Welcome to STA 101!

Recap: hypothesis testing in seven questions

What is an hypothesis test trying to do?

Silly example: flipping an unfamiliar coin.

Competing claims (hypotheses):

$$\begin{aligned} H_0 &: p = 0.5 && \text{(coin is fair)} \\ H_A &: p \neq 0.5 && \text{(coin is unfair)} \end{aligned}$$

What if: we flip the coin a bunch of times and get 51% heads, which is not exactly equal to 50%. So what?

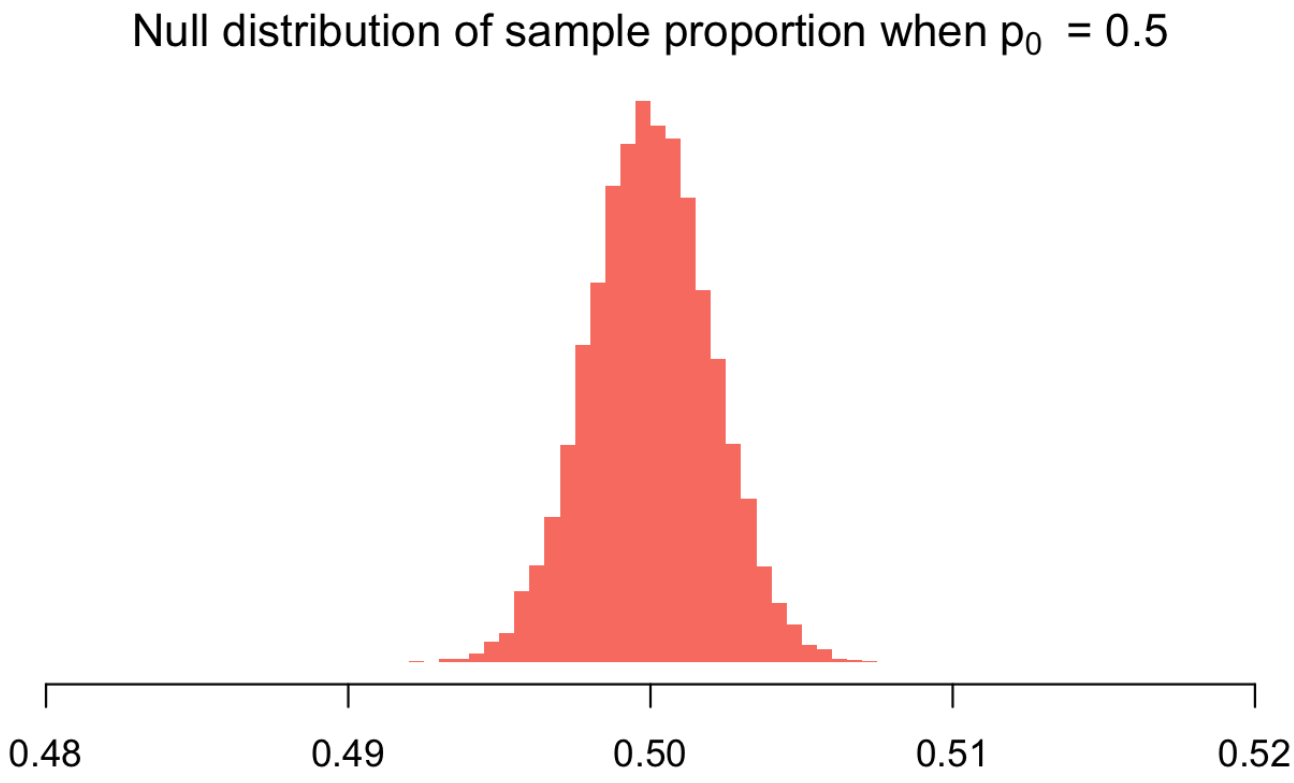
Two possibilities:

- Fair coin. We just got 51% as a quirk of the random sampling;
- Unfair coin. $51\% \neq 50\%$. Anomaly detected! Case closed!

An hypothesis test is trying to tell the difference between these.

If the null were true, what would the world look like?

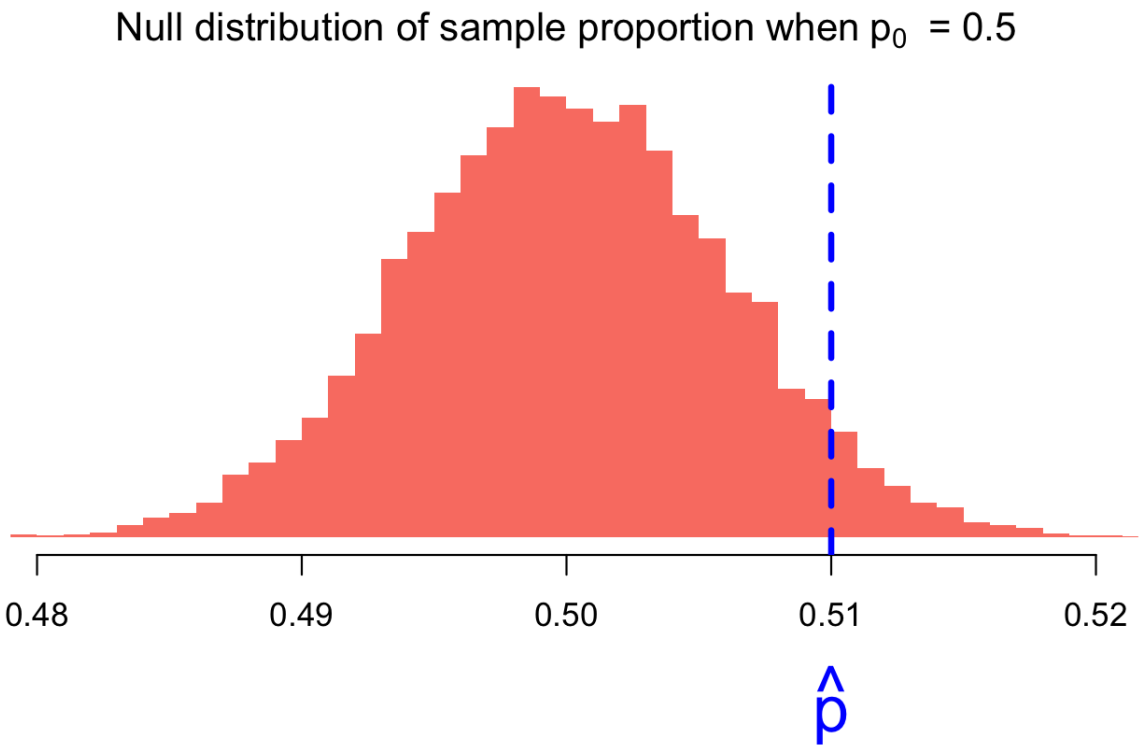
The null distribution is a *hypothetical* sampling distribution that visualizes how your estimates would vary across different random samples if the null were true.



If the null were true and random sampling was all that was going on, this would be the menu of options.

What does the world actually seem to look like?

