Decision Error

Grinnell College

November 9, 2026

Strength of Evidence

So far, our process has been as follows:

1. Being with a null hypothesis, $H_0 : \mu = \mu_0$
2. Collected data and comptute statistic, i.e, $\overline{x}$
3. Compare our statistic against the null distribution, i.e., $t = \frac{\overline{x} - \mu_0}{\hat{\sigma}/\sqrt{n}}$
4. Derive a p-value based on the statistic and the distribution

We found that we could use our p-value to quantify the strength of evidence against our null: the smaller the p-value, the less likely our observed data if the null were true

Decision Making

Based on the evidence we have collected, we must ultimately decide between one of two decisions:

1. There is sufficient evidence to reject H_0
2. There is *not* sufficient evidence to reject H_0

Decision Making

Just as our confidence intervals were correct or incorrect, so to may be our decision regarding H_0. In this case, however, there are two distinct ways in which our decision can be incorrect:

1. H_0 is *TRUE*, yet we reject anyway
   - Example: our sample was outlier and resulted in large t-statistic
2. H_0 is *FALSE*, yet we fail to reject it
   - Example: our confidence interval was too large and contained the incorrect null hypothesis

Decision Making

These two types of errors are known as Type I and Type II errors, respectively:

1. H_0 is *TRUE*, yet we reject anyway
   - Type I error
   - "False positive"
   - Evidence leads to wrong conclusion
2. H_0 is *FALSE*, yet we fail to reject it
   - Type II error
   - "False negative"
   - Not enough evidence to conclude

Decision Making

Test Result	True State of Nature	
	H_0 True	H_0 False
Fail to reject H_0	Correct	Type II Error
Reject H_0	Type I Error	Correct

Type I Errors

A Type I error describes a situation in which we incorrectly identify a null effect:

- Conclude that an intervention works when it does not
- Conclude that there is a relationship between two variables when there are not

A Type I error will occur, for example, when our null hypothesis is correct but our constructed confidence does not contain the true parameter (e.g., $\mu_0 = \mu$)

Type I Error Rate

We can control the rate at which we commit Type I errors with adjusting the *level of significance*, denoted α.

This is also called the *Type I error rate*

The Type I error rate has a *one-to-one* correspondence with our confidence intervals: a 95% confidence interval will permit a Type I error 5% of the time, corresponding to $\alpha = 0.05$

We *reject* our null hypothesis when p-value $< \alpha$

Checking a Previous Result

Previous researchers reported a mean crawdad claw length of 4.60 cm. We collect new data to check this claim.

$$H_0 : \mu = 4.60 \qquad H_A : \mu \neq 4.60$$

For this example, assume a normal population with a **known** population SD of $\sigma = 0.55$ cm.

Our sample: $n = 25$, $\bar{x} = 4.842$ cm.

Another researcher's sample: $n = 100$, $\bar{x} = 4.732$ cm.

One Decision Rule

With the population SD known, we use a z-score:

$$SE = \frac{\sigma}{\sqrt{n}}, \qquad z = \frac{\bar{x} - 4.60}{SE}.$$

At $\alpha = .05$, reject if $|z| < 1.96$

Our 95% confidence interval is $\bar{x} \pm 1.96\, SE$.

Changing n changes SE. The critical values stay at ± 1.96.

Two Samples, Two Rejections

Our sample: $n = 25$

$$SE = \frac{0.55}{\sqrt{25}} = 0.110$$

$$z = \frac{4.842 - 4.60}{0.110} = 2.20$$

95% CI:

$4.842 \pm 1.96(0.110)$

$(4.626,\ 5.058)$ cm

Other sample: $n = 100$

$$SE = \frac{0.55}{\sqrt{100}} = 0.055$$

$$z = \frac{4.732 - 4.60}{0.055} = 2.40$$

95% CI:

$4.732 \pm 1.96(0.055)$

$(4.624,\ 4.840)$ cm

Both reject H_0. Neither interval includes 4.60 cm.

But the Null Is True

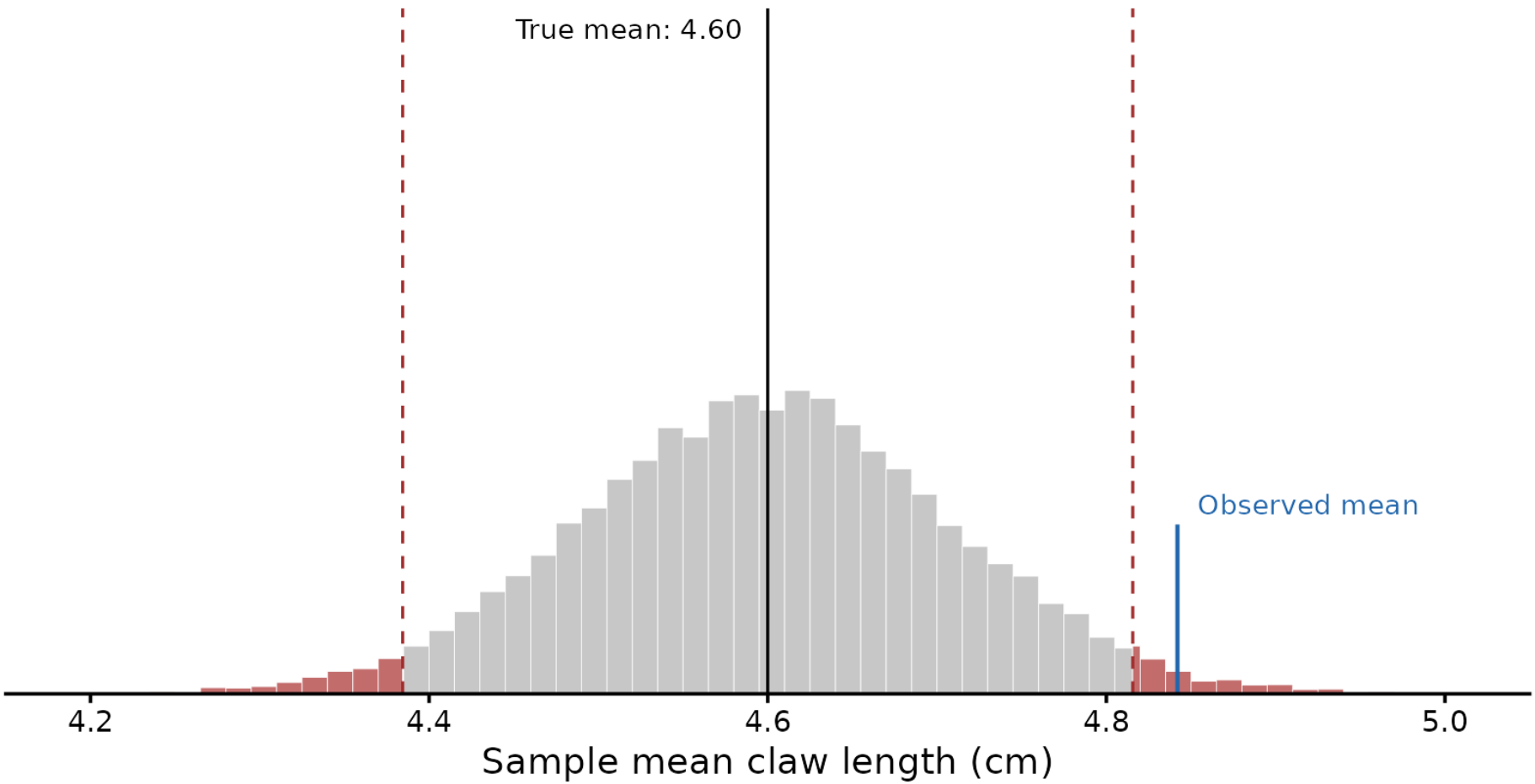
Suppose the true population mean really is 4.60 cm.

