ECON 3818

Chapter 18

Kyle Butts

27 September 2021

Chapter 18: Inference in Practice

Making Inferences

So far we have discussed two ways to make inferences about the parameter using our estimate

- Confidence intervals
- Hypothesis testing

Cautions about Confidence Intervals

Important to note that the margin of error doesn't cover all errors

- Address only the randomness due to grabbing a *random* sample
- Does not address issues such as undercoverage, nonresponse, etc.

Choosing Sample for Confidence Intervals

A researcher can determine the number of observations required in the sample in order to achieve a desired margin of error.

$$m = z^* \frac{\sigma}{\sqrt{n}} \implies n = \left(\frac{z^* \sigma}{m} \right)^2$$

where m is the desired margin of error, and z^* is the z-score associated with the confidence interval level

Example

Say we are recording tip size of patrons when a waiter writes a message on the receipt. We know $\sigma = 2$. We want to estimate the mean percentage tip μ for patrons who receive the message within ± 0.5 with 90% confidence. How many patrons must we observe?

In other words we want $m = 0.5$:

$$n = \left(\frac{z^* \sigma}{m}\right)^2 \implies n = \left(\frac{1.645 \cdot 2}{0.5}\right)^2 = 43.3$$

Cautions about Hypothesis Testing

These tests of significance depend on:

- The alternative hypothesis (left-tail, rigth-tail, two-tail)
- The sample size, n
- The level of significant, α

Planning for Hypothesis Testing

How do we choose α?

Our choice of level of significance, α, depends on whether we REALLY want not wrongly reject H_0 or if we REALLY don't want to fail to reject H_0

- *Example:* Are you NASA trying to land someone on the moon? small α!!!
- *Example:* Are you a business trying to figure out if an A/B test on your website went well? can have a larger α

Types of Error

In any statistical test there are four possible outcomes:

	H_0 TRUE	H_a TRUE
Reject H_0	Type I Error	Correct
Fail to Reject H_0	Correct	Type II Error

Type I Error

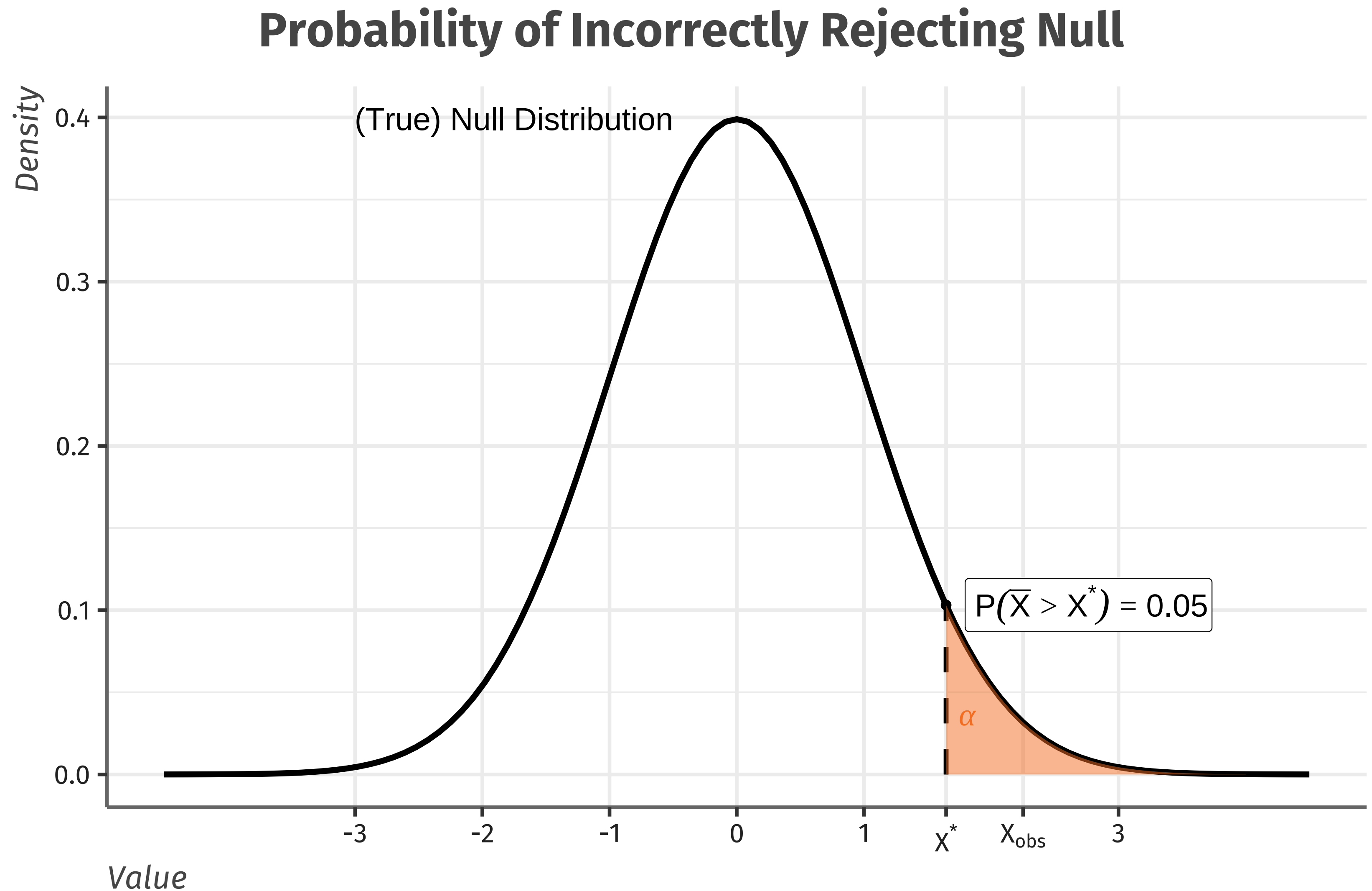
False Positive

Type I Error: We reject H_0, even though H_0 is true

- False-positive on a covid test
  - H_0: You do not have covid

Denote the probability of a type I error as α

Since our null hypothesis is *typically* that there is no effect, a type I error *typically* says there is an effect when in reality there is not



