Statistical inference in theory

PSCI 2301: Quantitative Political Science II

Prof. Brenton Kenkel

brenton.kenkel@gmail.com

Vanderbilt University

February 3, 2025

Recap

Last week — randomized experiments and causal inference

1. The importance of randomized treatment assignment
    - Randomization $\rightsquigarrow$ treatment/control groups broadly representative
    - Treatment status uncorrelated w/everything $\rightsquigarrow$ independence holds
    - Difference in group means $\approx$ Average treatment effect
2. Experiments in practice
    - The Gerber, Green, Larimer study of social pressure and turnout
    - Designing treatments to isolate the right effects
    - Working through external validity and ethics concerns

Today's agenda

Inference: How much do we learn about a population from sample data?

Key to the statistical approach — quantifying how wrong we could plausibly be

1. Law of large numbers
    - Sample mean converges to population mean
    - Key question for inference: How much data do we have?
2. Central limit theorem
    - Sample mean is approximately normally distributed across samples
    - Lets us calculate "margin of error" given sample size
    - ...or sample size we'd need for a given margin of error
3. Brief refresher on philosophy of hypothesis testing

Law of large numbers

Philosophy of inference

Starting simple: Inference about the mean

The mean gives us the easiest illustration of basic inferential principles.

Similar procedures for other stats like correlation, regression coefficient, treatment effect.

We want to know the population mean $E[X_i]$

→ e.g., average opinion of Trump on 0–100 scale among all US adults

But we can only calculate the sample mean $\mathbf{avg}[X_i]$

→ e.g., average opinion of Trump on 0–100 scale among survey respondents

How far off could the sample mean plausibly be?

Random sampling

Standard inference procedures assume a probability sample

- *Ex ante*, every unit in population has same probability of being sampled
- Ways this might fail
  - → Convenience sample like an online poll or asking whoever's around
  - → Differential non-response: some pop. members more prone to refuse

Also typically assume independence across observations

- Loosely: X_i above mean doesn't predict whether X_j is above mean
- May fail if we sample sets of units instead of individual units
  - → e.g., households, classrooms
- But we can use advanced methods to make corrections in these cases

Accuracy of the sample mean

Law of Large Numbers (LLN): As sample size N increases, $\mathbf{avg}[X_i] \approx \mathbb{E}[X_i]$

The Law of Large Numbers: Formal statement

For any difference $\delta > 0$ from the population mean and any probability $p > 0$, there is a sample size N such that there's a probability p or lower of drawing a sample of size N where

$$\left| \mathbf{avg}[X_i] - \mathbb{E}[X_i] \right| > \delta.$$

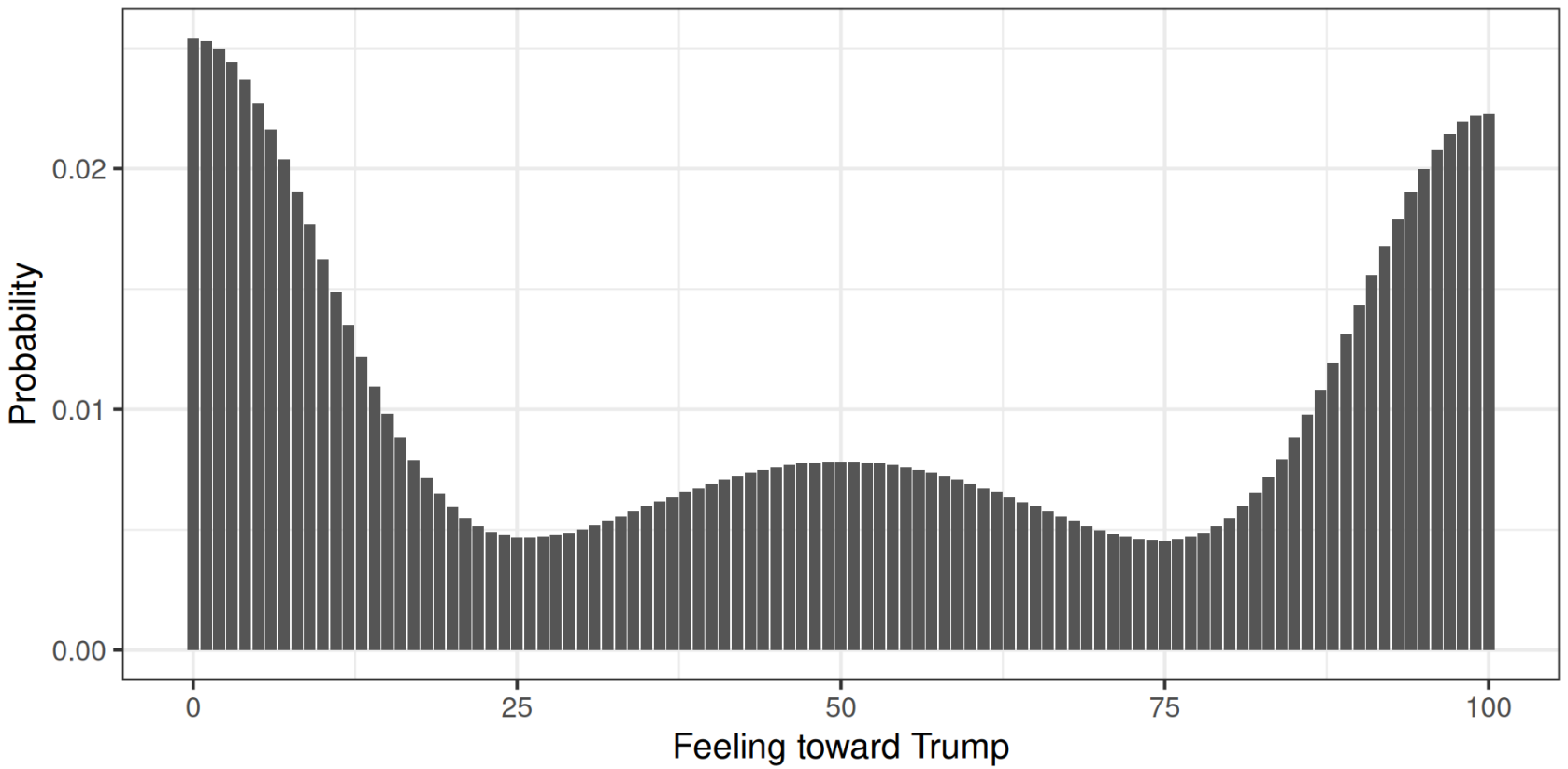
Any *given* sample might have a sample mean far from the population mean

...but the chance of a big difference gets smaller and smaller with N

This is true even when the population of units is infinite

Law of large numbers, illustrated

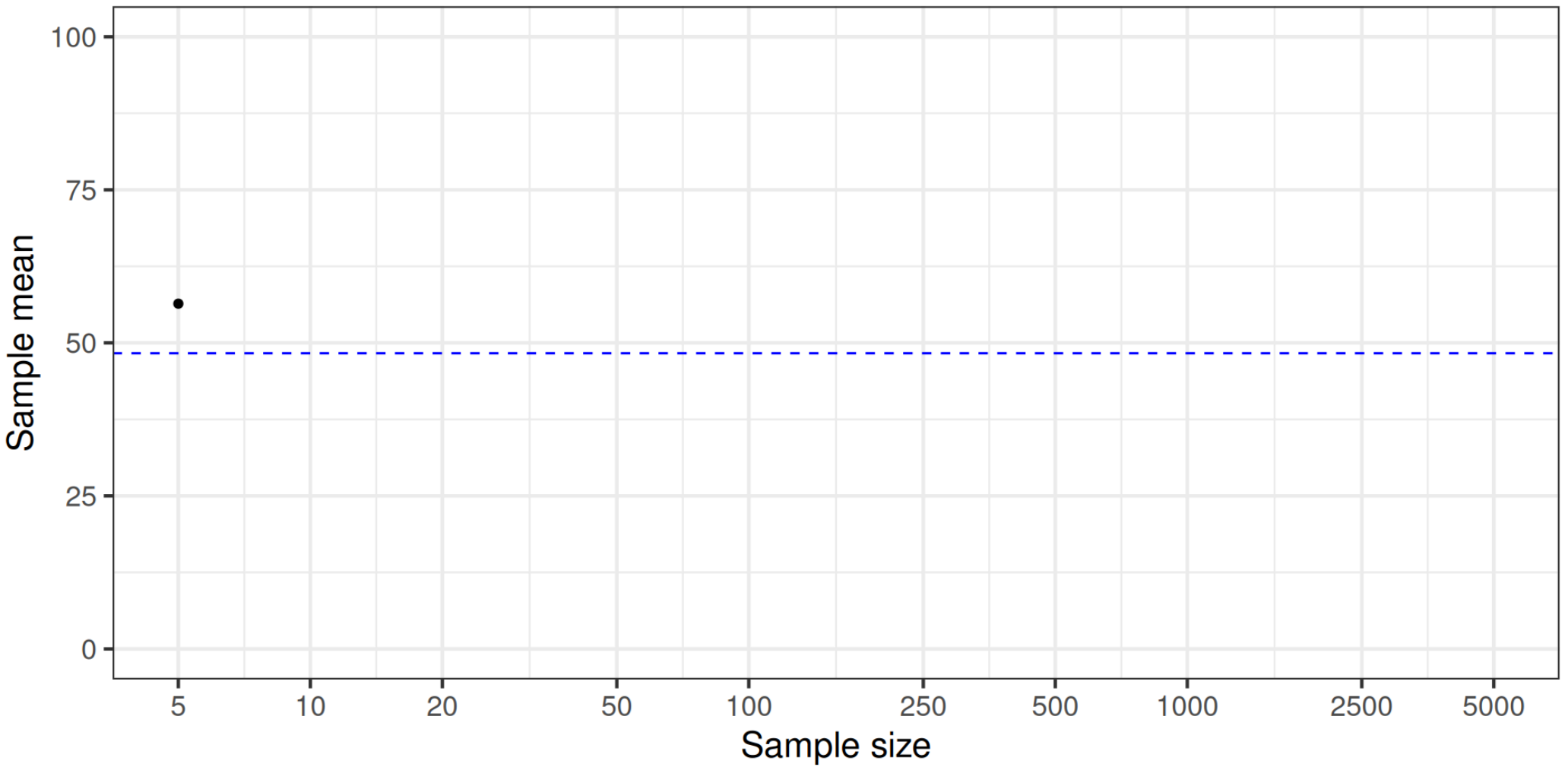
Imagine an infinite population with this distribution of opinions toward Trump