Both researchers have made a **Type I error**.

How often would this happen if we repeatedly collected new samples?

The next plots show 20,000 sample means at each sample size. The shaded tails mark samples that reject the true null.

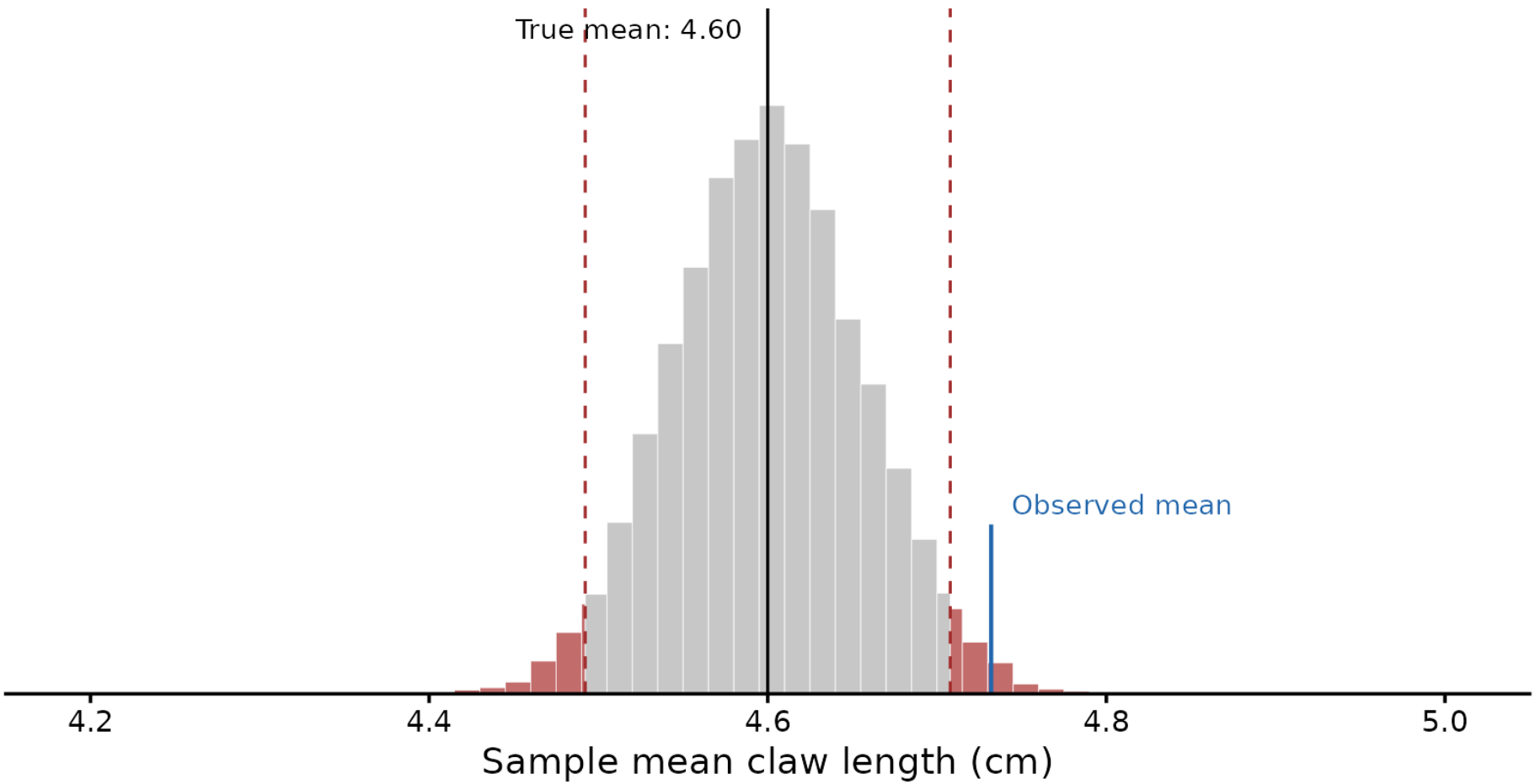
Repeated Samples: $n = 25$



$SE = 0.110$ cm. Reject outside $4.60 \pm 1.96(0.110)$.

Shaded tails: 5% in theory; 4.89% in this simulation.

Repeated Samples: $n = 100$



$SE = 0.055$ cm. Reject outside $4.60 \pm 1.96(0.055)$.

Shaded tails: 5% in theory; 5.08% in this simulation.

More Precision, the Same Error Rate

- Larger n gives a smaller SE and a narrower sampling distribution.
- The rejection boundaries move closer to 4.60 cm.
- The proportion beyond those boundaries stays at 5%.

Under our assumptions, α sets the Type I error rate.

Increasing n gives greater precision, but it does not make us any more or less likely to commit a Type I error.

Type II Errors

A Type II error describes a situation in which the null hypothesis is false, yet based on the evidence gathered we fail to reject it:

- An intervention has a clinical effect, but it is not detected
- An email is considered spam, but the filter does not detect it

At a fixed significance level, Type II errors are more likely when:

- There are too few observations in our sample
- The population has large variability
- The effect size is small

Effect Size

An important concept in identifying statistical differences is **effect size**: how far the population quantity is from its null value, or how strong the association between two variables are.