Probability of Incorrectly Rejecting Null
(True) Null Distribution
$P(\overline{X} > X^*) = 0.05$
α
Density
0.4
0.3
0.2
0.1
0.0
-3
-2
-1
0
1
X^*
X_{obs}
3
Value

Type II Error

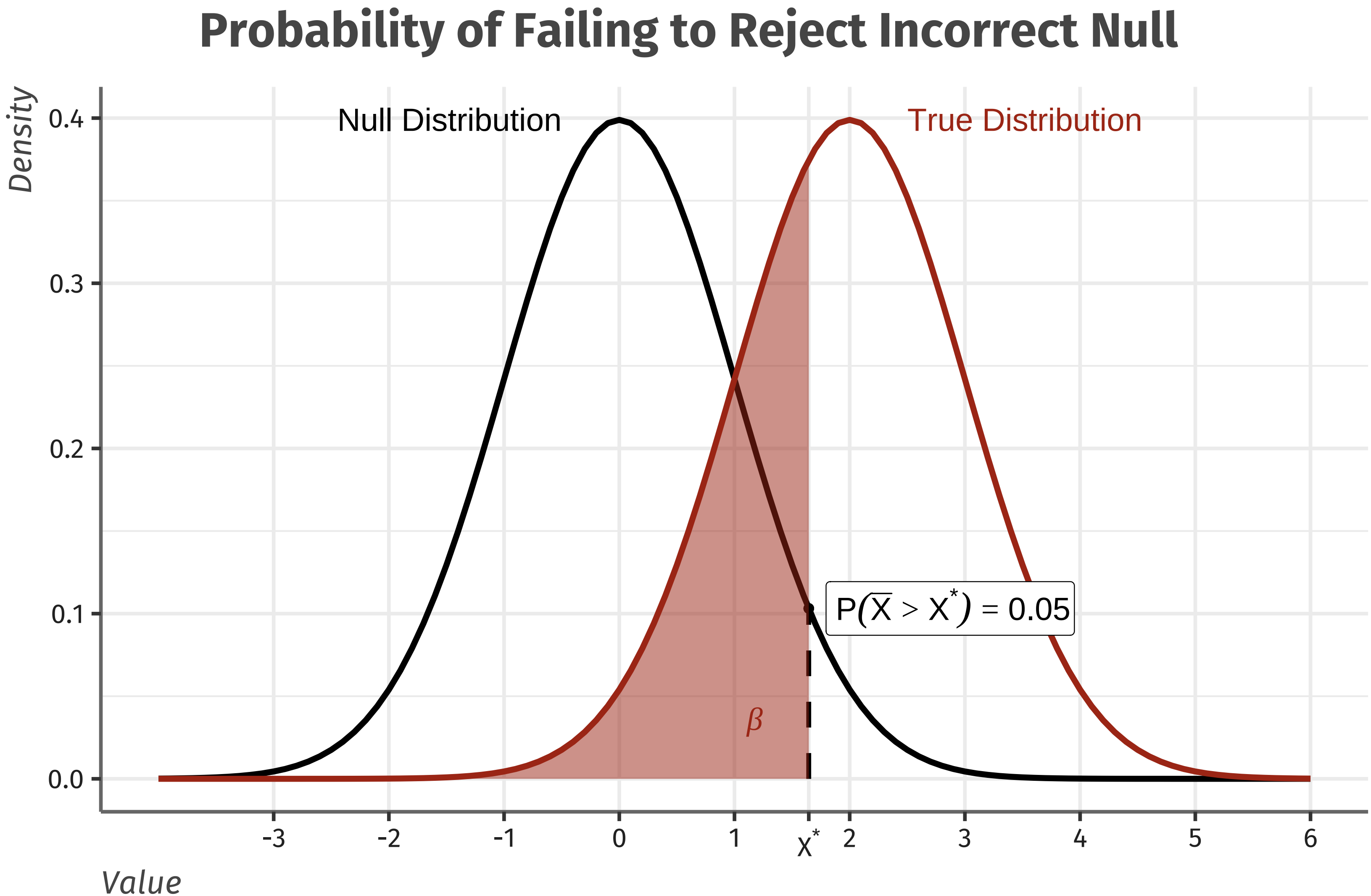
False Negative

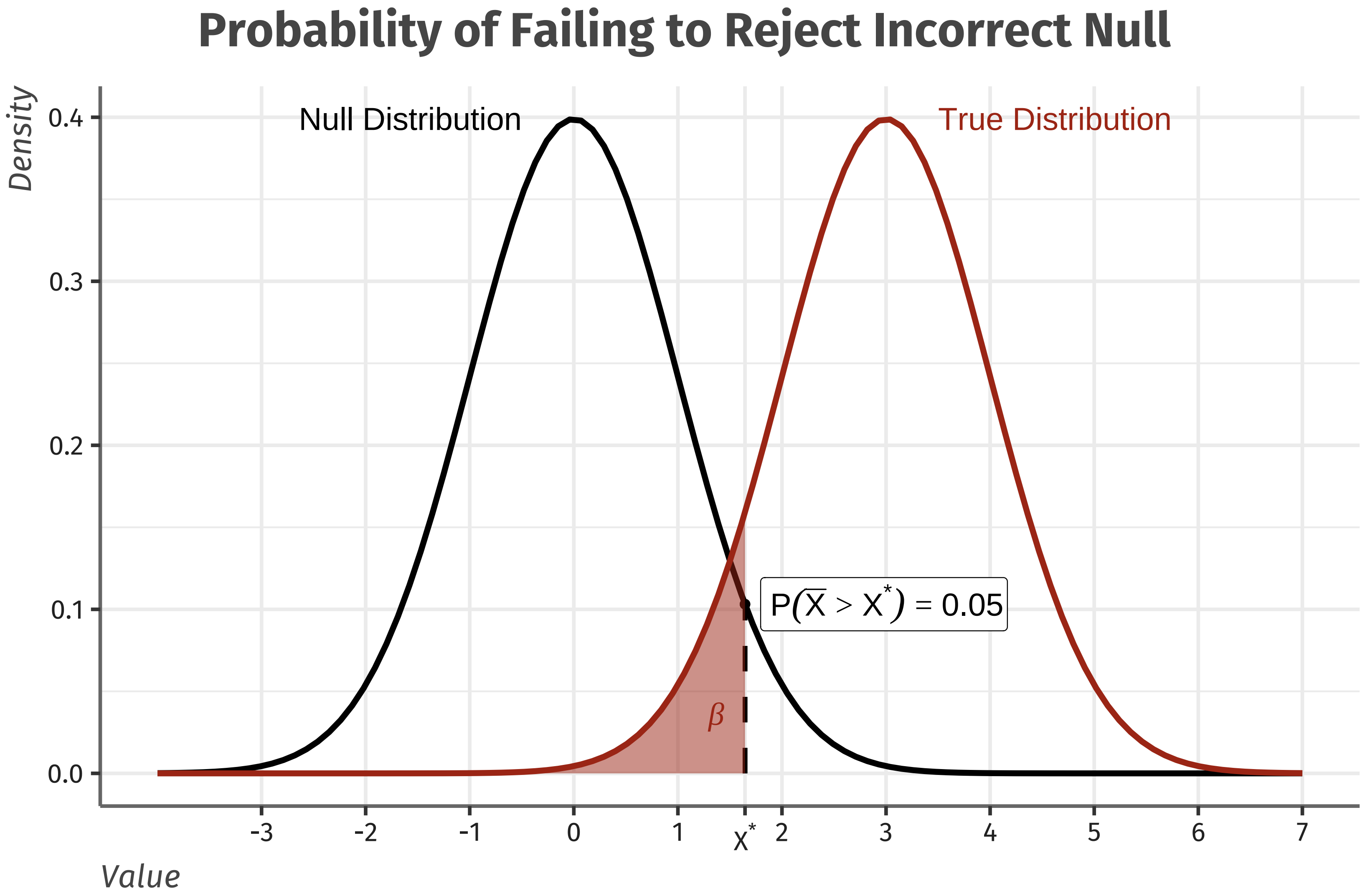
Type II Error: We fail to reject H_0, even though H_0 is false