Population mean: $\mathbb{E}[X_i] = \sum_{x=0}^{100} x \cdot \Pr(X_i = x) \approx 48.3$

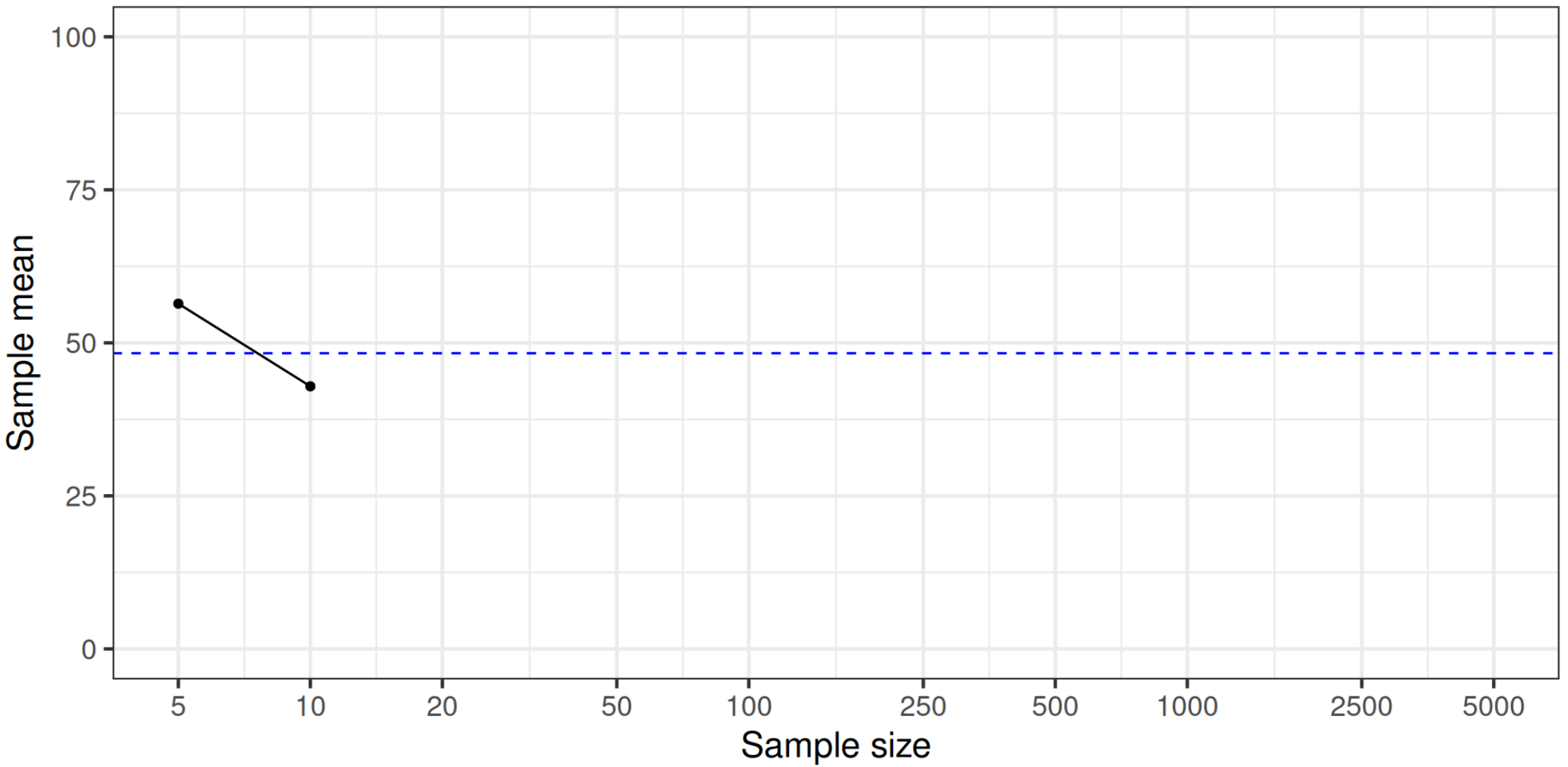
Law of large numbers, illustrated

Start with a sample of 5 units, calculate the sample mean



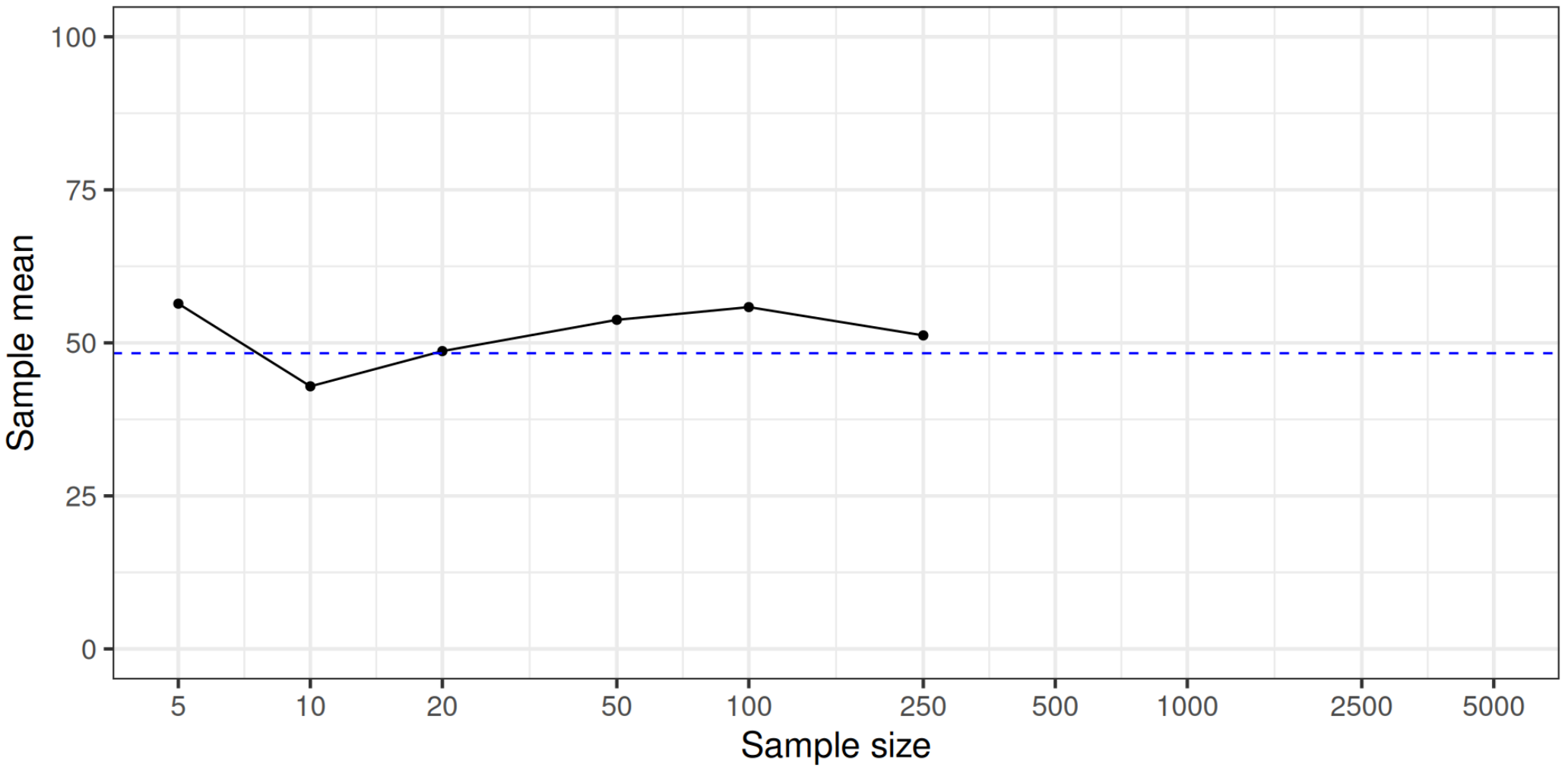
Law of large numbers, illustrated