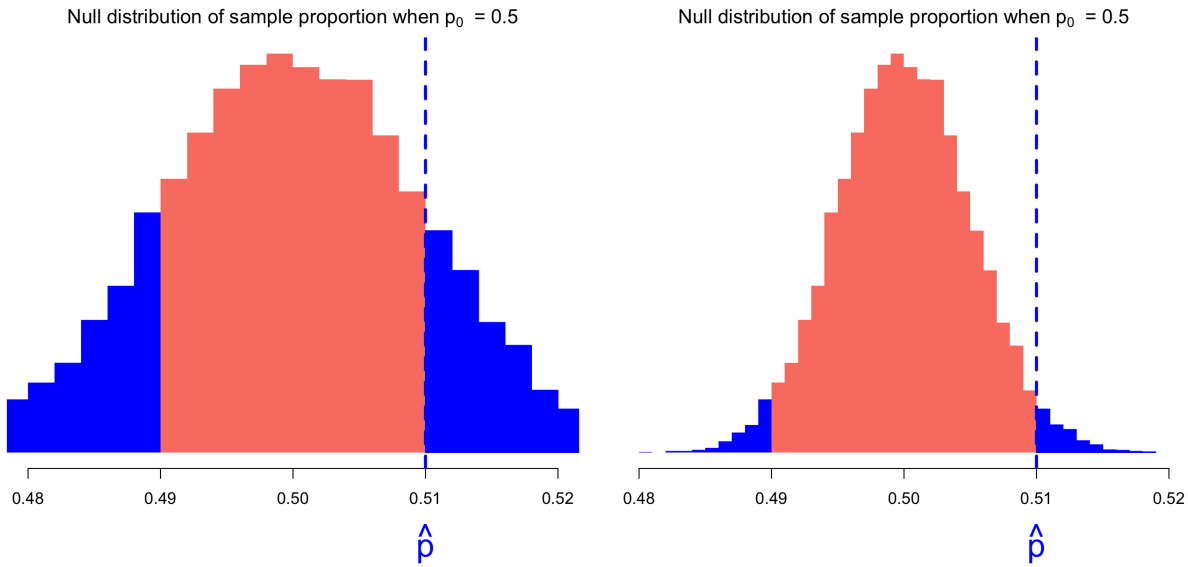
That's your point estimate. The actual answer from actual data:



- If the actual estimate is toward the middle of the null distribution, maybe you cannot rule out the null;
- If the actual estimate is out in the tails of the null distribution, maybe the null is totally bogus and should be rejected.

How do we quantify the difference between the hypothetical and reality?

Calculate the *p*-value: the probability, assuming the null is true, of an estimate *even more extreme* than the one you actually got.



- **"big" *p*-value**: your estimate is not out of the ordinary in a world where the null is true;
- **"small" *p*-value**: your estimate is unlikely if the null is true.

How do we decide if the *p*-value is small enough to reject?

Set a cut-off $0 < \alpha < 1$ called the discernibility level and do this:

- if *p*-value $< \alpha$, Reject H_0;
- if *p*-value $>= \alpha$, Fail to reject H_0.

It's very crude and unglamorous.

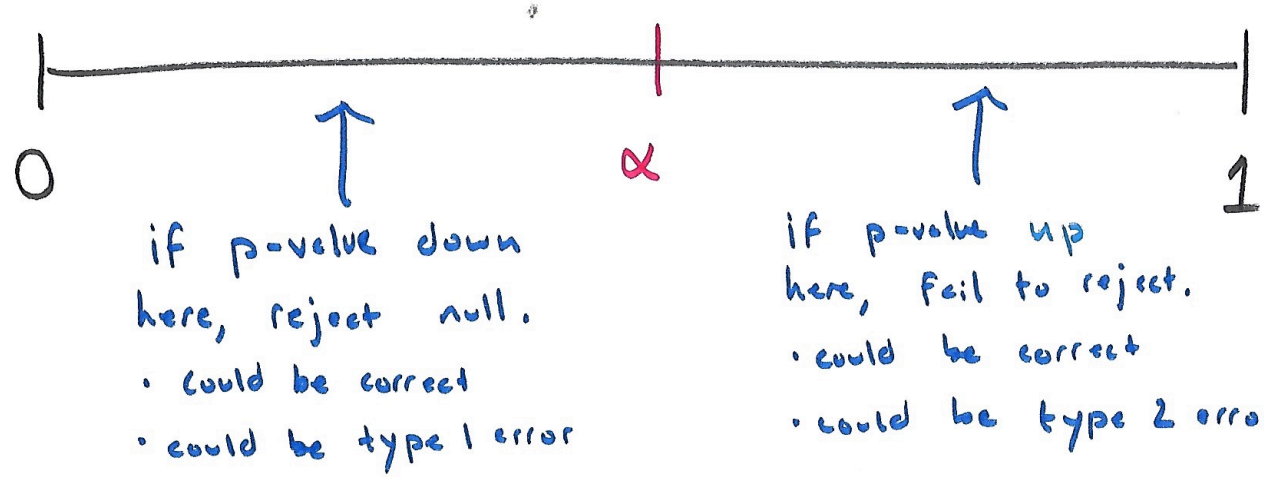
How do we pick the discernibility level?

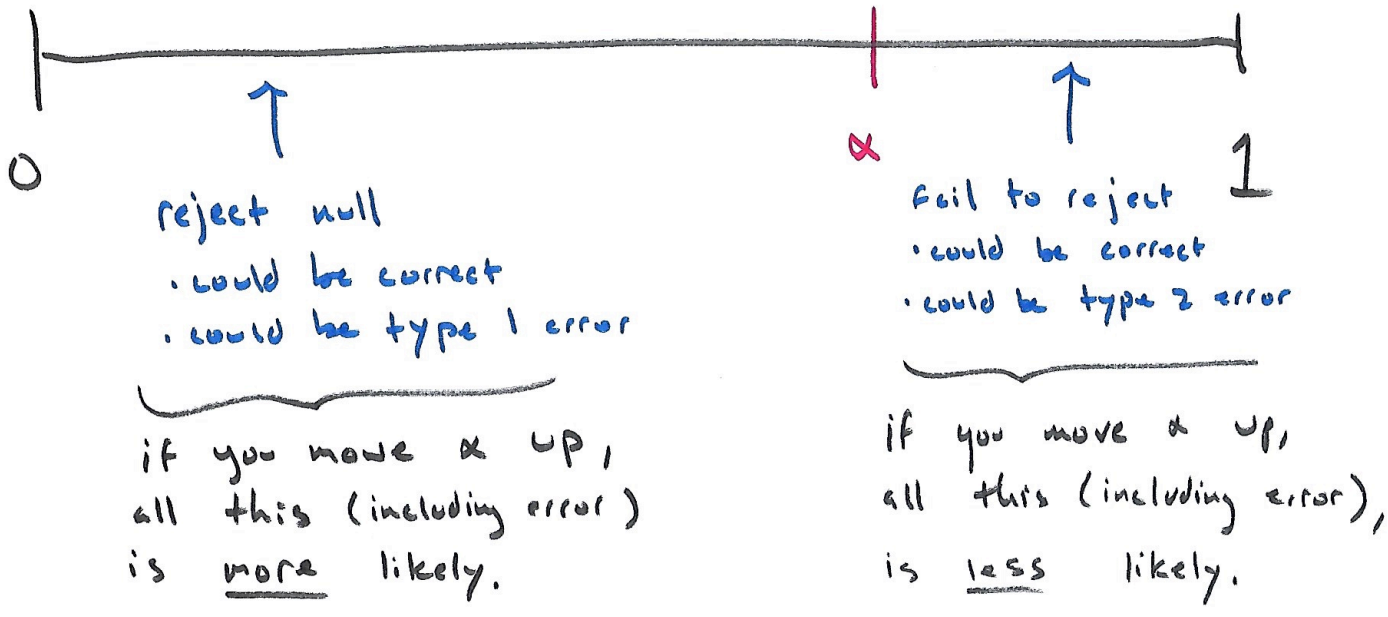
Pick α to balance the risk of two types of errors:

		Your decision: Reject H_0	Your decision: Fail to reject H_0
The truth	H_0 true	Type 1 error	Correct!
	H_0 false	Correct!	Type 2 error

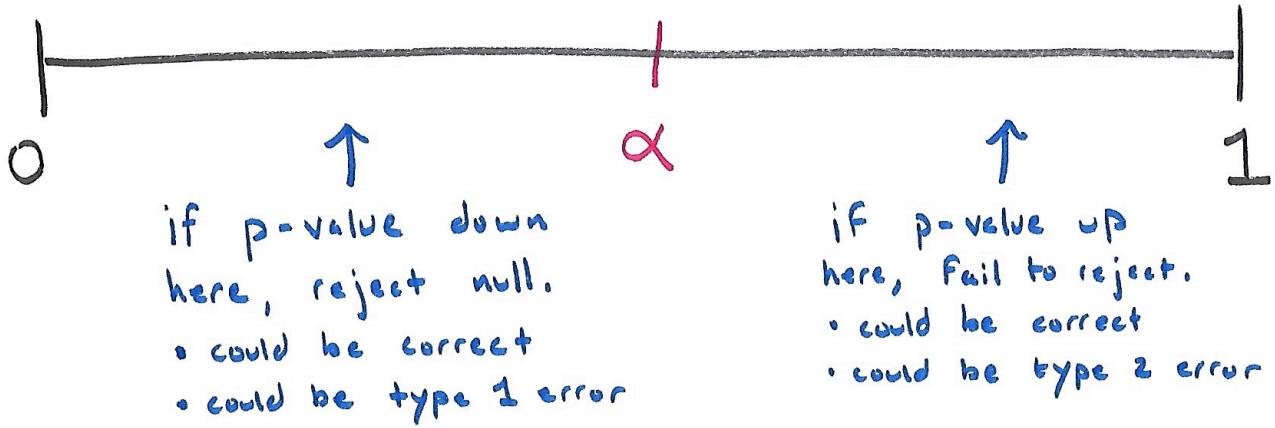
- $\alpha \uparrow \implies$ easier to reject H_0 $\implies$ Type 1 $\uparrow$ Type 2 $\downarrow$
- $\alpha \downarrow \implies$ harder to reject H_0 $\implies$ Type 1 $\downarrow$ Type 2 $\uparrow$
- **Typical choices**: $\alpha = 0.01, 0.05, 0.10, 0.15$.

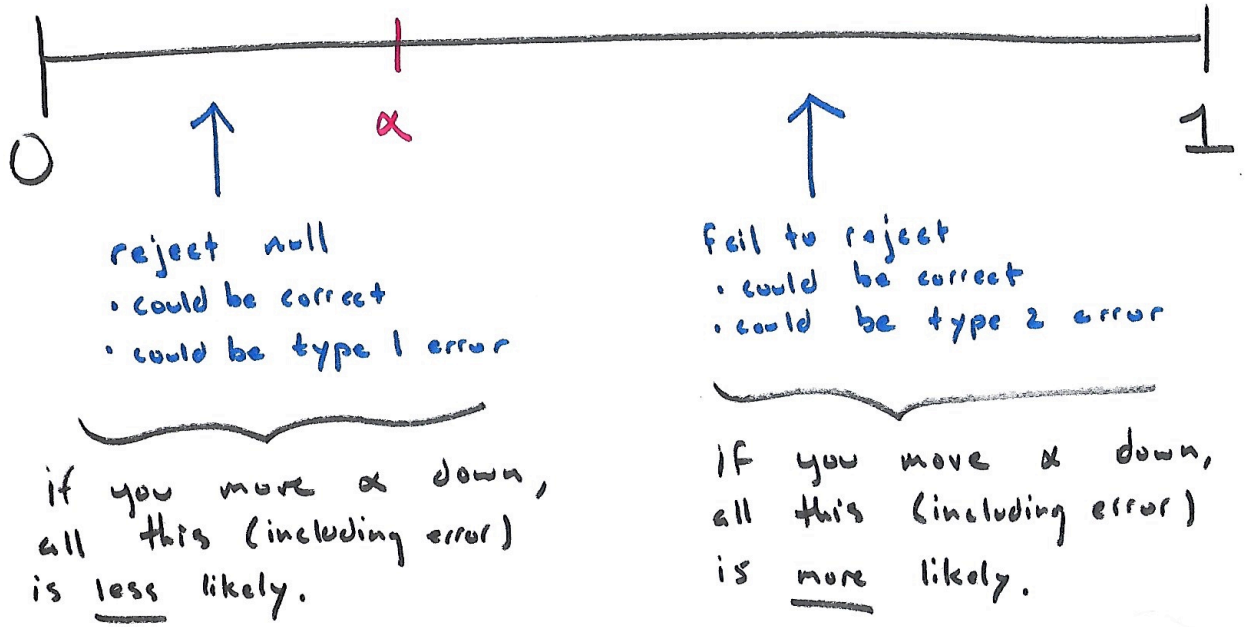
How does adjusting α change the error rates?





How does adjusting α change the error rates?





One pathetic slide about power

Power is the probability of rejecting the null hypothesis when it is false (i.e. of avoiding a Type II error):

$$\text{Power} = \text{Prob}(\text{reject } H_0 \,|\, H_0 \text{ is false}).$$

It is the chance that a study will detect a deviation from the null if one really exists. We want this to be as big as possible.

Power is a function of

- Sample size;
- Deviation from the null one hopes to detect;
- Variability in your data;
- The discernibility level you choose.

Big ol' question: subject to constraints like budget, how should I design my study, and how much data should I collect, to make power as big as possible? *Very important, but beyond our course...*

Cardinal Sins in Statistics, Part 2 of 91

Thou shalt not interpret the p-value as the probability that the null hypothesis is true. It is the probability of an extreme result *assuming the null is true*.

Cardinal Sins in Statistics, Part 3 of 107

Thou shalt not confuse statistical discernibility with substantive importance.

What we say in STA 101	What you will hear elsewhere
discernibility level	"significance" level
statistical discernibility	statistical "significance"

Traditionally, if p-value $< \alpha$, we reject H_0 and call the result "statistically significant". But this wording often misleads people into thinking the results are just plain *significant*, in a substantive sense. Wroooooong.

Example

The truth: a coin flip comes up heads with probability 0.499.

Hypotheses:

$$H_0 : \text{Prob(heads)} = 0.5$$
$$H_A : \text{Prob(heads)} \neq 0.5$$

Fact: H_0 is literally false. $0.5 \neq 0.499$.

Result: you flip the coin 10,000,000 times and get a *p*-value that's practically zero, and *correctly* reject H_0. So what?

Punchline: the machinery of statistics *cannot* tell you if your results are "meaningful" and "important". It can only tell you if the results are likely or not under random sampling.

statistical "significance" $\neq$ importance

Venial Sins in Statistics, Part 1 of 284

Thou shalt not *accept* the null hypothesis, even if the p-value is huge. You only "fail to reject" the null hypothesis.

Example: when a verdict is read out in court, it isn't "guilty" or "innocent." It's "guilty" or "not guilty," which is very different.

Notice a pattern?

Statistical questions...

- **Q**: Does the linear model fit well?

  A: Look at the spread of the residual distribution.

- **Q**: Is the unknown parameter reliably estimated?

  A: Look at the spread of the sampling distribution.