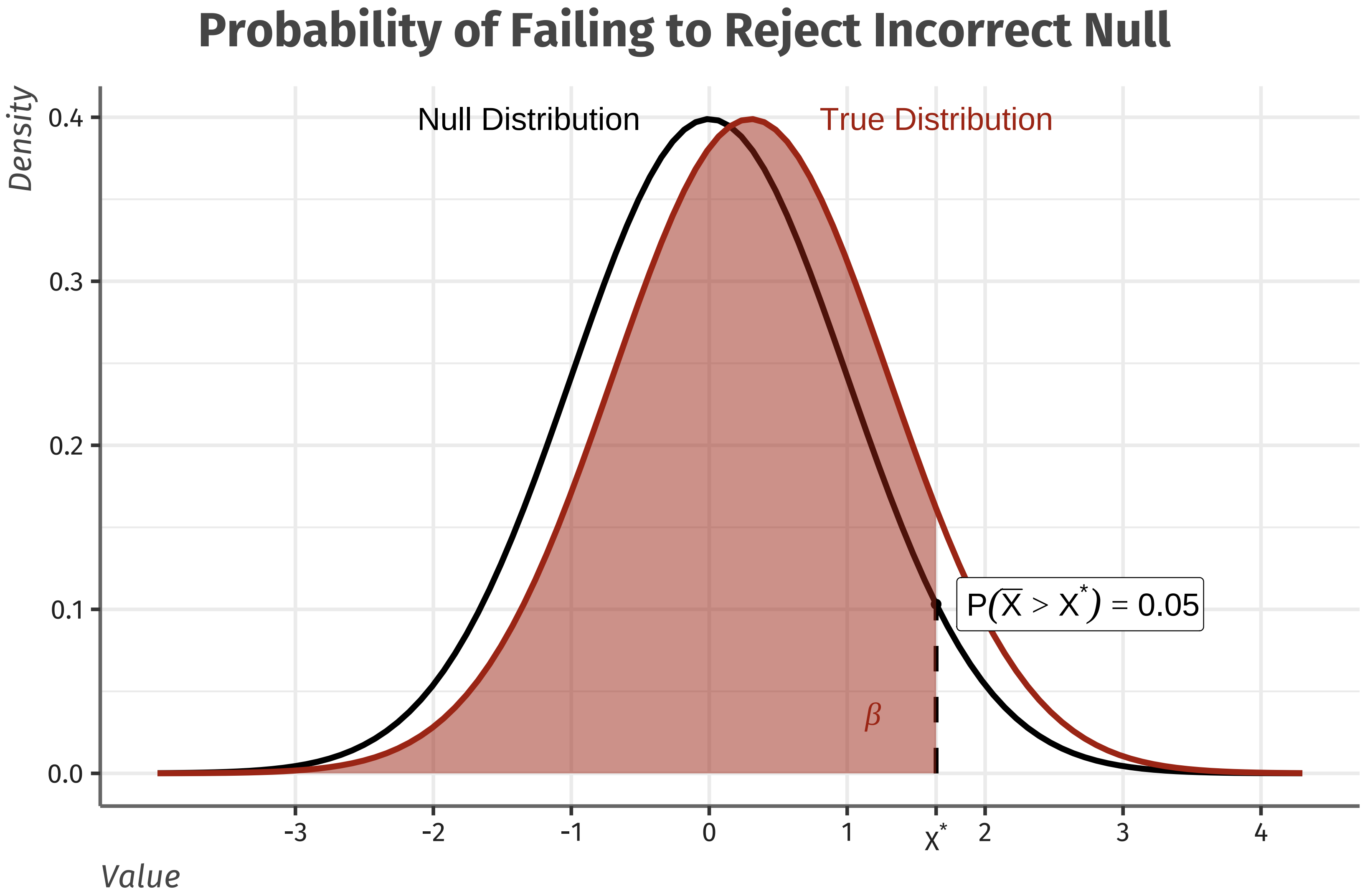
- False-negative on covid test
    - H_0: You do not have covid

Denote the probability of type II error as β

Since our null hypothesis is *typically* that there is no effect, a type II error *typically* says there is not an effect when in reality there is something different going on


Probability of Failing to Reject Incorrect Null
Density
0.4
0.3
0.2
0.1
0.0
Null Distribution
True Distribution
P(X̄ > X*) = 0.05
β
-3
-2
-1
0
1
X*
2
3
4
5
6
Value


Probability of Failing to Reject Incorrect Null
Density
0.4
0.3
0.2
0.1
0.0
Null Distribution
True Distribution
P(X̄ > X*) = 0.05
β
-3
-2
-1
0
1
X*
2
3
4
5
6
7
Value



Probability of Failing to Reject Incorrect Null
Density
Value
Null Distribution
True Distribution
P(X̄ > X*) = 0.05
β
0.4
0.3
0.2
0.1
0.0
-3
-2
-1
0
1
X*
2
3
4

How to remember

> When the boy cried wolf, the village committed Type I and Type II errors, in that order

There is no wolf

- Village rejects correct null (Type I)
- Village incorrectly fails to reject false null (Type II)

Clicker Question

Suppose we have the following hypothesis test:

- H_0: Taking multivitamins does not impact your running speed
- H_1: Taking multivitamins *will increase* your running speed

If we make the claim "Taking vitamins in the morning will increase your running speed" and it is not true, we have committed a:

