Hypothesis Tests
Part 3: Sample Size and Power

DASC 512

Risk-Based Experimental Design

Recall that we discussed Type I and Type II error earlier

We can design an experiment to balance the risks of each type of error.

		Truth	
		H_a False	H_a True
Researcher Concludes	H_a False	Correct	Type II Error
	H_a True	Type I Error	Correct

Risk-Based Experimental Design

- α is set to a value depending on the acceptable risk of a False Positive. We say we are $100(1-\alpha)\%$ confident in the result.
- β can also be set to a value depending on the acceptable risk of a False Negative specific to an alternative value of the parameter of interest.

		Truth	
		H_a False	H_a True
Researcher Concludes	H_a False	Correct $(p = 1-\alpha)$	Type II Error $(p = \beta)$
	H_a True	Type I Error $(p = \alpha)$	Correct $(p = 1-\beta)$

Consider…

- You run a dowel rod factory. You want to sample products from the line, measure them their diameter with calipers, and determine whether the mean diameter is different from 5mm. The variance is 1mm.
- Recalibrating the machines is time-consuming and expensive. You will only accept a 5% chance of unnecessarily doing that.
- However, selling faulty rods can be damaging to your reputation. You want to correctly identify a change greater than 1mm 95% of the time.
- How many rods must you sample?

Breaking down the problem

What we know:

$$\mu_0 = 5.0$$
$$\sigma = 1.0$$
$$\alpha = 0.05$$
$$\beta = 1 - 0.95 = 0.05$$
$$\delta_\mu = 1.0$$

Beta is the probability of failing to reject H_0 when H_a is true. Those hypotheses are:

$$H_0: \mu = 5.0$$
$$H_a: \mu \neq 5.0$$

Let’s think probabilistically

The sampling distribution is assumed to be

$$\bar{X} \sim \mu + \frac{\sigma}{\sqrt{n}} t(\nu = n - 1)$$

Let’s think probabilistically

The sampling distribution is assumed to be

$$\bar{X} \sim \mu + \frac{\sigma}{\sqrt{n}} t(\nu = n - 1)$$

The test statistic is

$$T = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

Let's think probabilistically

The sampling distribution is assumed to be

$$\bar{X} \sim \mu + \frac{\sigma}{\sqrt{n}} t(\nu = n - 1)$$

The test statistic is

$$T = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

The significance α represents

$$P\big((T < -t^*) \cup (T > t^*)\big)$$

Let's think probabilistically

The sampling distribution is assumed to be

$$\bar{X} \sim \mu + \frac{\sigma}{\sqrt{n}} t(\nu = n - 1)$$

The test statistic is

$$T = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

The significance α represents

$$P\big((T < -t^*) \cup (T > t^*)\big)$$

The chosen value of α defines the value of the critical value t^*.

Let’s think probabilistically

The alternative hypothesis represents an infinite number of possibilities. If the actual process mean is 5.000001mm, it’s very hard to tell. In this scenario, we are defining beta for a deviation of $\delta = \pm 1$mm.

Let’s think probabilistically

The alternative hypothesis represents an infinite number of possibilities. If the actual process mean is 5.000001mm, it’s very hard to tell. In this scenario, we are defining beta for a deviation of $\delta = \pm 1$mm.

Beta is then

$$\beta = P(-t^* < T + \delta < t^*) = P(-t^* - \delta < T < t^* - \delta)$$

Let's think probabilistically

The alternative hypothesis represents an infinite number of possibilities. If the actual process mean is 5.000001mm, it's very hard to tell. In this scenario, we are defining beta for a deviation of $\delta = \pm 1$mm.

