Welcome to STA 101!

Midterm 2: November 12 - 14

Three weeks from this coming Tuesday:

- **Same format**: 70% in-class (no tech); 30% take-home (tech);
- There will *not* be a page of free points;
- Same extra credit scheme:
  - +1 practice in-class
  - +1 practice take-home
  - +1 review sheet
  - +1 lab review attendance
  - +1 cheat sheet
- **Cumulative**: Ch. 1 - 17(?) all fair game
  - Ch. 11 - 17(?) emphasized
  - Ch. 9 deemphasized

(if applicable, make appointments in the testing center now.)

What we've said about statistics so far

- It's a confrontation with uncertainty;
- It confronts uncertainty by quantifying it;
- After data analysis delivers a "best guess" at the answer to a quantitative question, statistics shows up and supplements that guess with a margin of error;
- If the margin of error is large, uncertainty is high, and the guess is not so reliable;
- If the margin of error is small, uncertainty is lower, and maybe we can take the guess to the bank;
- The size of the margin of error is related to the sample size and how variable the data are. Small, variable datasets give less reliable conclusions than large, less variable ones.

Examples of questions we might ask

You ask a quantitative question:

- What is the "typical" lead level in the Flint MI drinking water?
- What is the probability that Cornel West wins the 2024 presidential election?
- How many jobs does a \$1.00 increase in the minimum wage create or destroy?

The answer would take the form of a single number.

All of (classical) statistics

Point estimation ("what's the best guess?")

- **Example**: what's my blood pressure?
- **Jargon**: point estimate

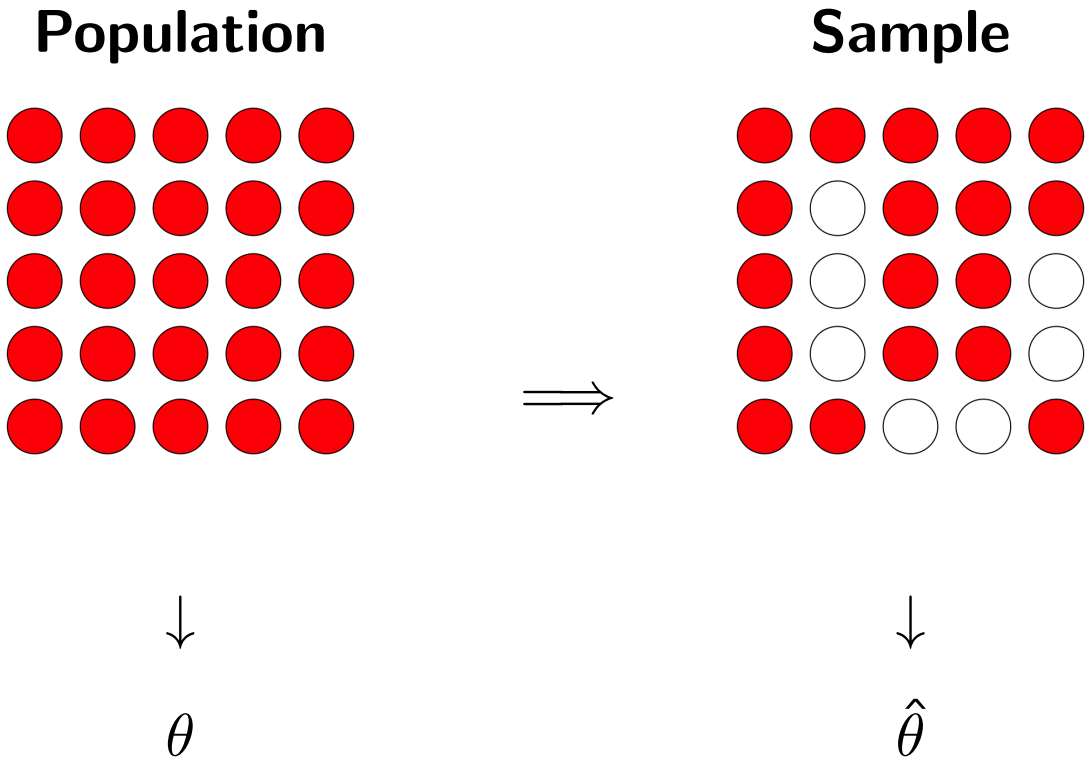
Interval estimation ("plus or minus what?")

- **Example**: what's a likely range for my blood pressure?
- **Jargon**: confidence interval, confidence level, ...

Hypothesis testing ("yeah, great, so...what can we conclude?")