a. Type I error
b. Type II error

Errors in Hypothesis Testing

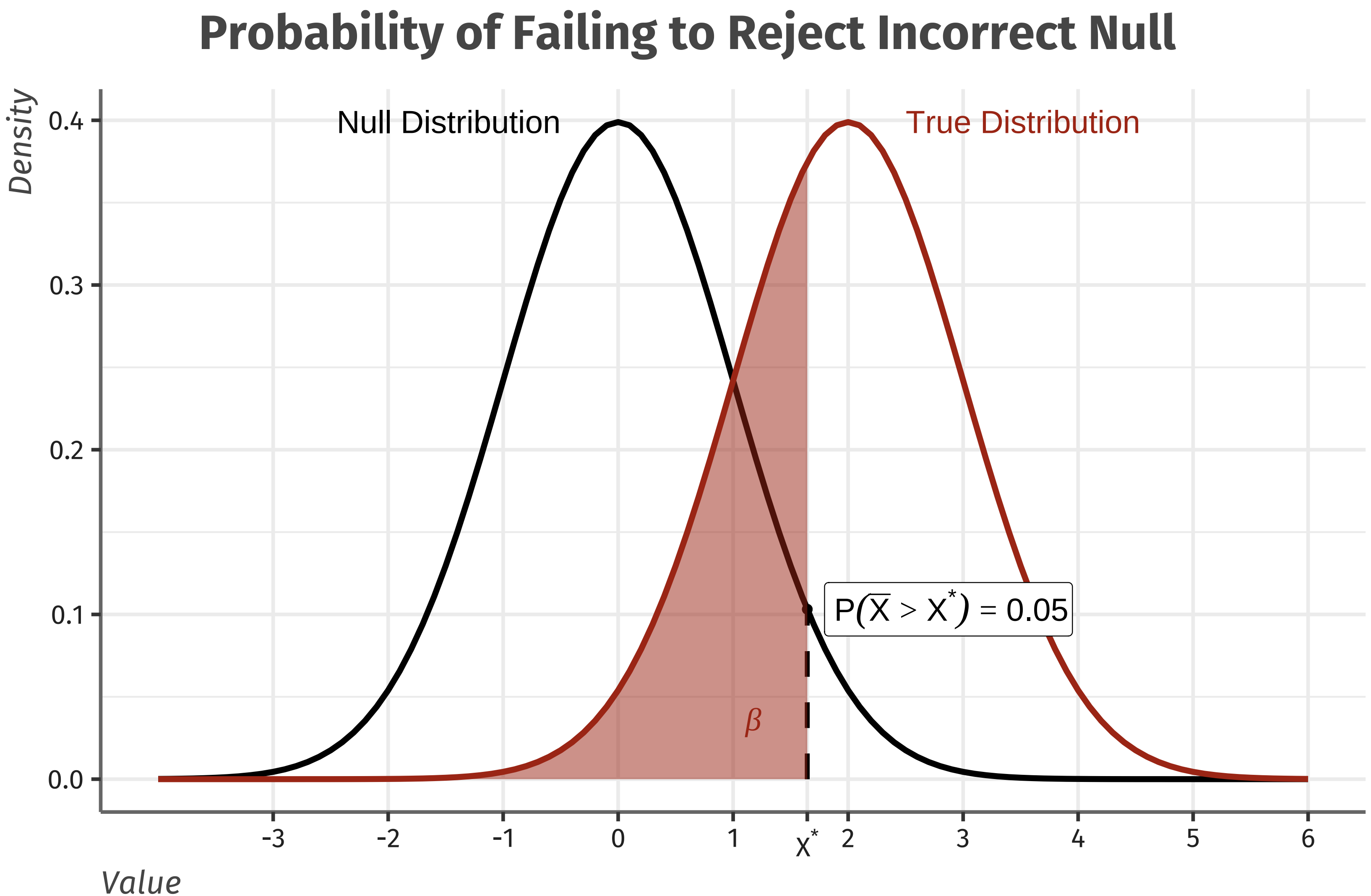
How do these errors happen?

- Our conclusions are based on sample data and probabilities
    - p-value tells us probability of observing it. The p-value is $ >0$ so it is possible to observe it
- We do not have enough information (sample size)
- We do not choose to be very rigorous (α)