- A difference in the original units, such as $\mu - \mu_0$
- A standardized difference, such as $(\mu - \mu_0)/\sigma$
- Other measures, such as correlations or odds ratios

Larger effects are easier to detect, even with more variability or fewer observations. Small effects require more precision to detect reliably.

What If the Null Is False?

We still test the previous researchers' claim:

$$H_0 : \mu = 4.60 \qquad H_A : \mu \neq 4.60.$$

Now suppose the **true mean is 4.80 cm**, with a known population standard deviation of $\sigma = 0.55$

A Type II error occurs if we **fail to reject** this false null.

Assuming we start with a sample of size $n = 25$, how often will our sample detect the difference?

The Size of the Effect

The effect compares the **true population mean** with the null value:

$$\mu - \mu_0 = 4.80 - 4.60 = 0.20 \text{ cm}.$$

Expressed in population SDs, the standardized effect is

$$\frac{\mu - \mu_0}{\sigma} = \frac{0.20}{0.55} \approx 0.36.$$

Both of these quantities are *independent of our sample* and describe the population and its difference from our null

Collecting multiple samples or including more observations in our sample does not change the true effect.

Power and Type II Error

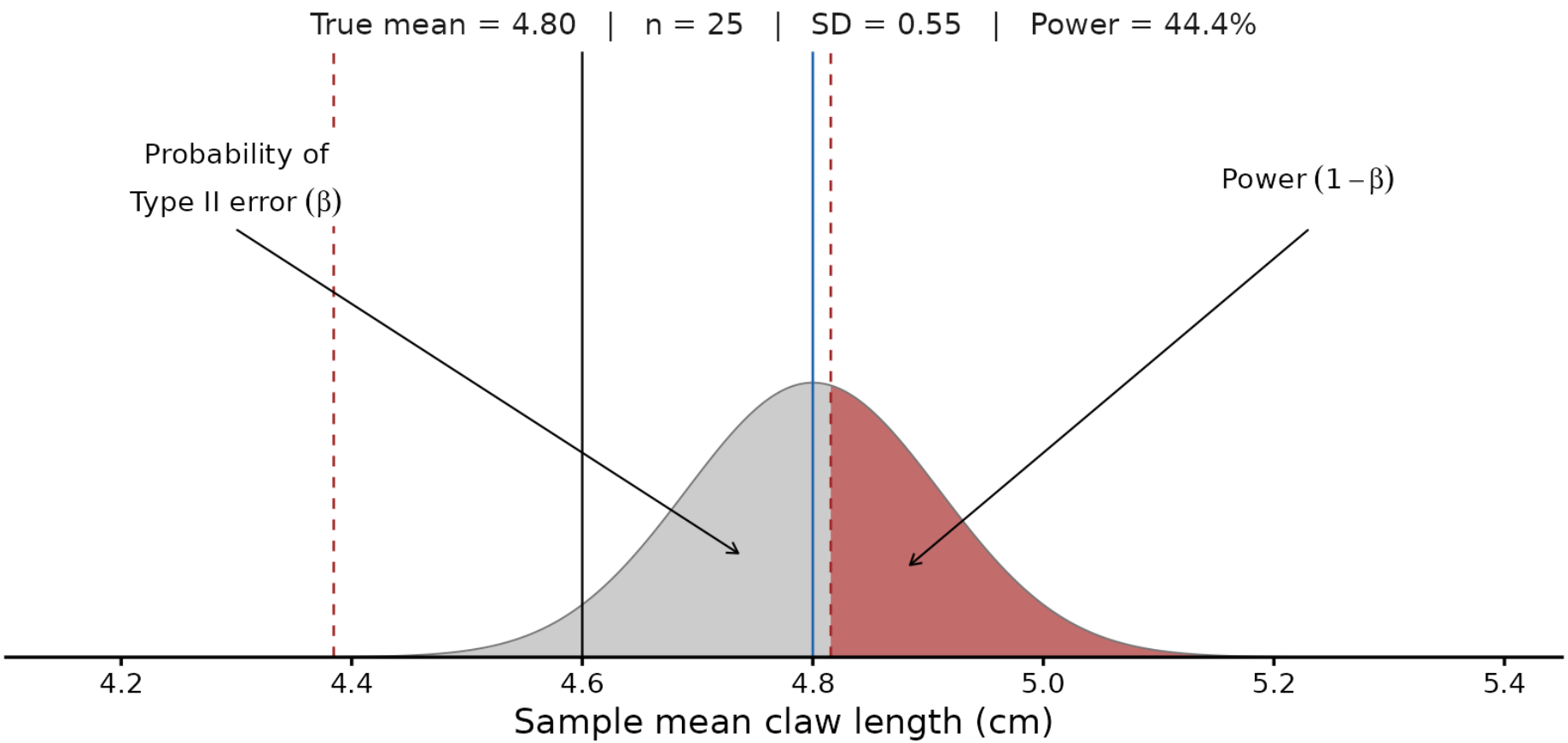
When the true mean is 4.80 cm, our sample means are centered at 4.80.

But our rejection boundaries still come from the **null value**, 4.60:

$$4.60 \pm 1.96\left(\frac{0.55}{\sqrt{25}}\right) = 4.60 \pm 0.2156 = (4.3844,\ 4.8156)$$

- Outside the boundaries: correctly reject. The probability is **power**, $1 - \beta$.
- Between the boundaries: fail to reject a false null. The probability is the **Type II error rate**, β.

