Understanding Uncertainty

Applied Regression in R

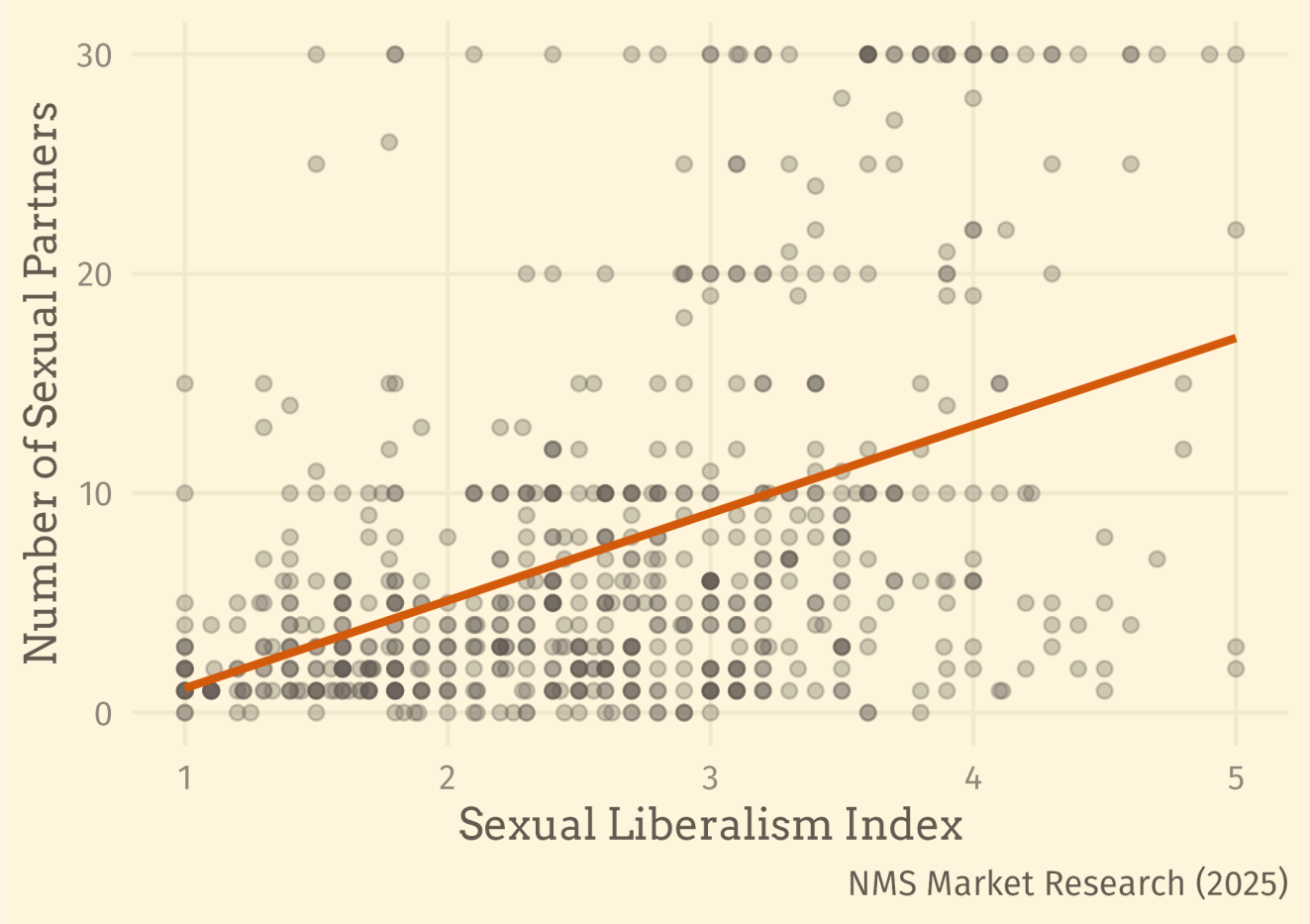
Aleš Vomáčka

02. 03. 2026

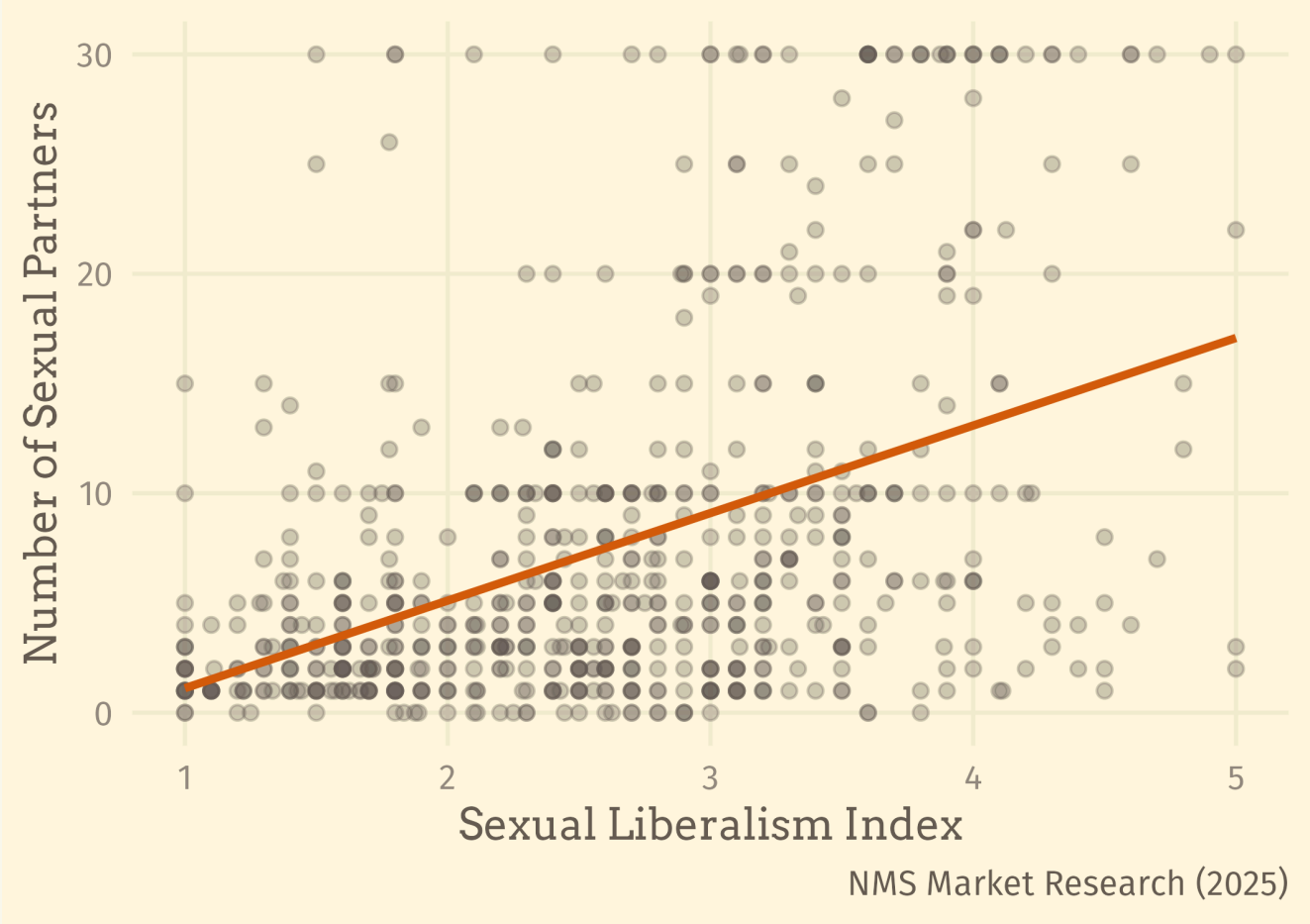
Faculty of Arts, Charles University

We know how to draw regression lines (curves, planes).

But the drawn line isn't necessarily the true line.

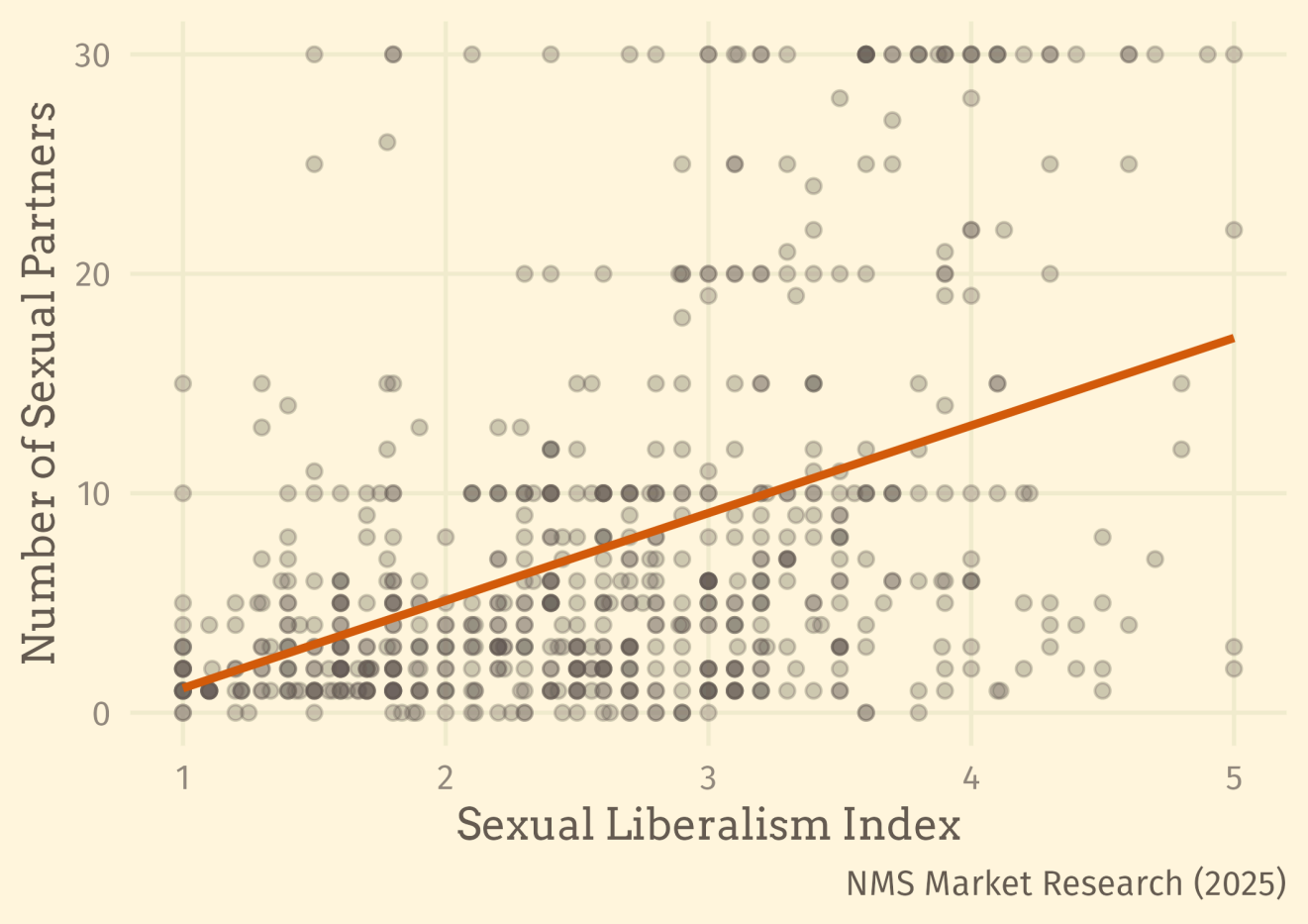


There are many sources of error/ **uncertainty**, both systemic and random - sampling error, measurment error, etc.



We will focus on **random sampling** error.

The general principles are true for sources.



Probability Crash Course

Probability Crash Course

Imagine it's shortly before presidential elections. We are conducting **pre election survey with 500 respondents**.

