Statistical theory: Making statistical decisions

Research Methods for Human Inquiry

Andrew Perfors

How do we control Type I errors?
(How do we enforce our significance level?)

How to build your own statistical test!

Basic idea: we need to figure out what kind of data we would expect to see if the null hypothesis were true

Then we compare it to the data we have

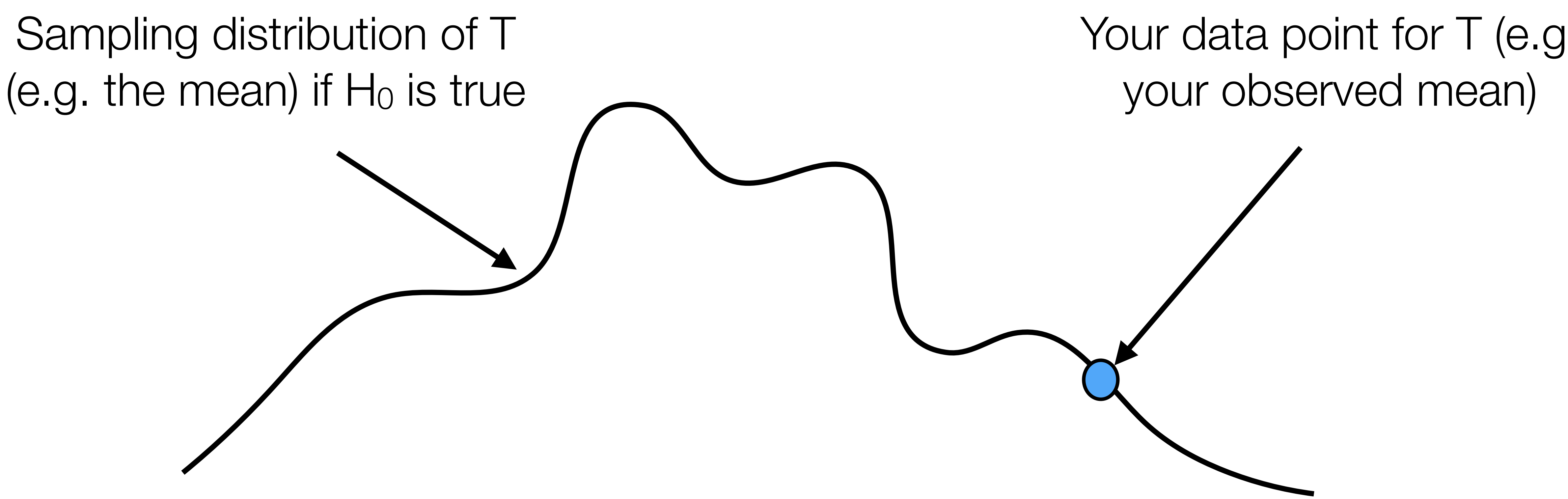
If we would expect to see our data **n**% of the time (or less) if the null were true, we reject the null

The choice for n is arbitrary and is known as the significance level.

How to build your own statistical test!

We need a few things…

1) A diagnostic test statistic, *T* (e.g., mean)
2) Sampling distribution of *T* if the null is true
3) The observed *T* in your data
4) A rule that maps every value of *T* onto a decision (accept or reject H0)



Does this data mean you should
accept H_0?

1) A ~~diagnostic~~ test statistic

- A test statistic *T* is…
  - A single number that you can calculate only from your observations
  - As long as it's just one number, it's a test statistic.

- These are all possible test statistics…
  - The mean of a set of observations
  - The standard deviation of a set of observations
  - The third-largest of a set of observations
  - The number of observations

The mean is very common because it usefully captures many datasets

- These are all not allowed:
  - The two largest numbers in the sample (two numbers!)
  - Your favourite number (not calculated from your sample!)

1) A diagnostic ~~test statistic~~

A test statistic is diagnostic if the null hypothesis and alternative predict different values

example: people agree that we’re running out of food (or not)

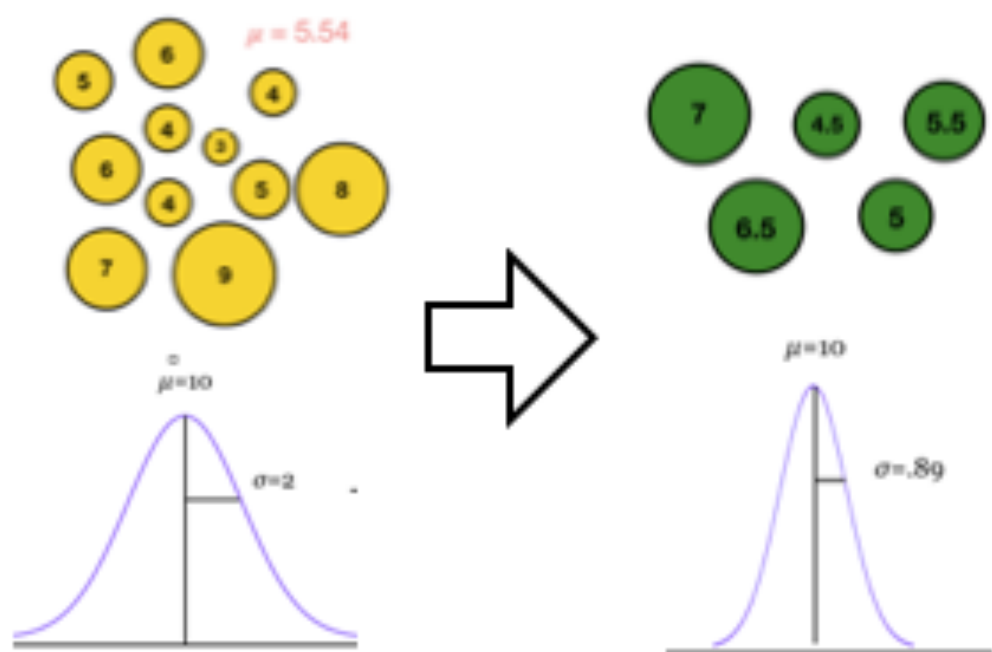
Test statistic: number of YES choices out of 100

H_0: Should be around 50
***H_1*:** Should *not* be around 50

yup, it’s diagnostic!

2) Sampling distribution if the null is true

- Sampling distribution of what, exactly???
- We've only talked about the sampling distribution for the <u>mean</u>



They are what you'd get if you took lots of different samples from the population, and calculated the **mean** of each of them

2) Sampling distribution if the null is true

- Sampling distribution of what, exactly???
- We've only talked about the sampling distribution for the <u>mean</u>
- But <u>any</u> statistic can have a sampling distribution
  - Sampling distribution for the standard deviation
  - Sampling distribution for the median
  - Sampling distribution for the 6th largest value in the sample

They are what you'd get if you took lots of different samples from the population, and calculated the **[whatever]** of each of them

2) Sampling distribution if the null is true (food example)

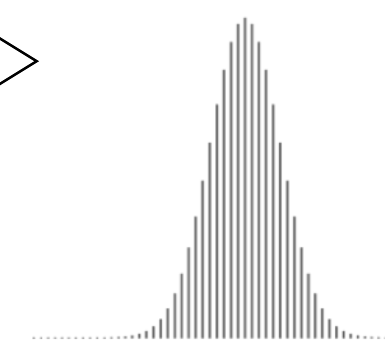
- To calculate this: assume H_0 is true
- Figure out what values of your test statistic you should expect

example: running out of food?