Now increase sample size to 10 units



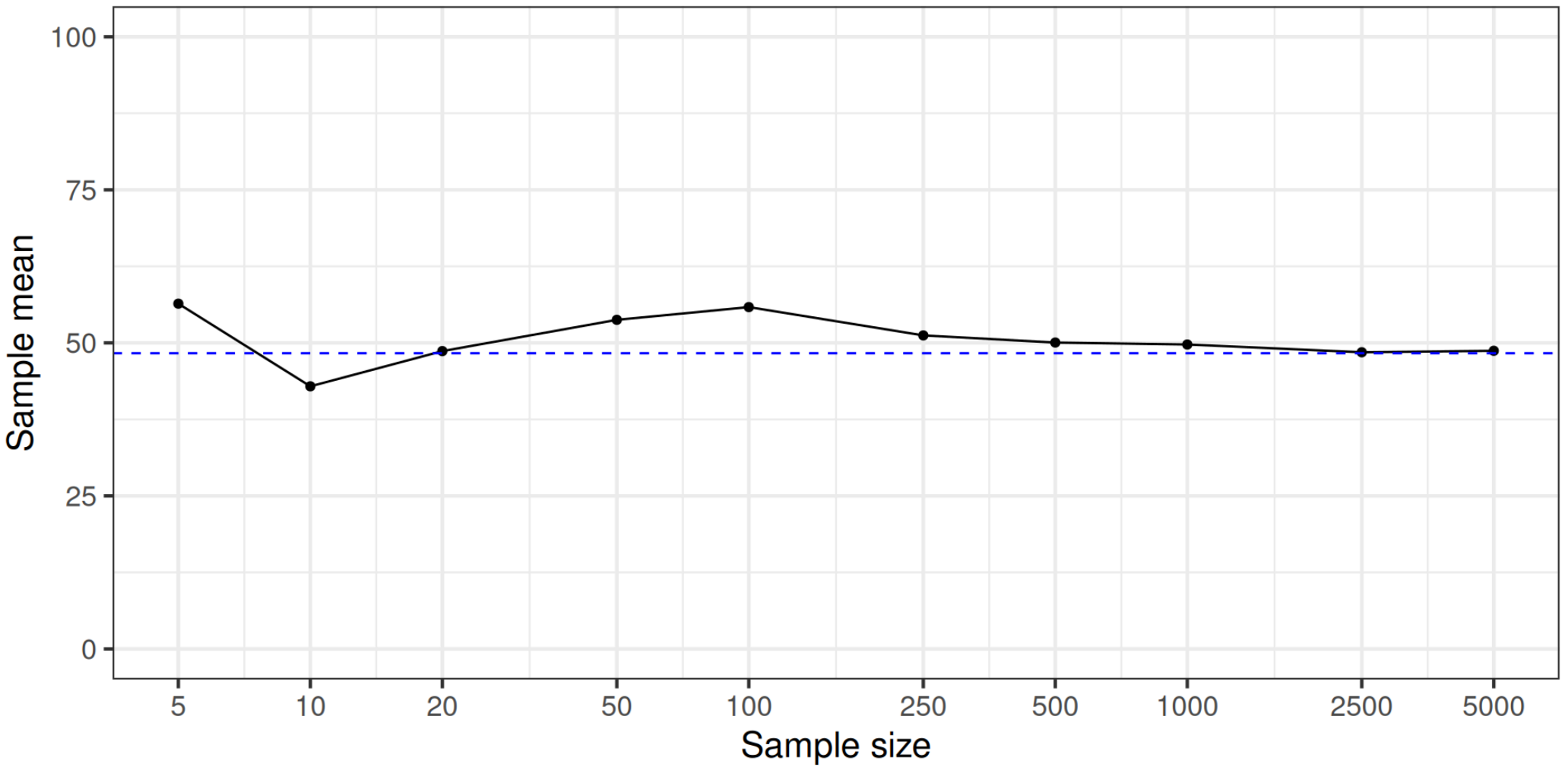
Law of large numbers, illustrated

Keep gradually increasing the sample size ...



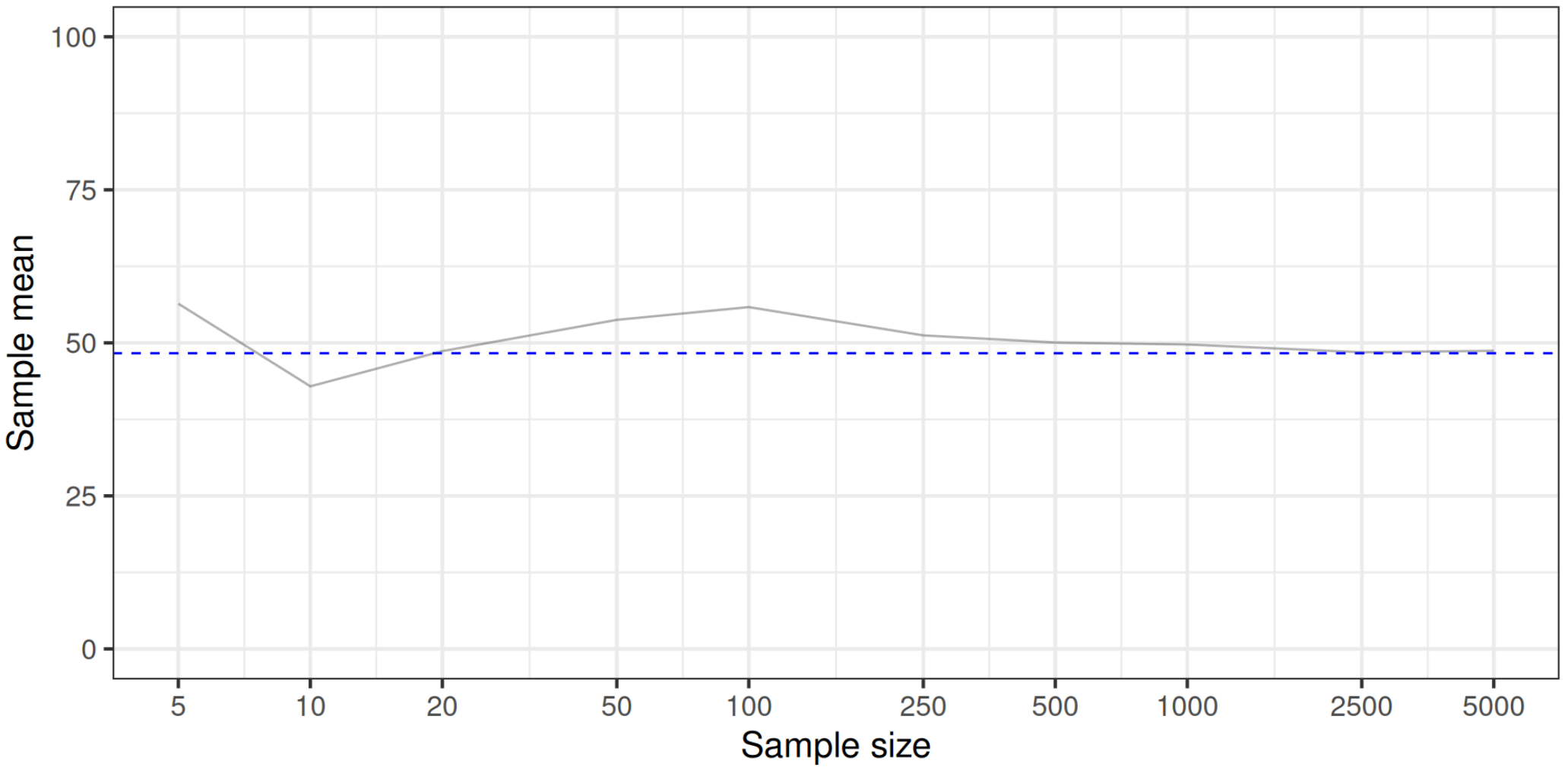
Law of large numbers, illustrated

Keep gradually increasing the sample size ...