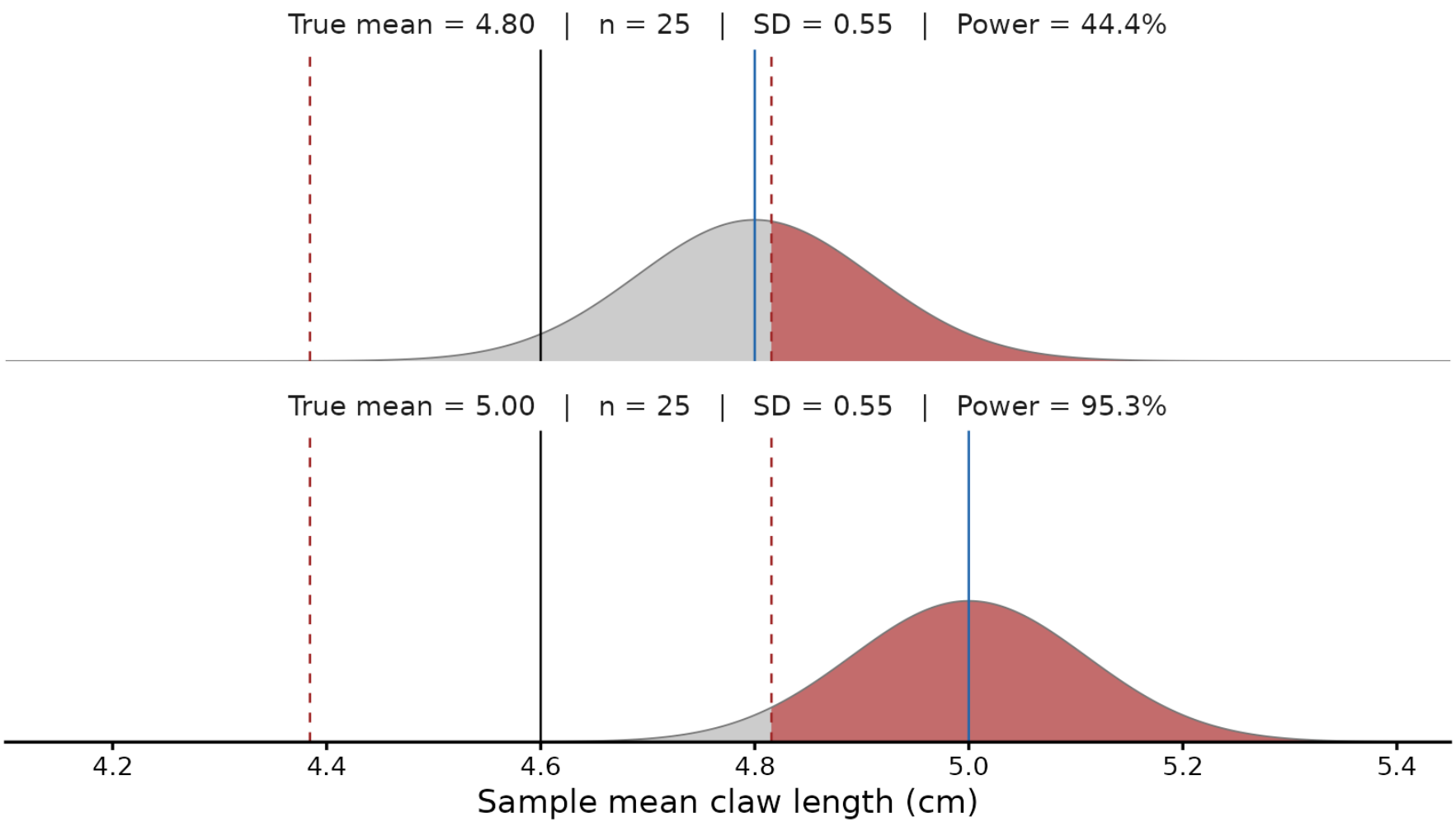
How Often Do We Detect the Effect?



Black: null mean. Blue: true mean. Dashed: rejection boundaries.

Red area: power $= 44.4\%$. Gray area: $\beta = 55.6\%$.

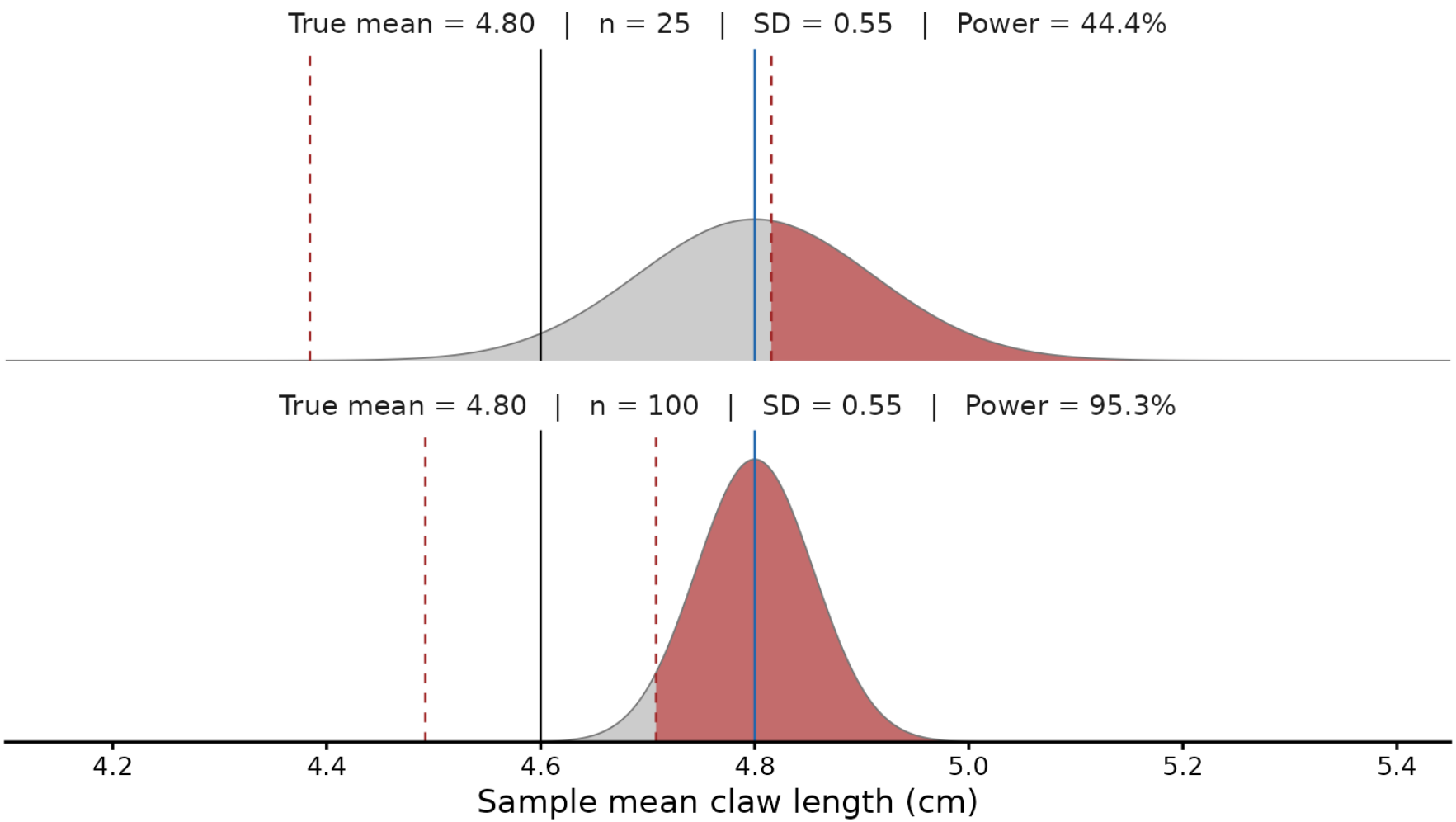
A Larger Effect



Hold n, σ, and α fixed. The boundaries stay put.