Test statistic: number of YES choices out of 100

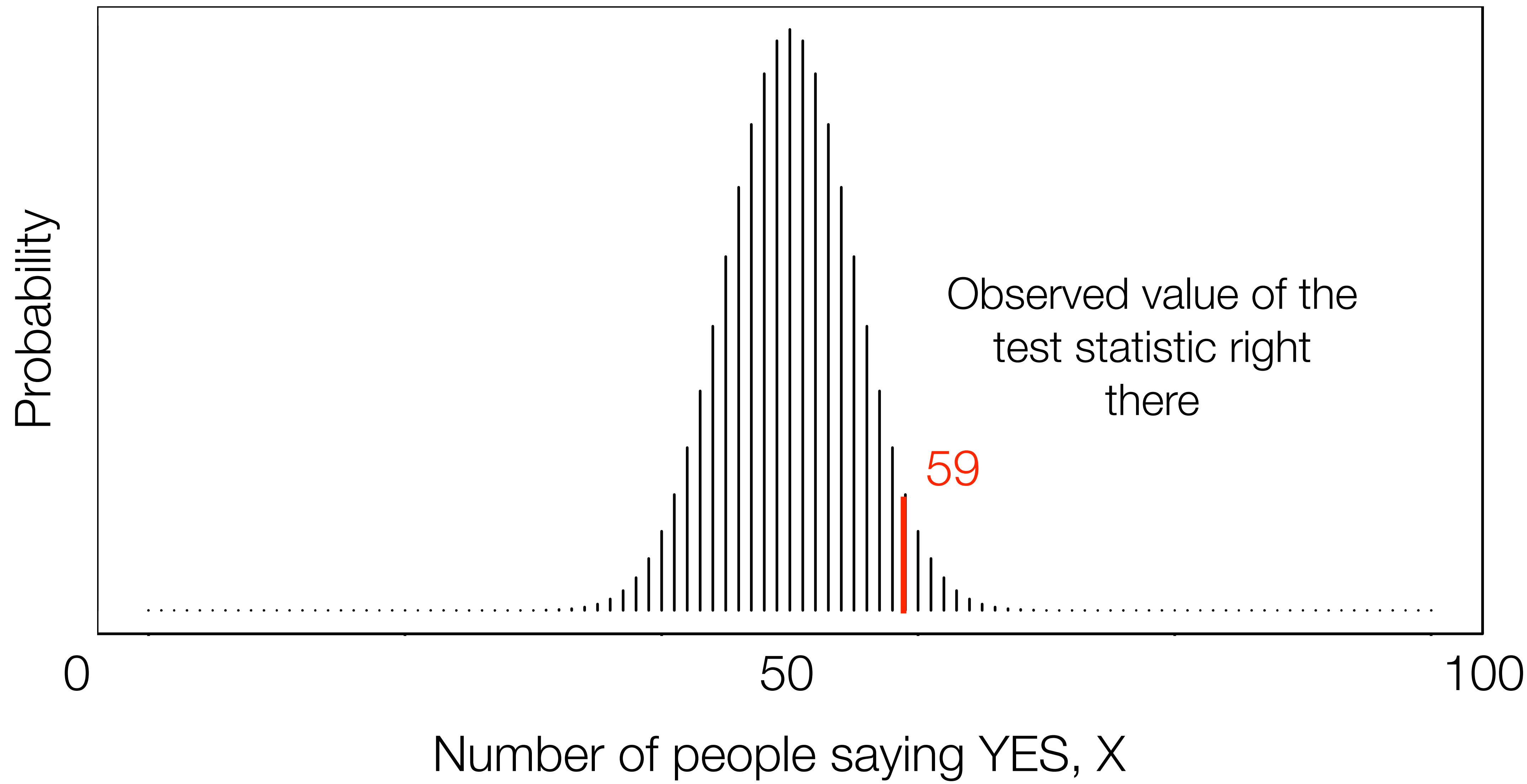
… count data of one of two possible events happening

binomial distribution!



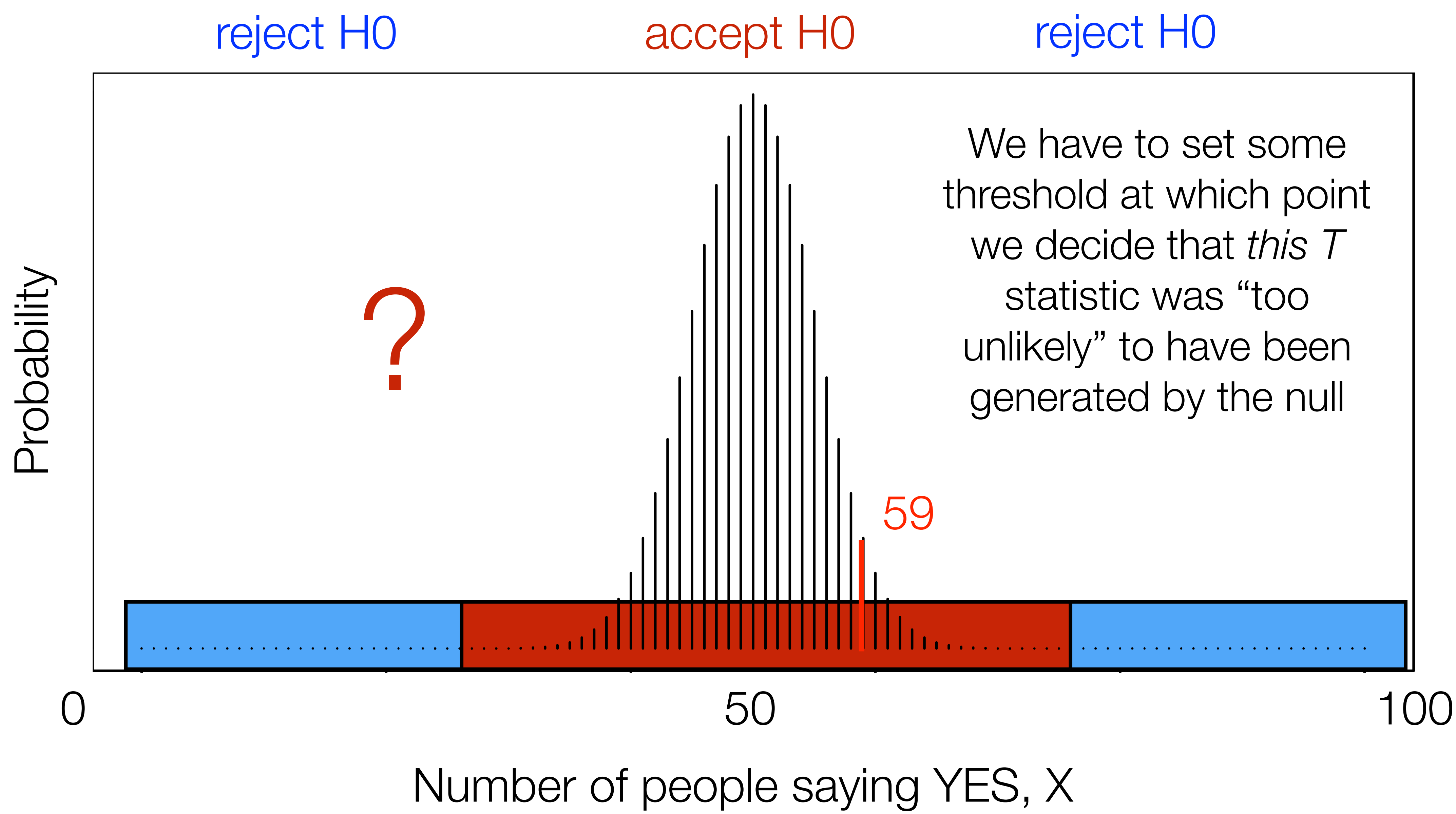
H₀: Out of 100 observations, should expect 50 to be YES (i.e., prob=0.5)

3) The observed *T* in your data (food example)