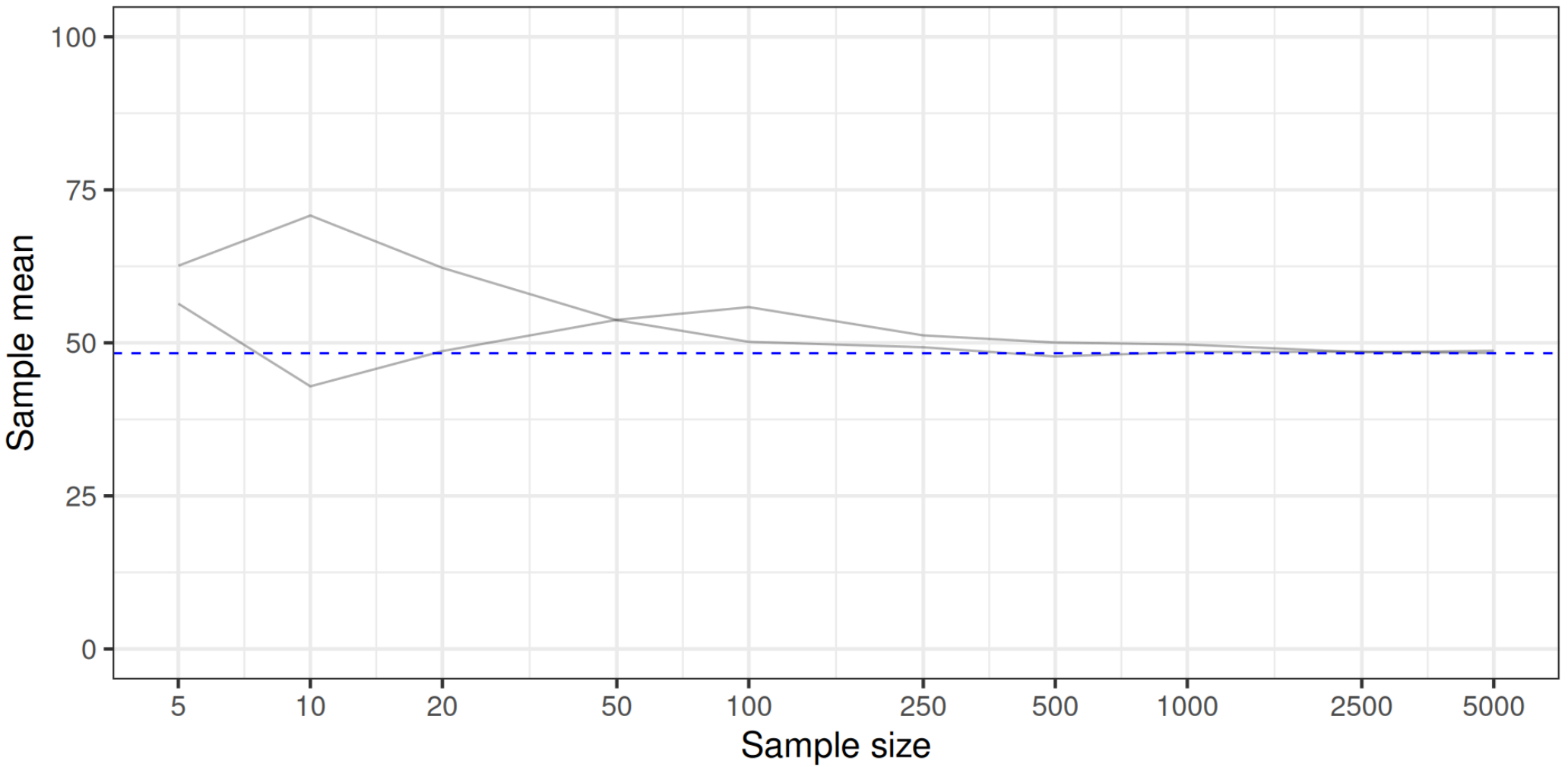
Law of large numbers, illustrated

...then repeat the process all over again with a totally new sample



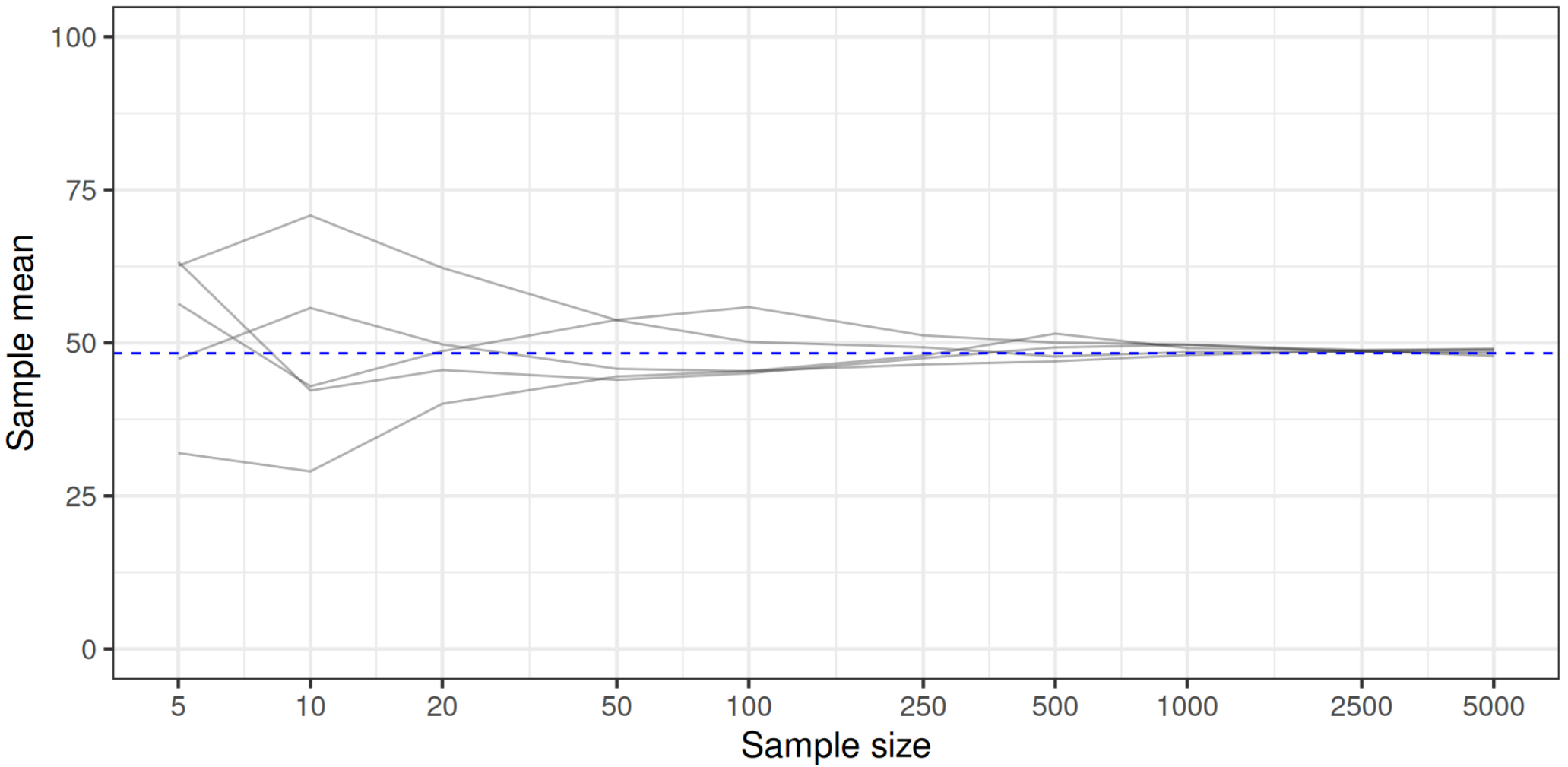
Law of large numbers, illustrated

...then repeat the process all over again with a totally new sample



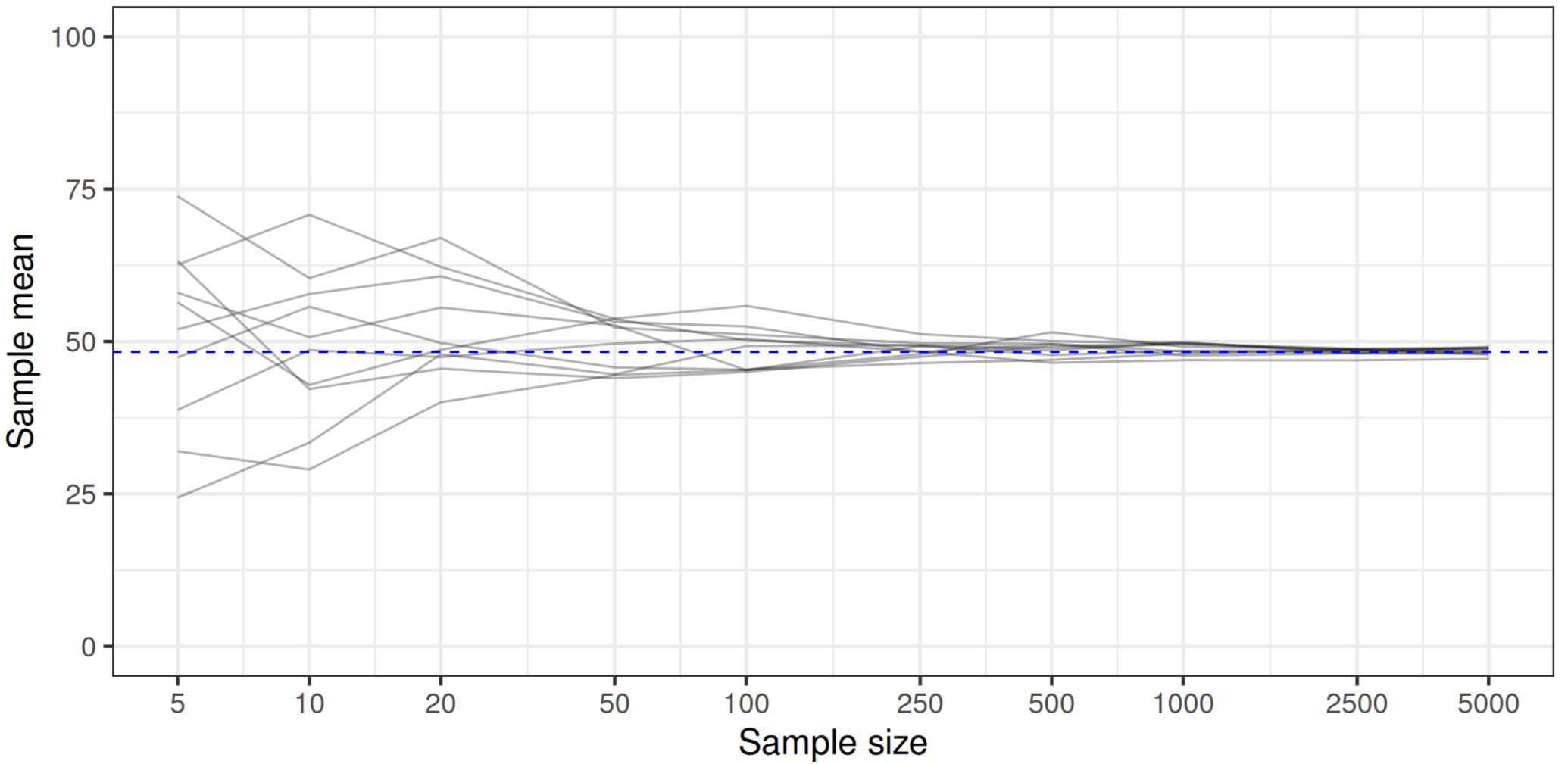
Law of large numbers, illustrated

...then repeat the process all over again with a totally new sample



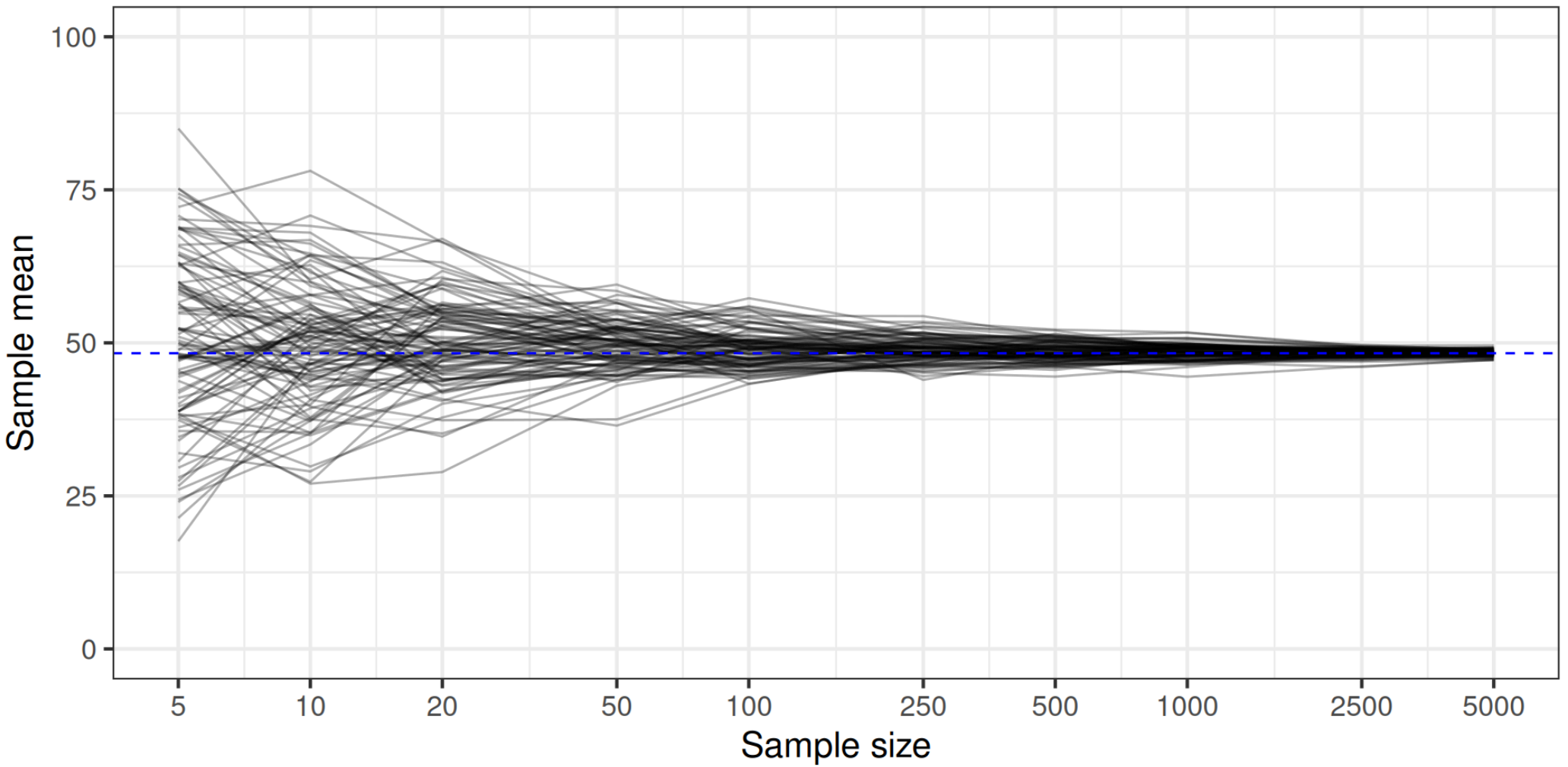
Law of large numbers, illustrated

...then repeat the process all over again with a totally new sample