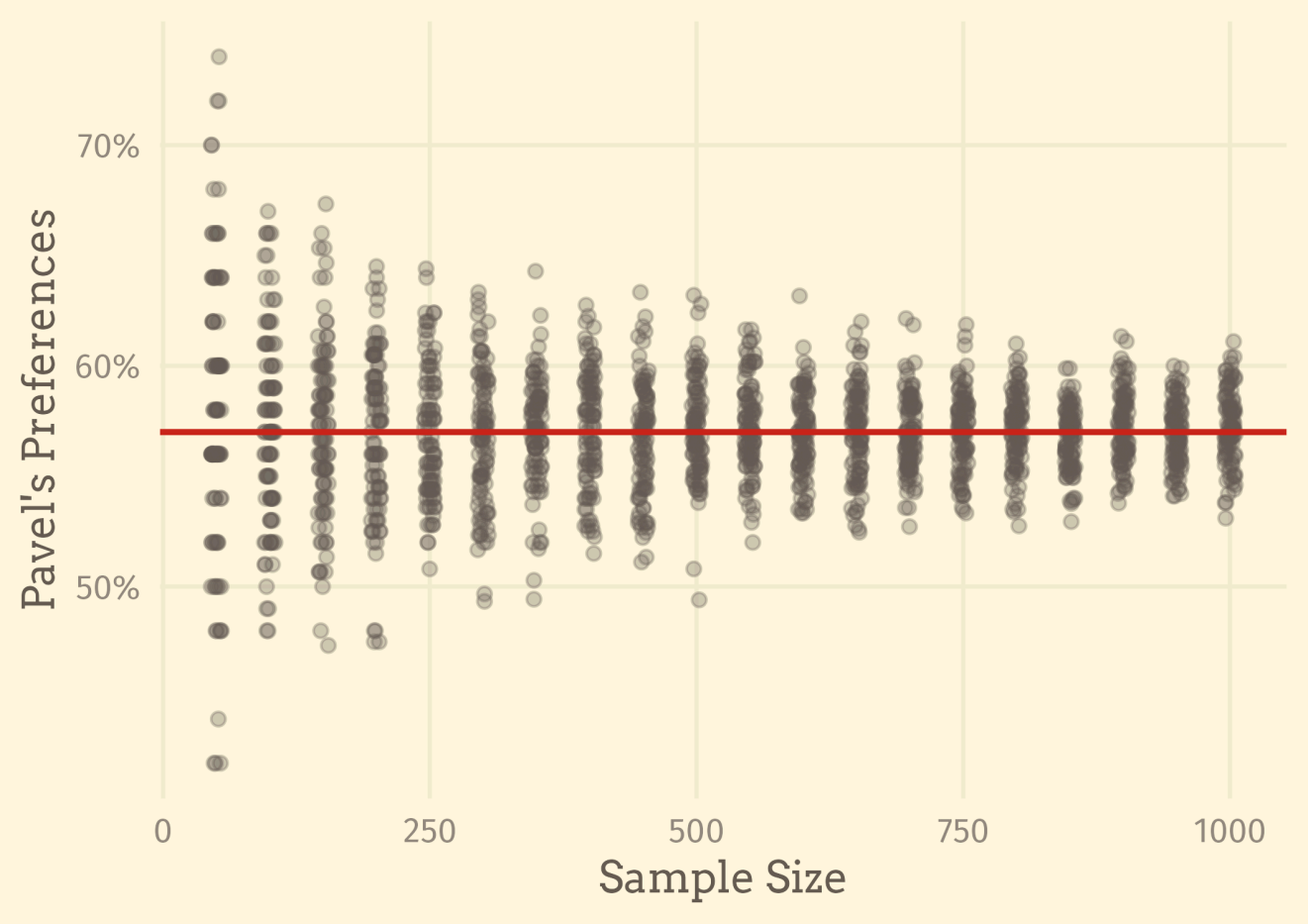
Our survey shows that **57% of voters** are going to vote for Petr Pavel.

How do we know we can trust the results? What is the theoretical reasoning?

Law of Large Numbers

Under some (mild) assumptions, the expected value of sample means is equal to the population mean.

In other words, **„on average“ the sample mean will be equal to the population mean**.



Probability Crash Course

Under some (slightly less) mild conditions, as the sample size approaches infinity, the distribution of sample means converges to normal distribution.

The standard deviation of this sample means is called **standard error**.

We even know what the standard deviation is: $\sqrt{\frac{\text{var}}{n}}$

In other words, while we don't know how big is the difference between sample mean and population mean in any specific sample, **we know the average difference (error) between them**.

Probability Crash Course

Combining the following:

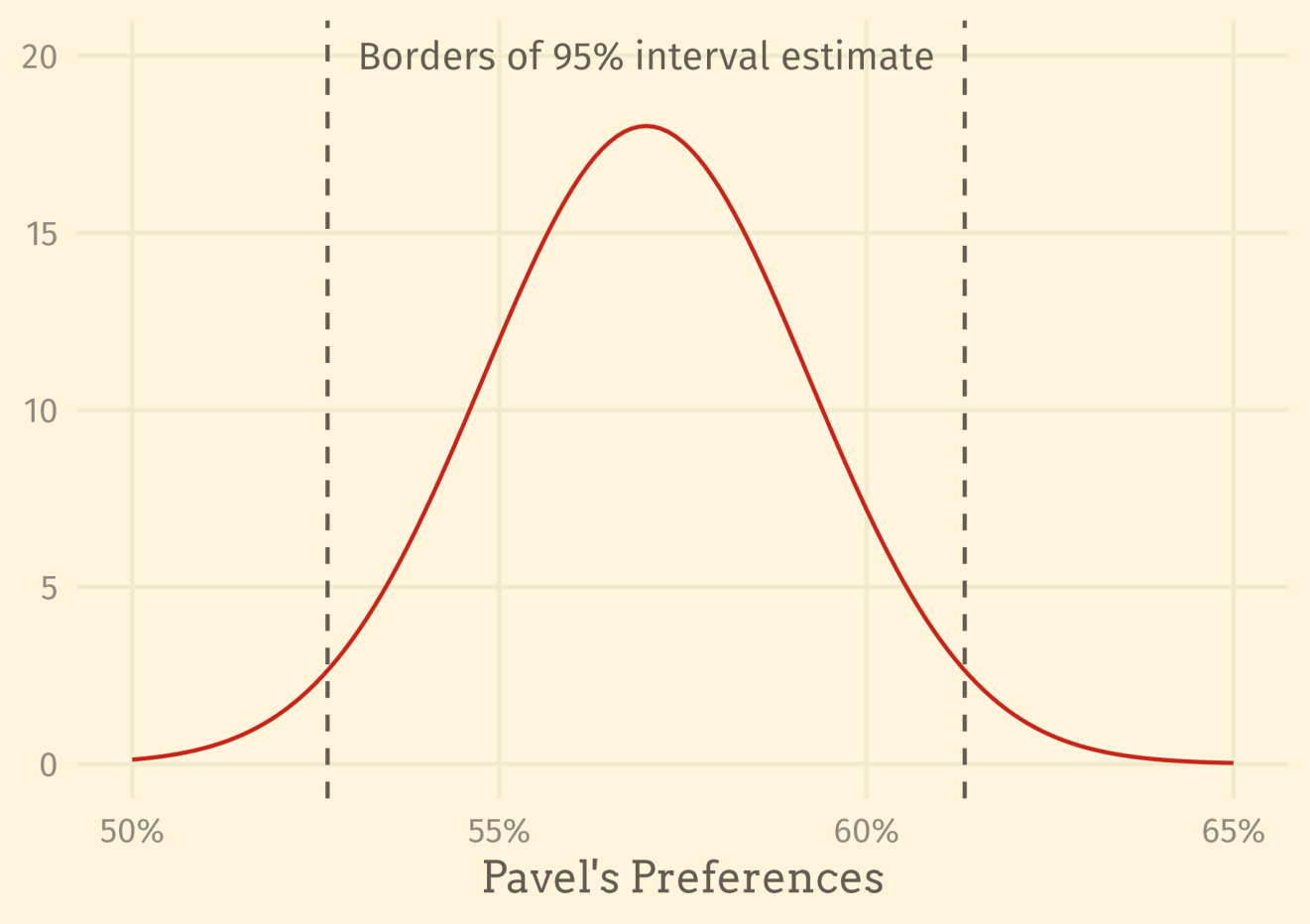
- On average, the sample mean is equal to the population mean.
- Distribution of sample means converges to normal distribution.
- The standard distribution of sample means is $\sqrt{\frac{\text{var}}{N}}$.