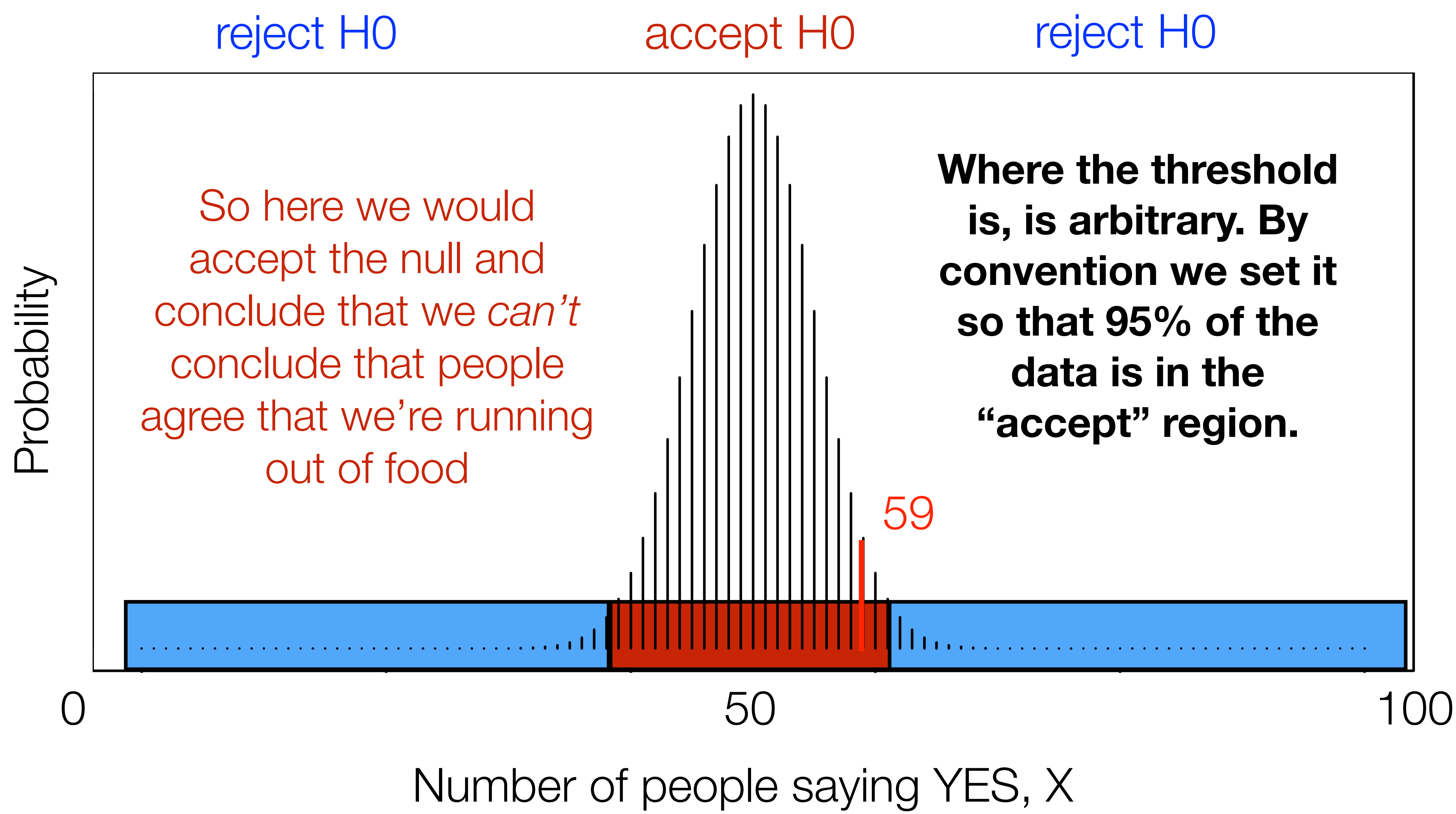
H_0: Out of 100 observations, should expect 50 to be YES (i.e., prob=0.5)

4) A rule that maps T onto a decision about H_0 (food example)



H_0: Out of 100 observations, should expect 50 to be YES (i.e., prob=0.5)

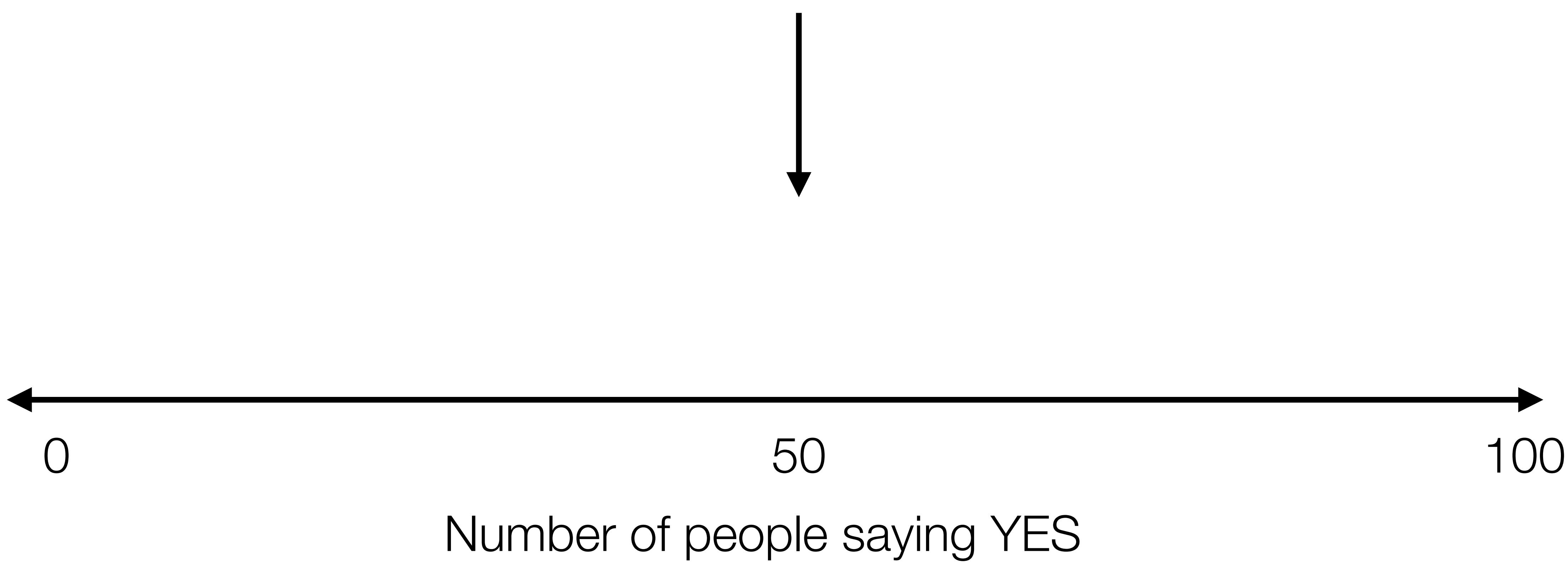
4) A rule that maps T onto a decision about H_0 (food example)