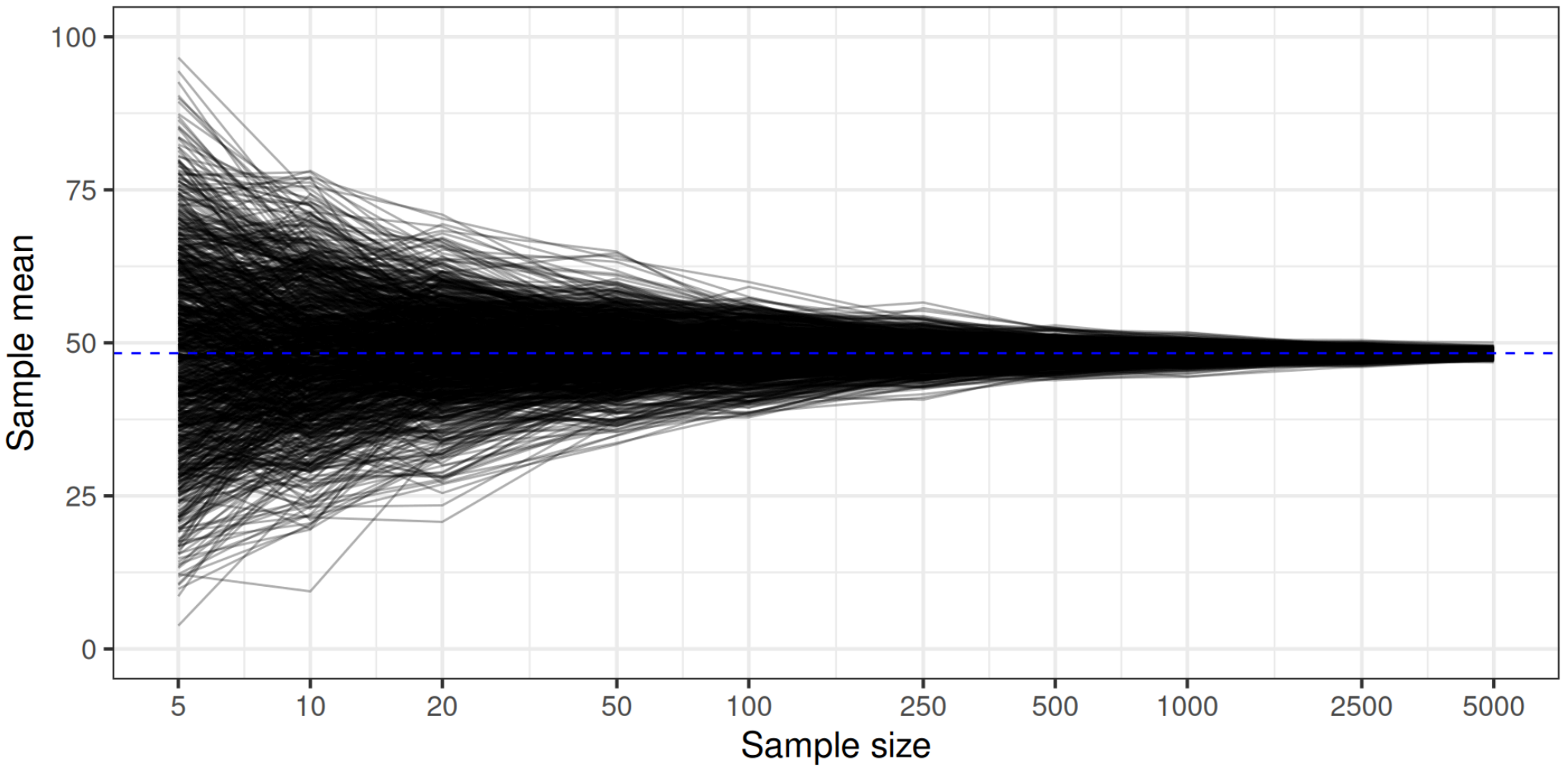
Law of large numbers, illustrated

...then repeat the process all over again with a totally new sample



Law of large numbers, illustrated

Lower proportion of big errors as sample size increases



Law of large numbers: Summary

Lessons to take from the law of large numbers:

- The sample mean is a good guide to the population mean
- Especially when the sample size is large

Unanswered questions:

- How far off is my sample mean from the population mean?
- How far off is a sample mean likely to be?