We can construct our **best guess of what the sampling distribution** looks like.

Probability Crash Course

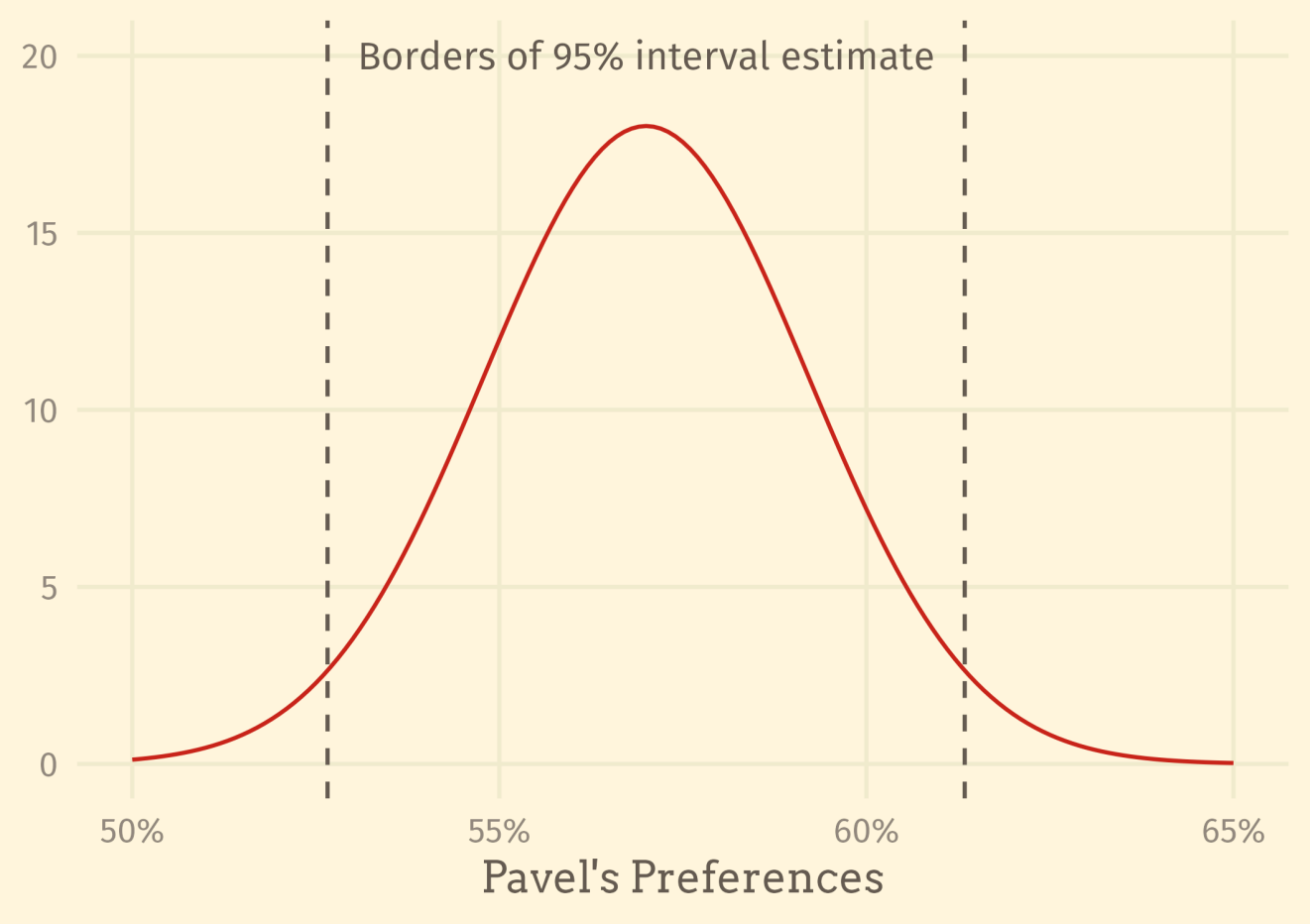
We use chunks of the estimated sampling distribution as **interval estimates**.

Most common is the 95% interval estimate.