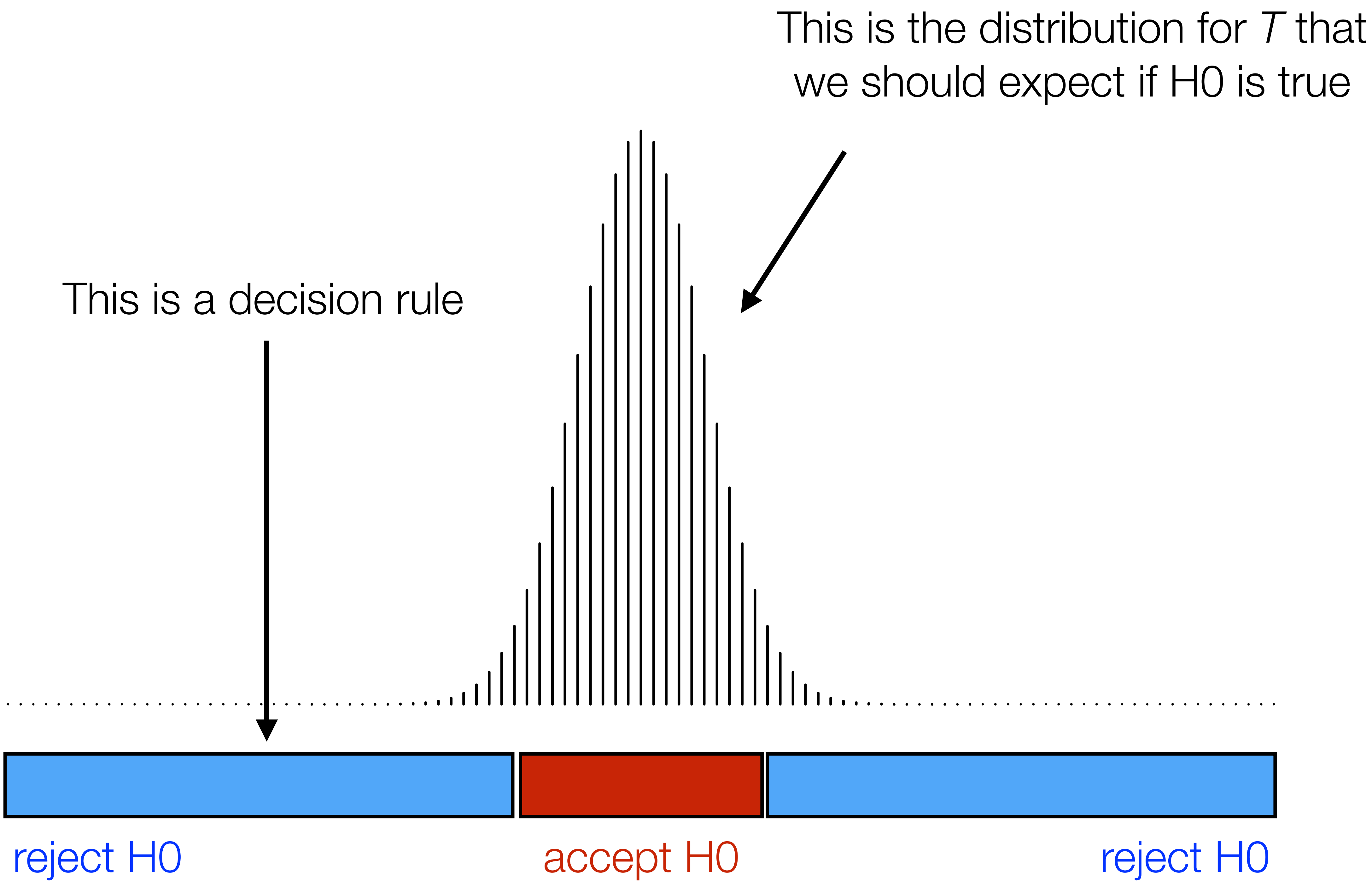


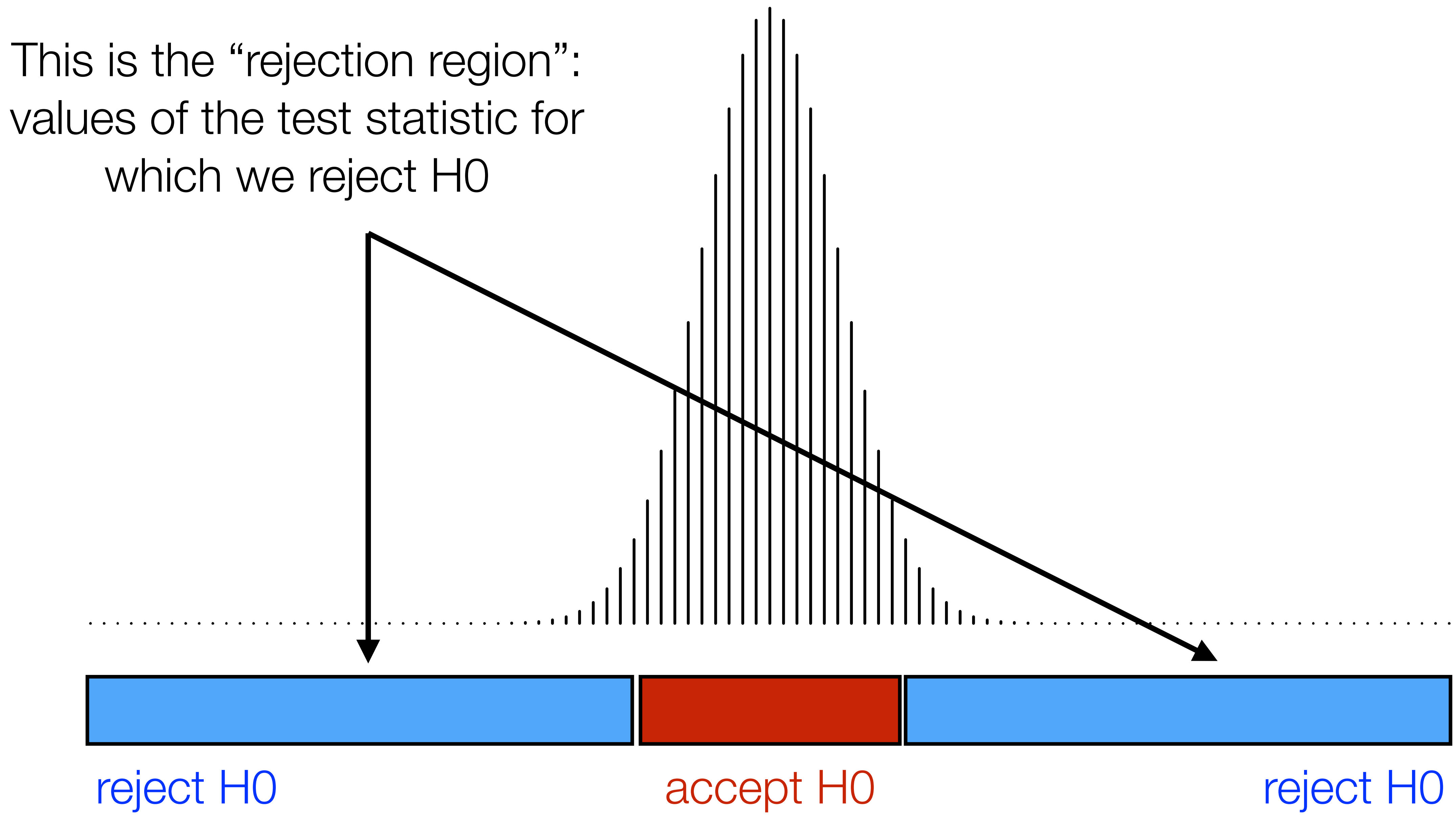
H_0: Out of 100 observations, should expect 50 to be YES (i.e., prob=0.5)