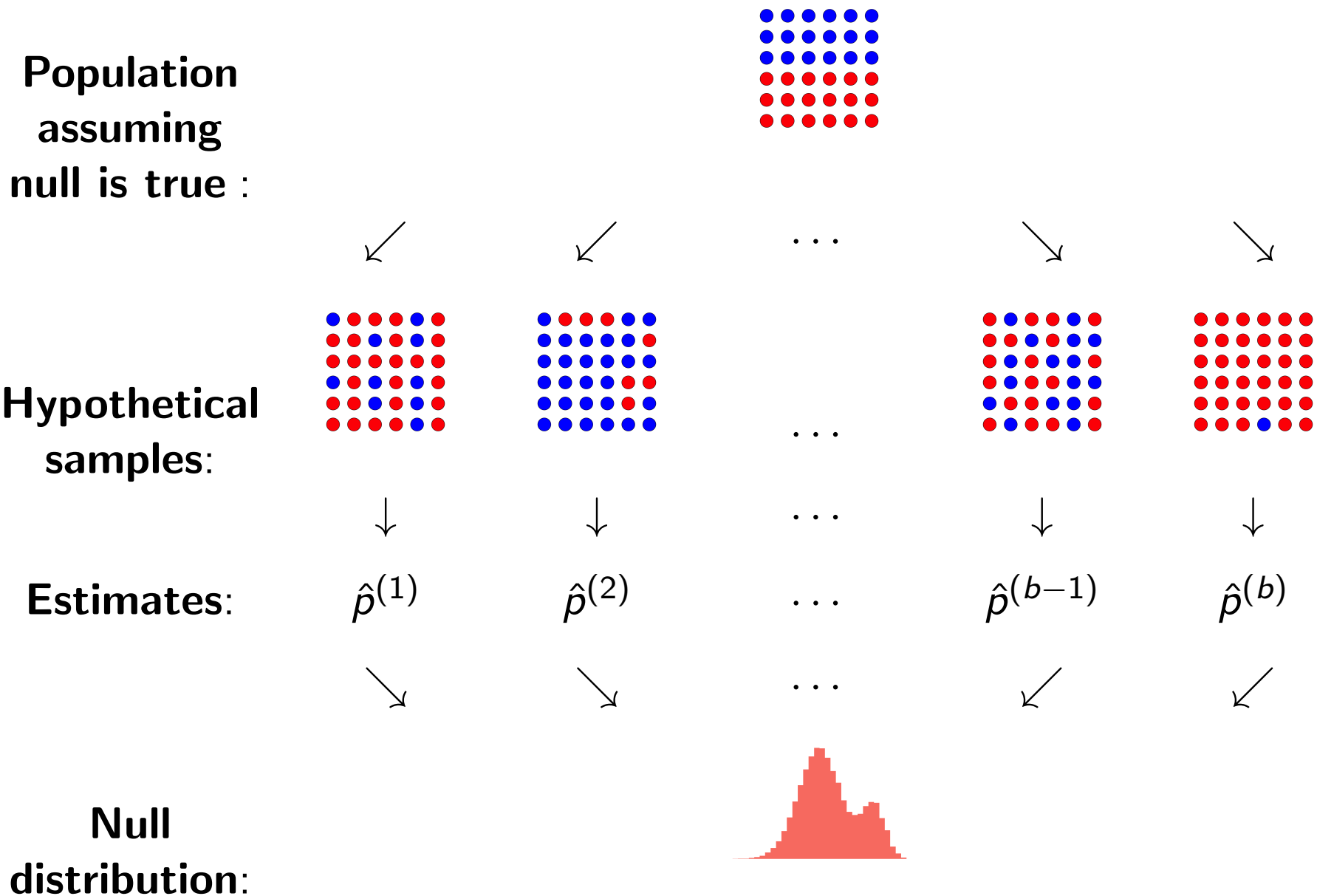
- **Q**: Do we have sufficient evidence to reject the null?

  A: Look at the spread of the null distribution.

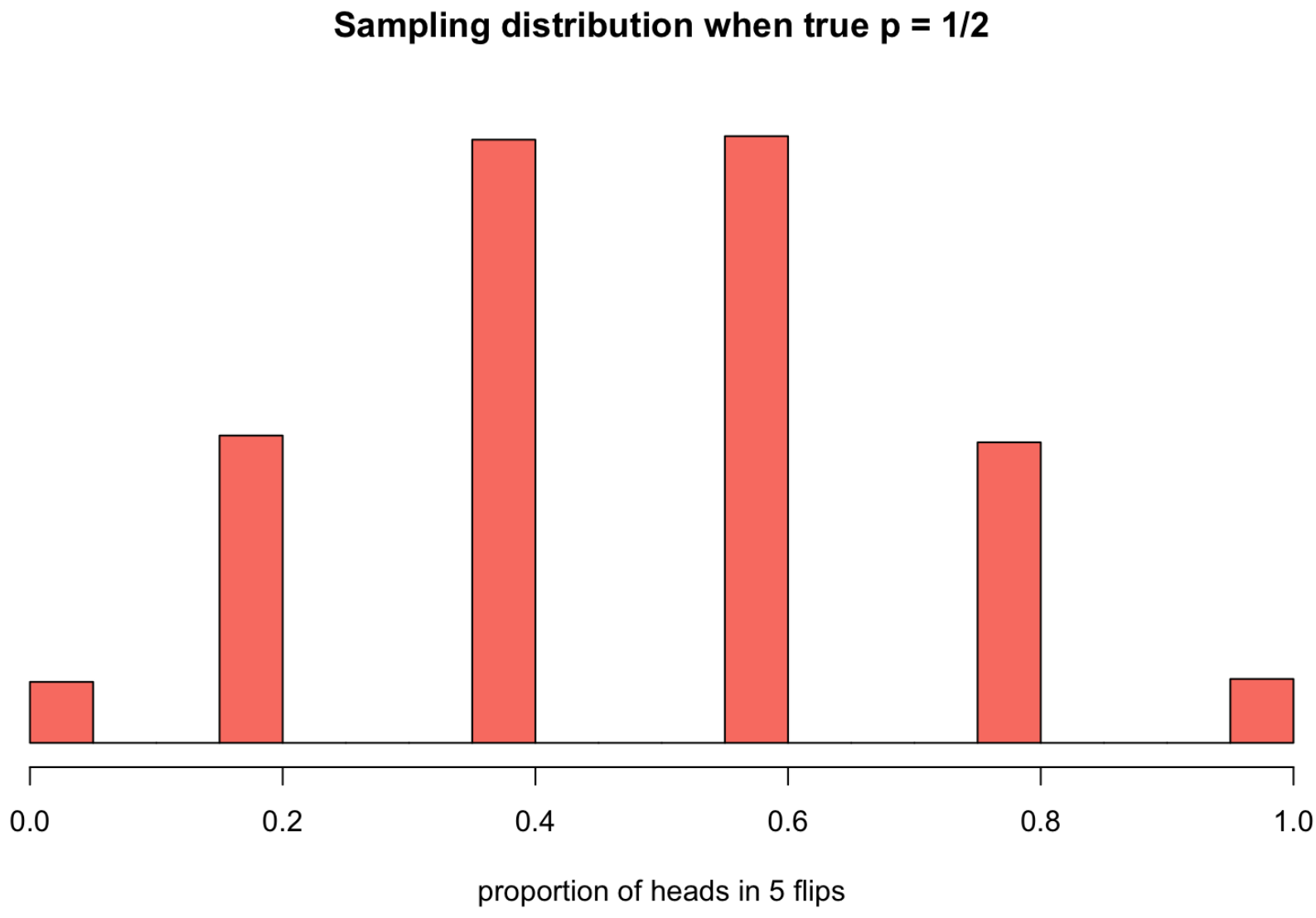
We typically represent a distribution with a histogram, and we measure spread with variance or standard deviation.