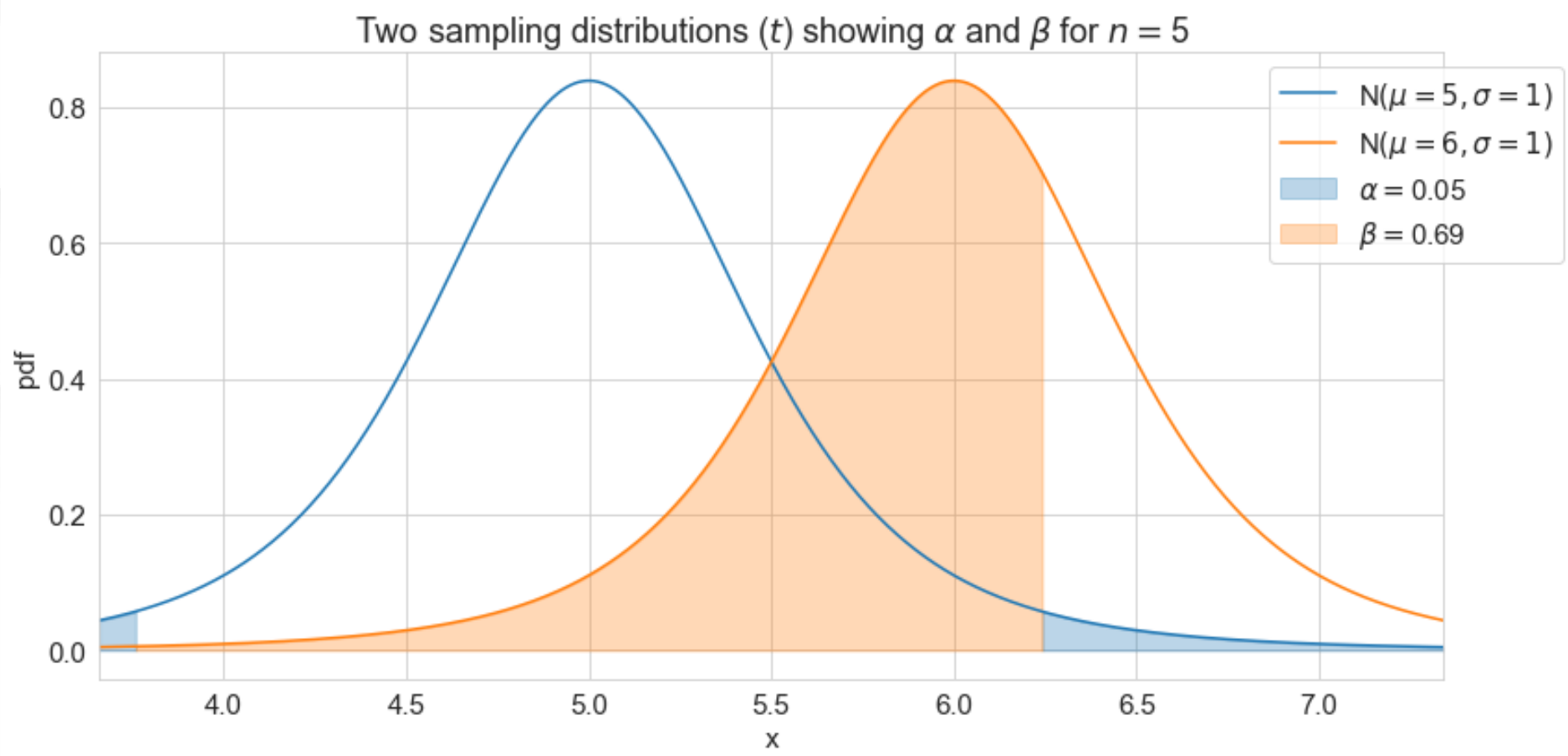
Beta is then

$$\beta = P(-t^* < T + \delta < t^*) = P(-t^* - \delta < T < t^* - \delta)$$

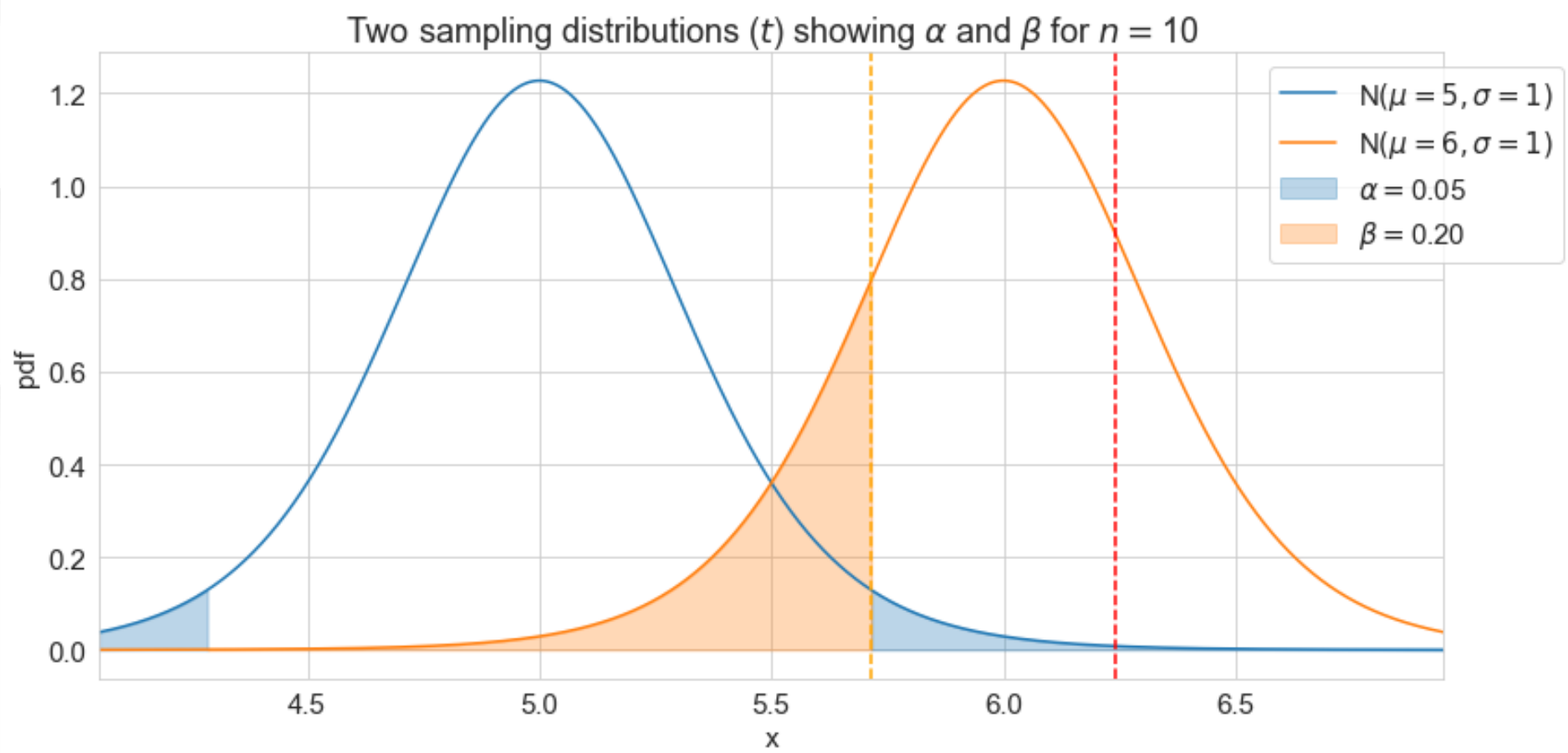
Or in functional terms

$$\beta = t_{CDF}(t^* - \delta) - t_{CDF}(-t^* - \delta)$$