- **Example**: soooo...am I at risk, or not?
- **Jargon**: test statistic, null distribution, significance level, p-value, Type 1 and 2 errors, power, ...

Population versus sample, again



Sampling uncertainty: how similar are sample and population?

Matters a lot for how close $\hat{\theta}$ will be to θ.

Sampling uncertainty

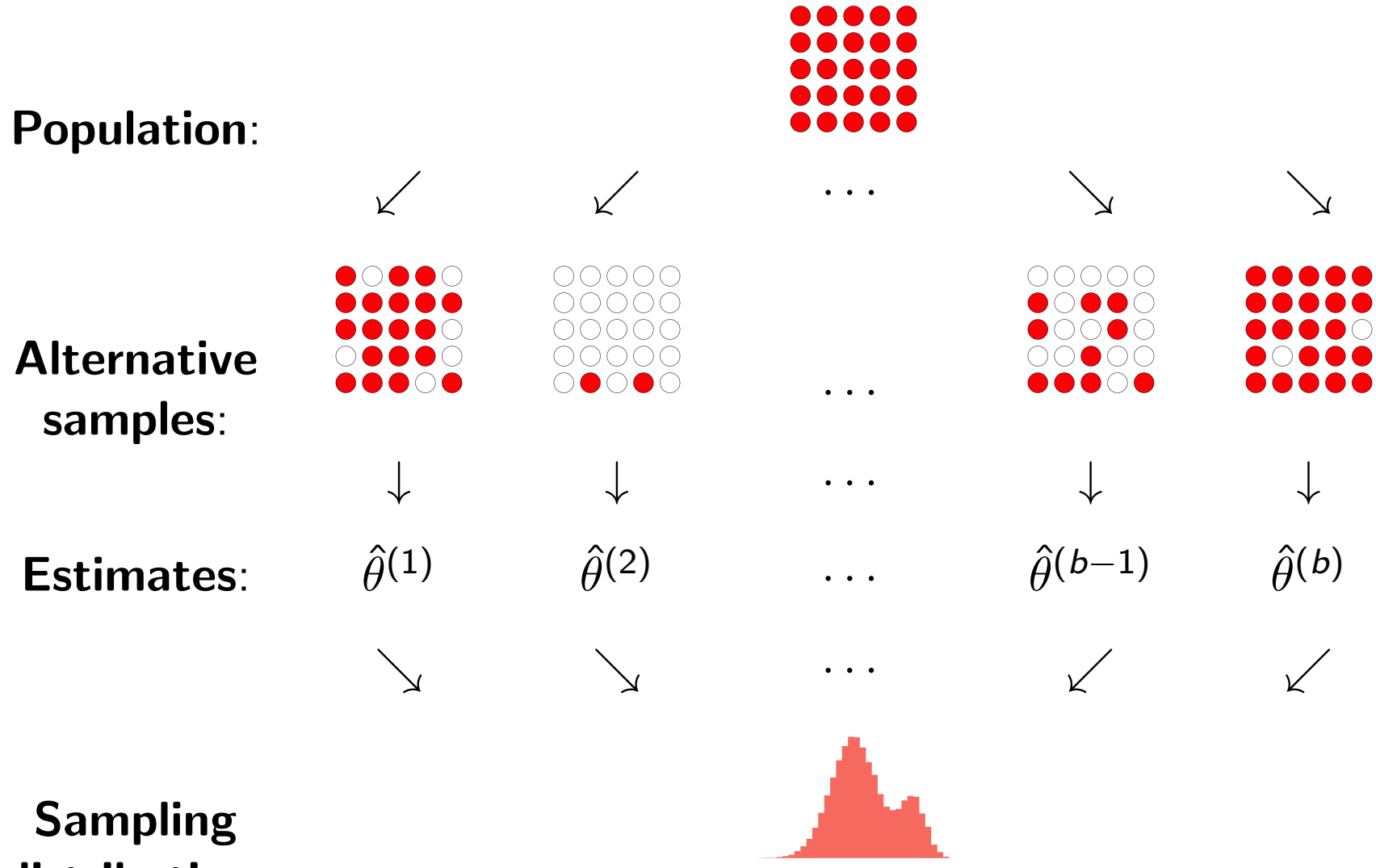
Reality: we only have one sample and one estimate;

Question: how reliable is that estimate?

Thought experiment: if, hypothetically, I had another sample, I would have a different estimate. How close would that estimate be to the one I actually have?

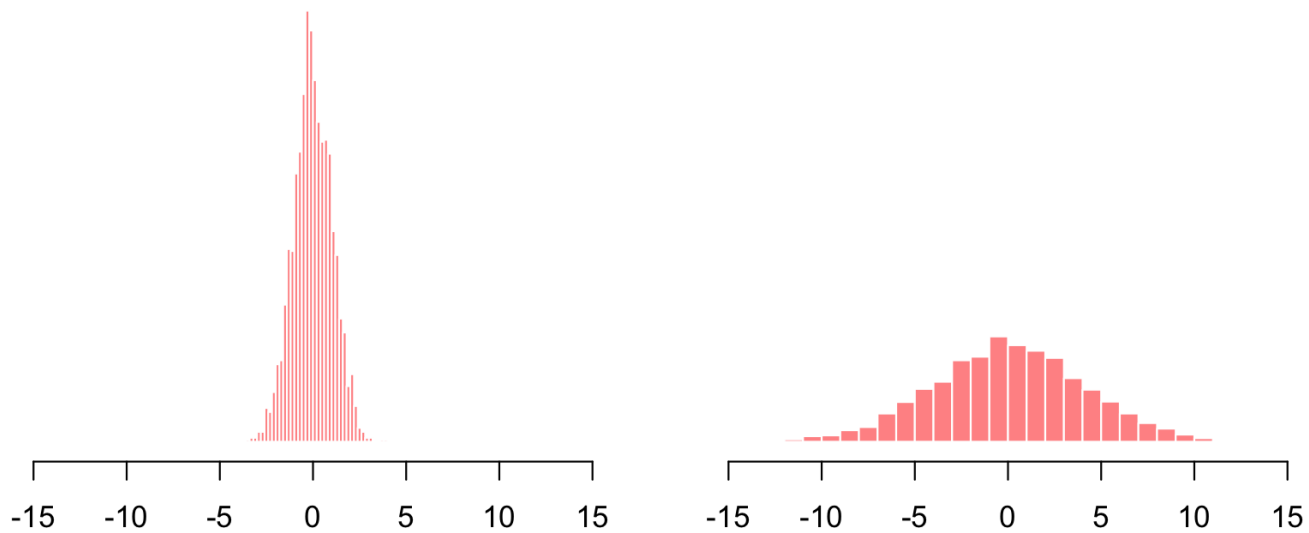
- if close, maybe my answer is pretty good;
- if far apart, can I trust the answer I have?

Sampling uncertainty

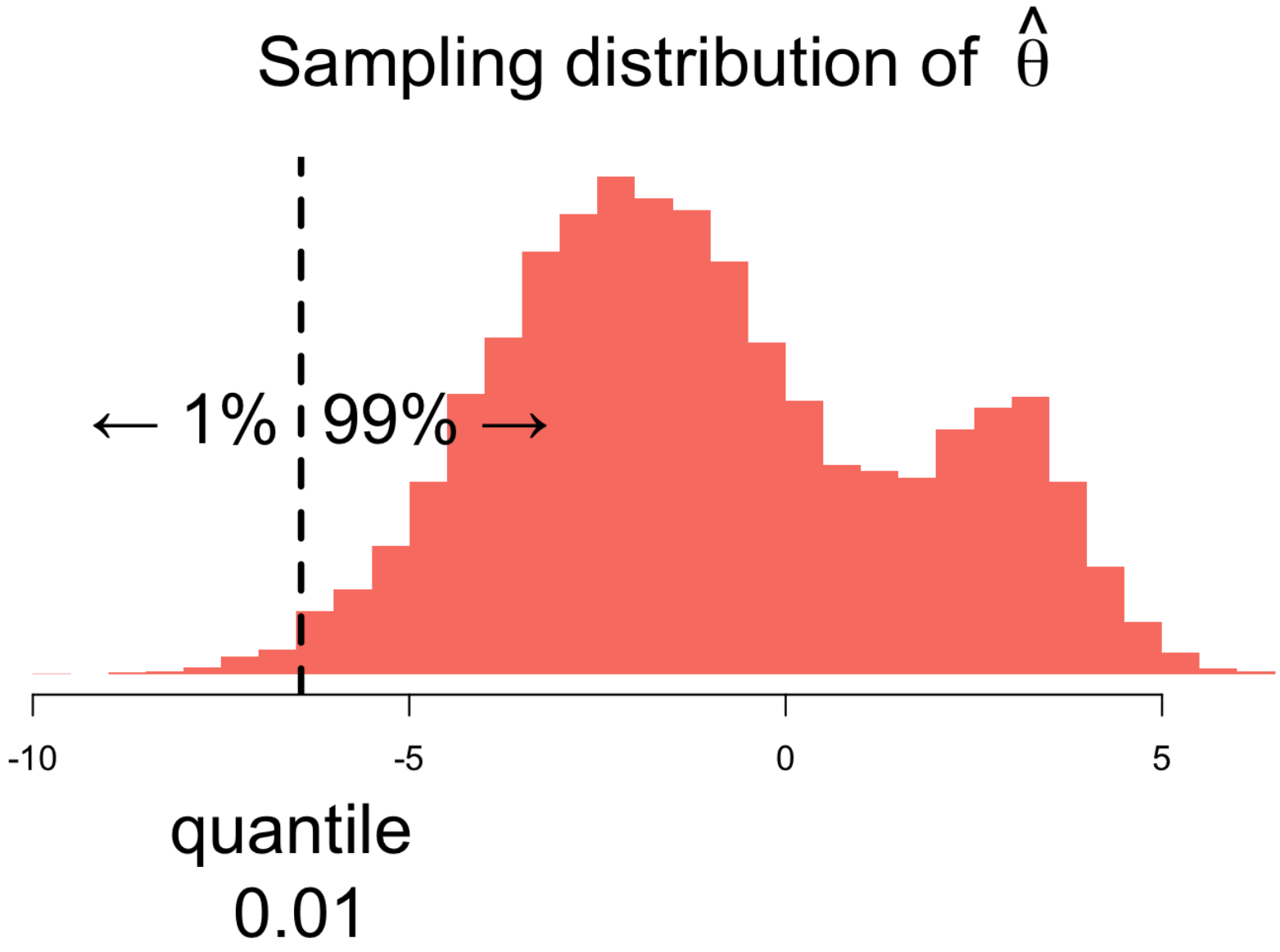


The sampling distribution

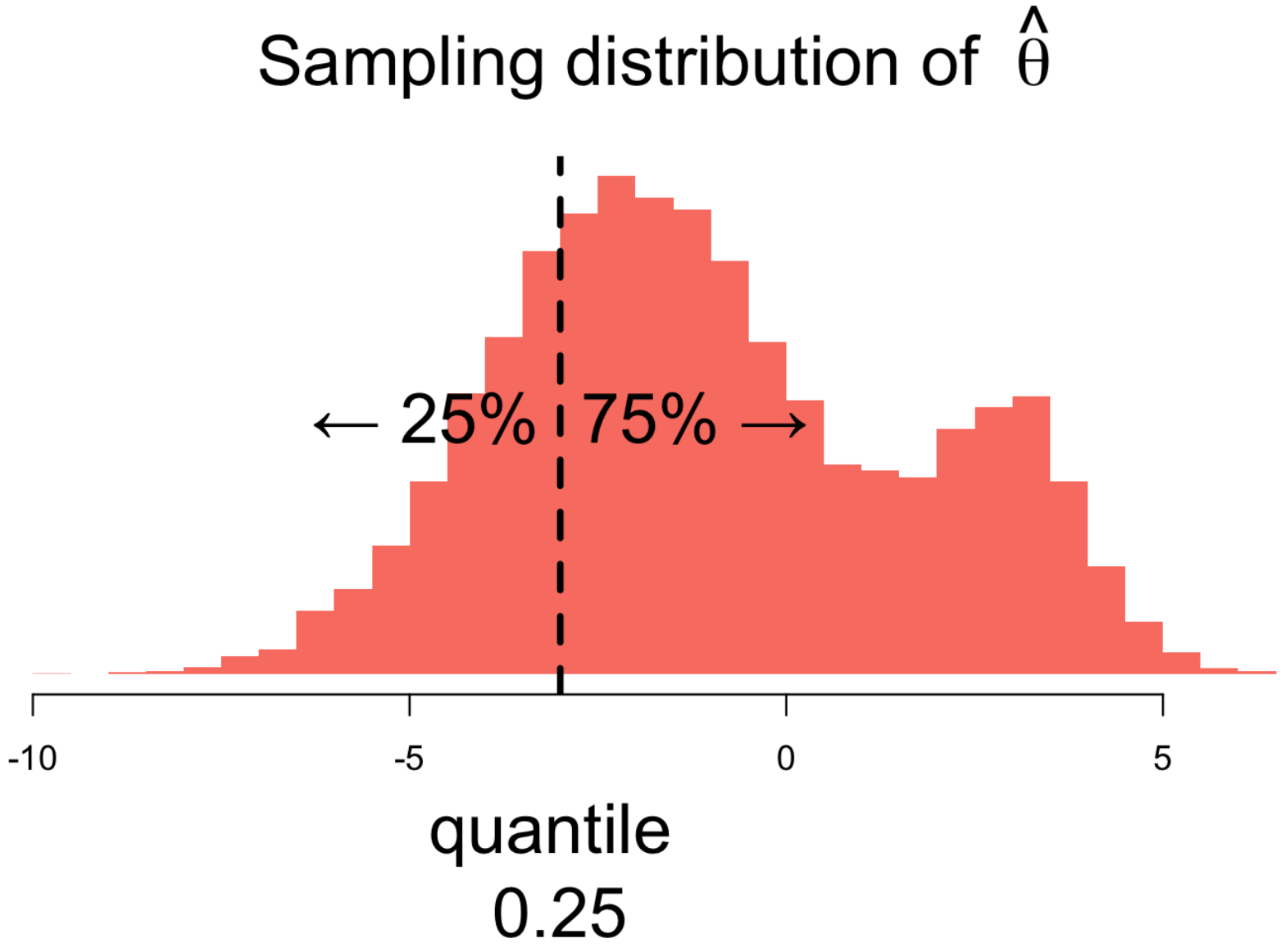
- The sampling distribution of an estimator describes the random variation you would observe in your guess if you based it on different, alternative samples from the population;
- If the *spread* of this distribution is low, *any* dataset would give basically the same answer, and uncertainty is low;
- If the spread of this distribution is high, your answer is highly sensitive to the sample, and uncertainty remains high.



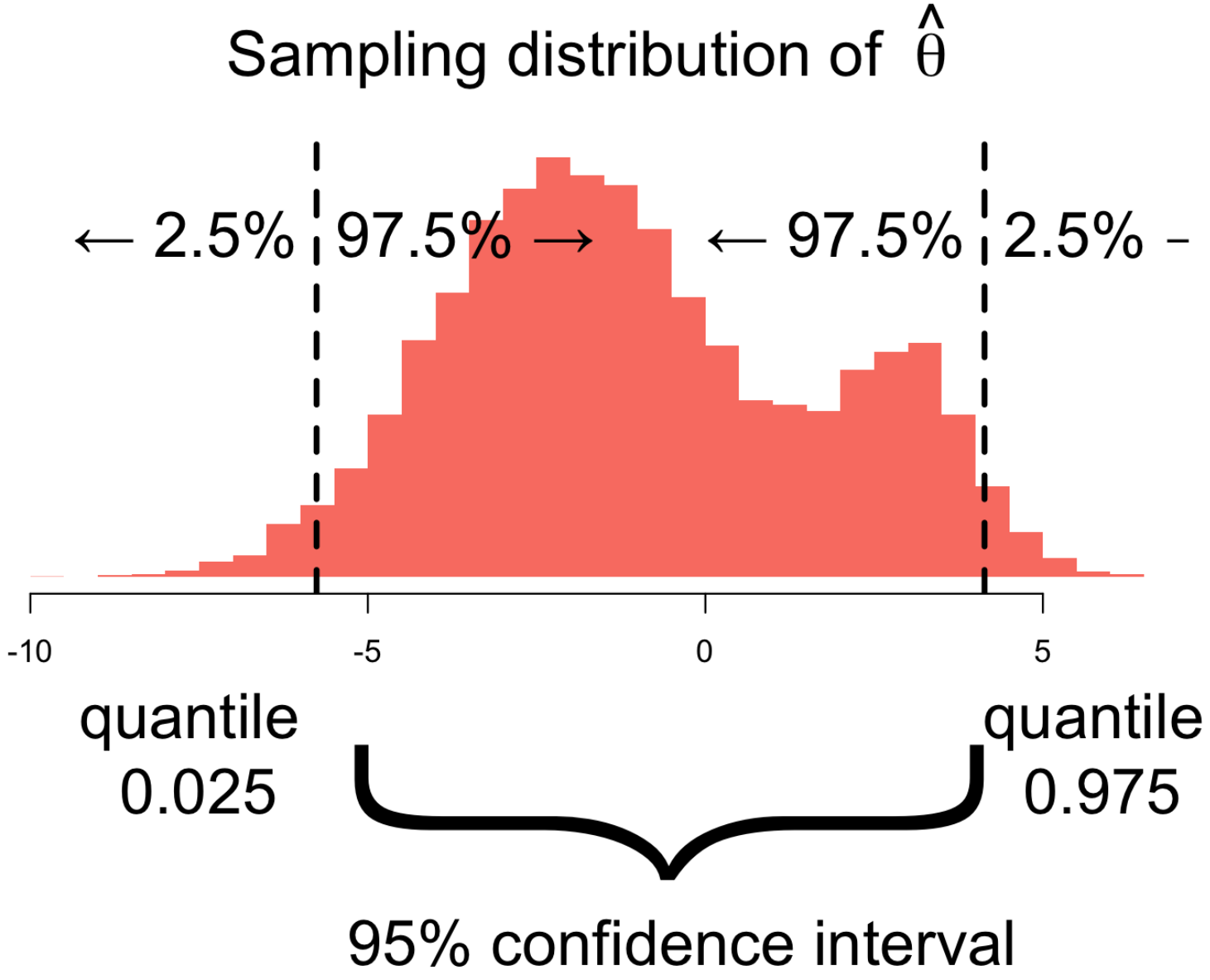
Quantiles of a distribution



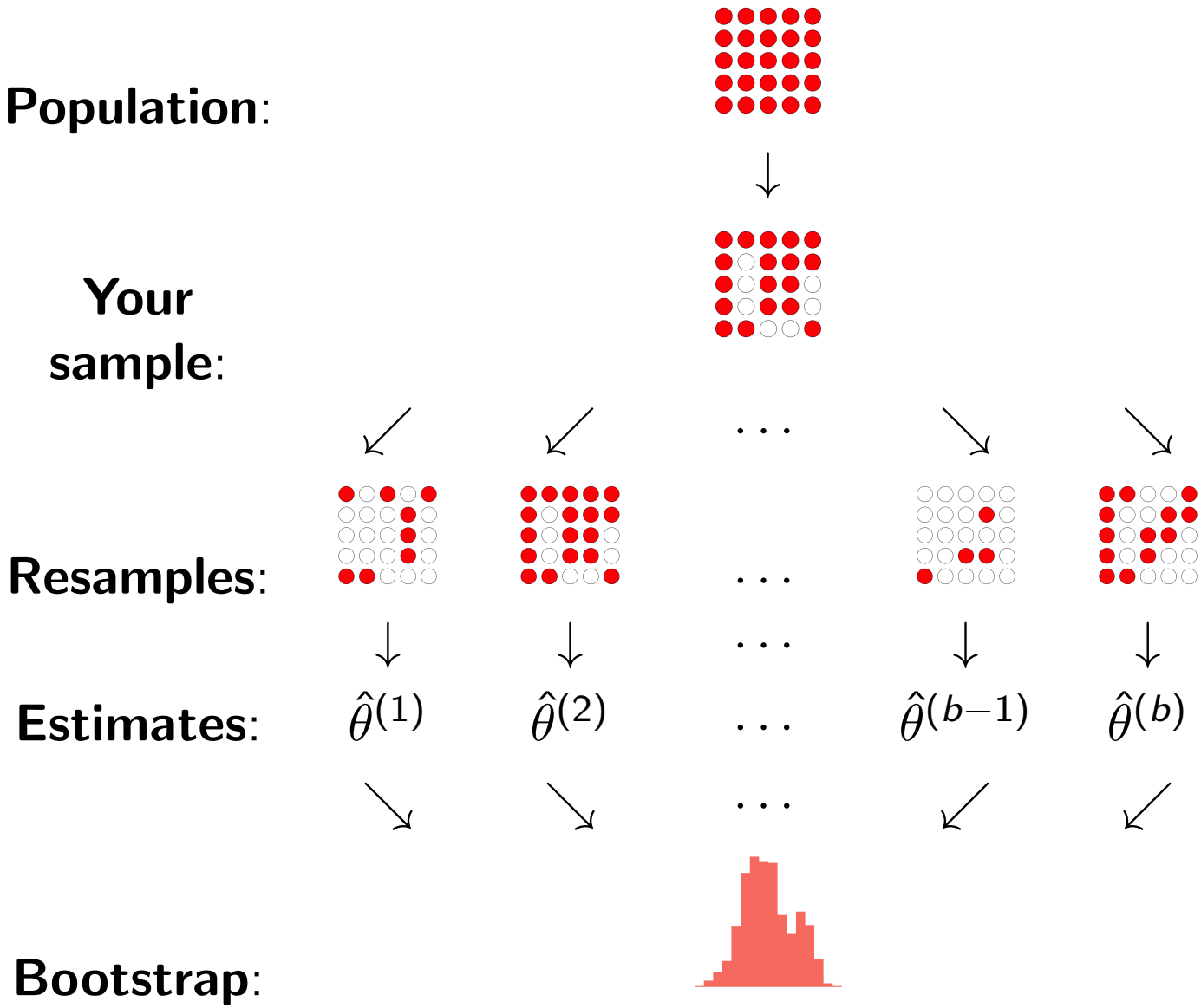
Quantiles of a distribution



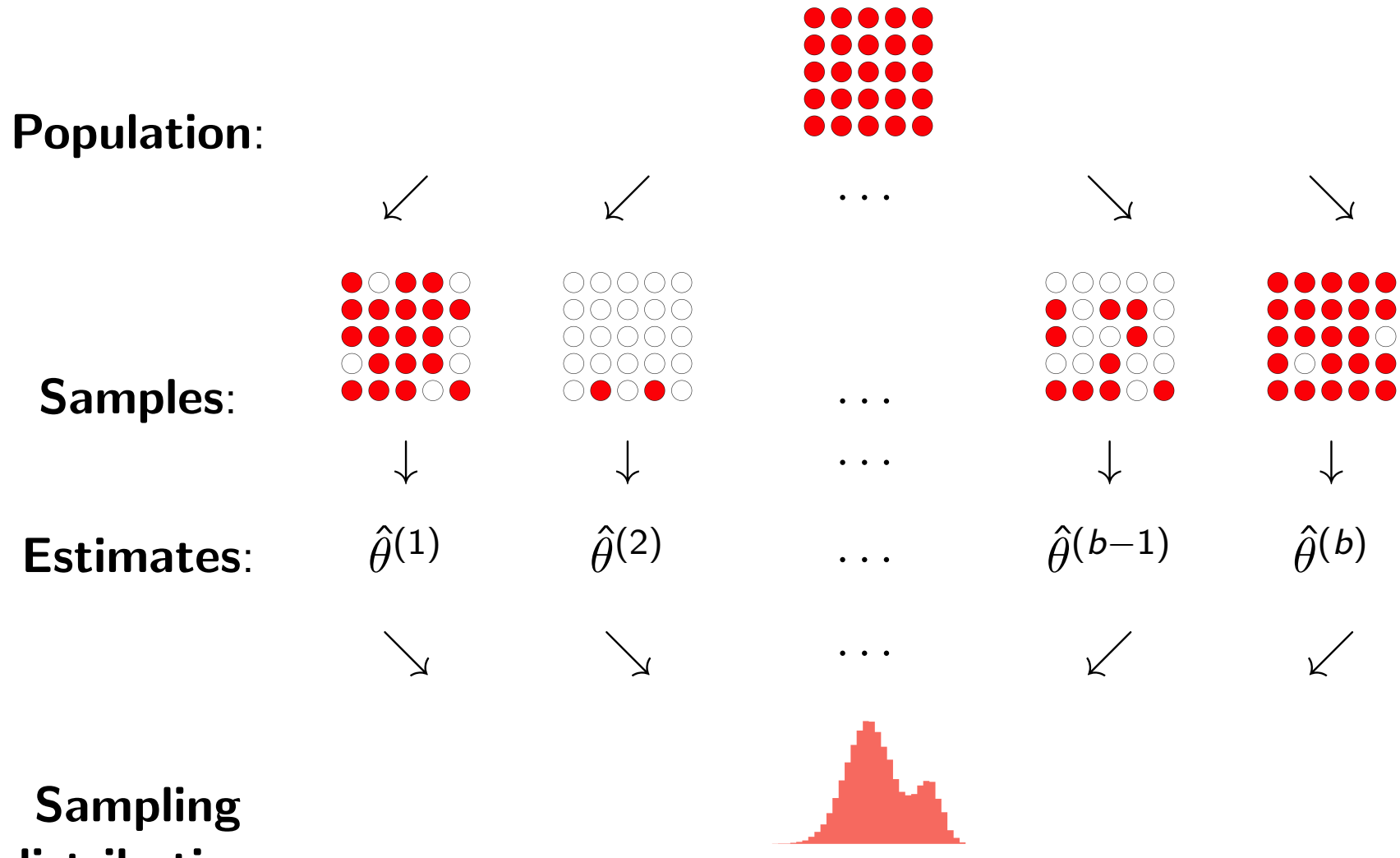
A 95% confidence interval based on quantiles