Make sure you understand these things! It's Chapter 5.

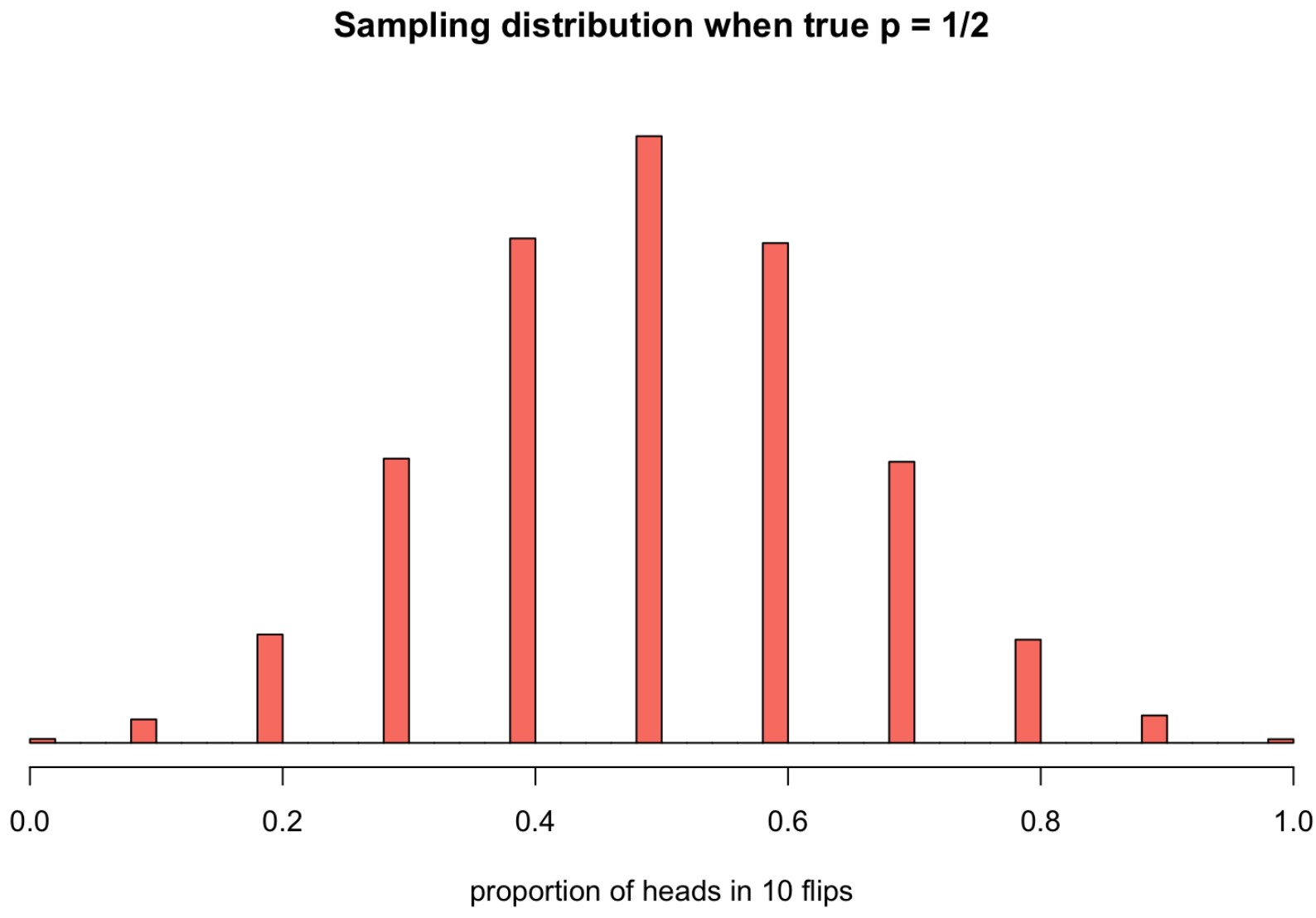
Into the weeds: how is the null distribution simulated?



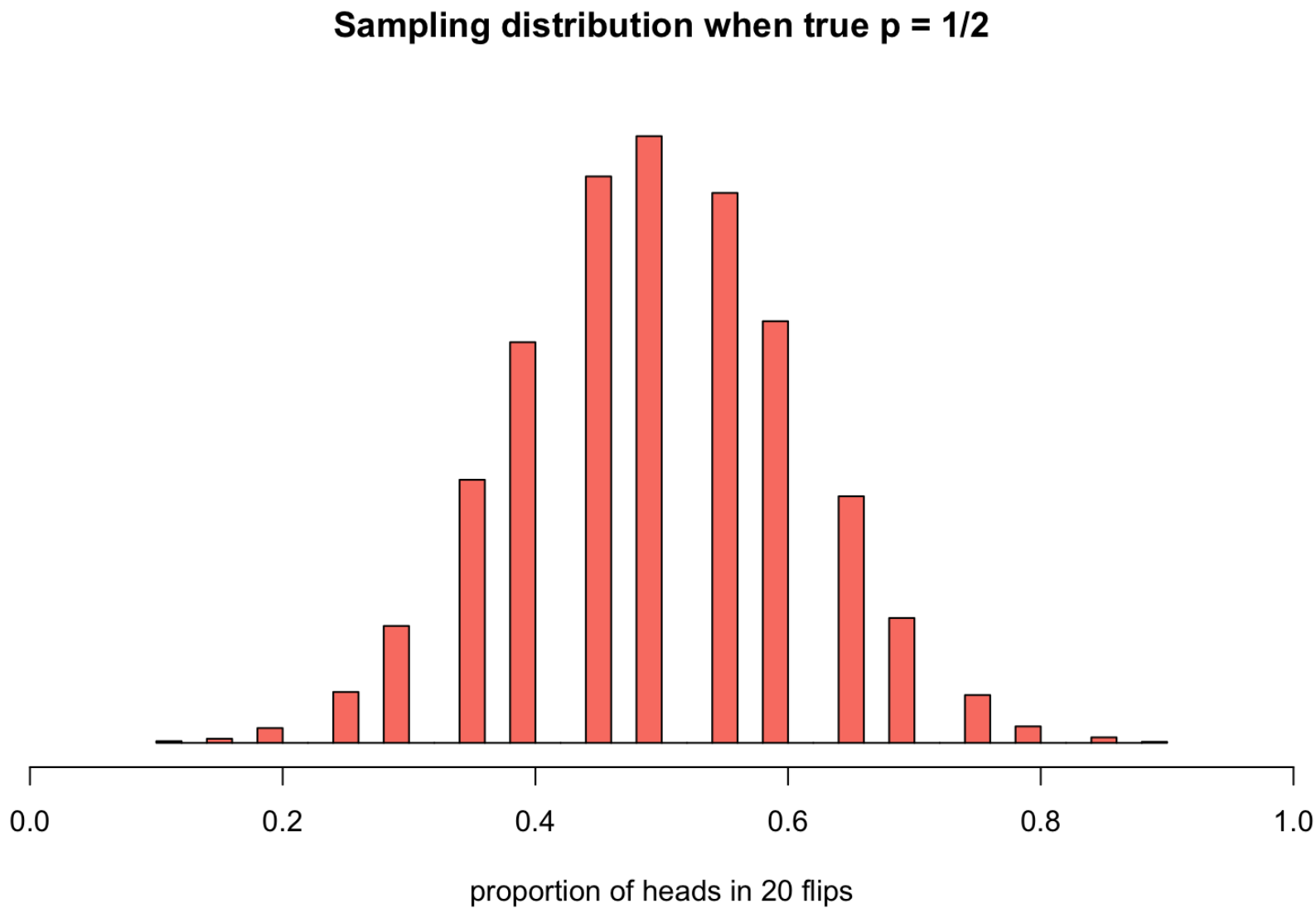
Sampling distribution of proportion from coin flip data