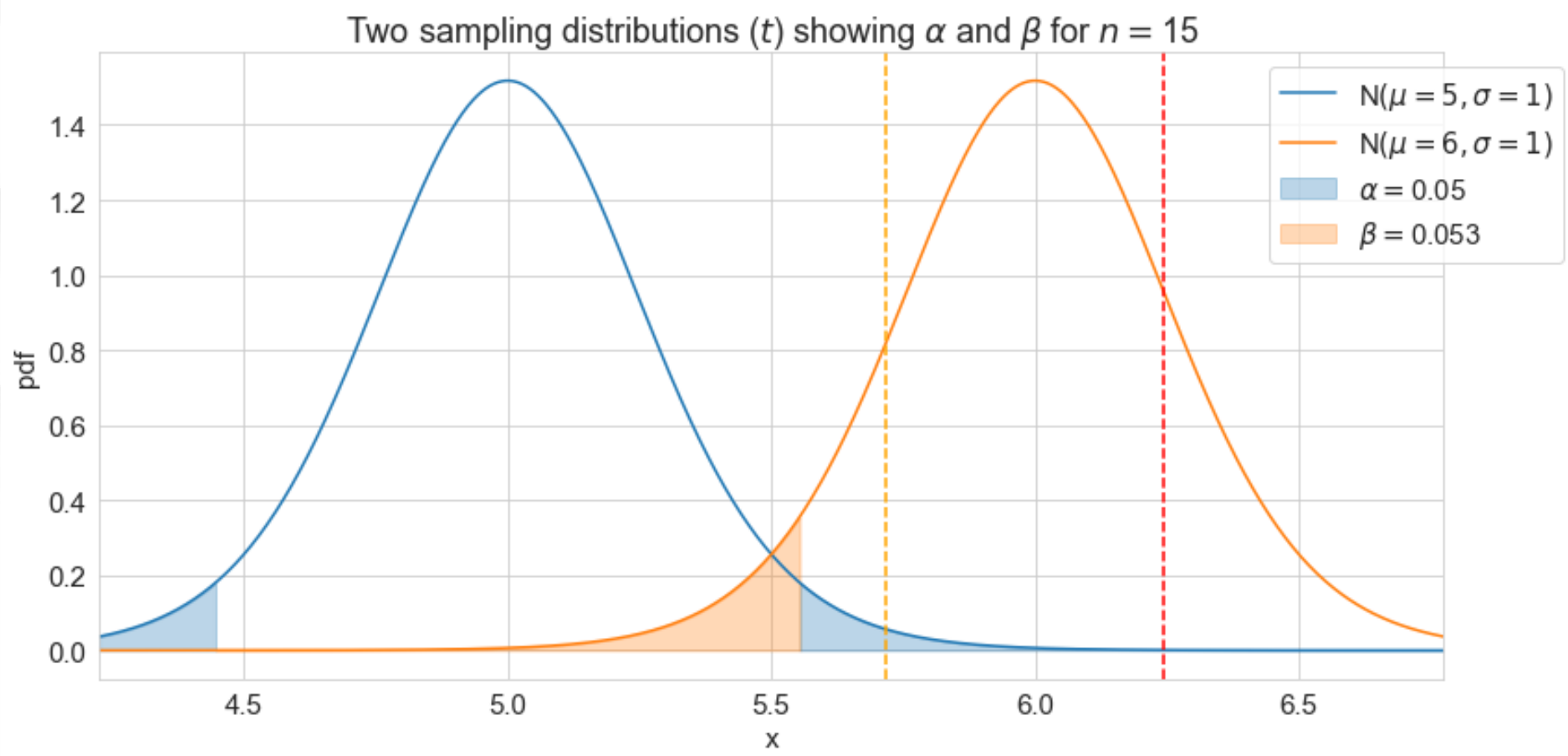
For a very small sample…



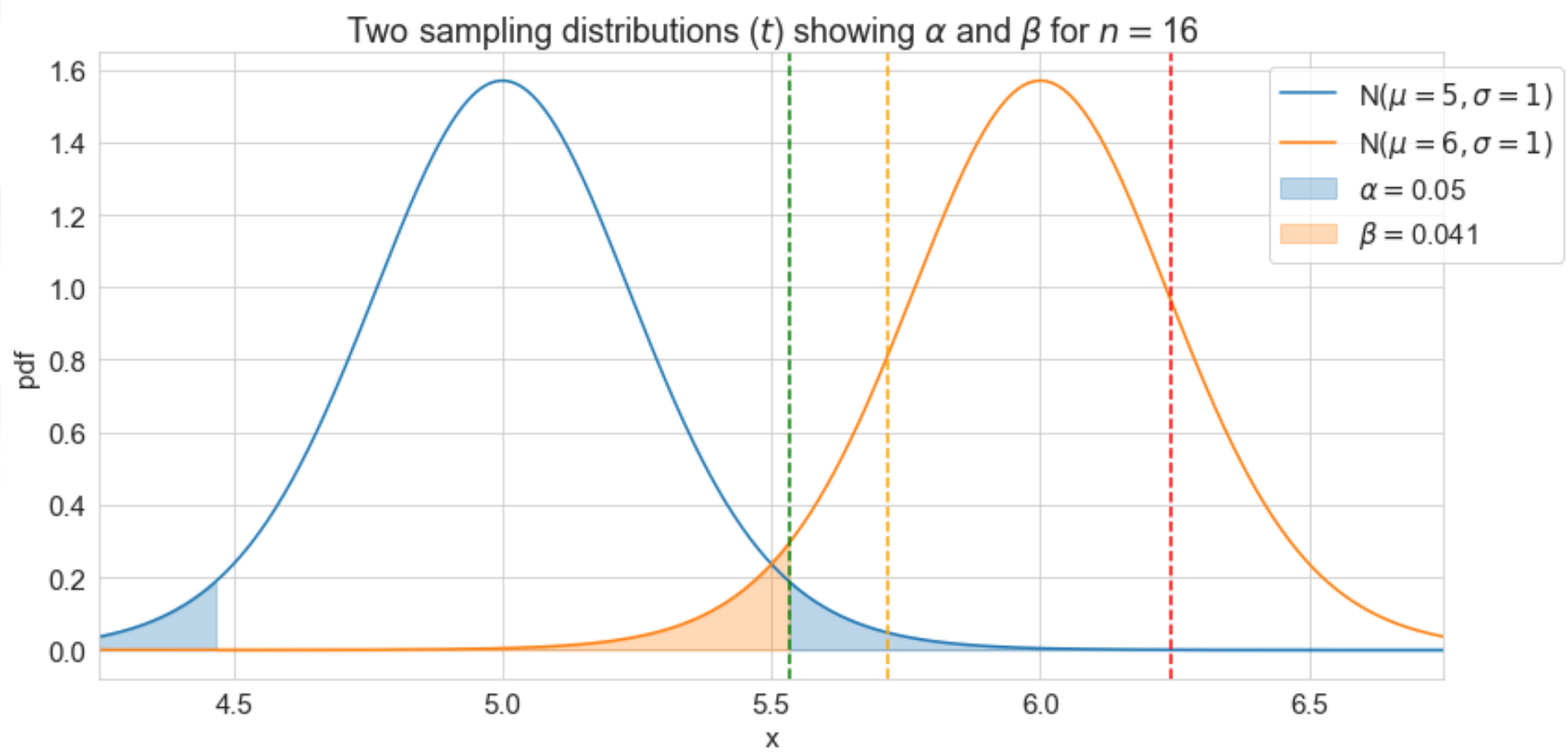
Doubling the sample size…



Almost there…



And now we have $\beta \leq 0.05$



Conclusion

You should sample 16 dowel rods to accomplish the task as specified.

So what is “power”?

Recall that α is significance and $(1 - \alpha)$ is confidence.

β doesn’t have a common name, but $(1 - \beta)$ is power.

Determining Sample Size

In the previous example, we iterated on n to find something that worked.