Central limit theorem

From LLN to CLT

LLN tells us *that* the sample mean $\approx$ the population mean in large samples

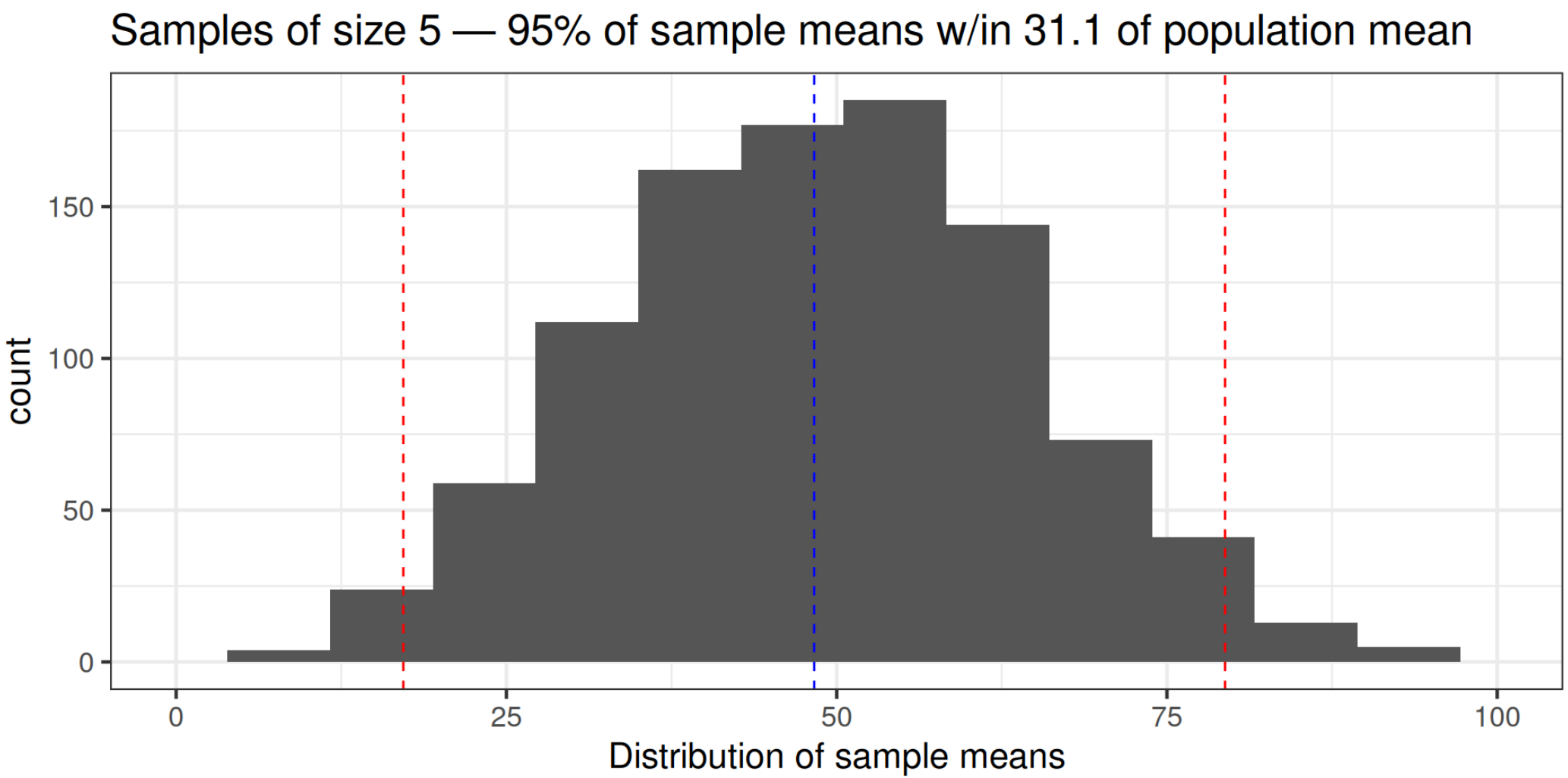
Central Limit Theorem (CLT) tells us *how* close we can expect it to be

Central Limit Theorem

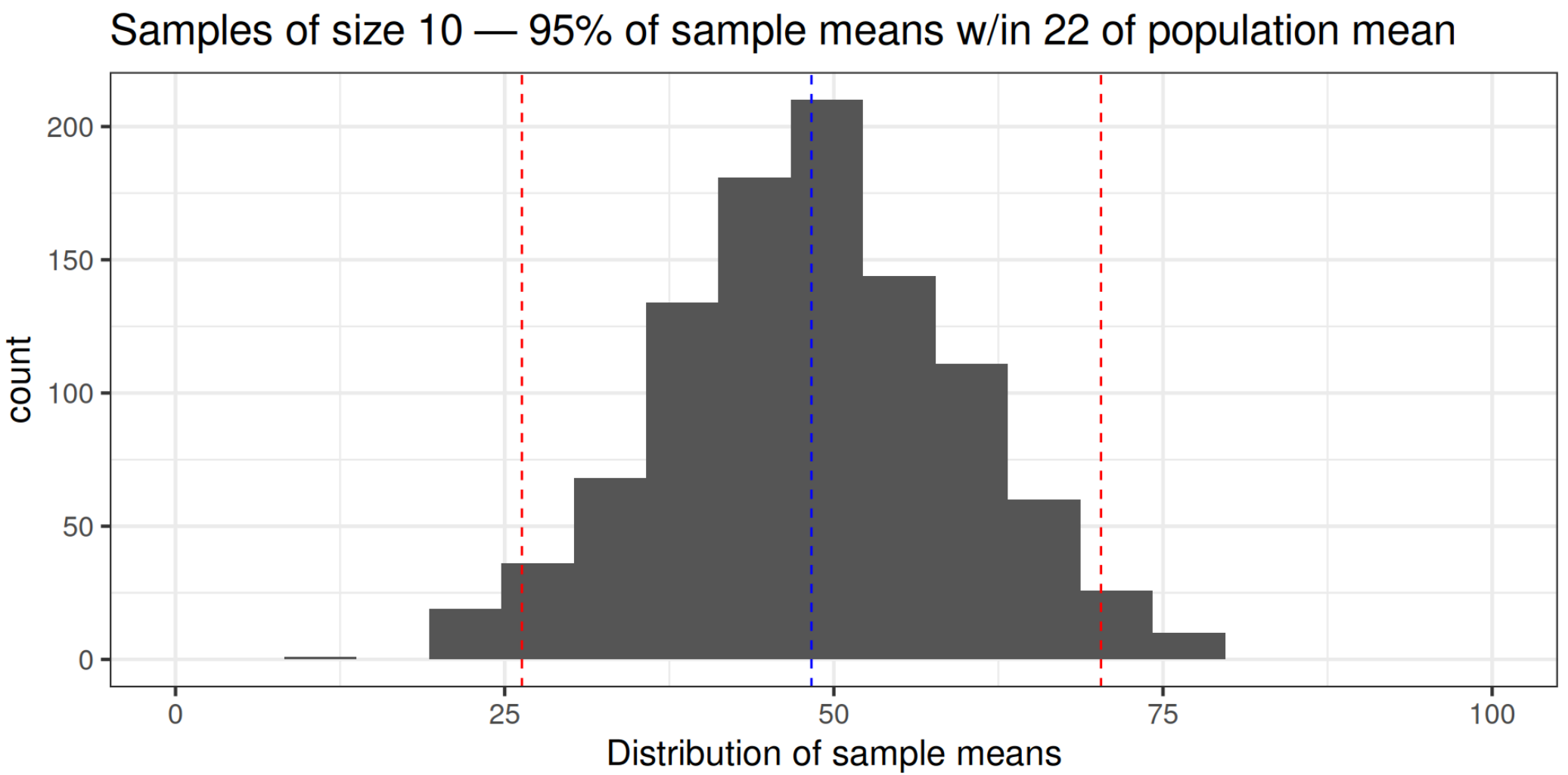
If the sample size N is large, then the distribution of the sample mean $\mathbf{avg}[X_i]$ across samples is approximately normal, with mean $\mathbb{E}[X_i]$ and standard deviation $\sqrt{\mathbb{V}[X_i]/N}$ (aka the standard error).