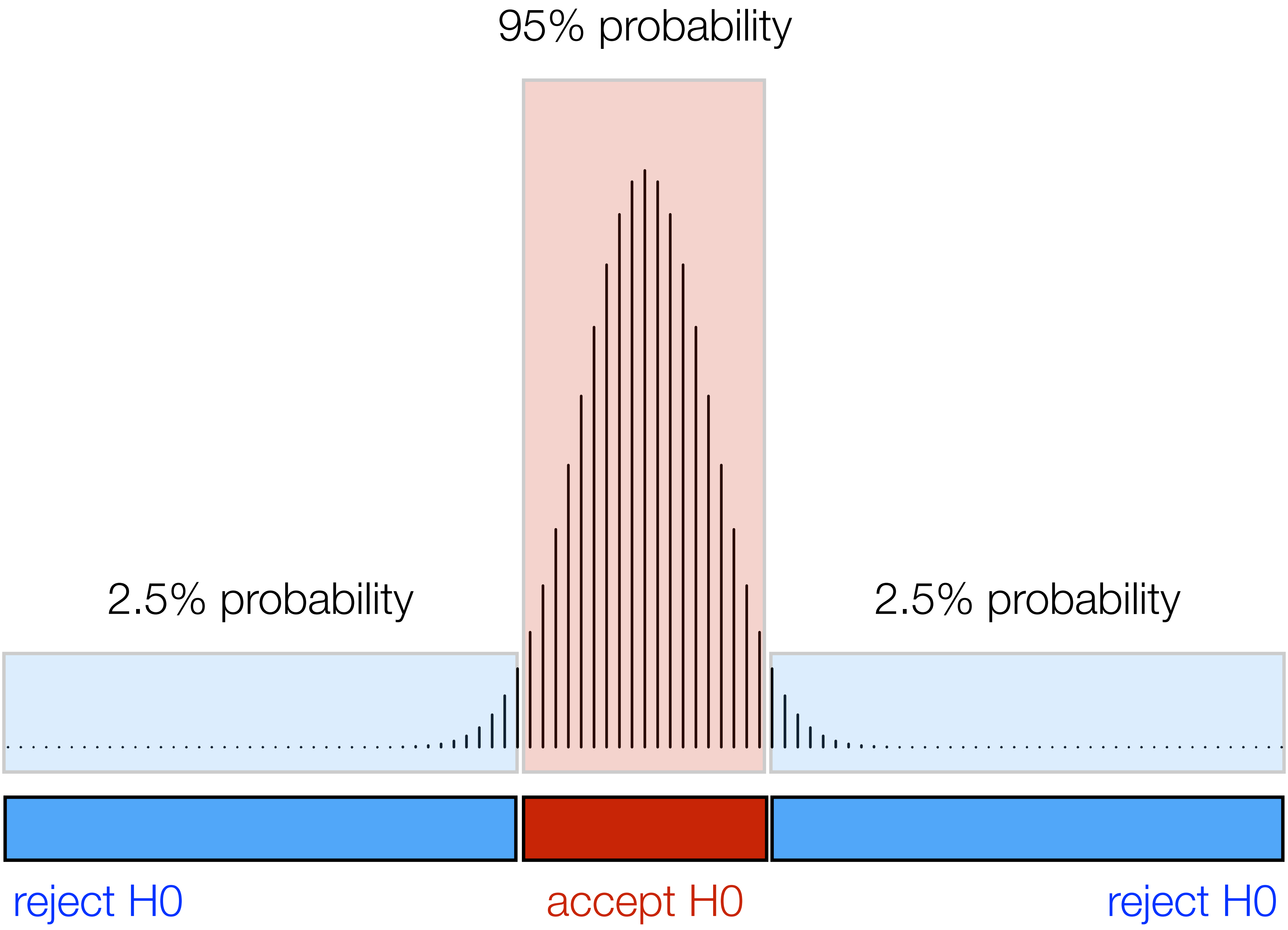
Let’s reprise

This is our diagnostic test statistic *T* and the range of values it can take on

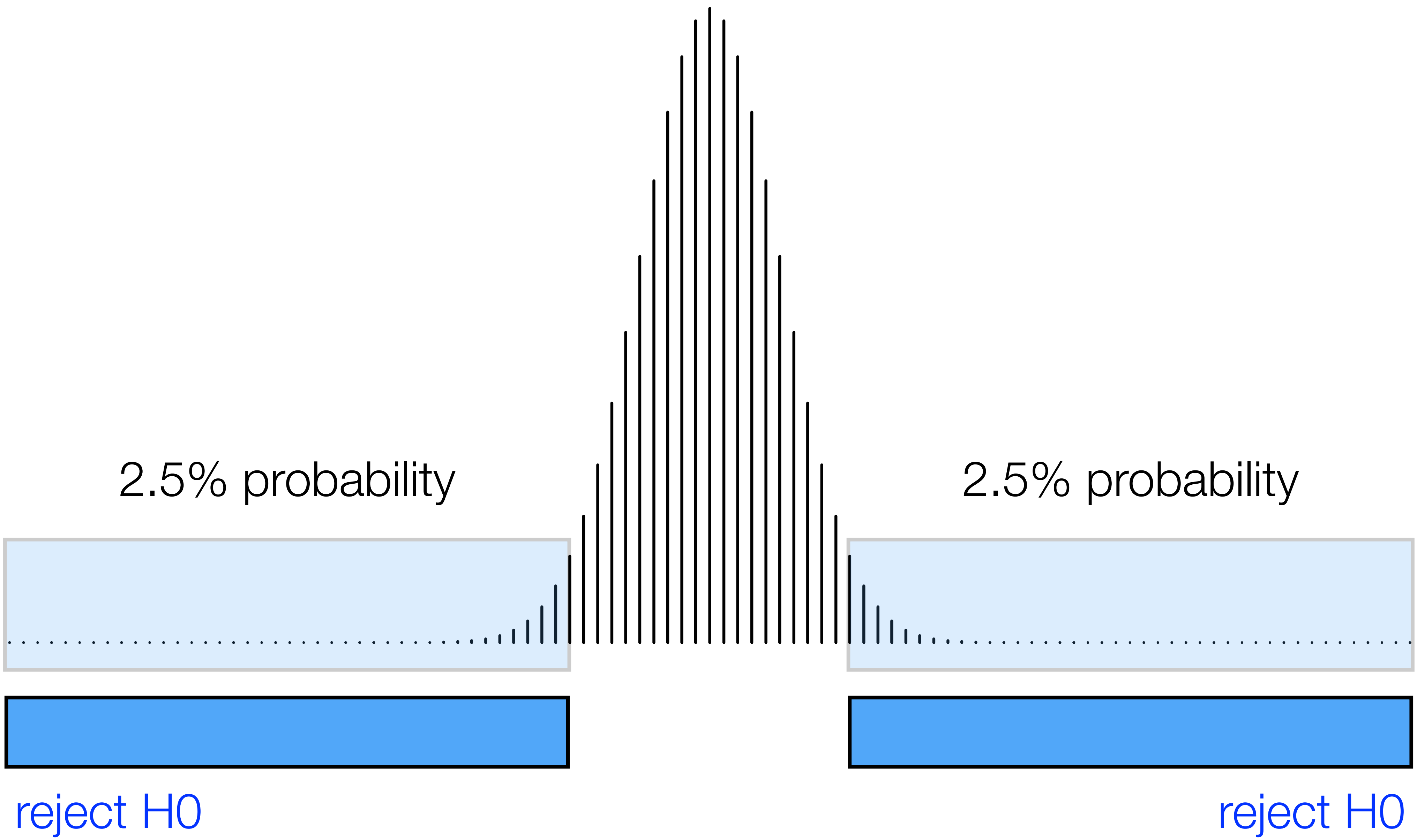



This is the distribution for T that we should expect if H0 is true
This is a decision rule
reject H0
accept H0
reject H0