Probability Crash Course

The 95% interval estimate for Pavel's preferences is between 52% and 61%

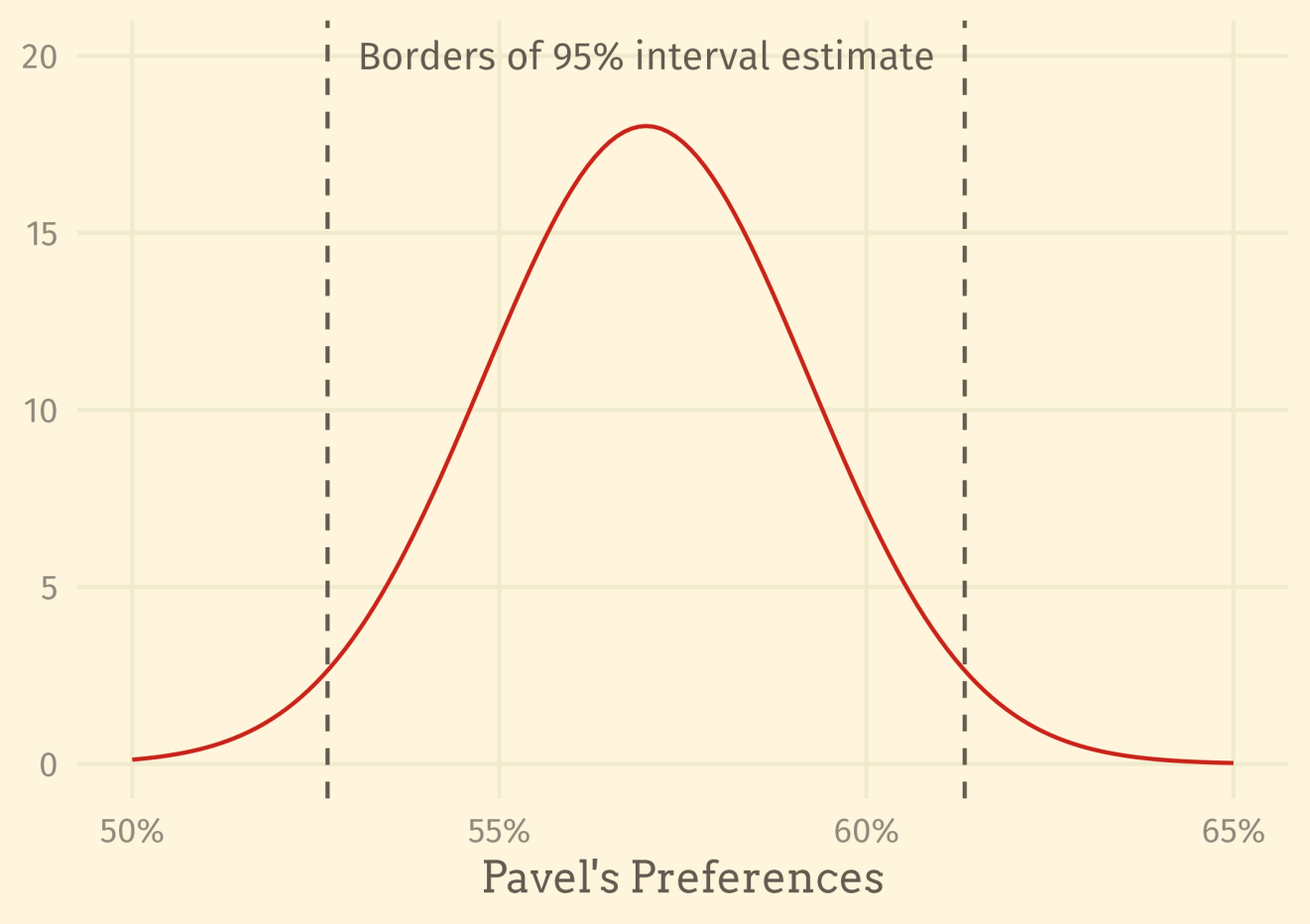


Questions?

Probability Crash Course

The 95% interval estimate for Pavel's preferences is between 52% and 61%

But what does that mean?



The Tale of Two Paradigms

The Tale of Two Paradigms

In every field of science, there are several competing worldviews.

In statistics, the two major competitors are **frequentism** and **bayesianism**.

The main of contention is a seemingly basic question:

What is probability anyway?

The Tale of Two Paradigms

Frequentist statistics is currently the most popular approach, dominating since the 1920s.

Probability is defined as the **number of successes across infinite number of trials**.

This definition of probability is sometimes called objectivist.

The Tale of Two Paradigms

Bayesian statistics is currently less popular, but historically older.

Probability is defined as the **strength of subjective belief that an event will happen**.

Often interpreted as betting odds - if believe Pavel has 60% of winning of the elections, fair betting odds are $\frac{0.6}{0.4} = 1.5$.

This definition of probability is sometimes called subjectivist.

The Tale of Two Paradigms

Example: The probability of fair coin landing on head is 0.5. What does it mean?

Frequentists: If we kept flipping the coin **infinitely many times**, it would land on head half of the time.

Bayesians: We have no more reason to **believe** the coin will land on head than that it wil land on tail.

Questions?

The Tale of Two Paradigms

The big question: *Cool, but why should we care??*

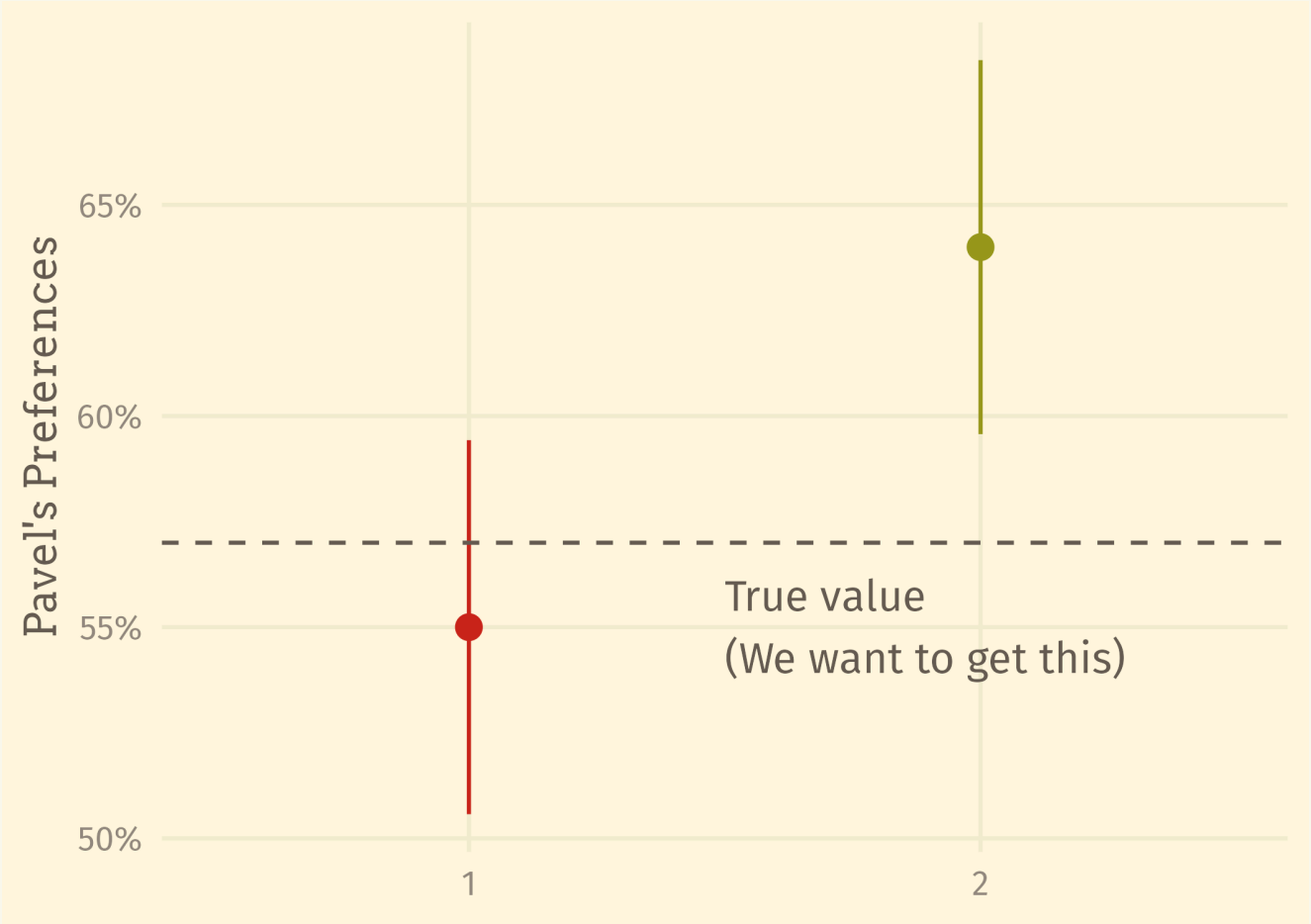
The answer: Which paradigm you adopt will have big impact on how you interpret the results of your analysis.