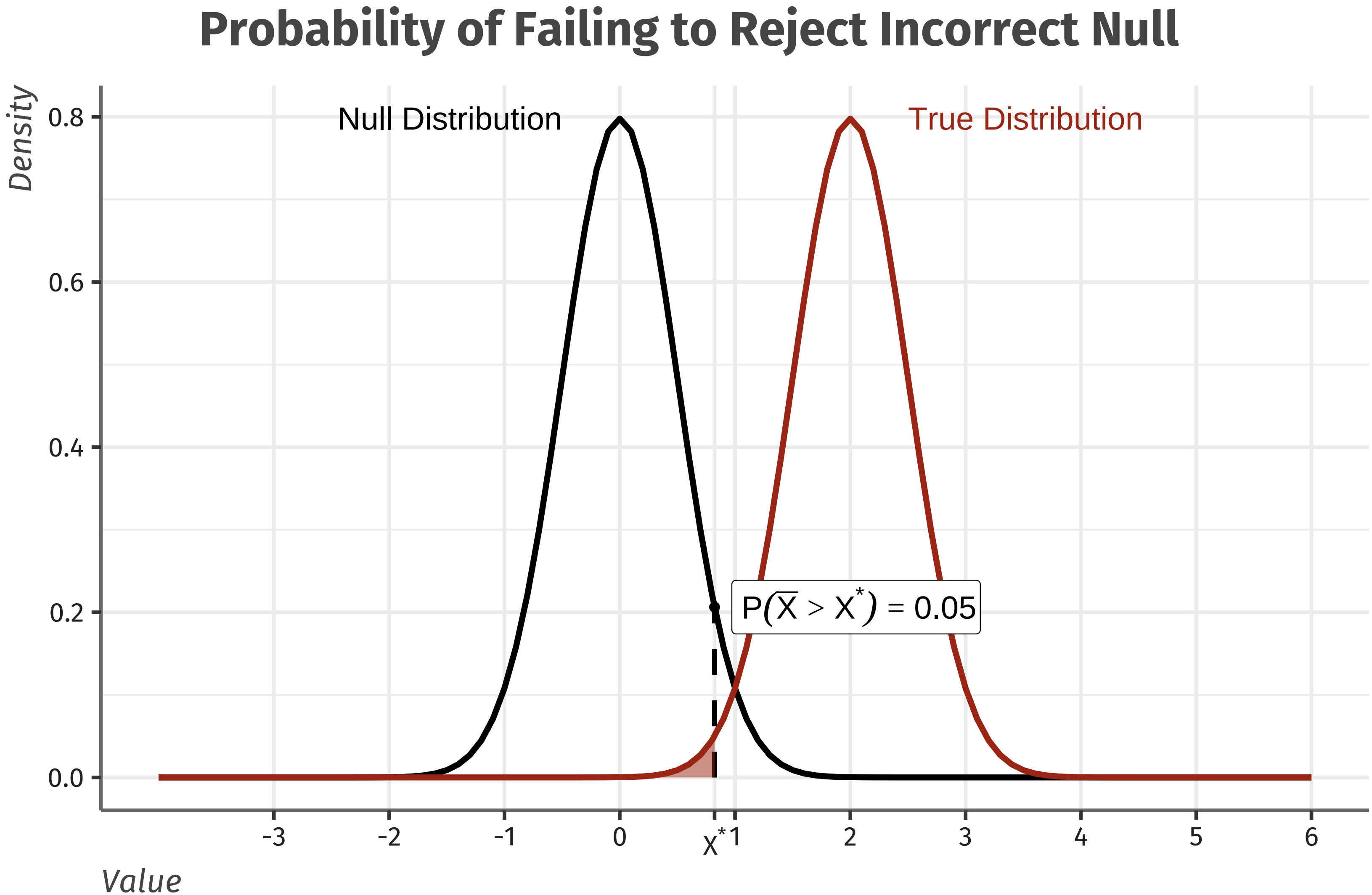
In particular we control

- Type I error is determined by the significance of the test α
- Type II error depends on the **true distribution** when the null is false
    - However, we can mitigate it by increasing the sample size

Improving power by increasing sample size



Improving power by increasing sample size



Size of a Test

Now that we've defined Type I error, lets define size:

The **size** of a test, α, is the probability of making a Type I error.

Given a null hypothesis $H_0 : \theta = \theta_0$; a test statistic $\hat{\theta}$; and a rejection region R,

The size is:

$$\alpha = P(\text{Type I Error}) = P(\hat{\theta} \in R \mid \theta = \theta_0)$$

Calculating the Size of a Test

How do we actually calculate α?

Let's suppose we have $n = 16$ and $\sigma = 1$, and we want to test H_0: $\mu = 3$ vs. H_a: $\mu > 3$.

Given a rejection region of $R = \{\bar{X} \mid \bar{X} > 3.41\}$, what is α?

$$\alpha = P(\hat{\theta} \in R \mid \theta = \theta_0) = P\left(\bar{X} > 3.41 \mid \mu = 3\right)$$

$$= P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} > \frac{3.41 - 3}{1/\sqrt{16}}\right) = Pr(Z > 1.64) = 0.05$$