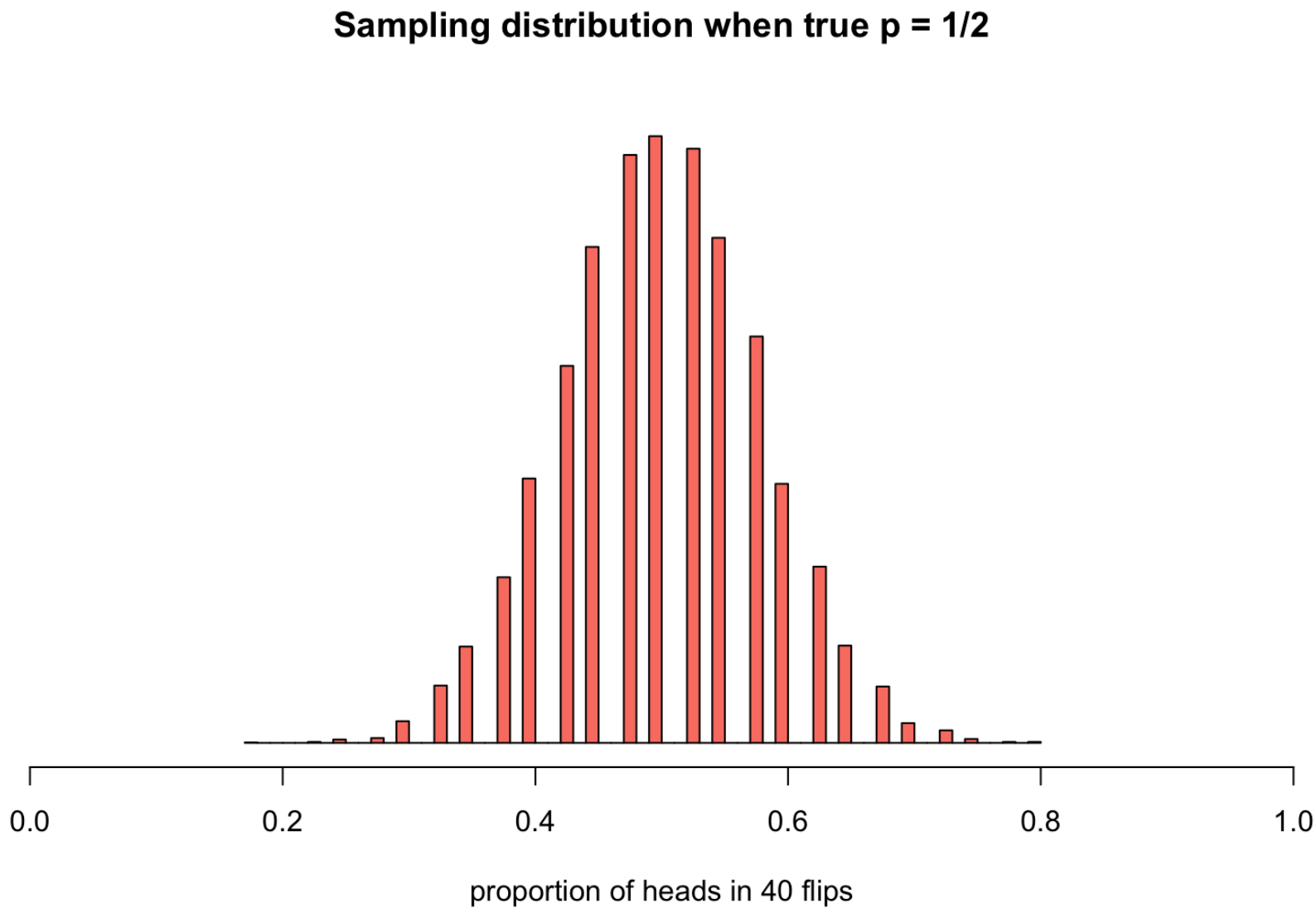
Sampling distribution of proportion from coin flip data



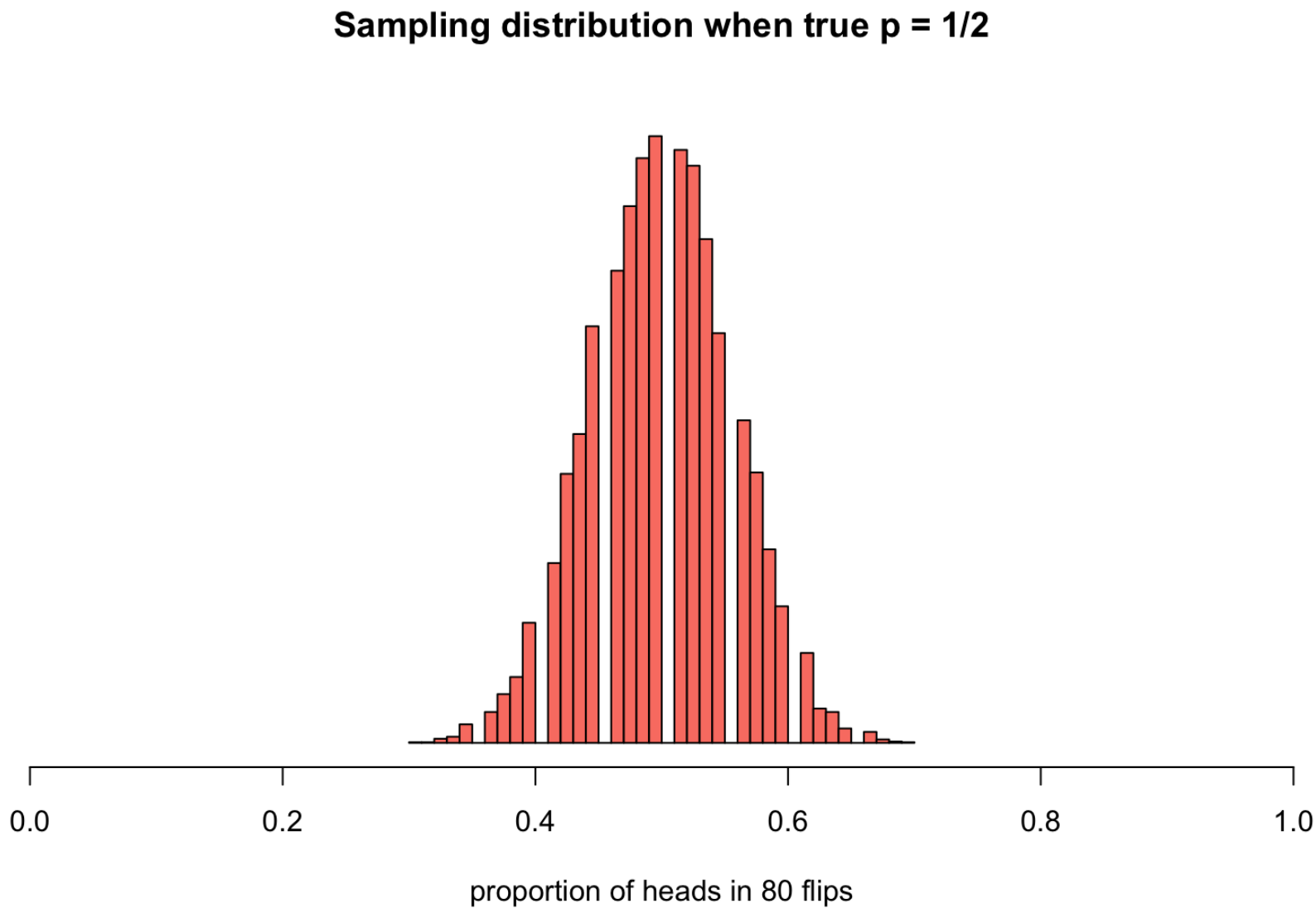
Sampling distribution of proportion from coin flip data