The Tale of Two Paradigms

The true value is 58%.

According the **frequentists**, what is the probability that...

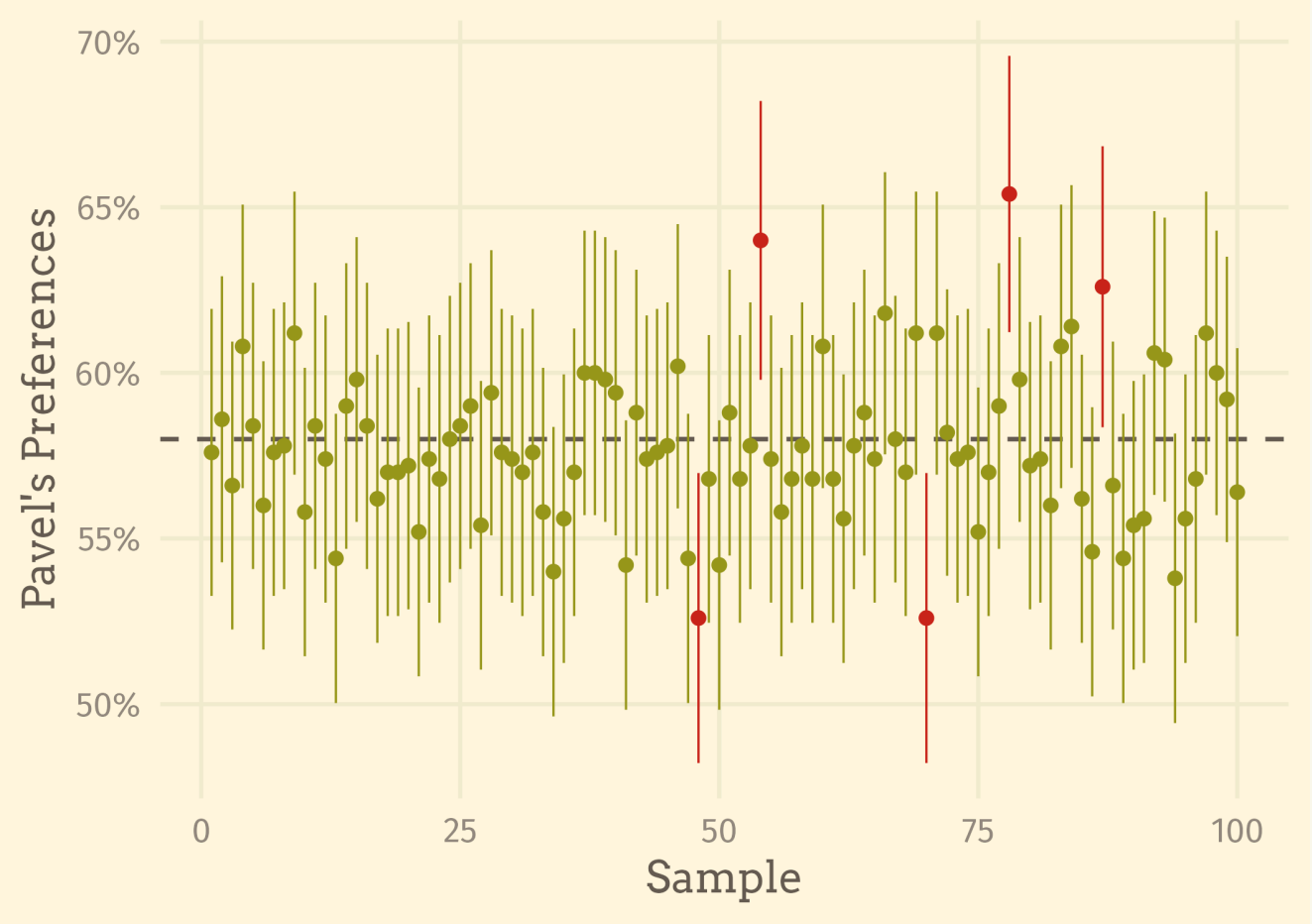
1. ... the first interval contains the true value?
2. ... the two interval contains the true value?



The Tale of Two Paradigms

Frequentist **confidence intervals** only tell us what happens across large number of trials.

They say nothing about individual samples.