Moving the true mean farther from the null increases power.

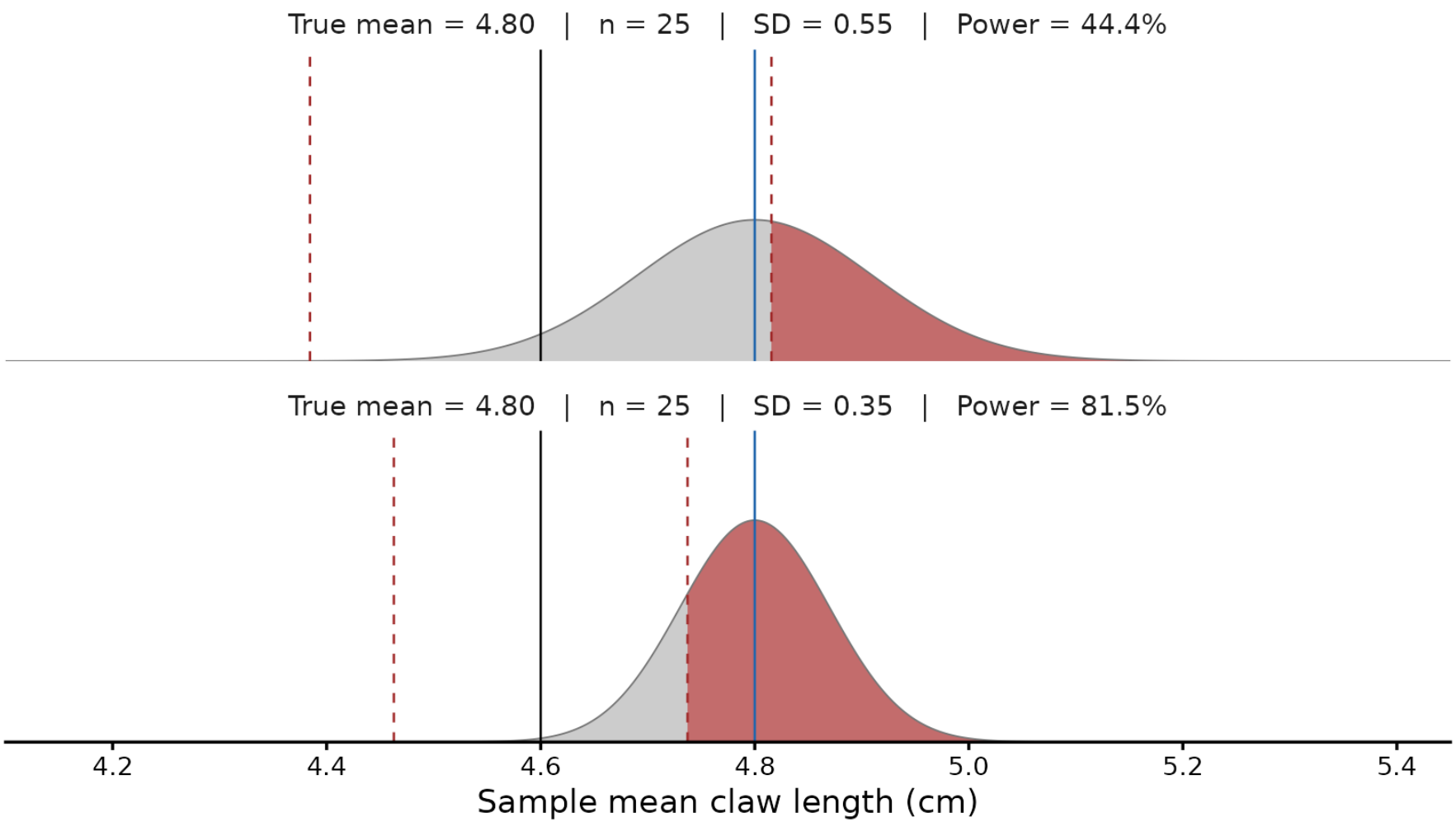
A Larger Sample



Hold the true effect and σ fixed. SE: 0.110 $\rightarrow$ 0.055 cm.

The distribution narrows and the boundaries move toward the null, while sample means fall closer to the true value, μ

Less Population Variability



Hold the true effect and n fixed. SE: 0.110 $\rightarrow$ 0.070 cm.

Smaller population SD also gives greater precision and power.

What Increases Power?

Throughout these comparisons, $\alpha = .05$.

- A larger true effect: more separation from the null value.
- A larger sample: smaller SE.
- Less population variability: smaller SE.

Increasing n does not change the population mean or population SD, but it *does* change the standard error associated with $\overline{x}$