Choosing Size

Note that we have to pick *either* the rejection region or the size

- We generally pick a size and calculate the rejection region based off that size
- Because the size is the probability of a rejecting a true null, by choosing α we are choosing how much we are willing to risk *incorrectly* rejecting the null hypothesis
- Higher α will mean more of the sample statistics are in the rejection region, meaning a higher risk of rejecting the null even though it's true

Power of a Test

While size deals with Type I Errors, power deals with Type II.

The **power** is the probability of correctly rejecting a false null, or 1 - P(Type II Error)

$$\text{Power} = 1 - P(\text{Type II Error})$$

Intuitively, power is the likelihood of detecting a false null using your test statistic.

Power and Probability of Type II Errors

A Type II error is the probability of failing to reject a false null

$$P(\text{Type II}) = P(\bar{X} \notin R \mid \mu = \mu_A)$$

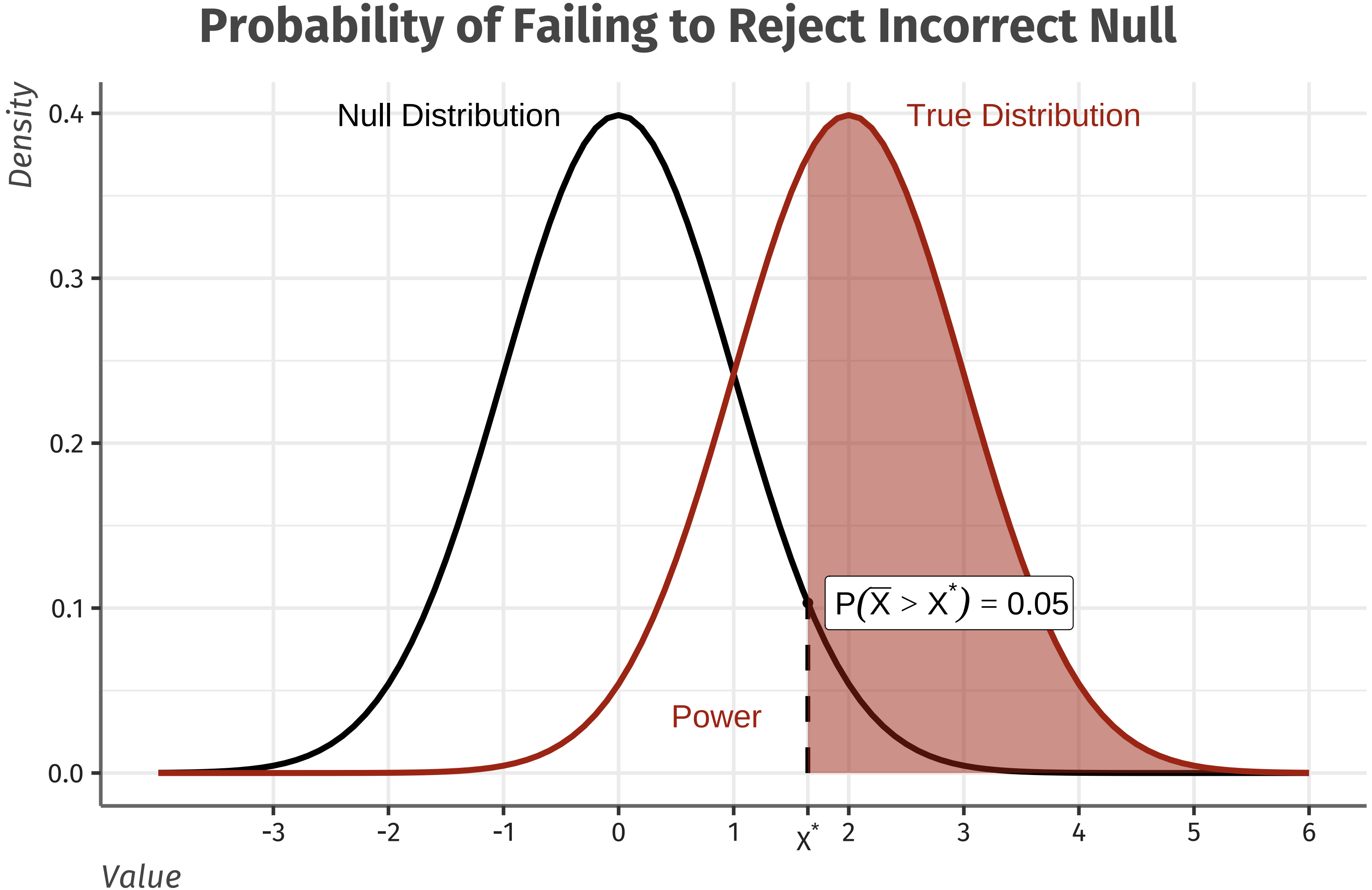
The **power** is the probability of correctly rejecting a false null