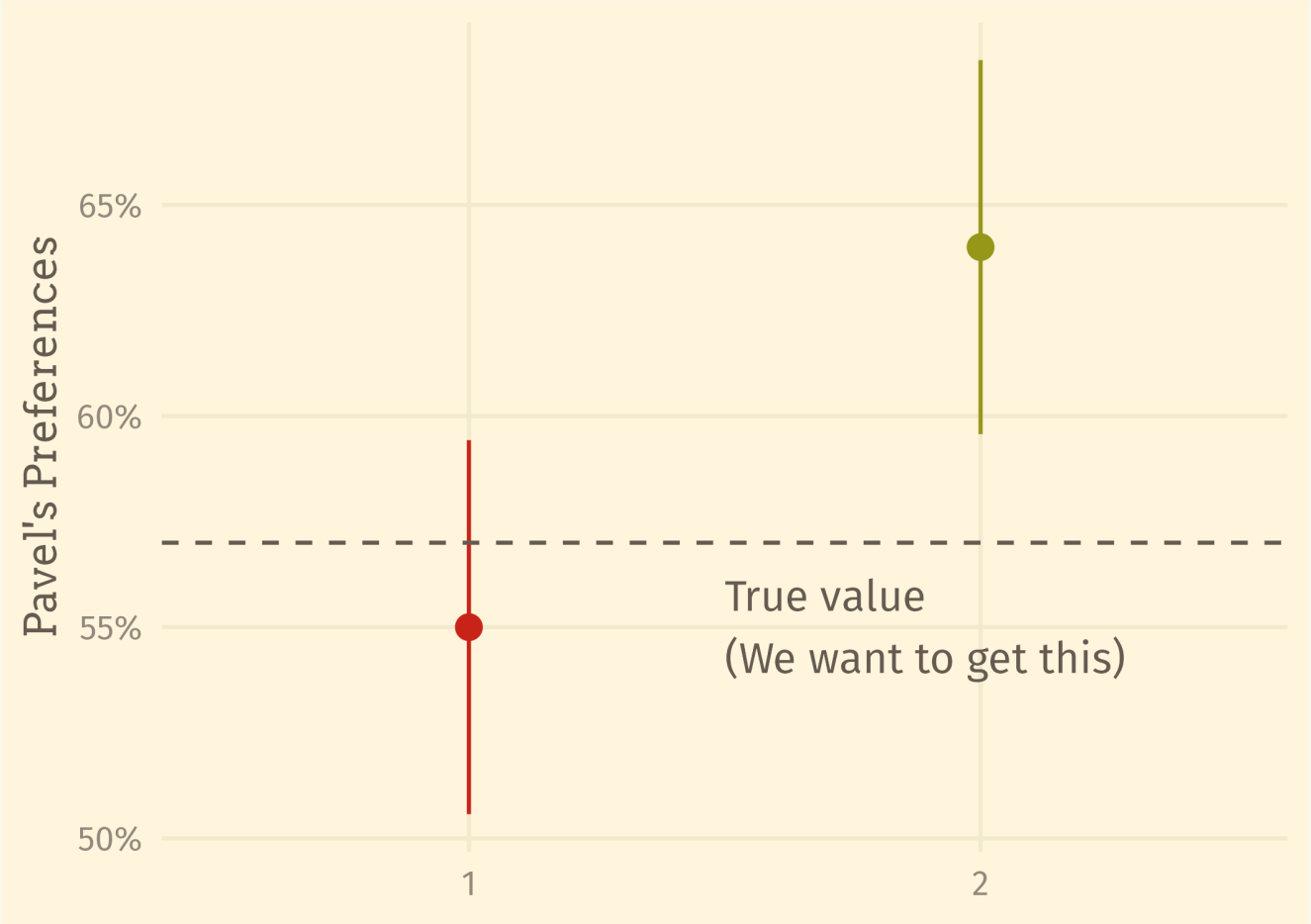


The Tale of Two Paradigms

The true value is 58%.

According the **bayesians**, what is the probability that...

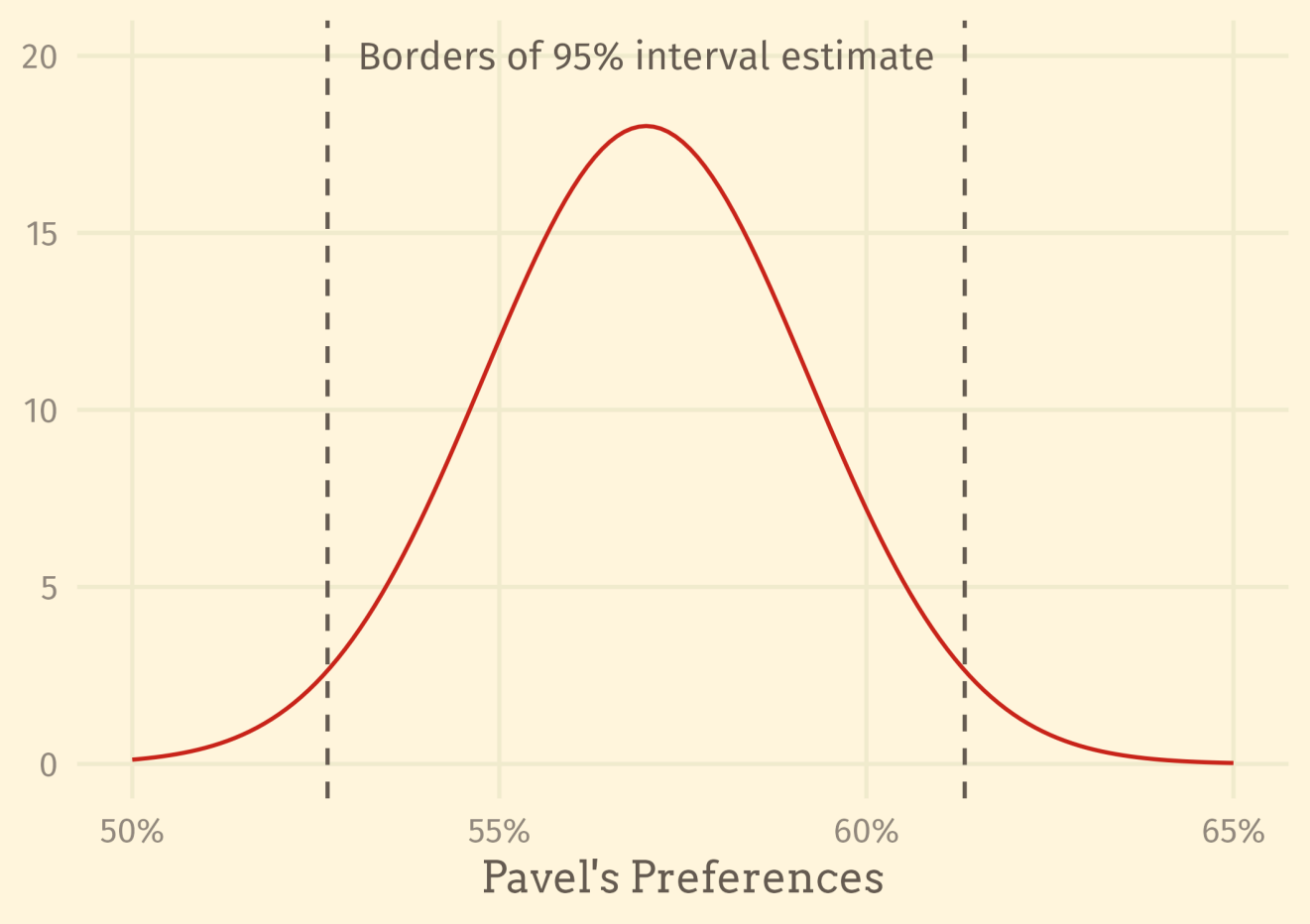
1. ... the first interval contains the true value?
2. ... the two interval contains the true value?



The Tale of Two Paradigms

Bayesian **credible intervals** provide range of value where we personally expect the true value to be.

They don't make promises about repeated studies.



The Tale of Two Paradigms

The 95% interval estimate for Pavel's voter preferences is

$< 0.58; 061 >$. What does it mean?

The Tale of Two Paradigms

The 95% interval estimate for Pavel's voter preferences is

$< 0.58; 061 >$. What does it mean?

Frequentists: If the elections happened under the same conditions **infinitely many times**, confidence interval with range of 0.03 ($= 0.61 - .058$) would include the true value 95% of time.