Making Rejection Harder

There is a tension, however, between α and β

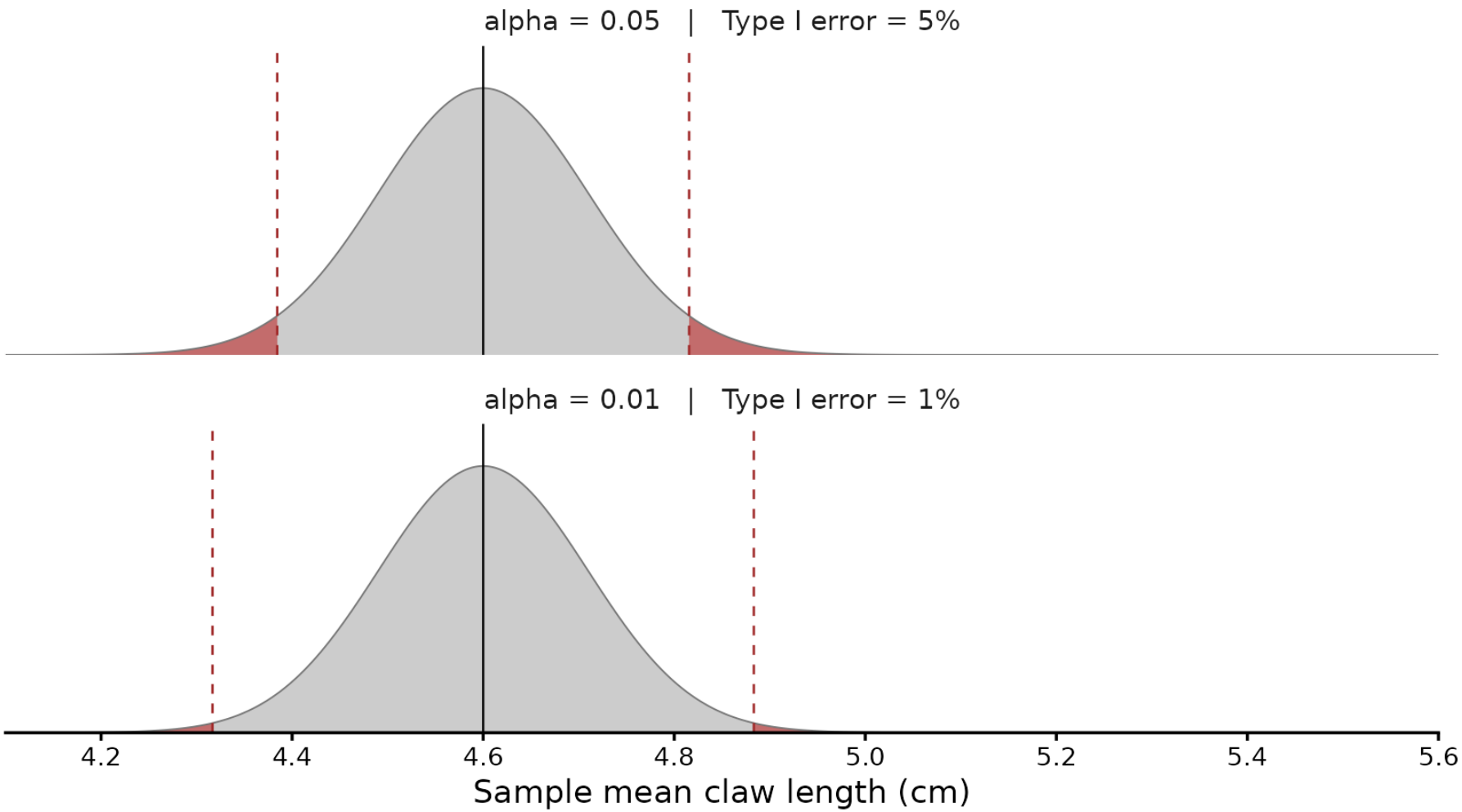
Lowering α requires stronger evidence to reject H_0, whether H_0 is true or false.

- If H_0 is true: fewer rejections means fewer Type I errors.
- If H_0 is false: fewer rejections means more Type II errors and less power.

With the population and sample size fixed, reducing α increases β.

How much power do we lose? That depends on the size of the effect.

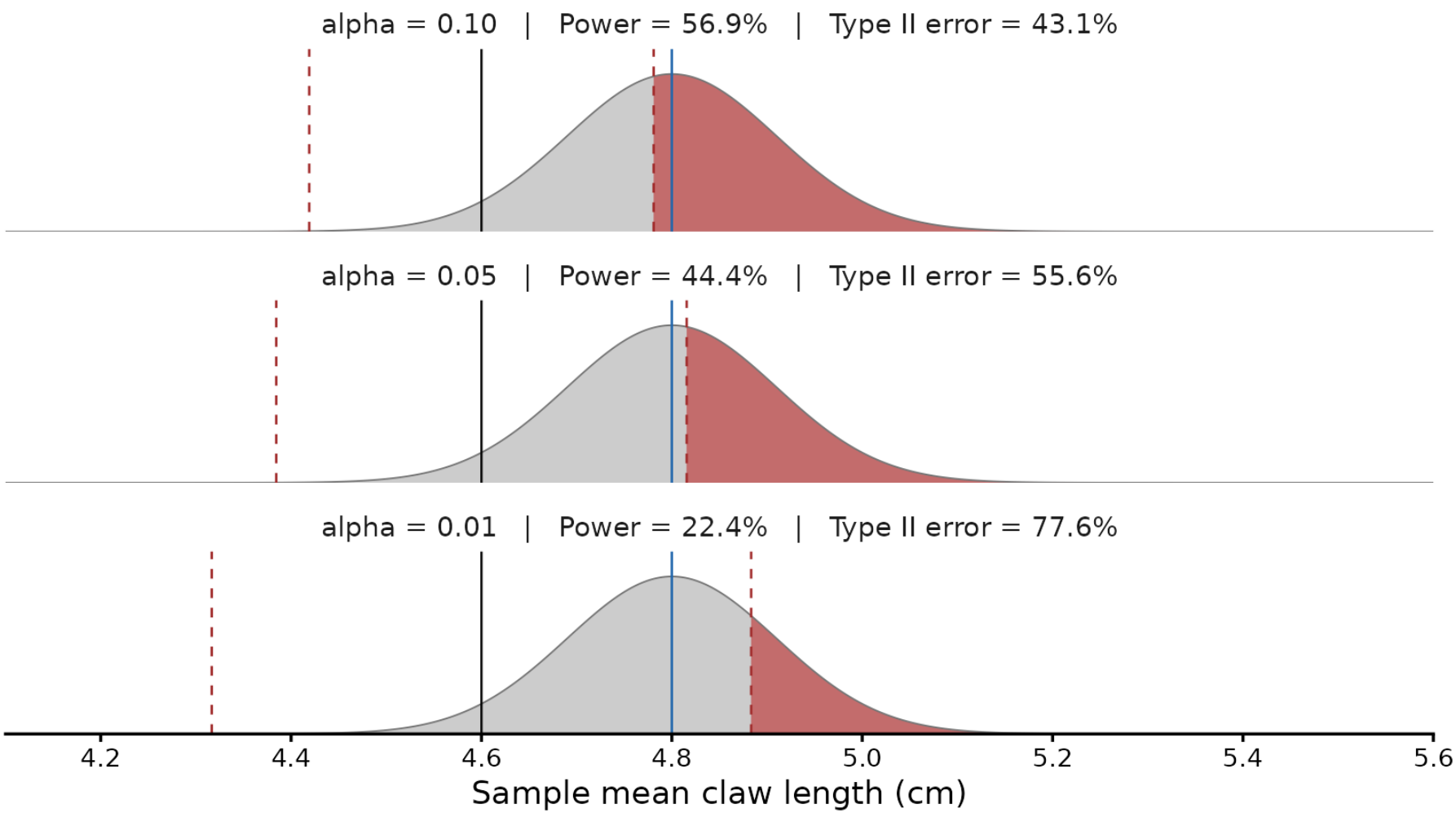
When the Null Is True



True mean = null mean = 4.60 cm. $n = 25$, $\sigma = 0.55$ cm.

Lower α: wider boundaries, fewer false positives.