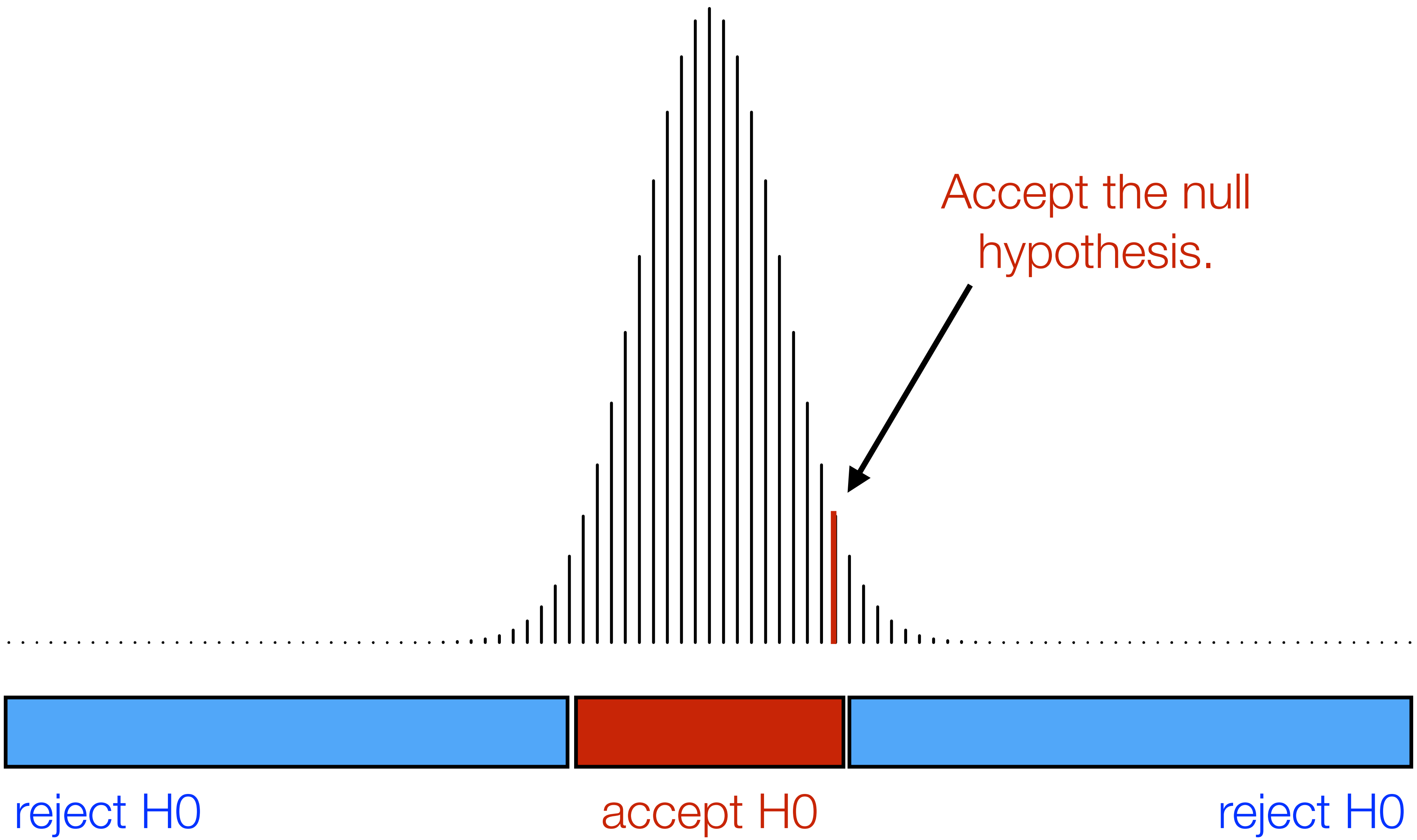

This is the "rejection region":
values of the test statistic for
which we reject H0
reject H0
accept H0
reject H0


95% probability
2.5% probability
2.5% probability
reject H0
accept H0
reject H0

If the null hypothesis is true, there is a 5% chance of falsely rejecting it. We have controlled our Type I error rate at 5%



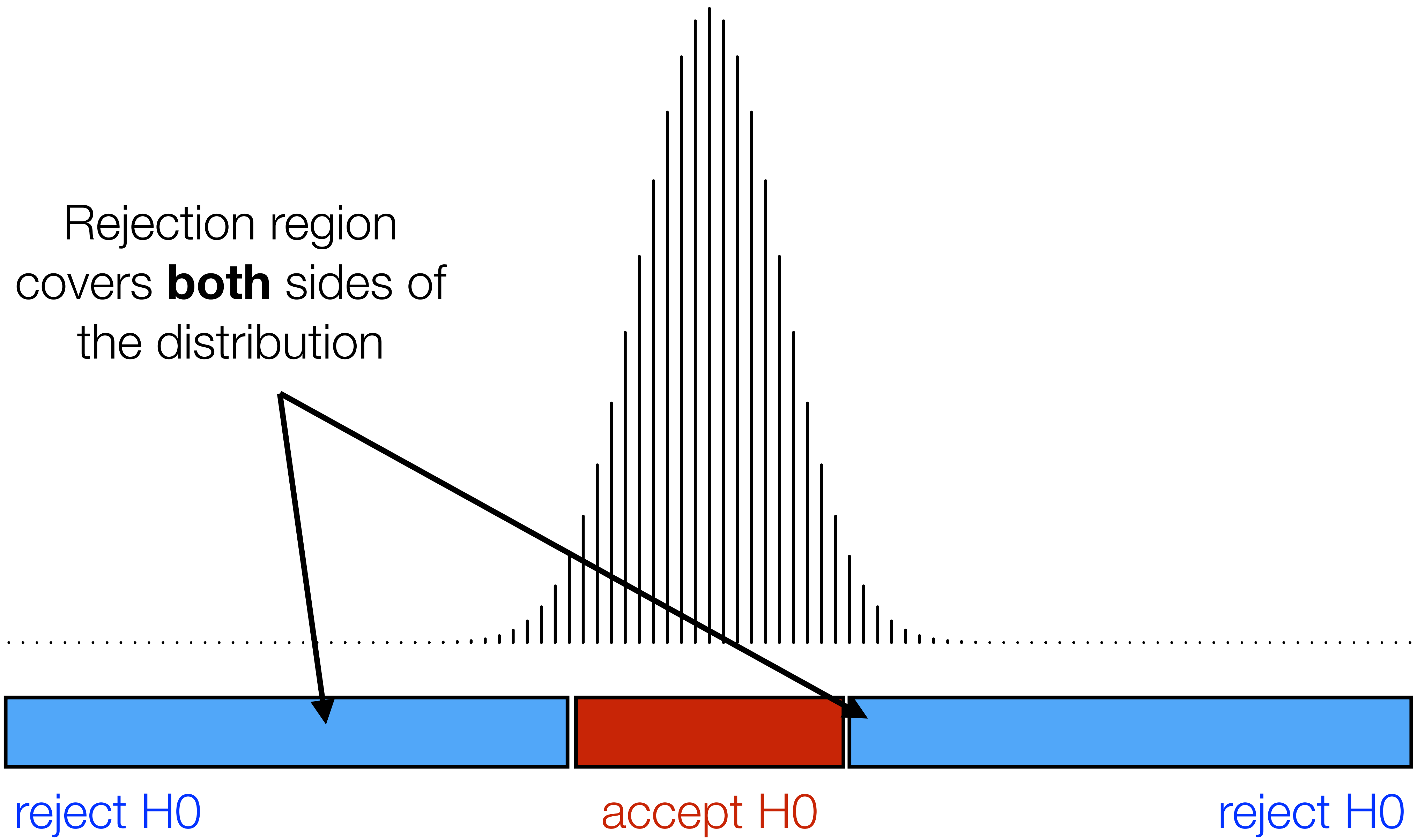
The last step is to see if your observed test statistic falls in the rejection region…



Note that this test has two sides. It is a two-sided test.