The Tale of Two Paradigms

The 95% interval estimate for Pavel's voter preferences is

$< 0.58; 061 >$. What does it mean?

Frequentists: If the elections happened under the same conditions **infinitely many times**, confidence interval with range of 0.03 ($= 0.61 - .058$) would include the true value 95% of time.

Bayesians: Based on our model and our data, **we believe** there is a 95% probability that between 58% and 61% voters will vote for Pavel.

The Tale of Two Paradigms

Picking between frequentis and bayesian approach was a huge deal back in the day. (There was *a lot* of arguing).

Today, most people opt for a more flexible approach, picking based on context.

And so should you.

The Tale of Two Paradigms

Example 1

You are a public opinion researcher and a political party wants to know what their chances in the upcoming elections are.

Frequentist approach bad: If the upcoming elections were **repeated infinitely** many times under the same conditions, we'd be able to tell you how often our results would be correct.

Bayesian approach good: Based on our model, this is the result **we believe** you are most likely to see.

The Tale of Two Paradigms

Example 2

You are public worker and your job is to screen school children for abnormaly high levels of anxiety.

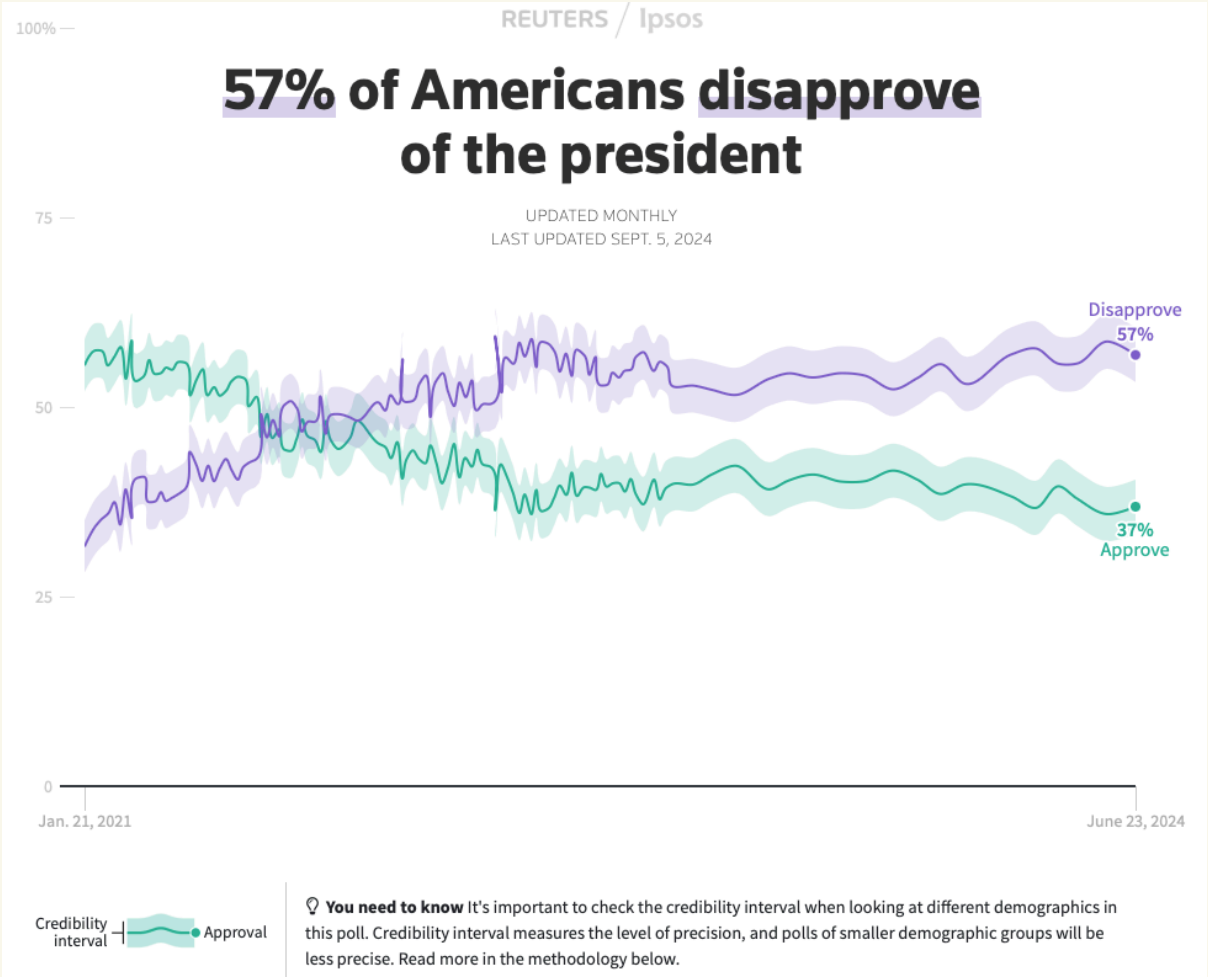
Bayesian approach bad: We can tell you what we think the anxiety level of each child is.

Frequentist approach good: We know we will estimate the anxiety level correctly for 95 out 100 children.

The Tale of Two Paradigms

Bayesian approach more common to the east from us, especially in survey and political science.

From Reuters: You need to know It's important to check the **credibility interval** when looking at different demographics in this poll. Credibility interval measures the level of precision, and polls of smaller demographic groups will be less precise. Read more in the methodology below.



Questions?

Caveats and Conclusion

Caveats and Conclusion

Bayesian statistics require specifying **priors** - your beliefs before seeing the data.

This makes it too subjective to some (results are partially dependent on what you believe they should be).

However, you can use **non-informative priors**, i.e. proclaim you have no expectations.

Caveats and Conclusion

Approach	(Intercept)	sexlib_index
Frequentist	-2.87 (-4.55;-1.19)	3.99 (3.38;4.6)
Bayesian	-2.86 (-4.6;-1.17)	3.99 (3.37;4.59)

As long as you have 1) large-ish samples, 2) simply-ish models and 3) non-informative priors, the results will be numerically (virtually) identical. It's all about interpretation.

Caveats and Conclusion

People used to think frequentist and bayesian approaches are incompatible.

Turns out, they very often converge to the same answer, the only difference is interpretation.

The whole debate is kinda „boomer problem“.

So why spend so much time talking about them?

Caveats and Conclusion

It's **important to recognize uncertainty** inherent to every analysis - overpromising leads to disappointment (look at you, pre election survyes...)

But thinking about uncertainty is hard, especially for lay people.

Choosing the right world view for the occasion can make it more approachable.

Questions?

InteRmozzo!