One method would be to assume a z-test and iterate up from that value

$$n = \left(z_{1-\alpha} + z_{1-\beta}\right)^2 \left(\frac{S}{\delta}\right)^2, \qquad \text{one-sided test}$$

$$n = \left(z_{1-\frac{\alpha}{2}} + z_{1-\beta}\right)^2 \left(\frac{S}{\delta}\right)^2, \qquad \text{two-sided test}$$

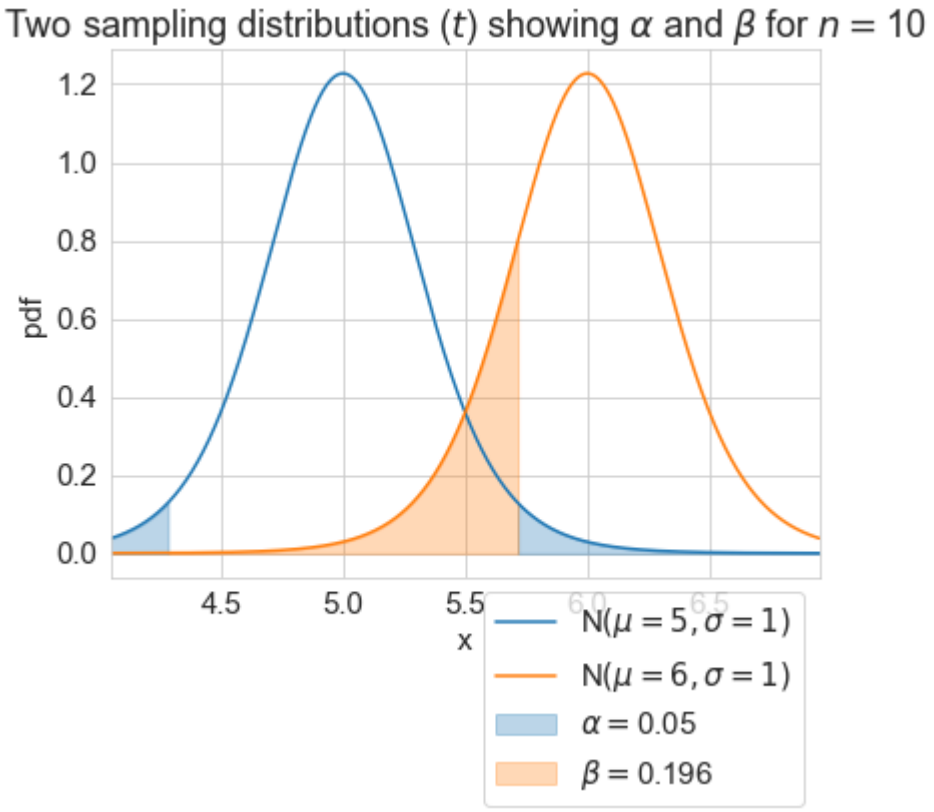
Determining Sample Size

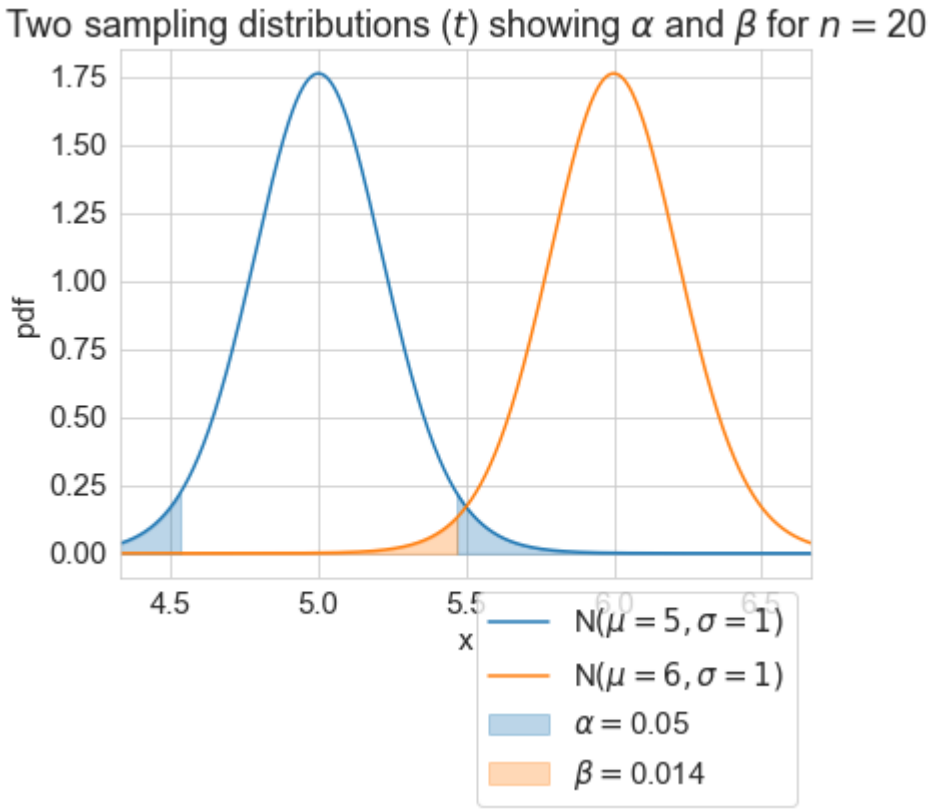
- Or… we can let Python do the hard part. Using statsmodels.stats we can specify three of four arguments and get the fourth as a response.
- Effect size (δ)
- Sample size (n)
- Significance (α)
- Power $(1 - \beta)$

How could I increase power?

How could I increase power?