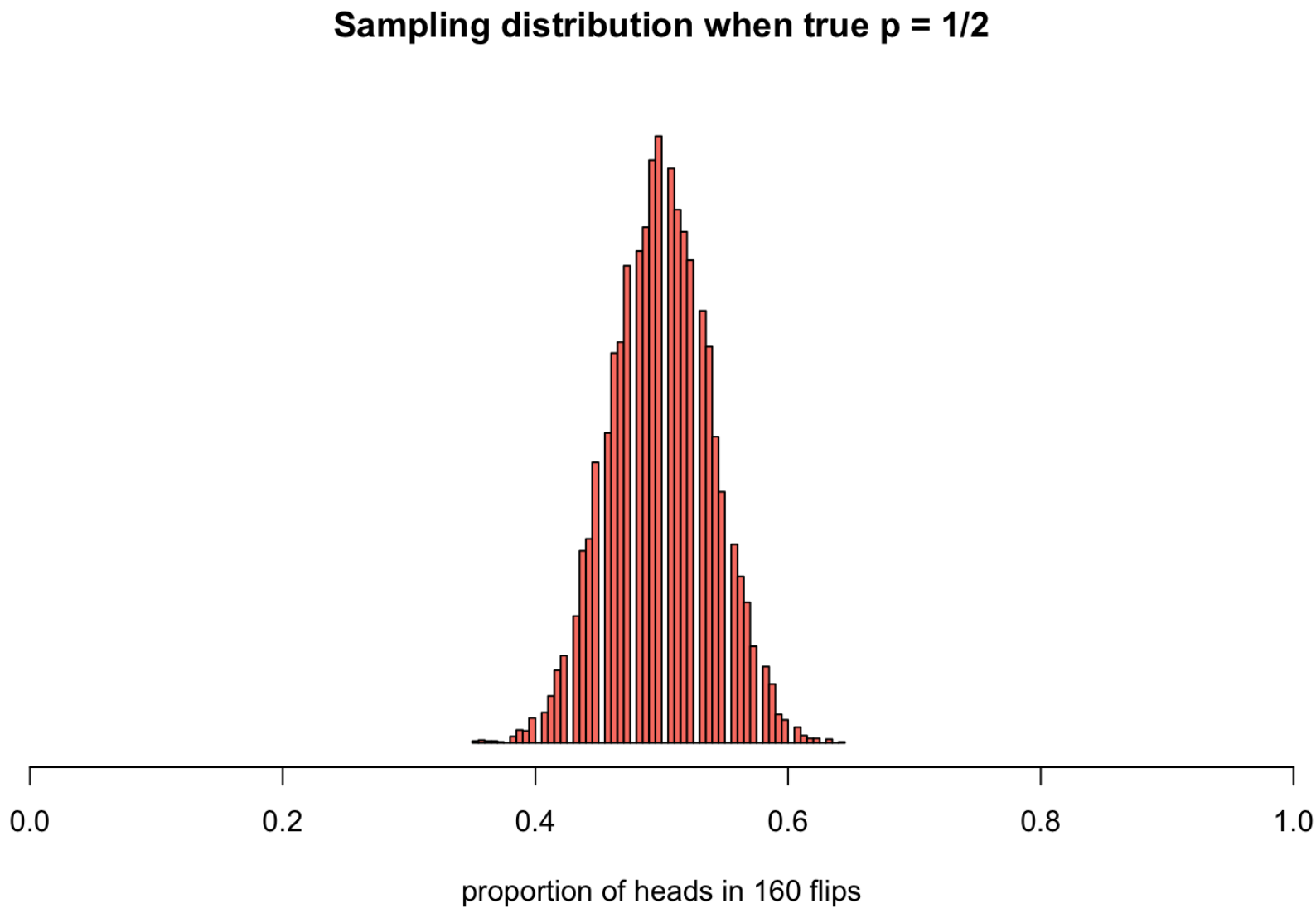
Sampling distribution of proportion from coin flip data



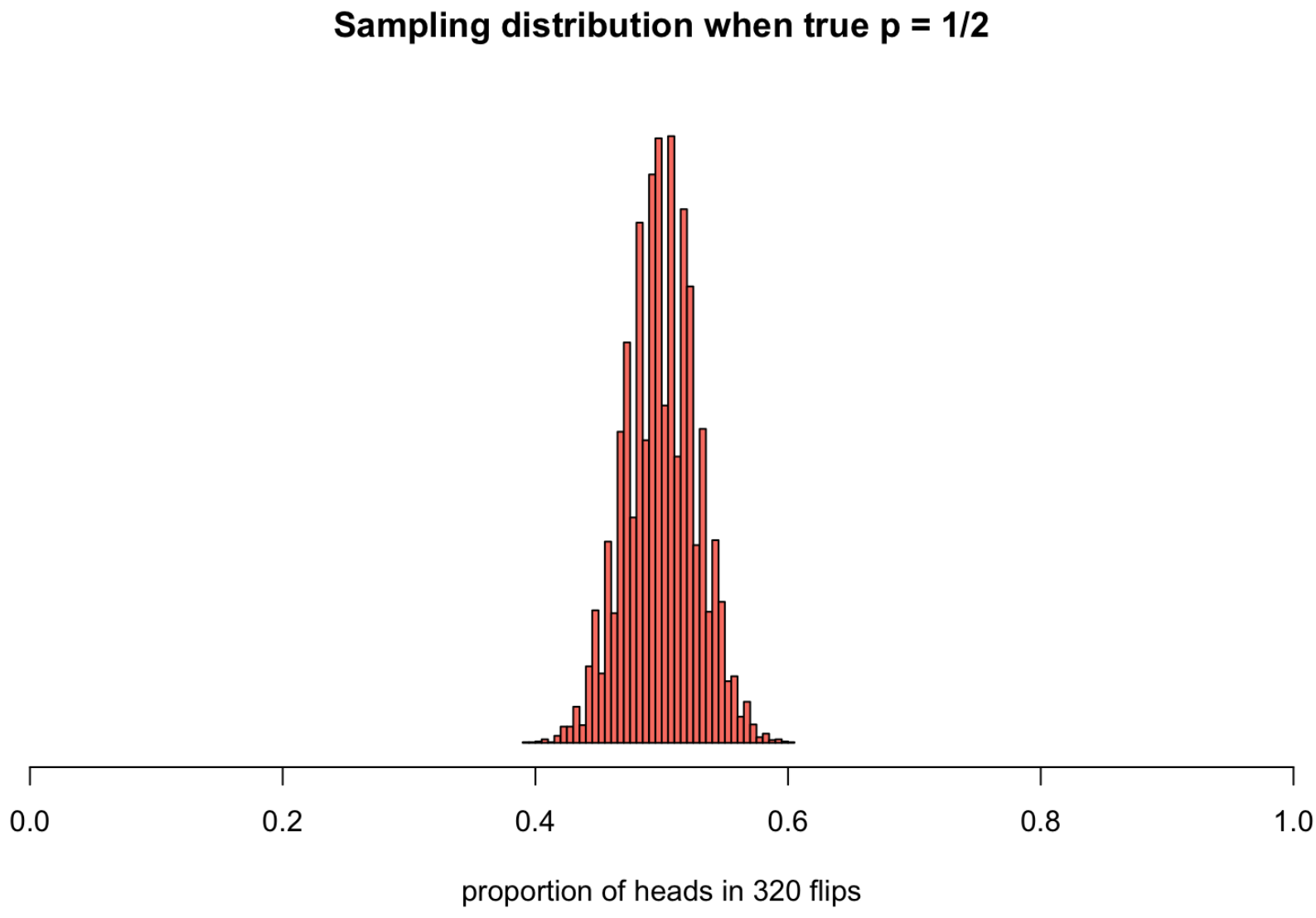
Sampling distribution of proportion from coin flip data