Null hypothesis: P("yes") = 0.5

Alternative hypothesis: P("yes") < 0.5 **or** P("yes") > 0.5



But what if you have a strong belief in the *direction* of your alternative hypothesis?

two-sided (directionless): “people agree about the food”

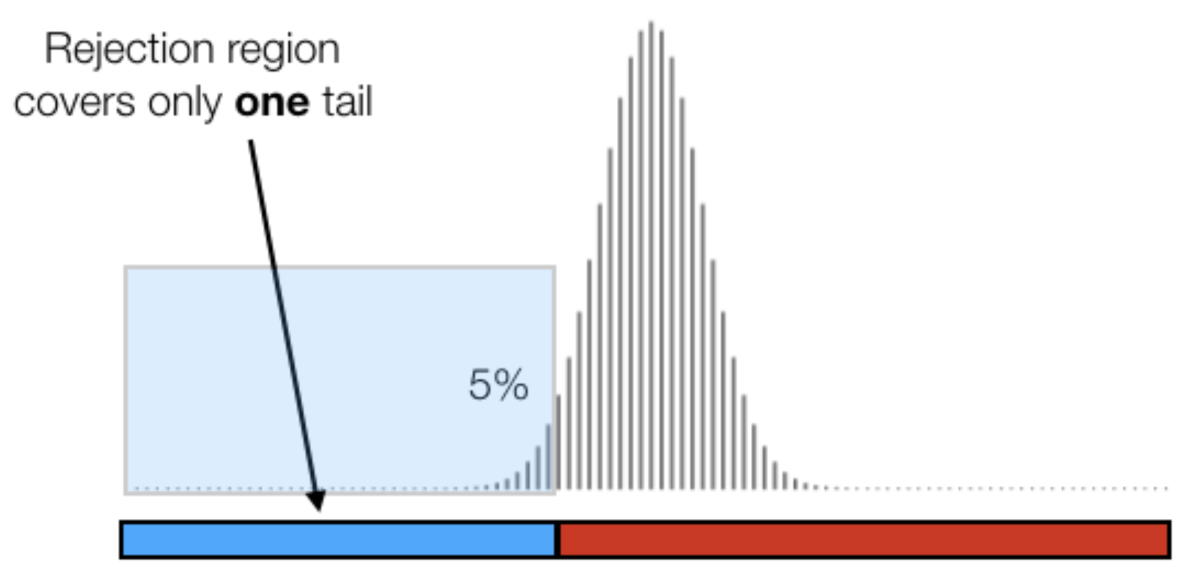
Null hypothesis: P(“yes”) = 0.5

Alternative hypothesis: P(“yes”) < 0.5 **or** P(“yes”) > 0.5

one-sided (directional): “people think we’re running out of food”

Null hypothesis: P(“yes”) <= 0.5

Alternative hypothesis: P(“yes”) > 0.5



Why 5%

- Why did we use 5% for our desired Type I error rate?
  - i.e., why do we have to have $\alpha = .05$?

- By convention
  - $\alpha = .05$ is the default significance level that we use in science
  - But people also use $\alpha = .01$ and $\alpha = .001$

What is a p-value?

p describes the Type I error rate you must be willing to tolerate if you want to reject H0

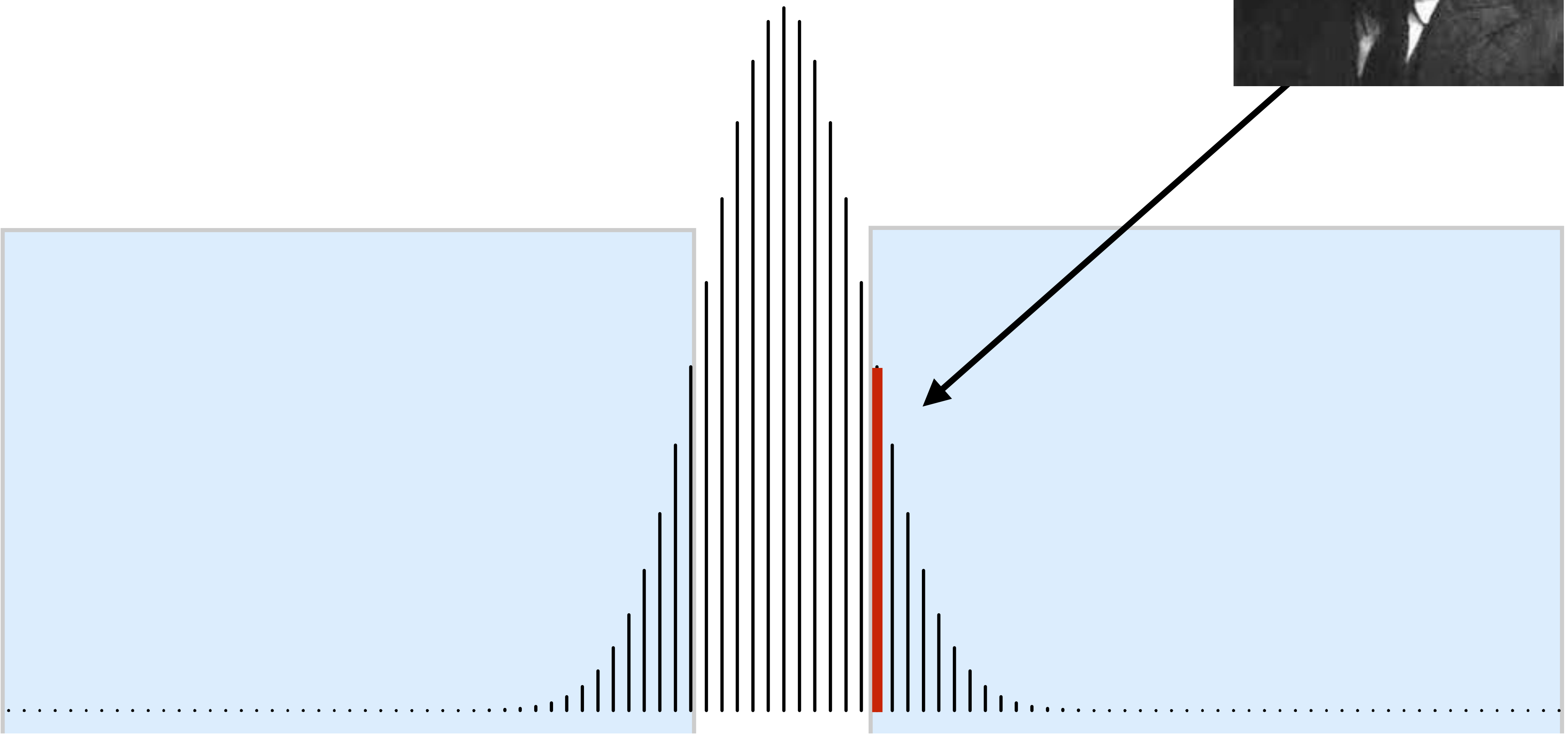
Neyman

What is a p-value?