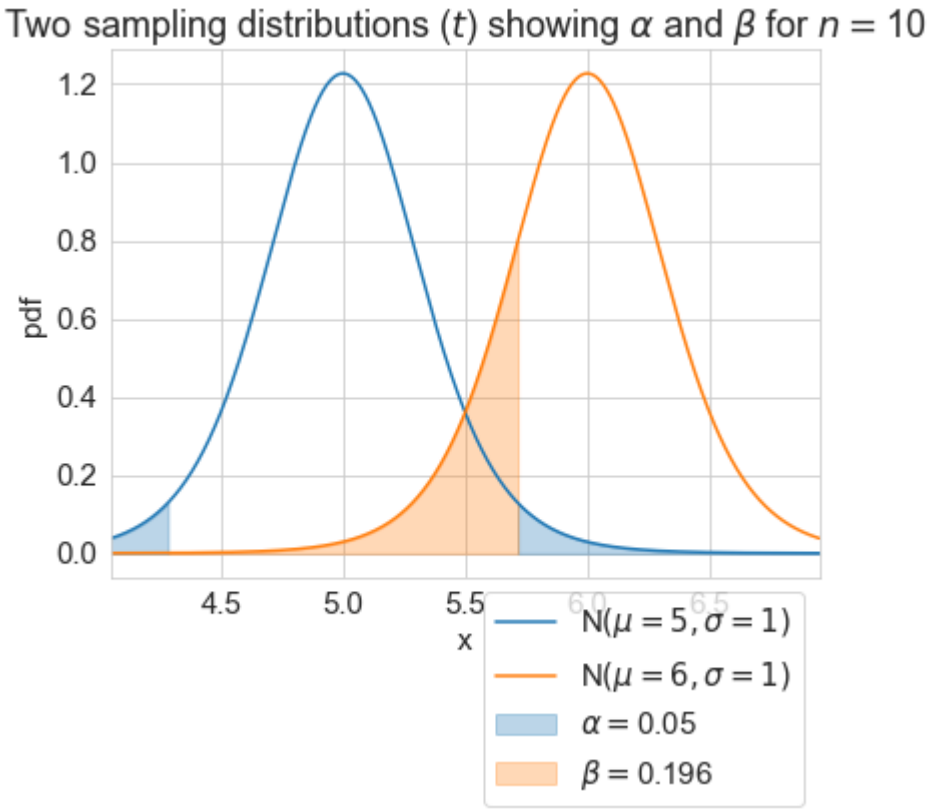
Increase Sample Size

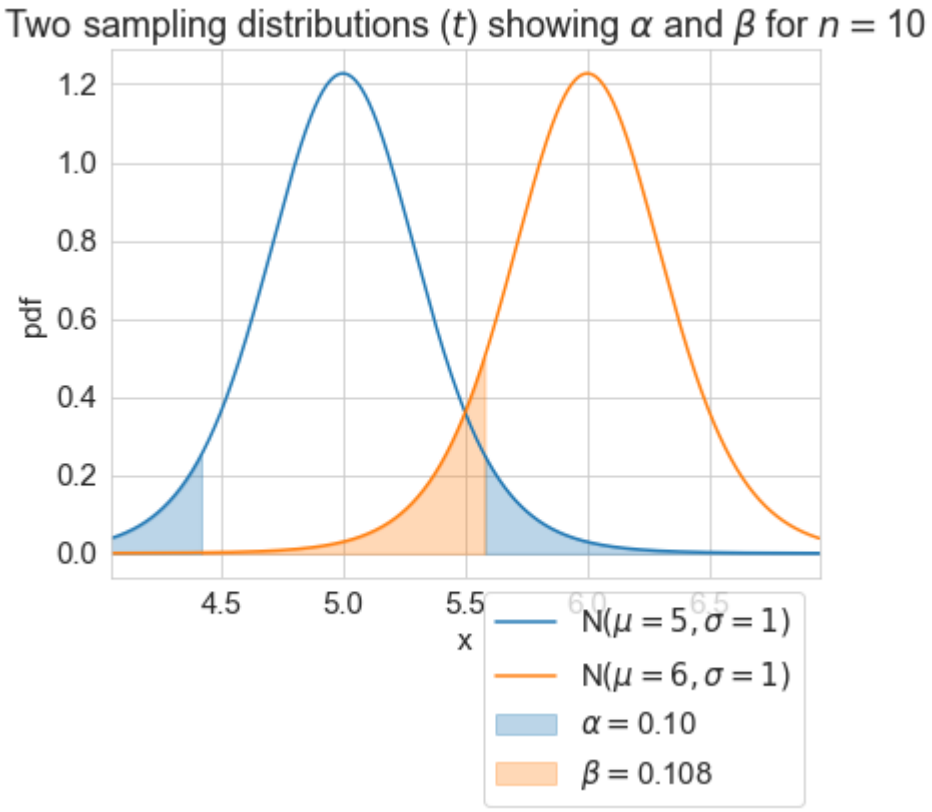




How could I increase power?