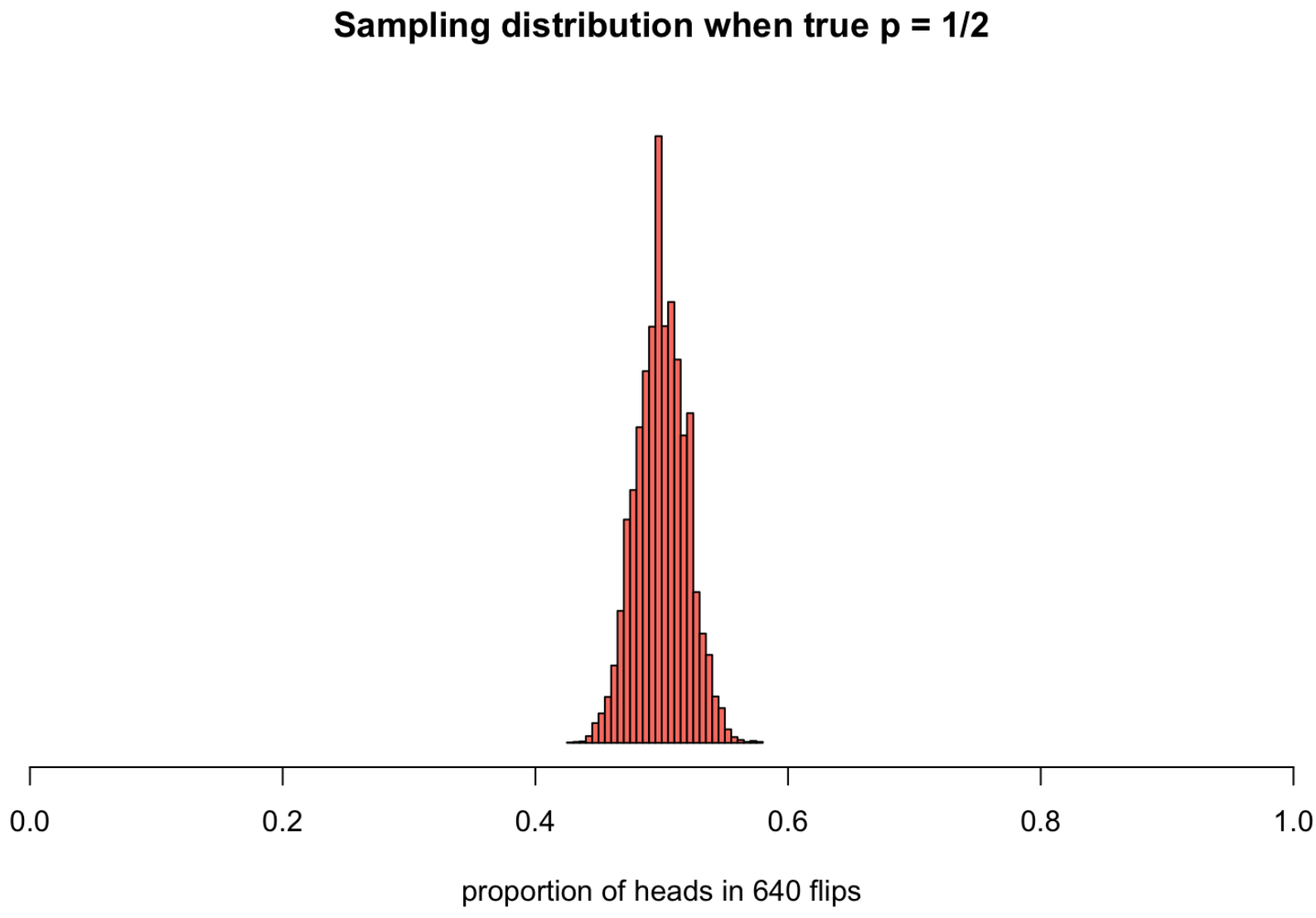
Sampling distribution of proportion from coin flip data



Sampling distribution of proportion from coin flip data



Sampling distribution of proportion from coin flip data

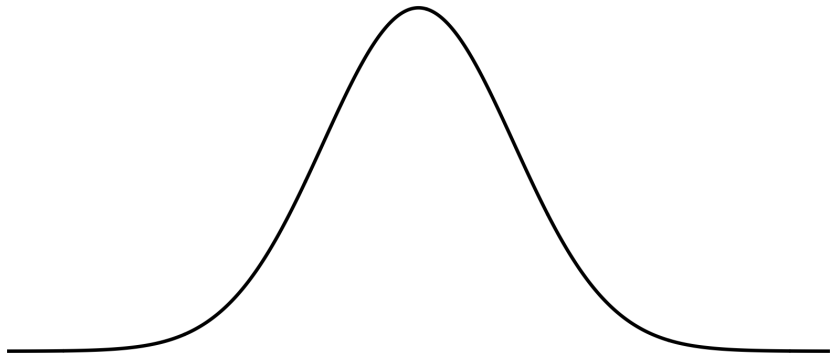


Three observations

- the sampling distribution is centered on the true value;
- as the sample size (number of flips) grows, the spread decreases;
- as the sample size grows, looks more and more like a bell curve.

This is not an accident.

The central limit theorem (CLT)



Theorem: As your sample size gets bigger and bigger, the sampling distribution of a statistic (sample mean, proportion, difference, etc) will look more and more like the **bell curve**.

Proof.

Take STA 240! □

Who cares?

In (classical) statistics, the sampling distribution is queen:

- (**Interval estimation**) we need to approximate the sampling distribution in order to compute a confidence interval;
- (**Hypothesis testing**) we need to approximate the null distribution in order to compute a p-value.

How do we approximate a sampling distribution?

- **Simulation**: bootstrap, permutation, `generate(...)`, etc.
- **Normal approximation**: assume the sampling distribution is roughly a bell curve, and use what we know about that.

The CLT gives theoretical justification for the second strategy.

The normal distribution (bell curve)

Notation: $X \sim \mathrm{N}(\mu, \sigma)$.

Two parameters:

- μ: "mu." The mean. Controls location of the middle;
- σ: "sigma." The standard deviation. Controls spread.

The 68-95-99.7 rule: