eight milk-tea cups
- 4 milk first, 4 tea first
- We arrange them in random order and ask lady to guess
- Lady knows there are 4 of each, but not which ones

Results

	True Order	
Lady's Guesses	**Tea First**	**Milk First**
Tea First	3	1
Milk First	1	3

- She gets it right $6/8$ times
- What can we conclude?

Problem

- How does “being able to discriminate” look like?
- We **do know** how a person *without* the ability to discriminate milk/tea order looks like
- Random guessing!
- This is our **null hypothesis** (H_0)
- Which lets us make **probability statements** about how the world look like **if the null hypothesis was true**

A person with no ability

Count	Possible combinations	Total
0	xxxx	
1	xxxo, xxox, xoxx, oxxx	
2	xxoo, xoxo, xoox, oxox, ooxx, oxxo	
3	xooo, oxoo, ooxo, ooox	
4	oooo	

- This is a symmetrical problem!

Ways of getting a number of tea-first cups right

A person with no ability

Count	Possible combinations	Total
0	xxxx	
1	xxxo, xxox, xoxx, oxxx	
2	xxoo, xoxo, xoox, oxox, ooxx, oxxo	
3	xooo, oxoo, ooxo, ooox	
4	oooo	

Ways of getting a number of milk-first cups right

A person with no ability

Count	Possible combinations	Total
0	xxxx	$1 \times 1 = 1$
1	xxxo, xxox, xoxx, oxxx	$4 \times 4 = 16$
2	xxoo, xoxo, xoox, oxox, ooxx, oxxo	$6 \times 6 = 36$
3	xooo, oxoo, ooxo, ooox	$4 \times 4 = 16$
4	oooo	$1 \times 1 = 1$

- A person guessing at random gets $6/8$ cups right with probability $\frac{16}{70} \approx 0.23$

Ways of getting a number of tea-first **and** milk-first cups right

A person with no ability

Count	Possible combinations	Total
0	xxxx	$1 \times 1 = 1$
1	xxxo, xxox, xoxx, oxxx	$4 \times 4 = 16$
2	xxoo, xoxo, xoox, oxox, ooxx, oxxo	$6 \times 6 = 36$
3	xooo, oxoo, ooxo, ooox	$4 \times 4 = 16$
4	oooo	$1 \times 1 = 1$