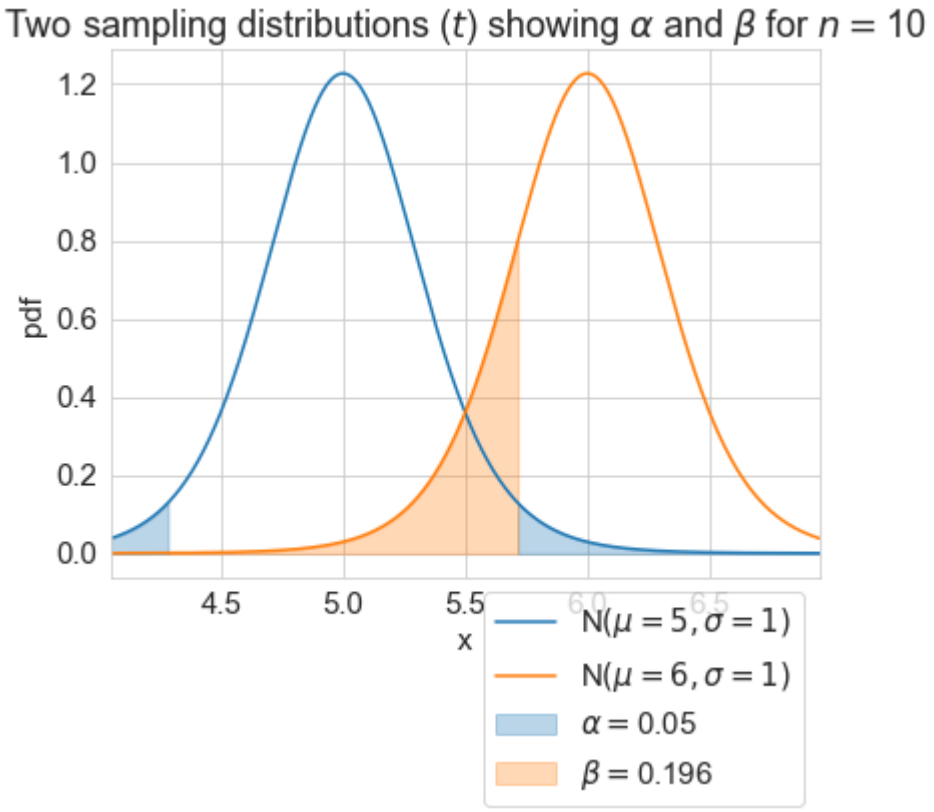
Decrease Confidence (Increase Significance)

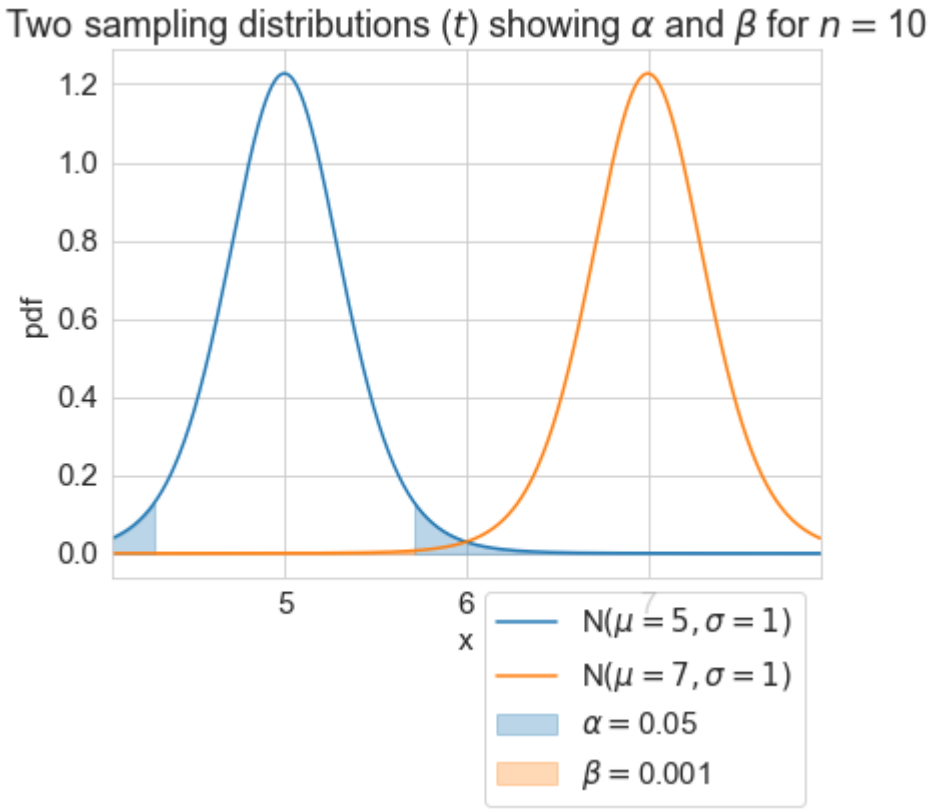




How could I increase power?