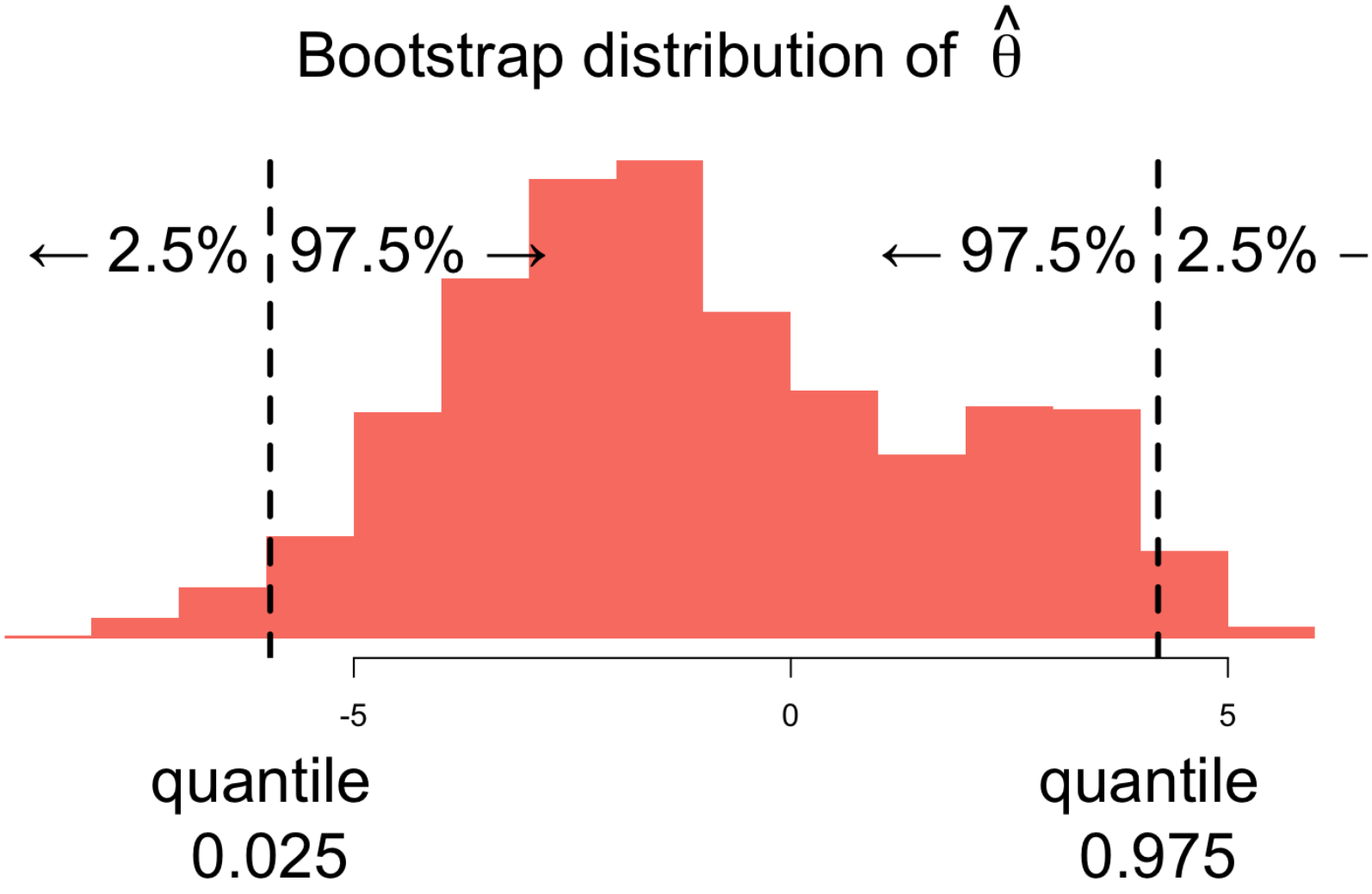
Approximating the sampling distribution with the bootstrap



Compare: ideal procedure we cannot do



Approximating confidence intervals with the bootstrap



Cardinal Sins in Statistics, Part 1 of 83

Interpreting a confidence interval:

- **NO**: there is a 95% chance that the true value is in the interval;
- **YES**: an interval *like this* will contain the true value in 95% of samples I could have seen.

Note: This is a subtle and difficult point that trips up even smart, experienced people that should know better.

(play with this: `https://rpsychologist.com/d3/ci/`)