- And **at least** $6/8$ cups with $\frac{16+1}{70} \approx 0.24$

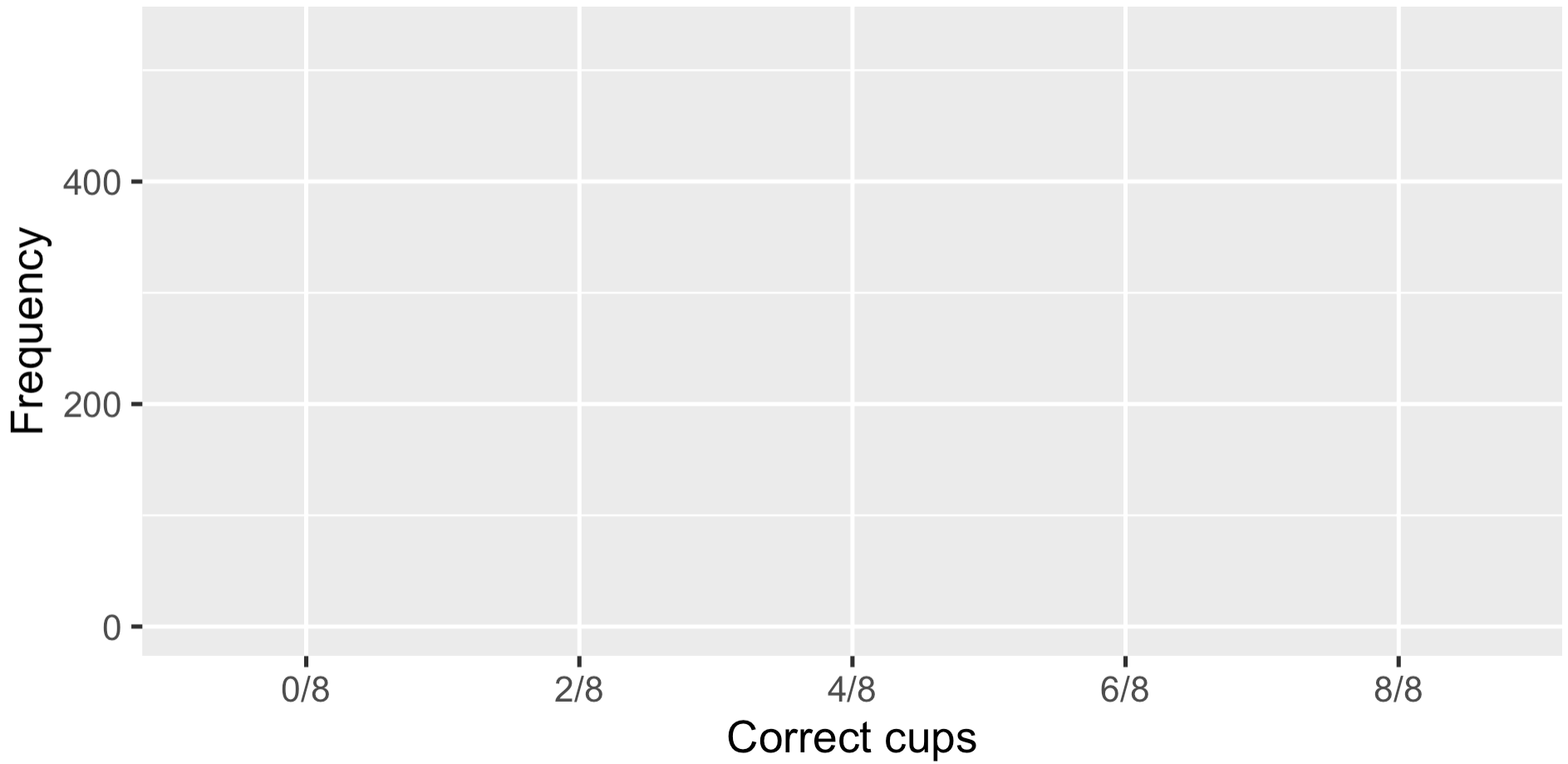
Ways of getting a number of tea-first **and** milk-first cups right

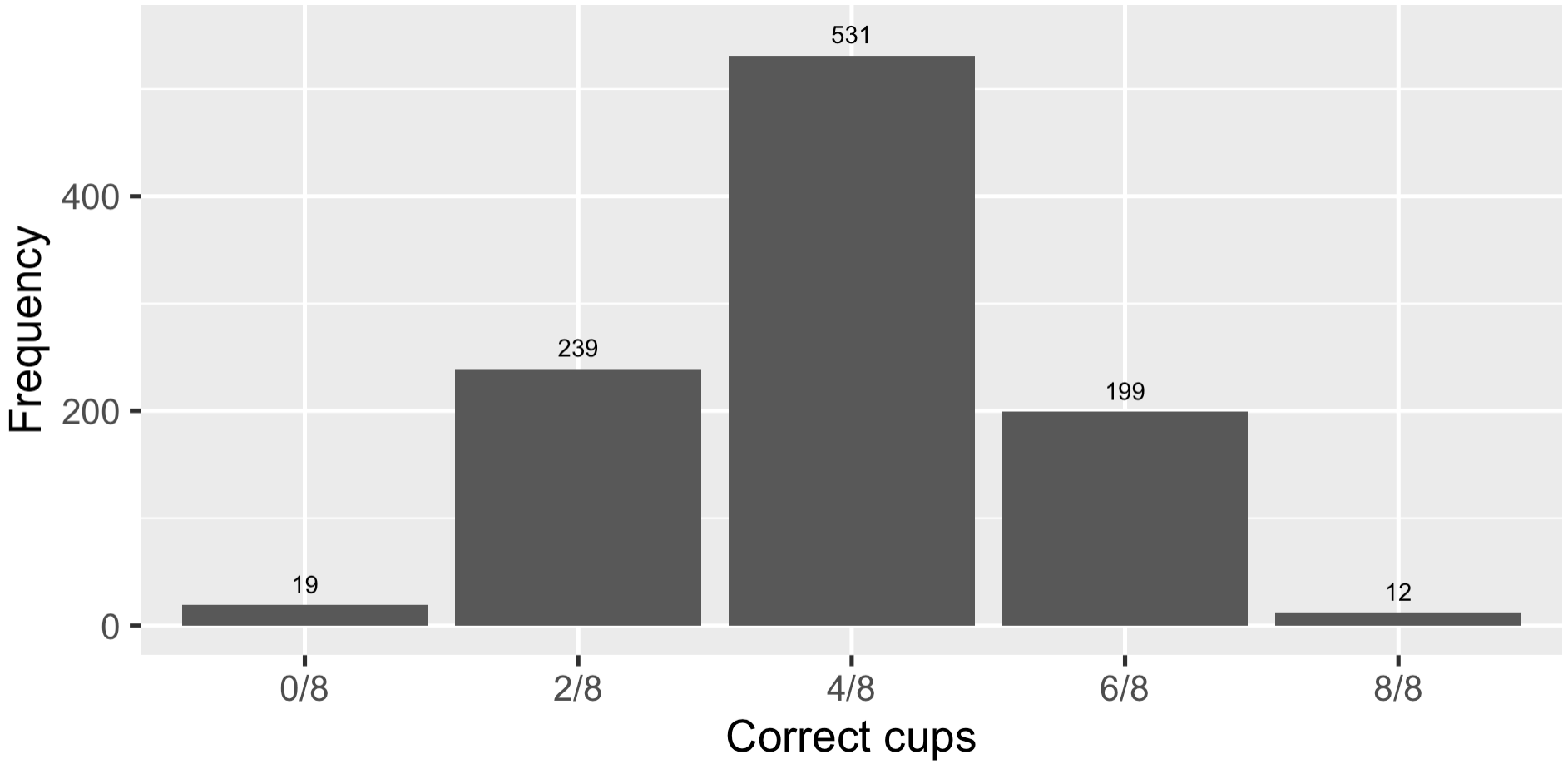
Another way to look at it

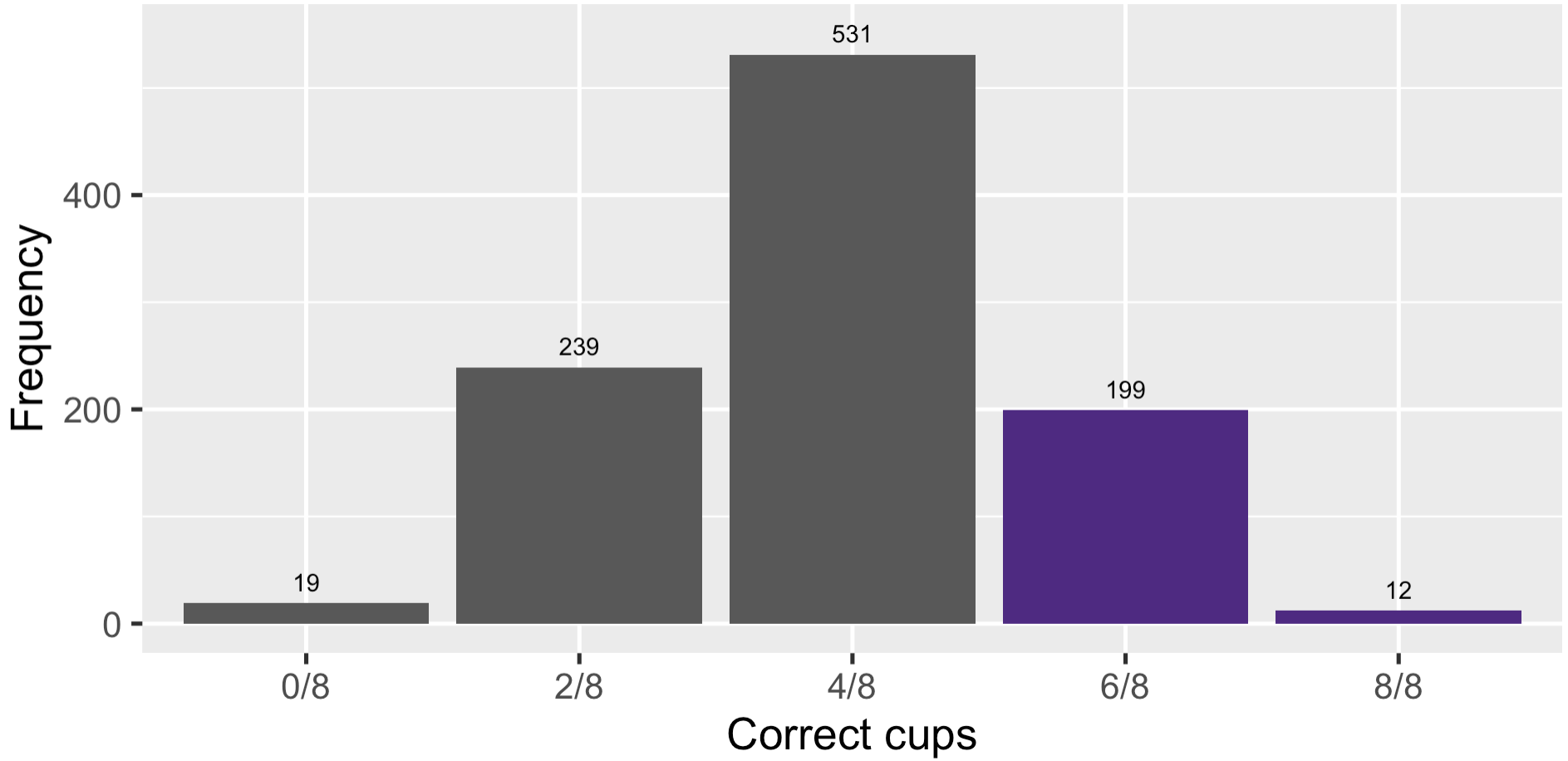
Count	Correct	Combinations	Probability
0	0/8	1/70	0.01
1	2/8	16/70	0.23
2	4/8	36/70	0.51
3	6/8	16/70	0.23
4	8/8	1/70	0.01

Random guesser: pick 0-4 right with corresponding probability

Simulate 1000 times to make a **probability distribution**


Frequency
400
200
0
0/8
2/8
4/8
6/8
8/8
Correct cups


531
239
199
19
12
400
200
0
Frequency
0/8
2/8
4/8
6/8
8/8
Correct cups



Random guesser gets *at least* 6/8 $\frac{(199+12)}{1000} \approx 0.21$

p-values

- If the lady is **not** able to discriminate milk-tea order, the probability of observing $6/8$ correct guesses or better is 0.24
- A **p-value** is the probability of observing a result *equal or more extreme* than what is originally observed…
- … *when* the **null hypothesis** is true
- Smaller p-values give more evidence **against** the null
- Implying observed value is *less likely to have emerged by chance*

This is Fisher's interpetation of p-values, which is the most common. Neyman and Pearson had different ideas

Rules of thumb

- A convention in the social sciences is to claim that something with $p < 0.05$ is *statistically significant*
- Meaning we have "enough" evidence to reject the null
- This is known as **Null Hypothesis Significance Testing** (NHST)
- Committing to a **significance level** α implies accepting that sometimes we will get $p < 0.05$ by chance
- This is a **false positive** result

No good reason for $\alpha = 0.05$ other than path dependency.

Types of error

	Unobserved reality	
Decision	**H_0 true**	**H_0 not true**
Don't reject H_0	True negative	False negative (type II error)
Reject H_0	False positive (type I error)	True positive