$$\text{Power} = P(\bar{X} \in R \mid \mu = \mu_A)$$

You can think of power as the probability of *not making a Type II error*

You can calculate power by doing $1 - P(\text{Type II})$ or by calculating the power directly.

Power



Probability of Failing to Reject Incorrect Null
Density
0.4
0.3
0.2
0.1
0.0
Null Distribution
True Distribution
P(X̄ > X*) = 0.05
Power
-3
-2
-1
0
1
X*
2
3
4
5
6
Value

Calculating the Power of a Test

Back to previous example, where $n = 16$, $\sigma = 1$, and $R = \{\bar{X} \mid \bar{X} > 3.41\}$. And we are testing $H_0 : \mu = 3$ vs. $H_1 : \mu > 3$:

Power can be calculated in two ways:

$$\text{Power} = P(\text{reject } H_0 \mid \mu_0 = \mu^*) = P(\bar{X} \in R \mid H_0 \text{ false})$$

$$\text{Power} = 1 - P(\text{type II Error}) = 1 - P(\bar{X} \notin R \mid H_0 \text{ false})$$

Calculating the Power of a Test

In order to calculate the power of a test, we must assume a specific true mean, μ_A.

For example, what is the power of the test if the true mean is $\mu_A = 4$?

$$\text{Power} = P(\bar{X} \in R \mid \mu = \mu_A = 4)$$

$$P(\bar{X} > 3.41 \mid \mu = 4) = P(Z > \frac{3.41 - 4}{1/\sqrt{16}}) = 0.9908$$

Calculating the Power of a Test

We can also calculate the power of a test by subtracting the probability of making a Type II error ((\beta) from 1.

$$\beta = P(\bar{X} < 3.41 \mid \mu = \mu_A = 4)$$

$$\implies P(Z < \frac{3.41 - 4}{1/\sqrt{16}}) = 0.0092$$

Meaning the power of the test is:

$$\text{Power} = 1 - \beta = 1 - .0092 = .9908$$

There is a 99.1% chance that in *repeated sampling* we reject the null that $\mu = 3$ if the true mean is equal to 4.

Group Question

Assume $X \sim N(\mu, 5^2)$. From a sample size of $n = 100$, we wish to test the following at the $\alpha = 0.05$ level

$$H_0 : \mu = 3$$

$$H_1 : \mu > 3$$

What is the power of your test if $\mu = \mu_A = 4$?