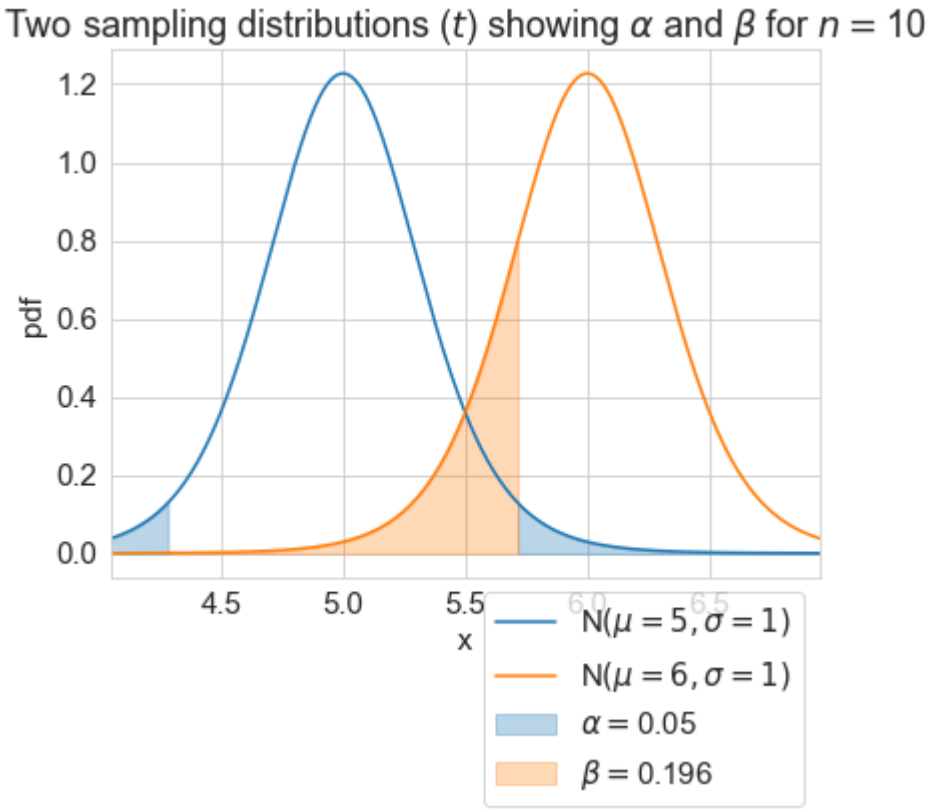
Increase Detectable Difference

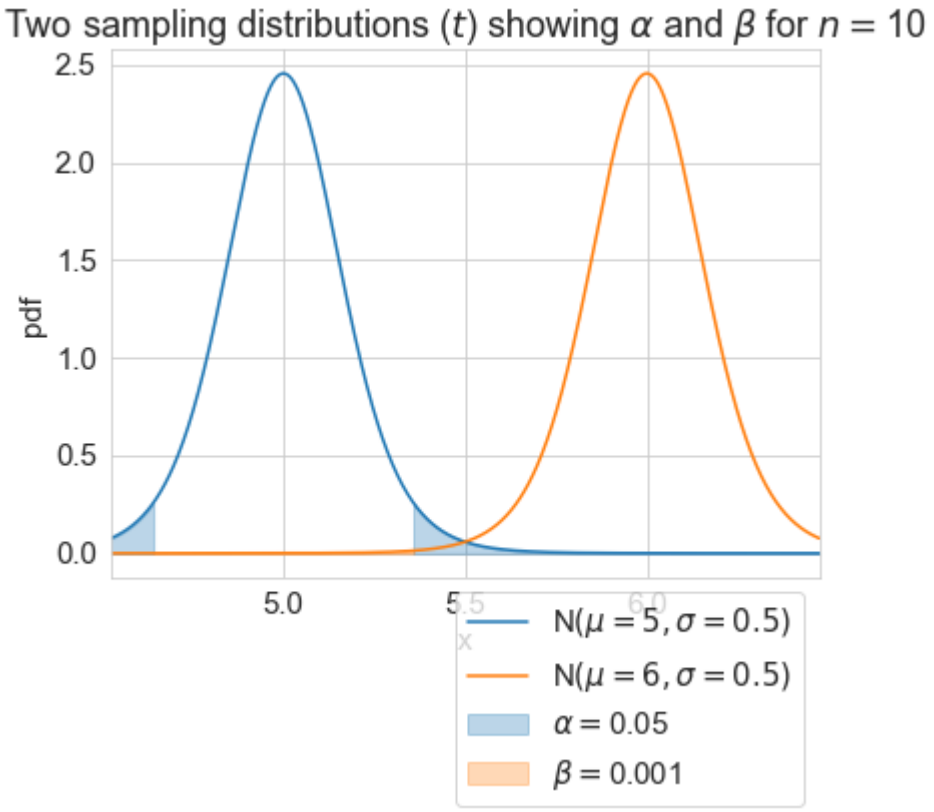




How could I increase power?

Reduce Process Variance





Recap

Power $(1 - \beta)$ is the probability of correctly rejecting H_0 when H_a is true for detectable difference δ

Power is a trade-off with:

- Sample Size
- Alpha
- Detectable Difference
- Variance

Python can make determining sample size easy for you.