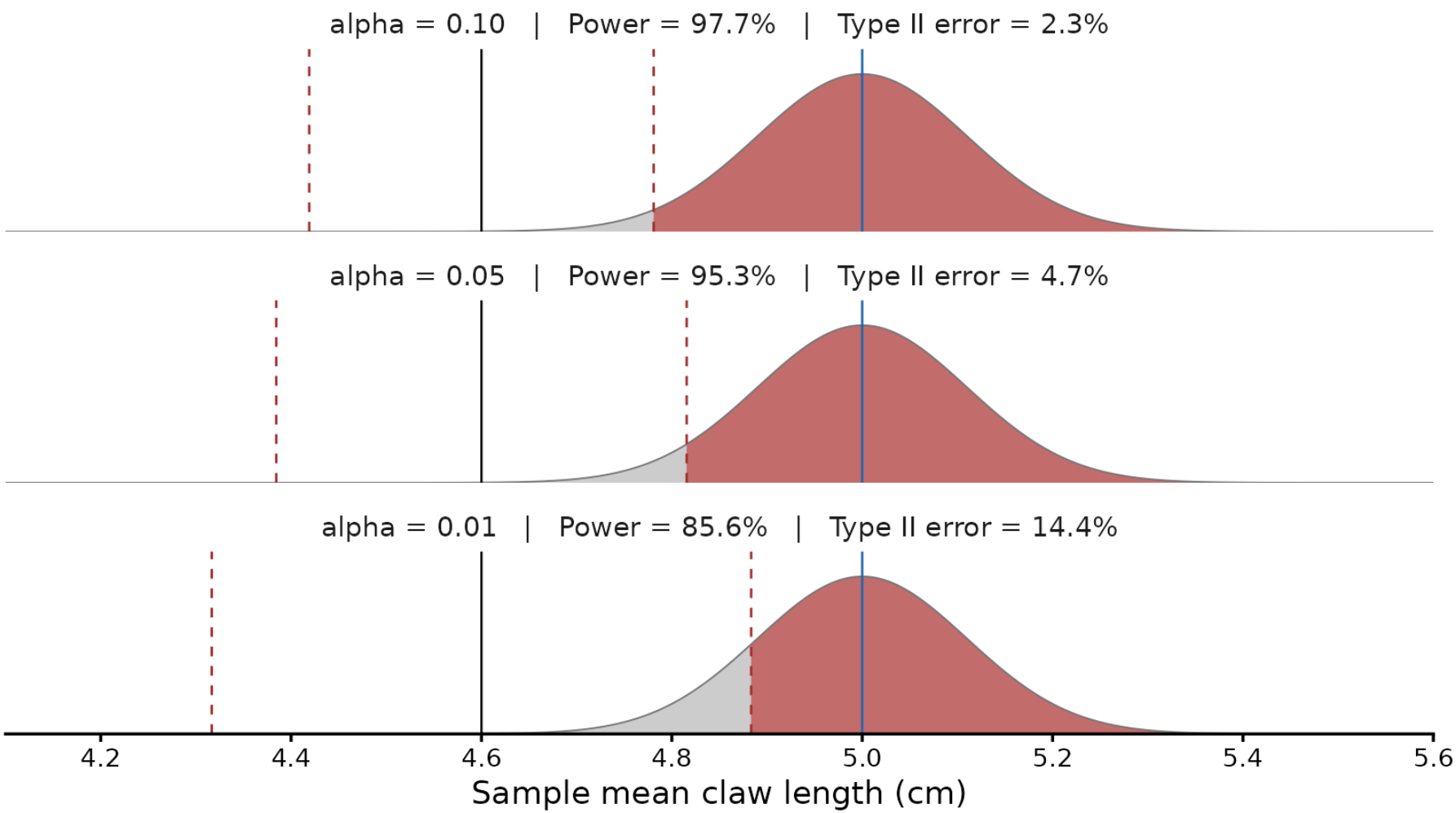
A Small Effect: Alpha Matters



True mean $= 4.80$ cm: effect $= 0.20$ cm. $n = 25$, $\sigma = 0.55$.

The distribution stays put. Wider boundaries leave more of it gray.