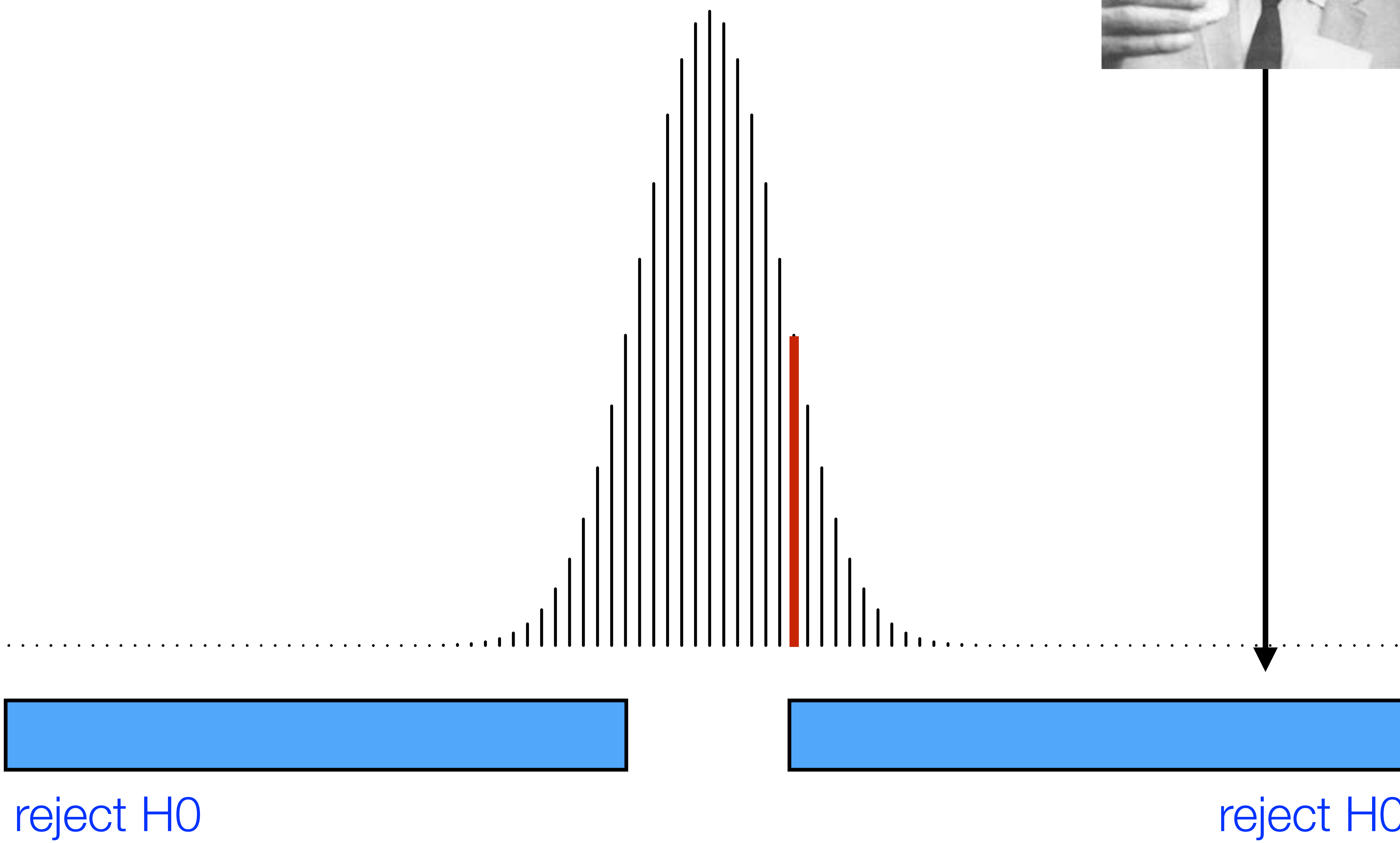
Fisher

p is the probability - if H0 is true - of observing a test statistic at least as extreme as the one that was actually found

The null hypothesis says there's a 27% chance of getting data more extreme than what we actually observed, so $p = 0.27$



In order to force our data into the rejection region, we had to be willing to tolerate a 27% Type I error rate, so p = 0.27



What to do in real life?

- When running a hypothesis test:
  - Adopt a "standard" significance level of $\alpha = 0.05$
  - If $p < \alpha$, reject the null hypothesis
  - Otherwise accept (or fail to reject) the null hypothesis

There is a <u>very</u> common trap!
Never, ever say the following…

so badness

wrong

evil

evil

“*p* is the probability that the null hypothesis is true”

evil

much wrong

evil

more wrong

still very wrong

wow

It <u>really</u> isn't

The p-value is a claim about how likely you were to see your data if the null hypothesis were true. This is not the same thing as a claim about whether the null hypothesis *is* true.

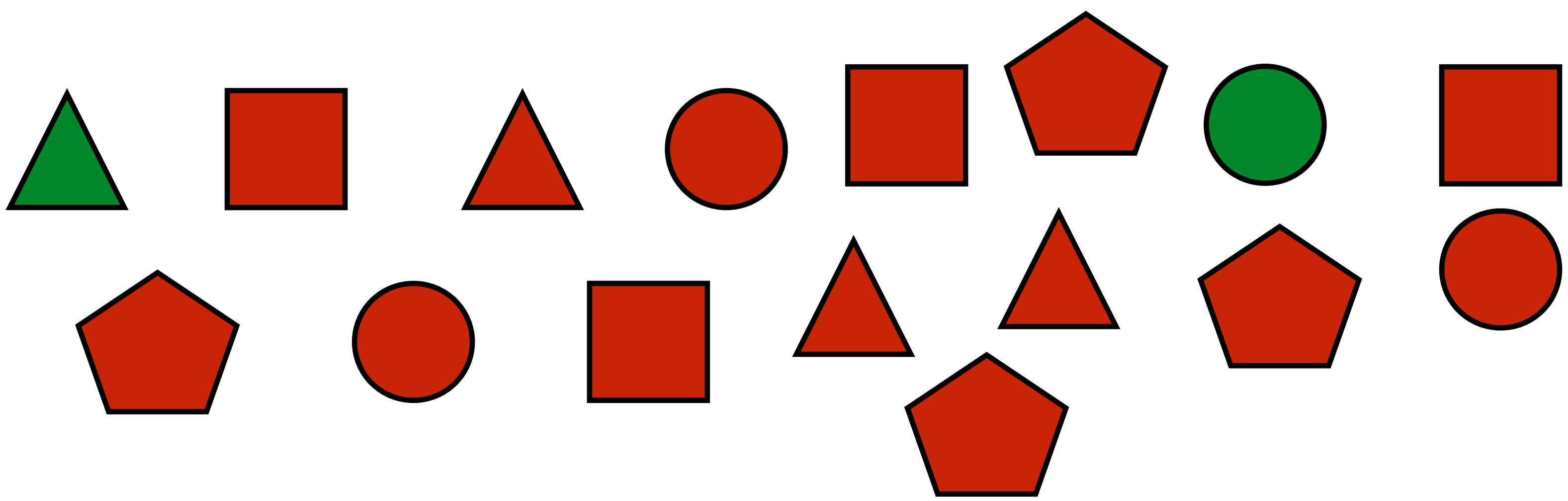
A claim about whether H0 is true depends on what other hypotheses you're considering. For that, you need to be able to evaluate them too (and we haven't done that!)

It <u>really</u> isn't

To see why: Suppose there are only two hypotheses in the world:

H0: I expect to see the same number of red and green things.
H1: I expect to see only blue things.

You observe:



This data is pretty unlikely if the null is true, but even *less* likely if H1 is true. So on balance even though we should reject the null this doesn't mean that H1 is therefore true.

Reporting your p-values

You often see something in a paper that looks like this:

(The details in the stats references differ according to the test)

This is the "stat reference". You can think of it like statistical citations: you're providing the *evidence* the reader needs in order to evaluate the claim written in the text

There was no significant difference in accuracy between conditions, $\chi^2(4)=5.19$, $p=.269$

This bit is about the test statistic for that test (don't worry about it for now)

This is the *p*-value. People differ on whether you should report exact *p*-values or just $p<.1$, $p<.05$, etc. I suggest you go with exact, to 3 or 4 decimal places.

More things you can’t say

More forbidden words

- Don’t say:
  - “The null hypothesis is true”
  - “The alternative hypothesis is false”
  - We have “proved” that BLAH

- Why not?
  - Because we haven’t…
  - These are all very definitive statements.
  - They imply we know the truth.
  - We don’t <u>know</u> the truth…

Some better phrasing

- If you retain H0:
  - “We retain the null hypothesis”
  - “We failed to reject the null hypothesis”
  - “The test was not significant”

- If you reject H0:
  - “We reject the null hypothesis”
  - “The test was significant”

People disagree about this one

- Some people don’t like these:
  - “Accept” the null
  - “Accept” the alternative

- I personally don’t mind much. But be careful.
  - “Accepting the null” does *not* imply evidence *for* H0
  - $p < .05$ means there is evidence against the null
  - $p > .05$ doesn’t necessarily mean there is evidence for it (because that requires a comparison to another hypothesis)
  - NHST isn’t built to find evidence *for* H0
  - (Bayesian methods can do that though)

See the `w5day2exercises.Rmd` file for the exercises!