- Normal distribution: bell curve shaped
  - → 68% of observations within 1sd of mean
  - → 95% of observations within 2sd of mean
- CLT does not assume the distribution of X_i itself is normal

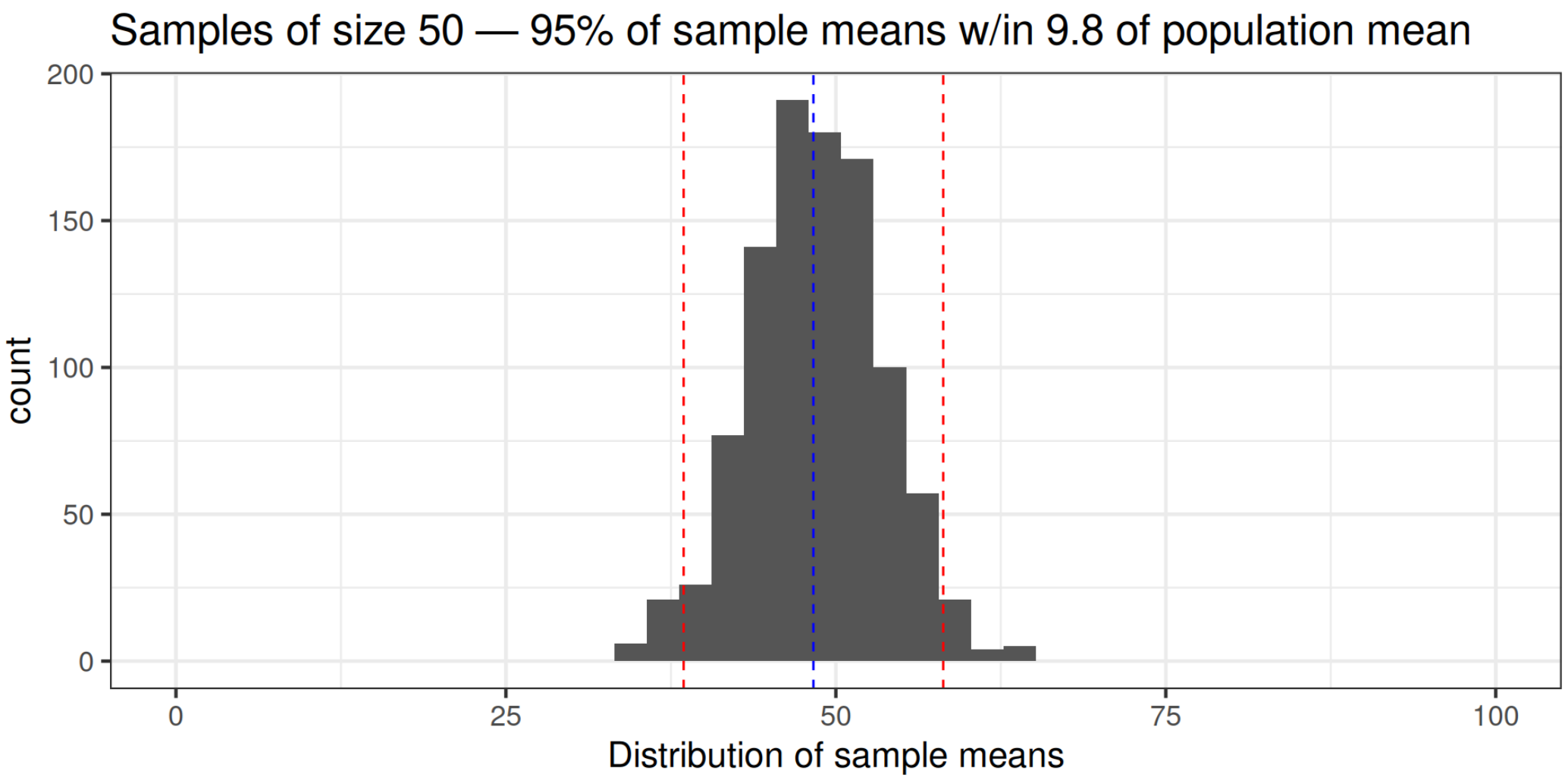
Central limit theorem, illustrated



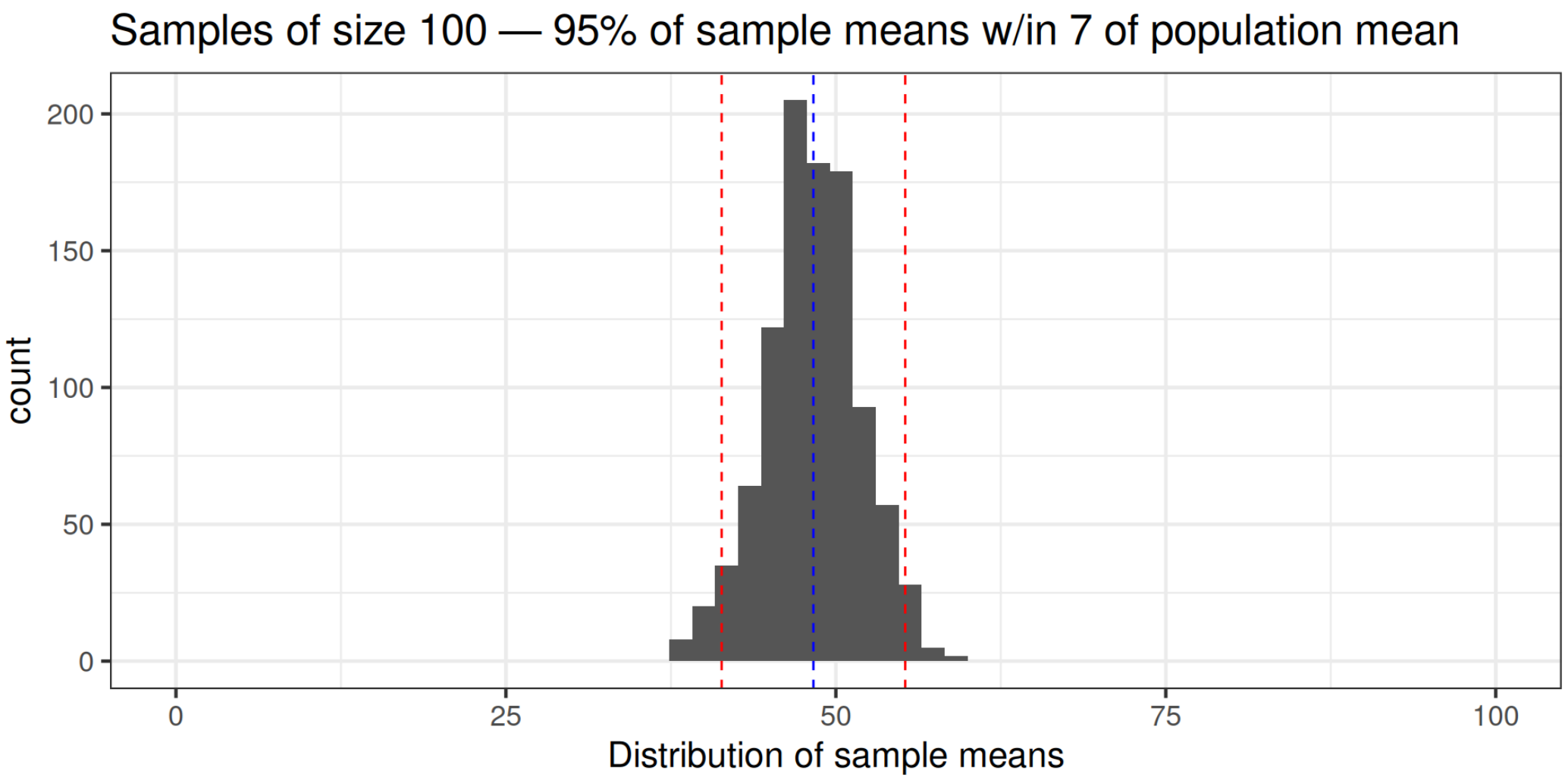
Central limit theorem, illustrated



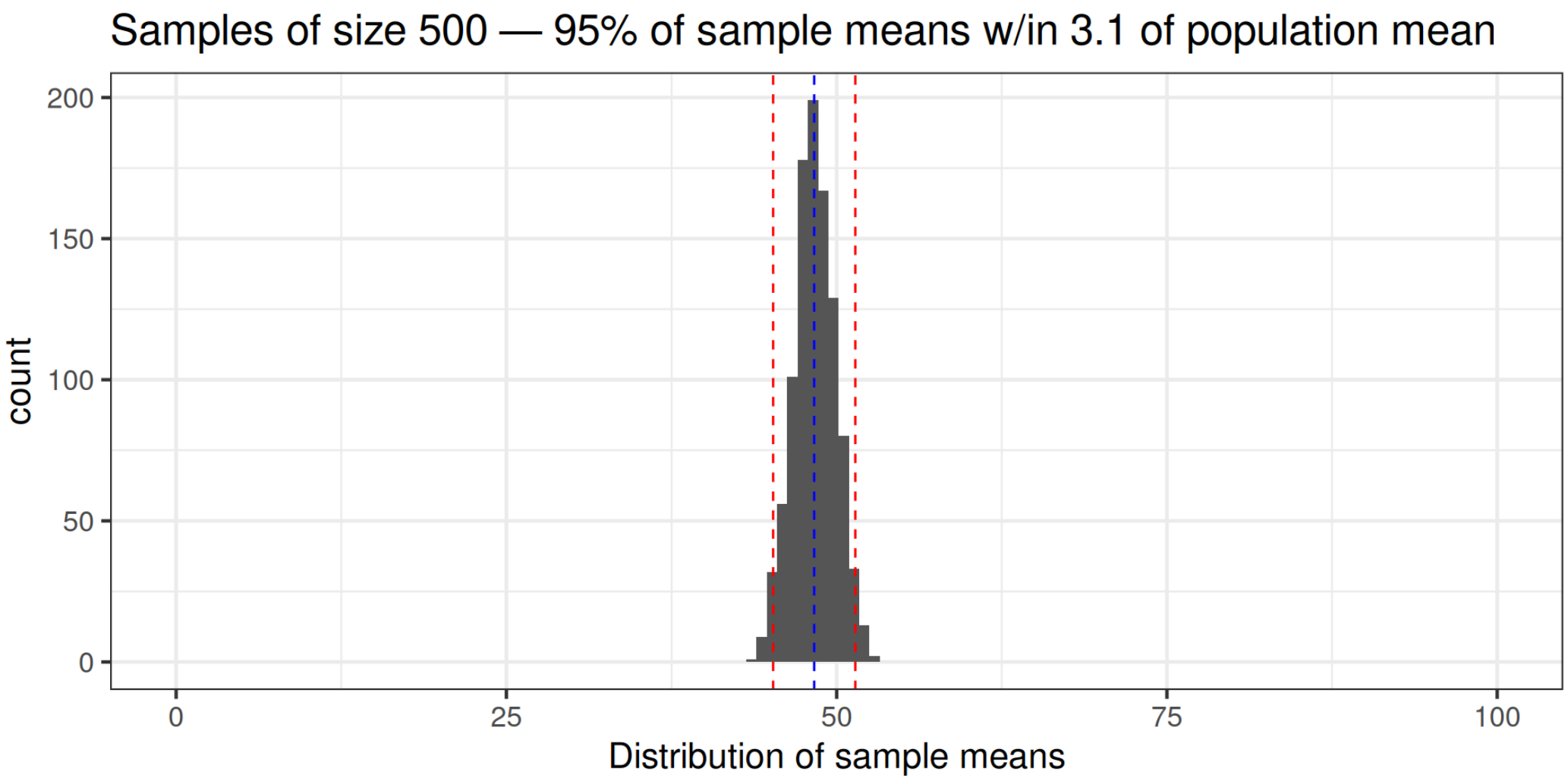
Central limit theorem, illustrated



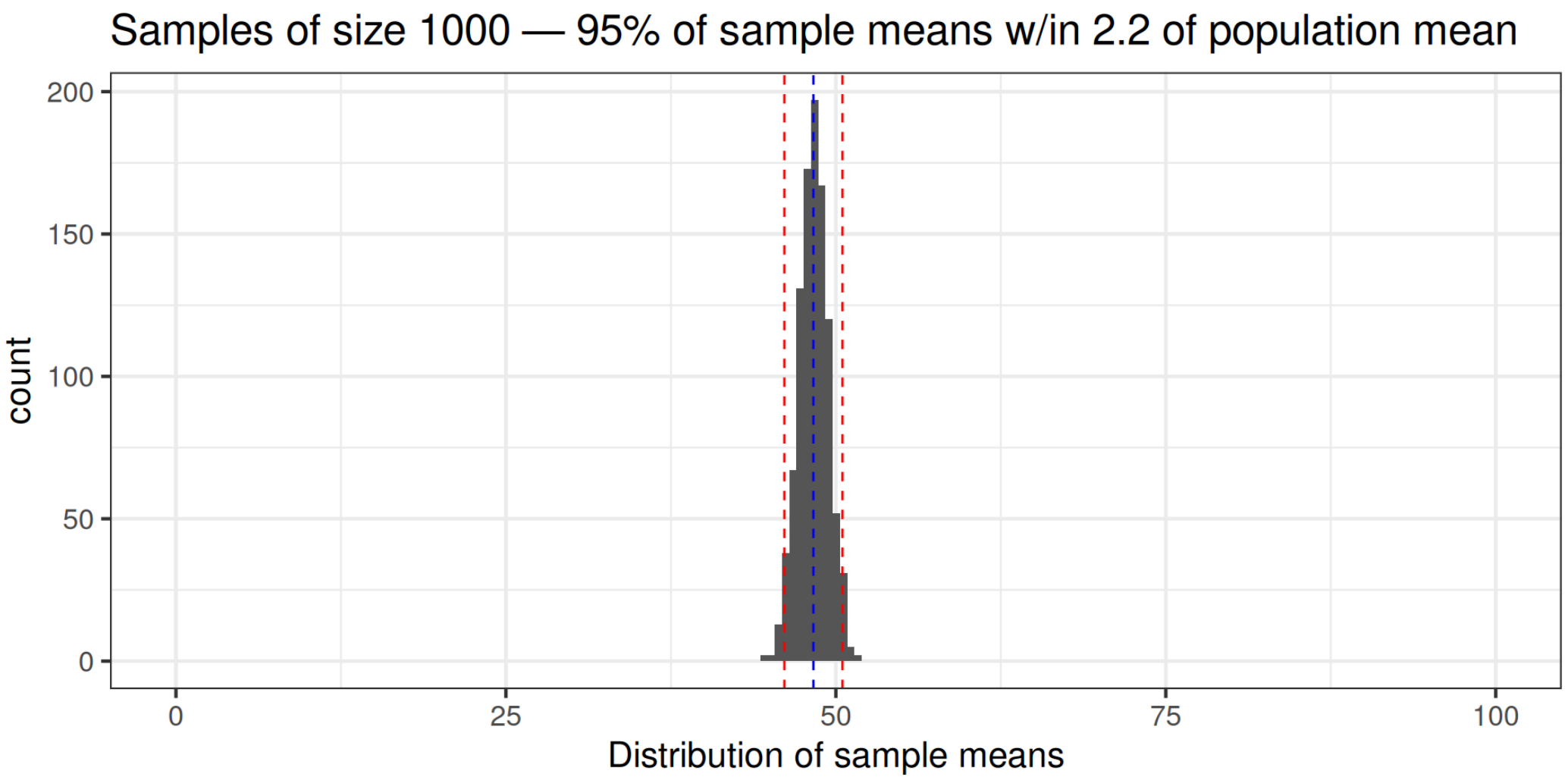
Central limit theorem, illustrated



Central limit theorem, illustrated



Central limit theorem, illustrated



Using the central limit theorem

95% of sample means within this margin of population mean:

$$\frac{2 \cdot \mathrm{sd}[X_i]}{\sqrt{N}}$$

Diminishing returns: Must quadruple sample size to halve the margin of error

If you want a 95% chance of sample mean within M of population mean, need sample of

$$N \geq \frac{4 \cdot \mathrm{sd}[X_i]^2}{M^2}$$

Central limit theorem with binary variables

One issue with calculating needed N: $\mathbf{sd}[X_i]$ possibly unknown in advance

But if X_i is binary, we know $\mathbf{sd}[X_i] = \sqrt{\mathbb{E}[X_i](1 - \mathbb{E}[X_i])} \leq 0.5$

Allows us to figure out "worst case" sample sizes

```
tibble(desired_margin = seq(0.07, 0.01, by = -0.01),
       needed_sample = ceiling(4 * 0.25 / desired_margin^2))
```

```
# A tibble: 7 × 2
  desired_margin needed_sample
           <dbl>         <dbl>
1         0.07             205
2         0.06             278
3         0.05             400
4         0.04             625
5         0.03            1112
6         0.02            2500
7         0.0100         10000
```

Hypothesis testing

Basics of hypothesis testing

Philosophy: Set up procedures with a known failure rate

Procedure:

1. Set up a "null hypothesis" — e.g., "50% of voters approve of Trump"
2. Identify the "test statistic" to calculate
    - Usually statistic of interest divided by its standard error
3. Calculate distribution of test statistic if null hypothesis is true
    - CLT is what makes this feasible!
4. Compare to "critical value" to reject null hypothesis with known failure rate
    - If null hypothesis is true, then will (incorrectly) reject in designated percentage of samples
    - Can't tell from any given sample if null is false, or if we were unlucky!

Wrapping up

What we did today

1. LLN: As N increases, sample mean is less likely to be too far off the population mean
2. CLT: For $N \geq 30$ or so, expect sample mean to be within $2\,\text{sd}[X_i]/\sqrt{N}$ of pop mean in 95% of samples
3. Hypothesis testing: Leverage the CLT to set up tests with a known "failure" rate in case null hypothesis is true

Next time — Inference for treatment effects