a. 0.85

b. 0.15

c. 0.64

d. 0.36

Interpreting Power

Power is the probability of correctly rejecting a false null hypothesis

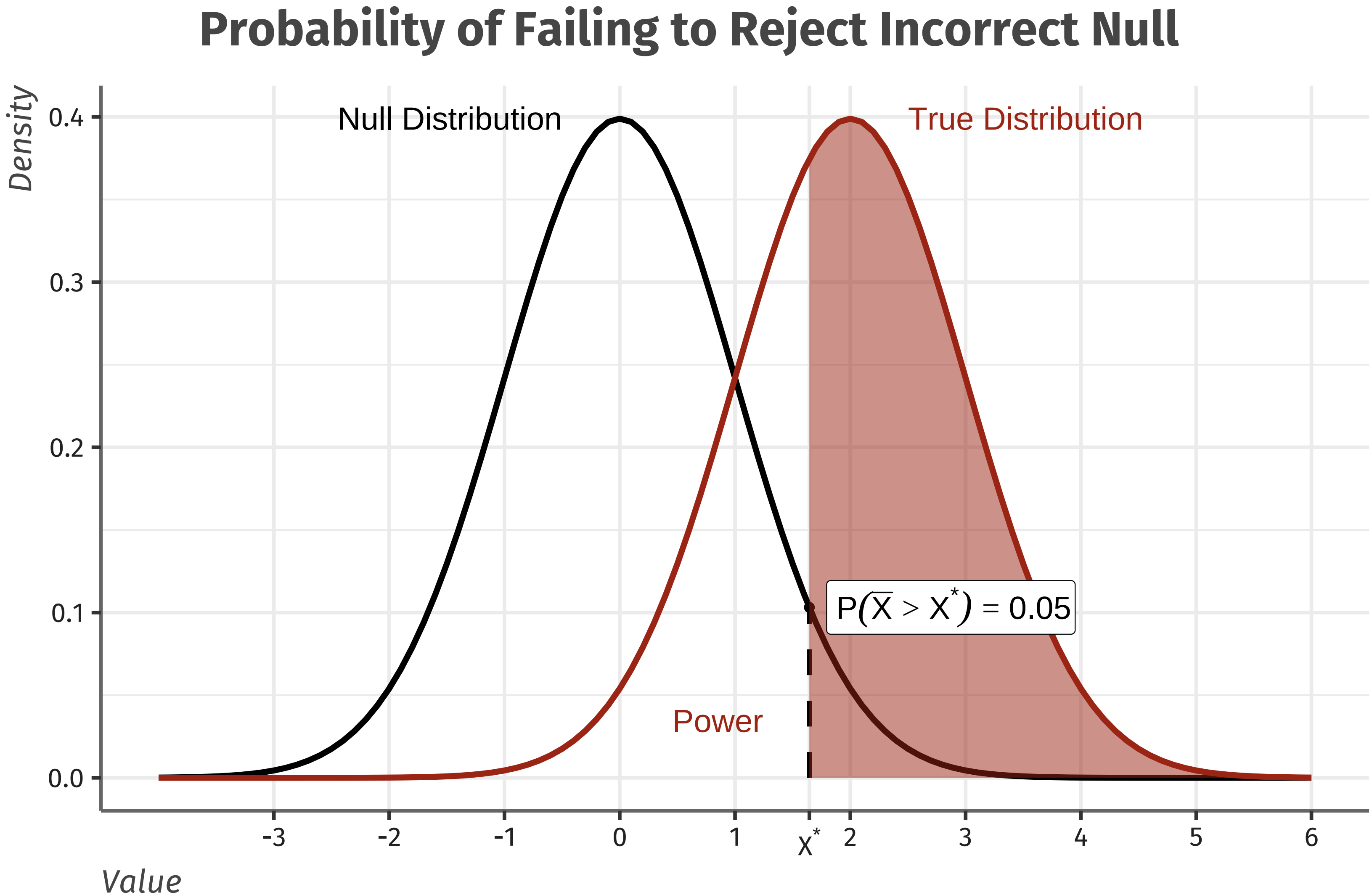
- Can be thought of as our ability to identify a true value from an alternative

In general, the power is a function of the true value*

- It changes as we try out different possible true values

* Must specify a specific true μ in order to calculate power

Power



Probability of Failing to Reject Incorrect Null
Null Distribution
True Distribution
$P(\bar{X} > X^*) = 0.05$
Power
Density
0.4
0.3
0.2
0.1
0.0
-3
-2
-1
0
1
X^*
2
3
4
5
6
Value

Visualizing Underpowered Estimates

Imprecise Estimates

Visualizing Underpowered Estimates

Small Relative Differences

Spotting Underpowered Estimates

How can we avoid underpowered estimates? There are two main root causes:

Imprecise estimates

- Low precision/high variance
- Large standard errors interpreted as "no effect"

Small relative differences between θ_0 and θ_A

- Precise estimates can detect small relative differences
- Imprecise estimates require large relative differences to detect the truth.

Watch for imprecise estimates! They are often interpreted as a true result when really they are underpowered.

Example

Underpowered Estimates

Suppose from 10 observations you estimate that raising the minimum wage by 1% would lead to only a 0.1% decline in employment on average with a standard deviation of 6%. Can you reject the null that employment wouldn't decrease at the 5% significance level?

$$p\text{-value} = Pr(\bar{X} < -0.1 \mid \mu = 0) = Pr\left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} < \frac{-0.1 - 0}{6/\sqrt{10}}\right)$$

$$= Pr(Z < -0.053) = 0.479$$

Since $p\text{-value} \nless \alpha$, we conclude there is not enough evidence to say that average employment reduction is not 0% (no effect of minimum wage).

Example

Underpowered Estimates

Great news! Raising the minimum wage has no statistically discernible effect on employment, right? Well.. hold on... If there is an effect on employment our statistic may be too underpowered to detect it. Let's calculate the power of this test....

Example

Underpowered Estimates

Calculate power by $P(\bar{X} \in R \mid \mu_0 = -0.5)$

This means we must first calculate the rejection region

If $\alpha = .05$, then the rejection region is $R = \{\bar{X} \mid \bar{X} < -3.12\}$.

Example

Underpowered Estimates

Let's assume a reasonable negative impact on employment of 0.5%. (So we're assuming the true $\mu = -0.5$).

Then the power is:

$$P(\text{Reject } H_0 \mid \mu = -0.5)$$

$$P(\text{Reject } H_0 \mid \mu = -0.5)$$

$$P(\text{Reject } H_0 \mid \mu = -0.5) = P(Z < -1.38) \approx 0.0836$$

Our power to detect a measurable effect is a measly 8.4%!