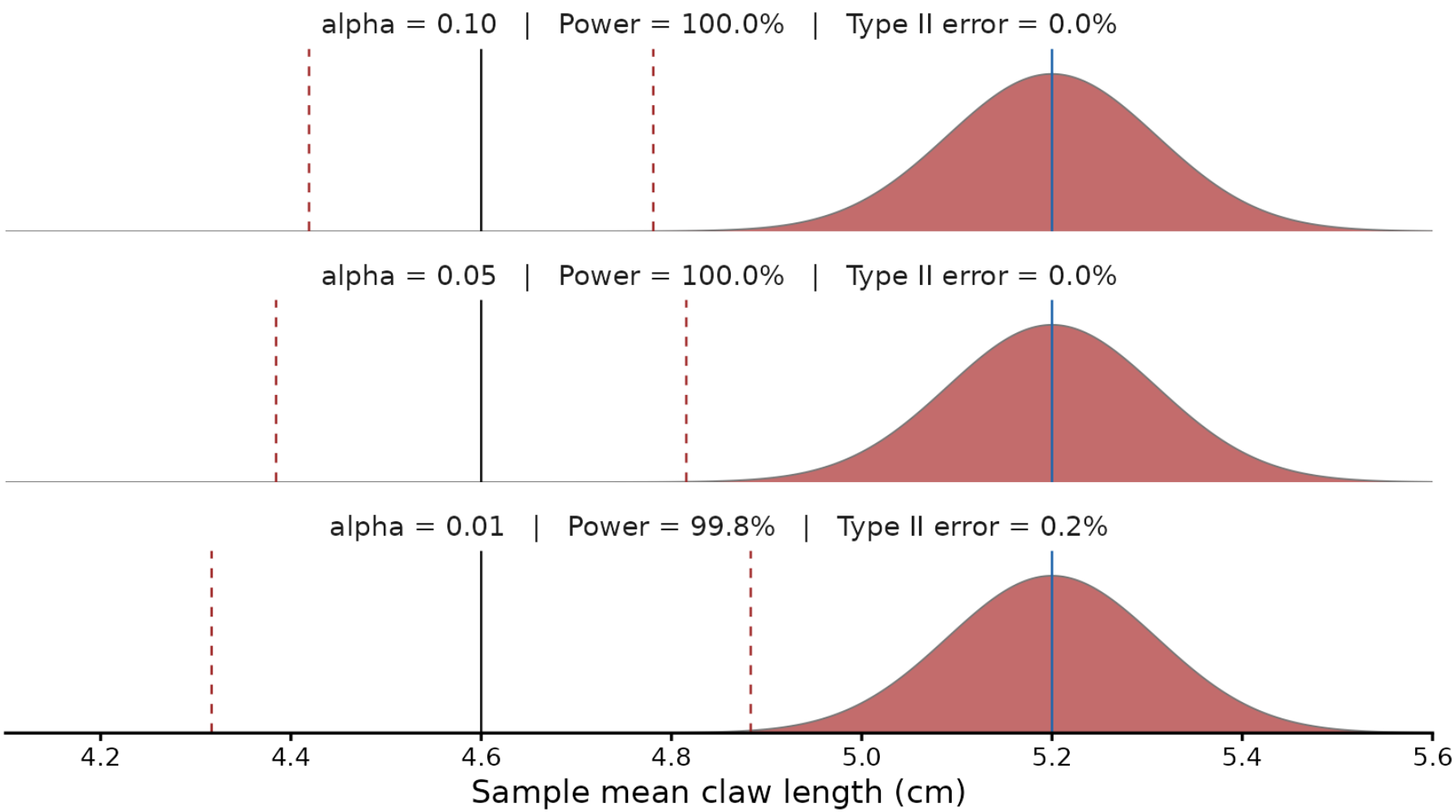
A Larger Effect: Less Sensitivity



True mean $= 5.00$ cm: effect $= 0.40$ cm. $n = 25$, $\sigma = 0.55$.

Power is higher throughout, but a stricter threshold still has a cost.

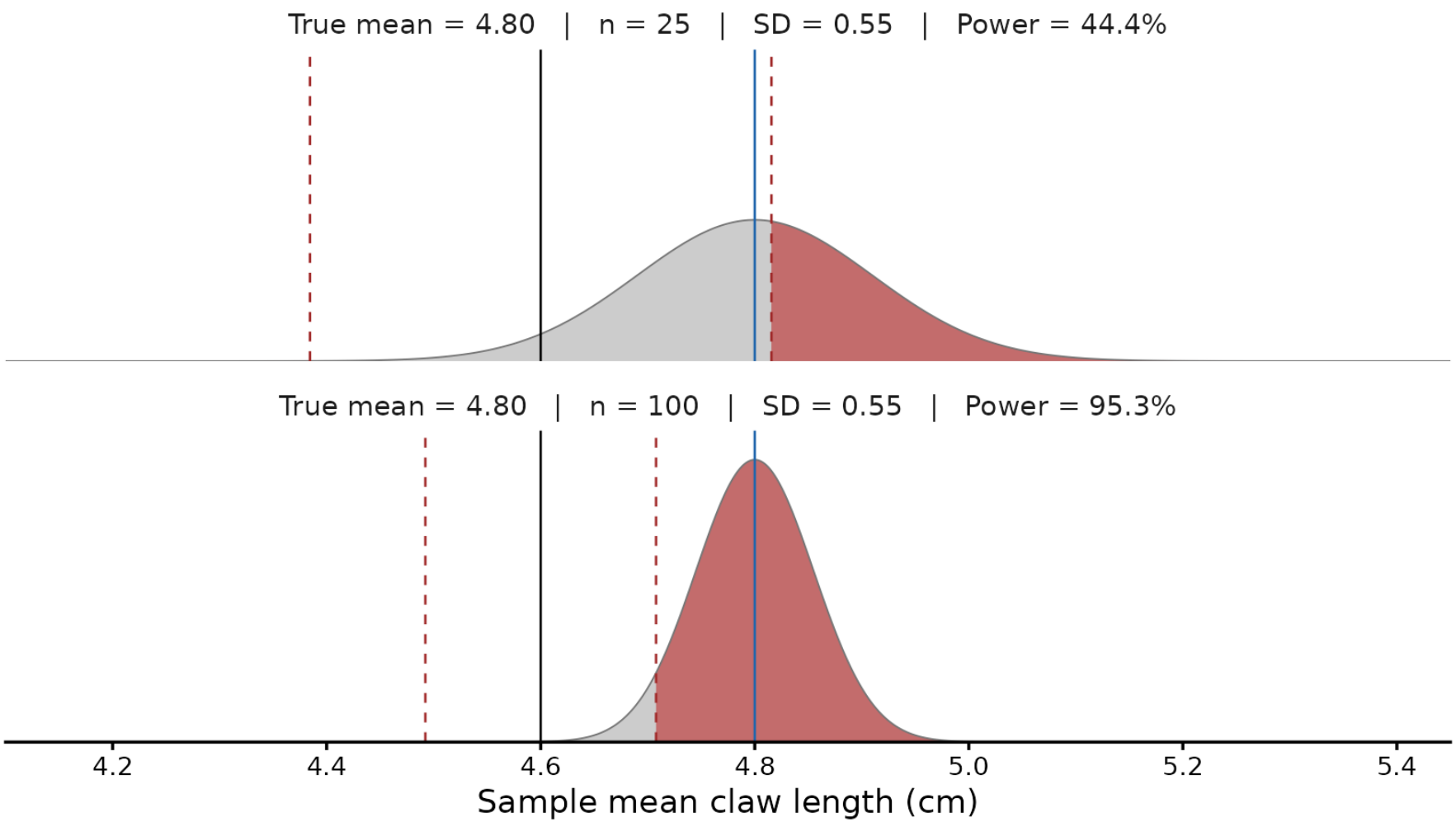
A Very Large Effect: Little Changes



True mean $= 5.20$ cm: effect $= 0.60$ cm. $n = 25$, $\sigma = 0.55$.

Nearly all sample means reject at every threshold shown.

Fix Alpha; Increase Sample Size



Keep $\alpha = .05$, the effect at 0.20 cm, and $\sigma = 0.55$ cm.

Larger n increases power and reduces β, while α stays .05.

Planning for Power

We cannot choose the true effect, but we can plan for an effect worth detecting.

1. Fix α to control the Type I error rate.
2. Identify an effect small enough to matter and estimate population variability.
3. Choose a sample size that gives adequate power for that effect.

Sample size is the quantity we can change to improve power without relaxing the standard for rejection.

Drawing Conclusions

As we never truly know whether H_0 is correct or not, we must simultaneously be prepared to combat both types of error

Test Result	True State of Nature	
	H_0 True	H_0 False
Fail to reject H_0	Correct $(1-\alpha)$	Type II Error (β)
Reject H_0	Type I Error (α)	Correct $(1-\beta)$

- Type I error $= P(\text{Reject } H_0 | H_0 \text{ true}) =$ false alarm
- Type II error $= P(\text{Fail to reject } H_0 | H_A \text{ true}) =$ missed opportunity