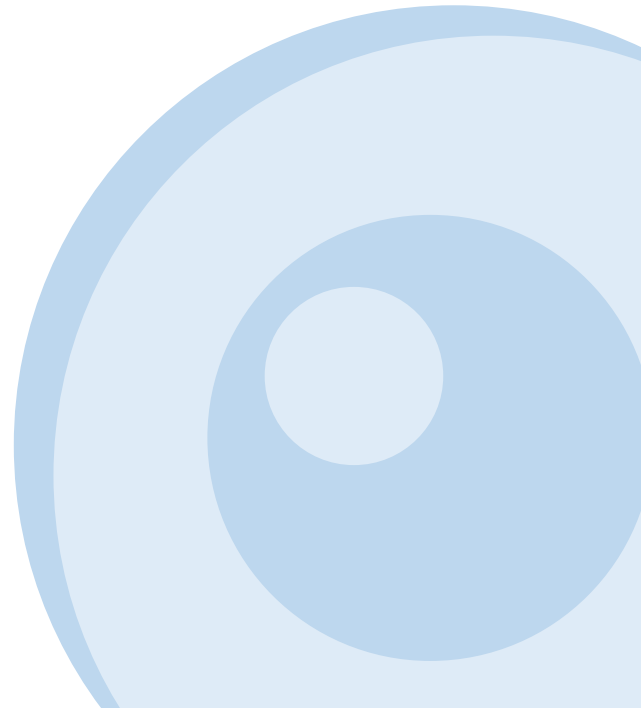


PSYC 193R:

Data Analysis & Visualization with R

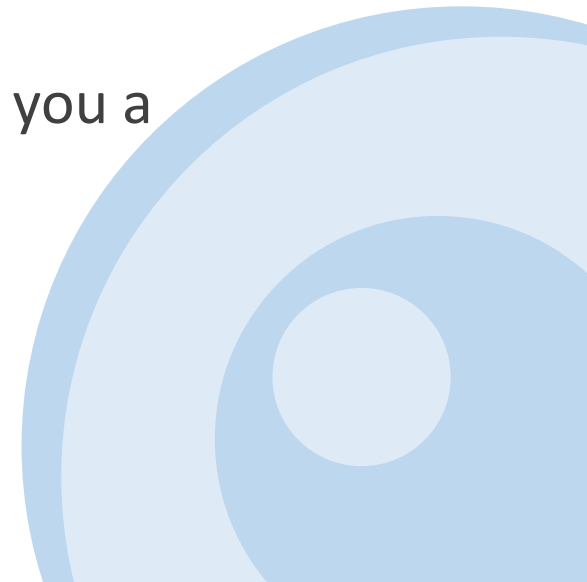
Lecture 5: Statistical Power

Jonas S.-H. Lau [silau -et- ucsd.edu]
psyc193r.ucsd.edu



Deciding # participants for a study

- Strategy 1:
pick a random number, e.g., $n = 100$, as long as the sample is representative of the population
- Strategy 2:
as long as $n \geq 30$
- Strategy 3:
get as many participants until your funding / time runs out
- Strategy 4:
run the statistical test repeatedly and add participants until it gives you a significant result
- Other Strategies:



Agenda

- Effect Size
- Confidence Interval
- Type I and II errors
- Statistical Power

If the experiment is repeated over and over, what is the range of sample means that we will get [95%] of the time



Scenario

- SAT is a standardized college entrance exam. It has a maximum score of 1600, a mean of 990 and standard deviation of 220. To test the effectiveness of “The Official SAT Practice”, The College Board (which administers SAT) teamed up with Khan Academy for a study. They recruited 49 high school students, providing them with 20 hours of free SAT practice. The group attained an average score of 1045 points. Test The College Board’s claims that “The Official SAT Practice” helps students for the SAT.
(with $\alpha = 0.05$, two-tailed test)

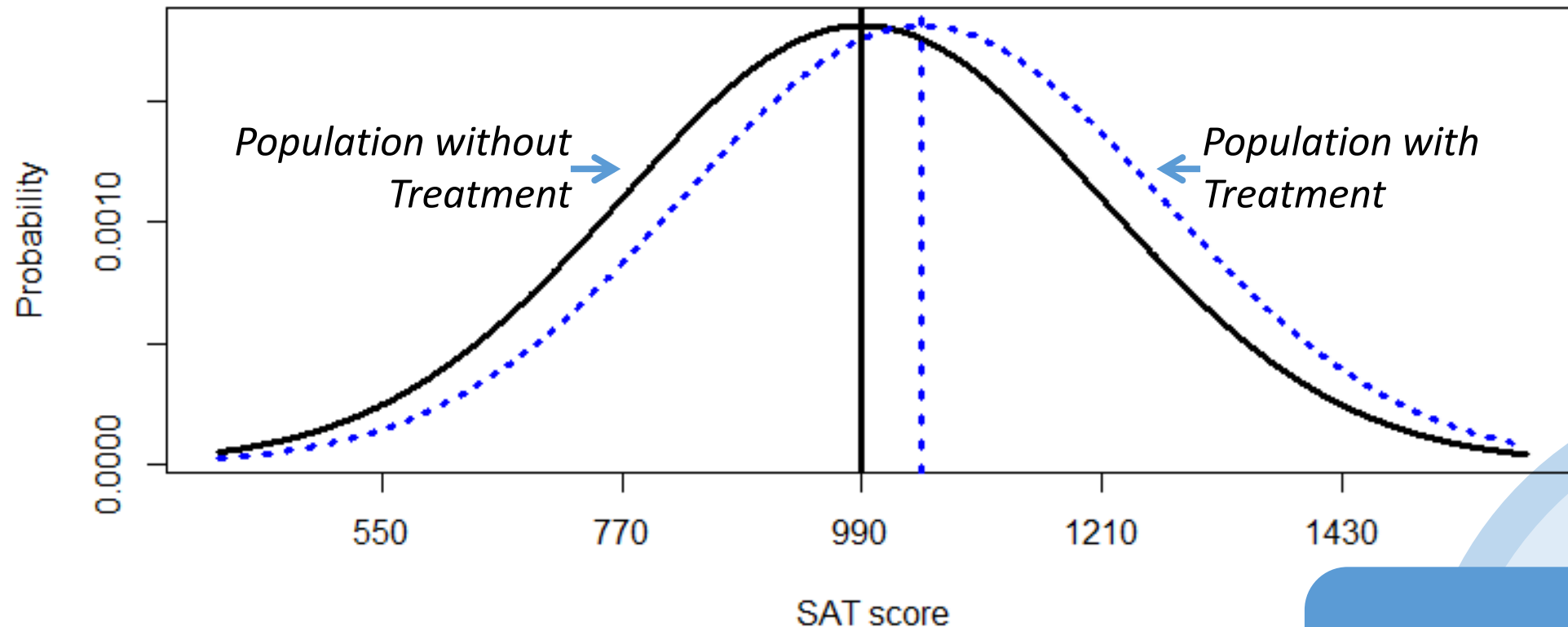
Effect size: Effect in Standard Deviations

- We want to describe the size of the effect with a unit-free measure
- So we don't have to know the scale of DV
- For example, what's the average improvement in SAT scores for an individual with 20 hours of the "The Official SAT Practice"

$$Cohen's\ d = \frac{|\mu_{treatment} - \mu_{baseline}|}{\sigma}$$

Cohen's d	Effect Size
0.2	Small
0.5	Medium
0.8	Large

Population distributions: Cohen's d



$$\text{cohens_d} = (1045 - 990) / 220$$

small / medium /
large effect?

Confidence Interval (C.I.): Reliability Measure

In an ideal world

- When we devise a study --
We are interested in the *population parameters* of the treated population (usually the mean)
- All high school students with “The Official SAT Practice”

$\mu_{treatment}$



Can never be known

In reality

- What we have is a *sample mean*

$M_{treatment}$



Not entirely accurate

Confidence Interval (C.I.): Reliability Measure

In an ideal world

- When we devise a study --
We are interested in the *population parameters* of the treated population (usually the mean)

$$\mu_{treatment}$$

In reality

- What we have is a *sample mean*

$$M_{treatment} \pm Z_{(1-\alpha/2)} \times \sigma_M$$

If the experiment is repeated over and over (same n and condition), for $(1 - \alpha)$ of the time, the sample means will fall within the interval

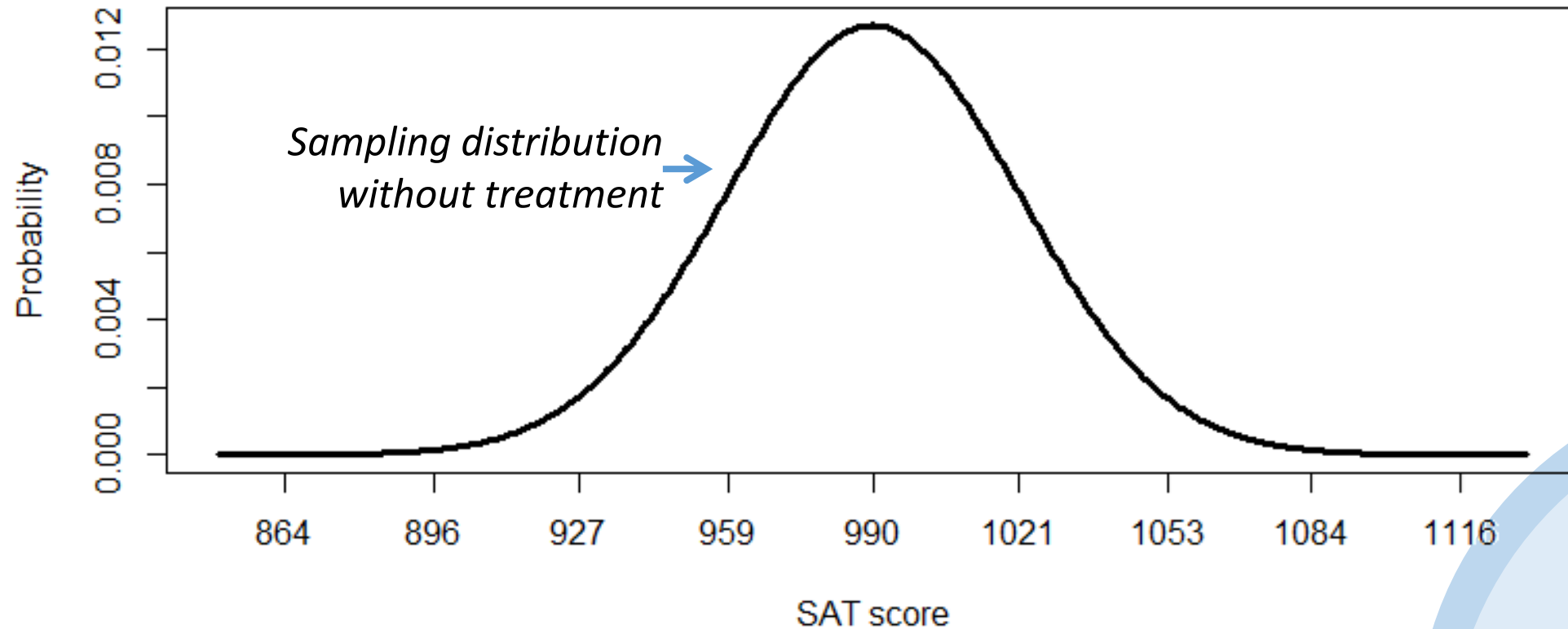
95% Confidence Interval

- SAT is a standardized college entrance exam. It has a maximum score of 1600, a mean of 990 and standard deviation of 220. To test the effectiveness of “The Official SAT Practice”, The College Board (which administers SAT) teamed up with Khan Academy for a study. They recruited 49 high school students, providing them with 20 hours of free SAT practice. The group attained an average score of 1045 points. Test The College Board’s claims that “The Official SAT Practice” helps students for the SAT.
(with $\alpha = 0.05$, two-tailed test)

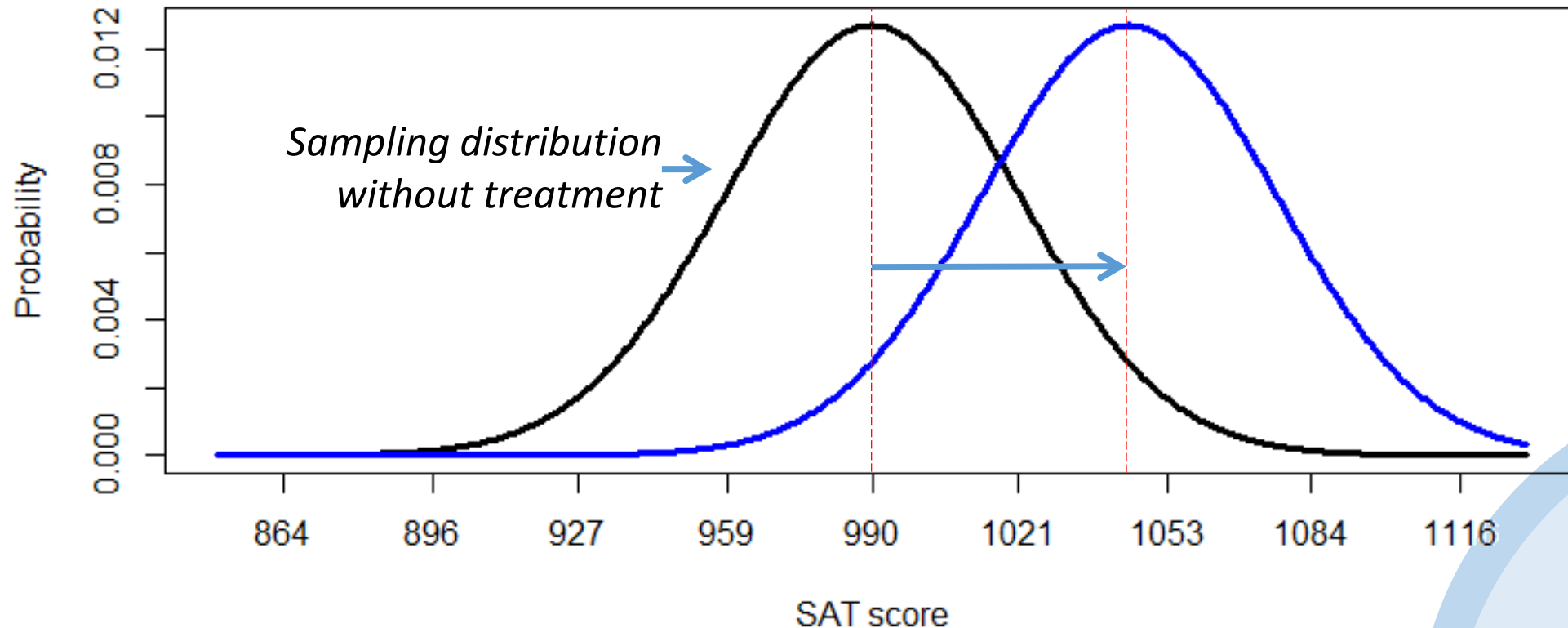
$$95\% \text{ C.I.} = M_{treatment} \pm Z_{(1-0.05/2)} \times \sigma_M$$

If the experiment is repeated over and over (same n and condition), for 95% of the time, the sample means will fall within the interval

Sampling distribution ($\mu_{\text{baseline}} = 990$, $\sigma_M = 220/\sqrt{49}$)



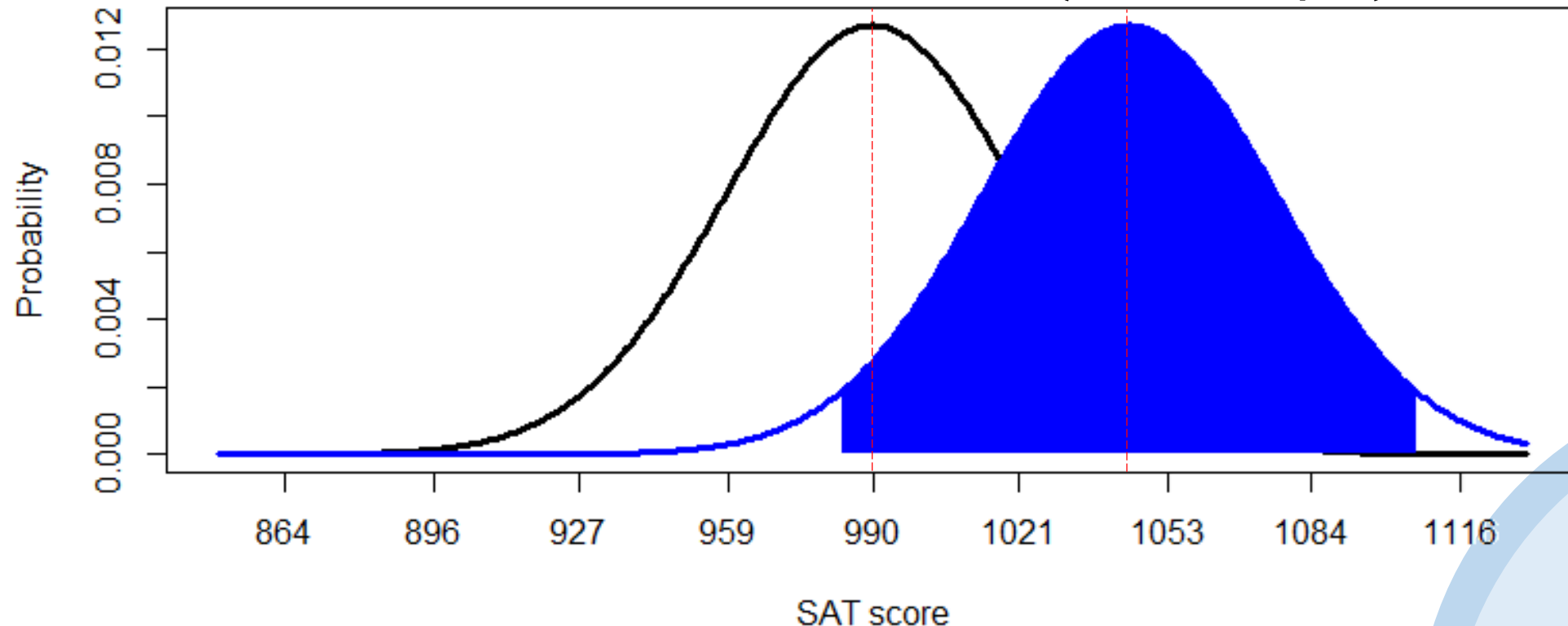
Sampling distribution ($\mu_{\text{baseline}} = 990$, $\sigma_M = 220/\sqrt{49}$)



We can assume that the mean shifts with treatment, but the dispersion does not change

95% Confidence Interval of the treatment means

$$95\% \text{ C.I.} = M_{\text{treatment}} \pm Z_{(1-0.05/2)} \times \sigma_M$$



If the experiment is repeated over and over (same n and condition), for 95% of the time, the sample means will fall within the interval

Type I error, Type II error, and Statistical Power

- Type I error (α):
Probability of rejecting H_0
when there is no treatment effect in reality (i.e., H_0 is true)
- Type II error (β):
Probability of *not* rejecting H_0
when there is *a real treatment effect* in reality (i.e., H_1 is true)
- Power:
Probability of rejecting H_0
when there is *a real treatment effect* in reality (i.e., H_1 is true)
In other words: power = $1 - \beta$

		Decision	
		Do not reject H_0	Reject H_0
Reality	H_0 is true	Correct decision	Type I error (α)
	H_1 is true	Type II error (β)	Power

Type I and Type II errors

- Type I error: Rejecting H_0 when it is true
- Type II error: Not rejecting H_0 when it is false

Do they make decision errors?

1. Patient A came to **Doctor D**, worrying that he had lung cancer. Doctor D looked at his x-ray photo, concluded that he was fine and discarded him. Patient A died 2 weeks later, autopsy confirmed that he died from lung cancer.
2. **Peter** was the TA for PSYC 1. In the first midterm, he suspected that Jonas cheated in the exam. It turned out that Jonas has great integrity (and he is the most honest student at UCSD). He would never cheat.

		Decision	
		Do not reject H_0	Reject H_0
Reality	H_0 is true	Correct decision	Type I error (α)
	H_1 is true	Type II error (β)	Power

Type I and Type II errors

- Type I error: Rejecting H_0 when it is true
- Type II error: Not rejecting H_0 when it is false

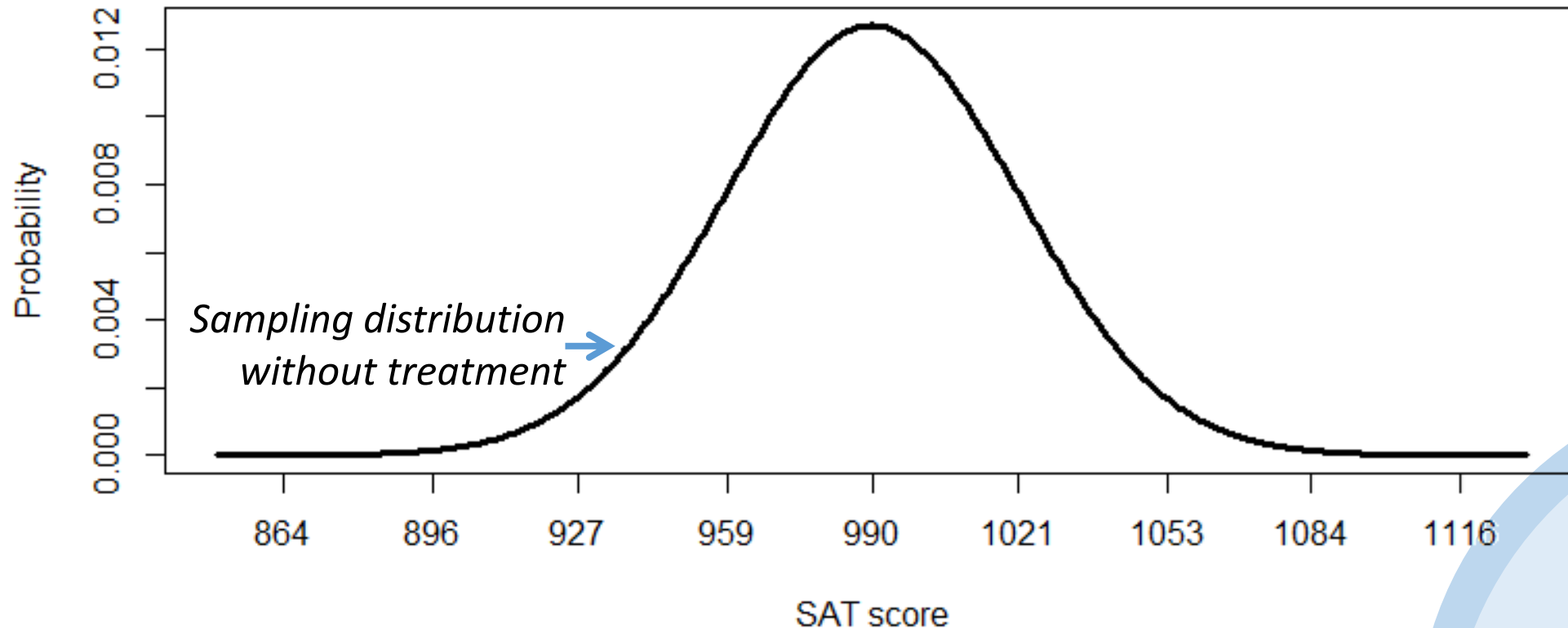
Do they make decision errors?

3. Patient B came to **Doctor D**, worrying that she had breast cancer. Doctor D looked at her mammogram, concluded that she had breast cancer, and ordered a biopsy. Biopsy confirmed that the lady is free from breast cancer.
4. Tim had so much fun at happy hours last Friday and he drank a lot. He was stopped by a **highway patrol officer** for a breath test. The machine was not calibrated well so it did not show that Tim had more than the law permits. The officer let Tim go.

		Decision	
		Do not reject H_0	Reject H_0
Reality	H_0 is true	Correct decision	Type I error (α)
	H_1 is true	Type II error (β)	Power

Type I error (α): Reality = No Treatment Effect

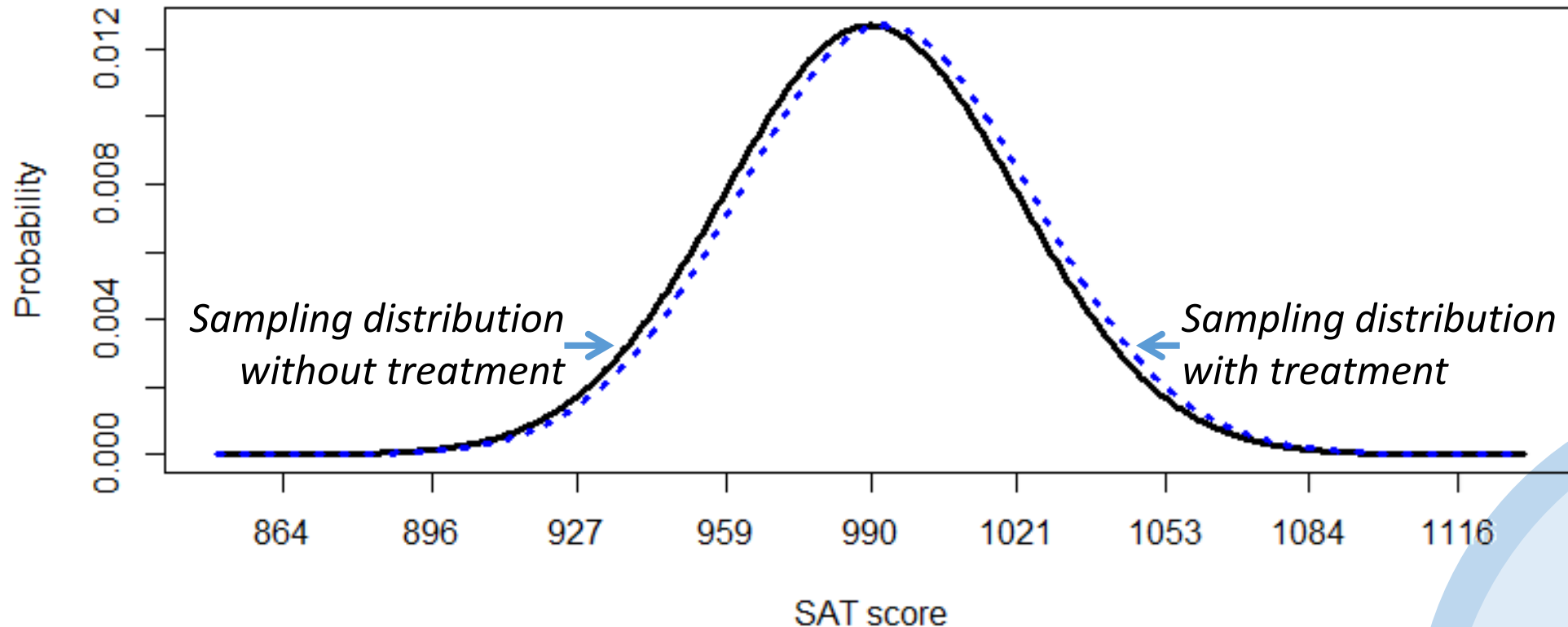
- Probability of rejecting H_0 when there is no treatment effect in reality (i.e., H_0 is true)



Sampling distribution (width of the curve denotes standard error)

Type I error (α): Reality = No Treatment Effect

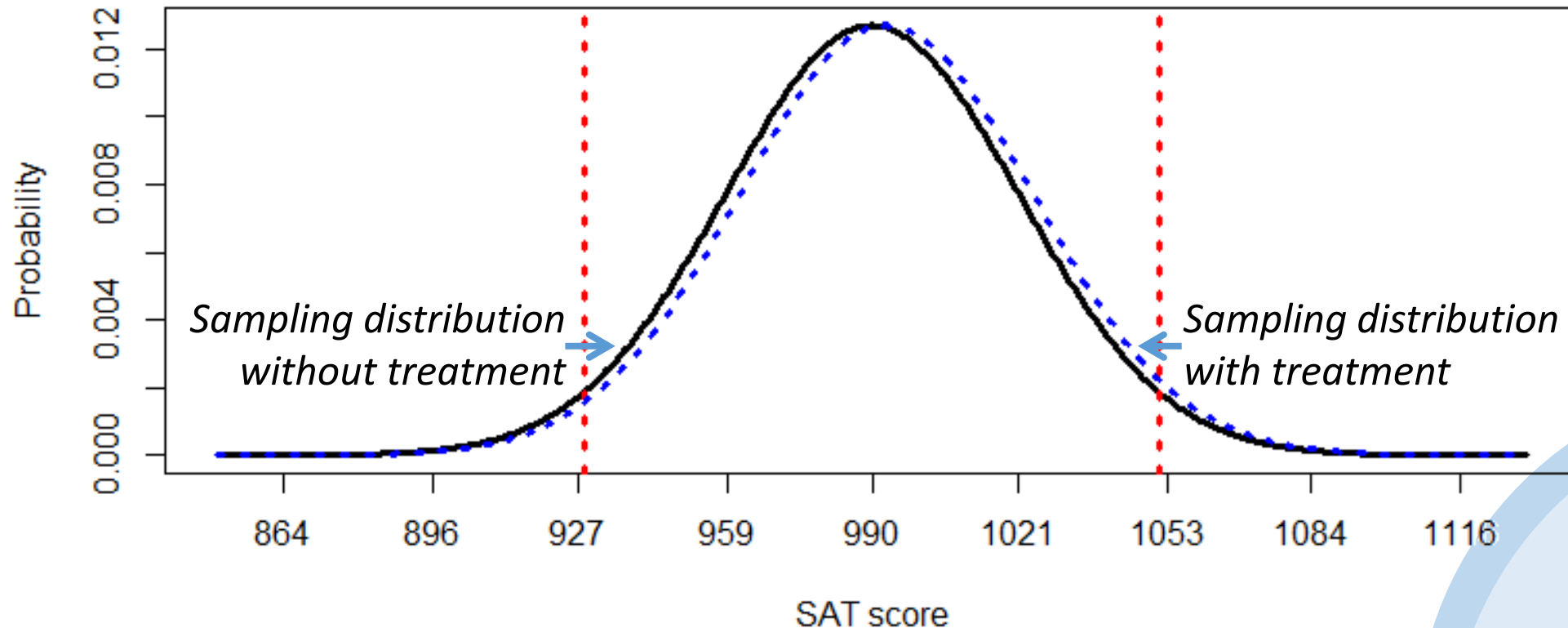
- Probability of rejecting H_0 when there is no treatment effect in reality (i.e., H_0 is true)



Sampling distributions (width of the curves denote standard errors)

Type I error (α): Reality = No Treatment Effect

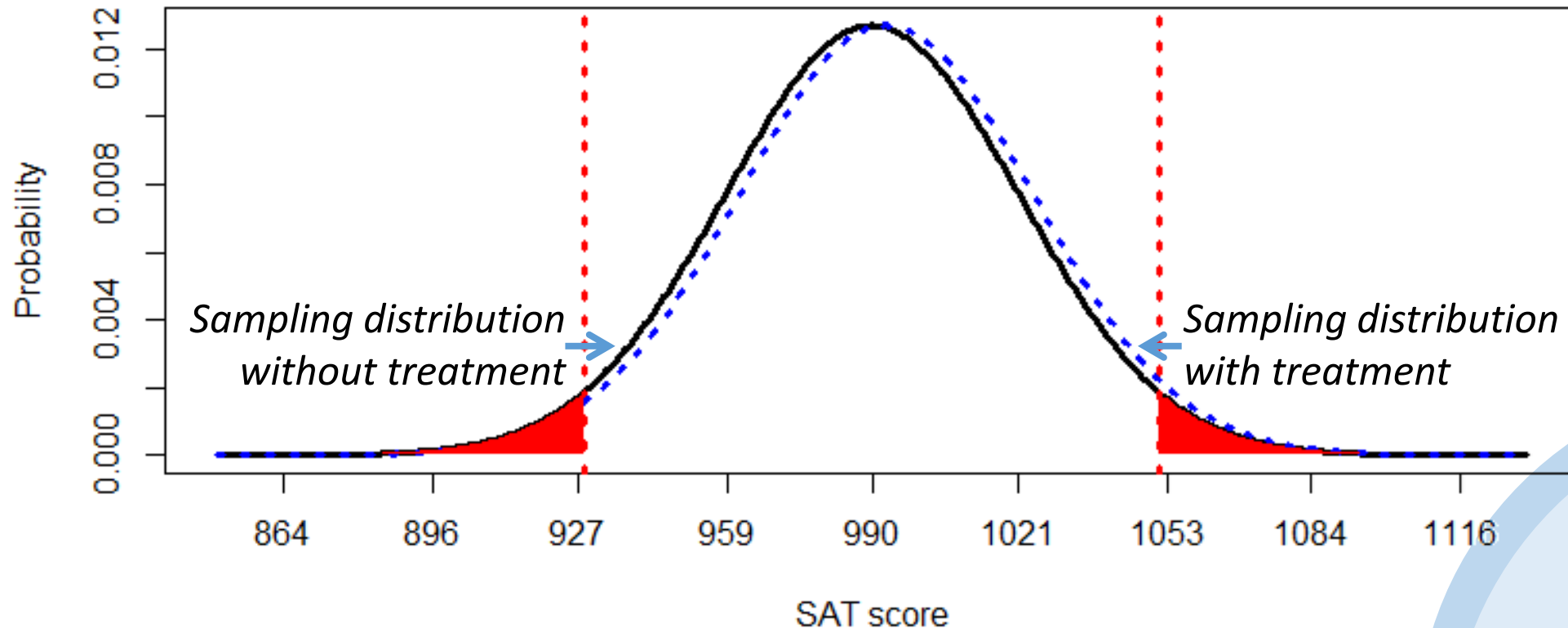
- Probability of rejecting H_0 when there is no treatment effect in reality (i.e., H_0 is true)



Sampling distributions (width of the curves denote standard errors)

Type I error (α): Reality = No Treatment Effect

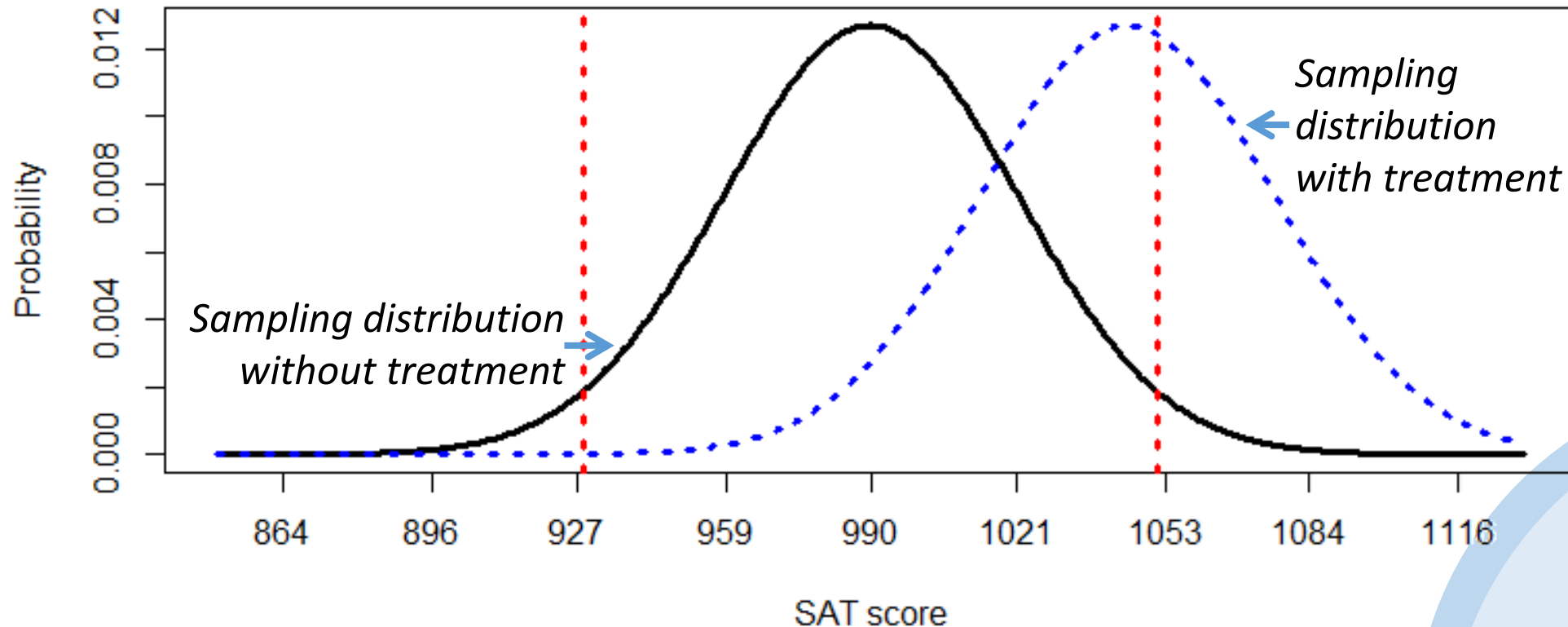
- Probability of rejecting H_0 when there is no treatment effect in reality (i.e., H_0 is true)



Sampling distributions (width of the curves denote standard errors)

Type II error (β): Reality = Real Treatment Effect

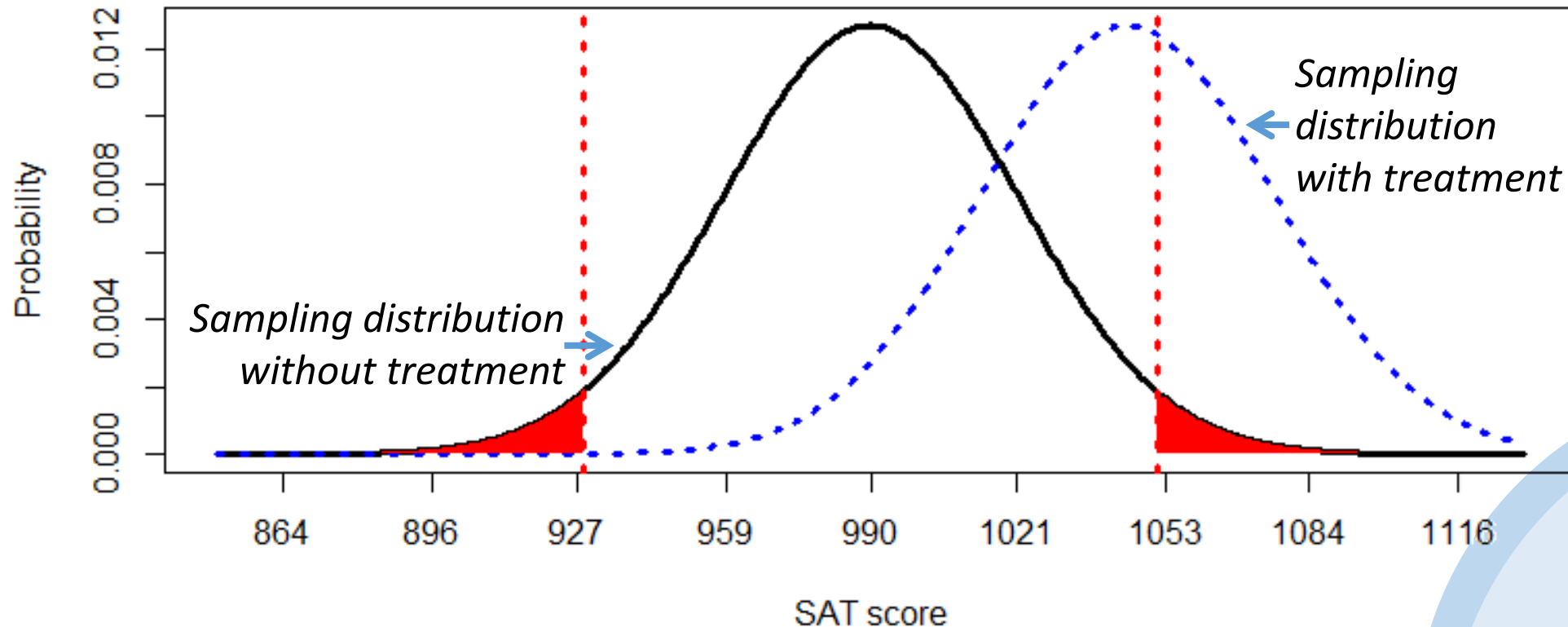
- Probability of not rejecting H_0 when there is a treatment effect in reality (i.e., H_1 is true)



Sampling distributions (width of the curves denote standard errors)

Type II error (β): Reality = Real Treatment Effect

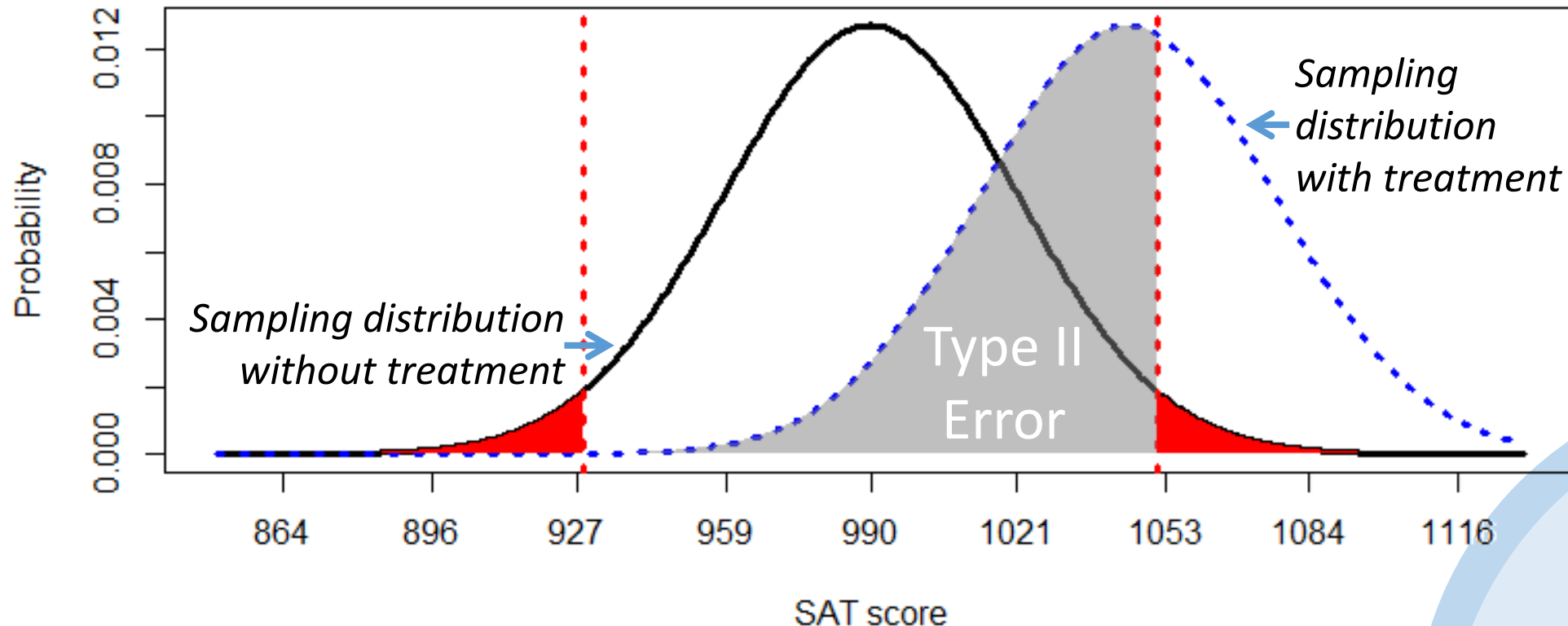
- Probability of not rejecting H_0 when there is a treatment effect in reality (i.e., H_1 is true)



Sampling distributions (width of the curves denote standard errors)

Type II error (β): Reality = Real Treatment Effect

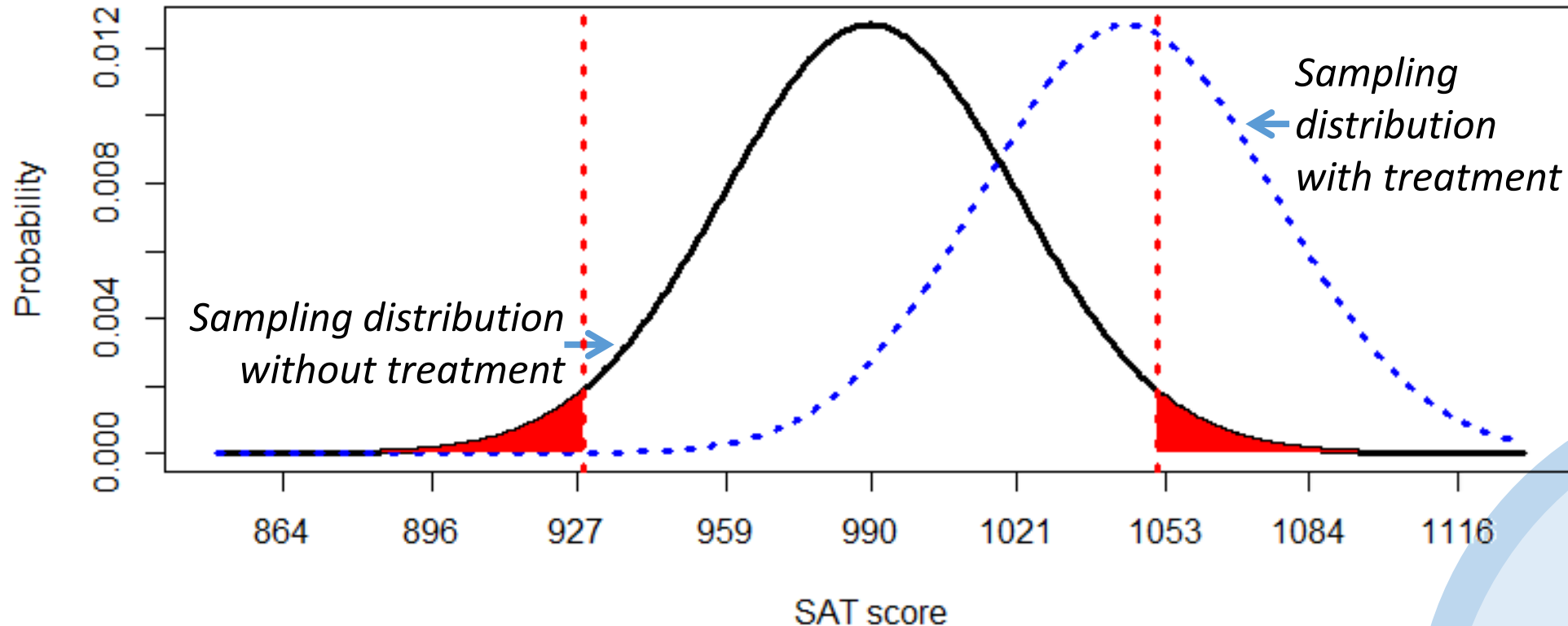
- Probability of not rejecting H_0 when there is a treatment effect in reality (i.e., H_1 is true)



Sampling distributions (width of the curves denote standard errors)

Statistical Power: Reality = Real Treatment Effect

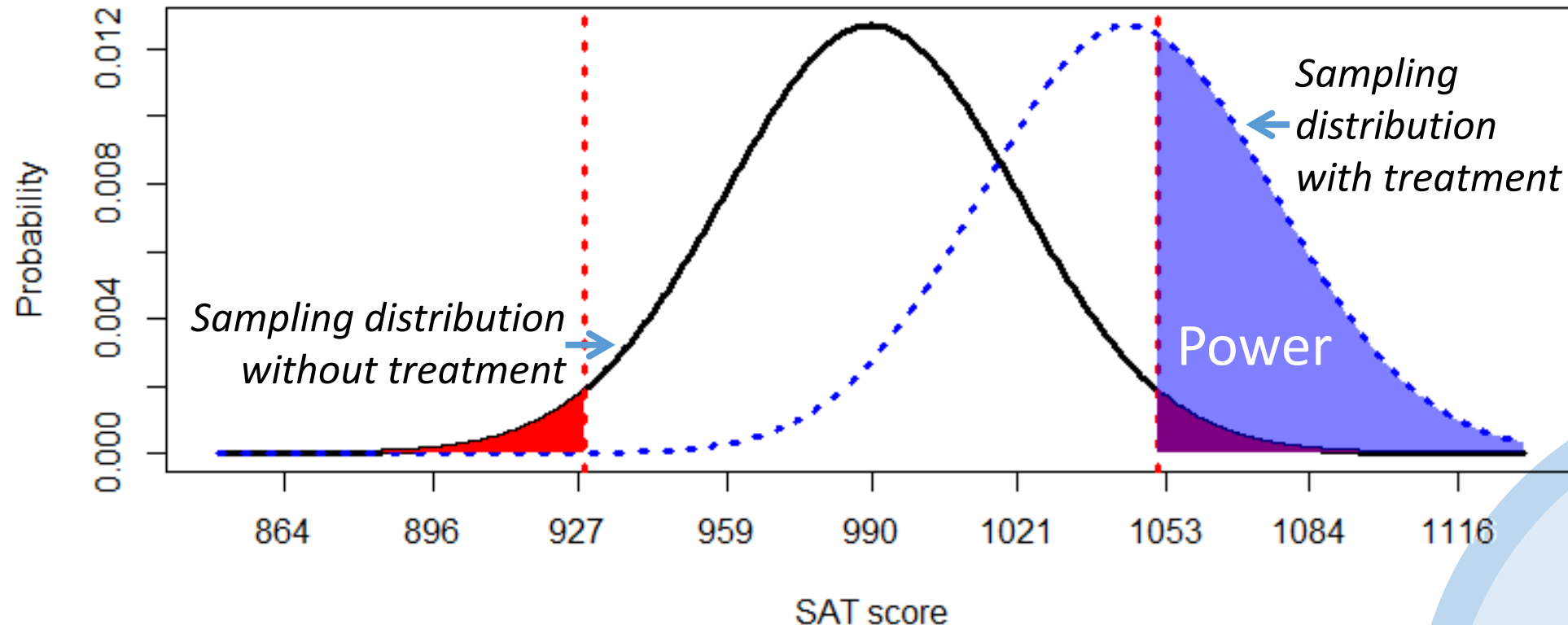
- Probability of rejecting H_0 when there is a treatment effect in reality (i.e., H_1 is true)



Sampling distributions (width of the curves denote standard errors)

Statistical Power: Reality = Real Treatment Effect

- Probability of rejecting H_0 when there is a treatment effect in reality (i.e., H_1 is true)



Sampling distributions (width of the curves denote standard errors)

So how many participants do we need for a study?

- Strategy 1:
pick a random number, e.g., $n = 100$, as long as the sample is representative of the population
- Strategy 2:
as long as $n \geq 30$
- Strategy 3:
get as many participants until your funding / time runs out
- Strategy 4:
run the statistical test repeatedly and add participants until it gives you a significant result
- Strategy 5:
do a power analysis to figure out the number of participants you need prior to running your study / experiment

Statistical Power: Why do we care?

- You want to estimate the probability of getting a significant result *before* you perform the study
 - If there is little chance to show a treatment effect (usually $> 80\%$ power)
 - You don't invest the money, time and other resources
- You want to estimate the number of participants you need to get a significant result
 - Data collection is expensive:
 - Time (investigators, researchers, participants)
 - \$ (researchers, participants)
 - Effects on participants (ineffective clinical trial, loss of other treatment opportunities)

In general, power is affected by all these factors:

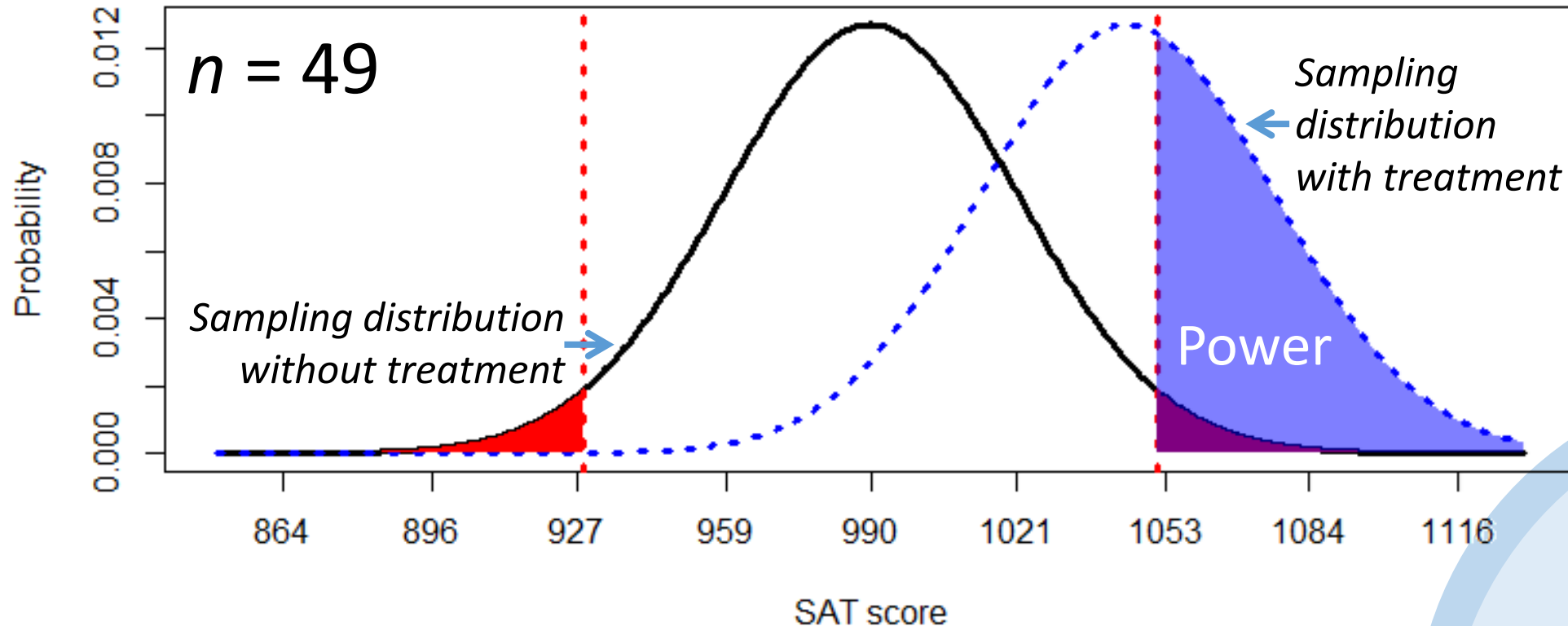
- Baseline population mean
- Baseline population standard deviation
- Sample mean with treatment
- Sample size
- α level

Often, these factors are unknown prior to the study, but you can estimate / guesstimate them:

- Sample mean with treatment
 - Population standard deviation
- in the literature or with a pilot study

Statistical Power: Effect of Sample Size

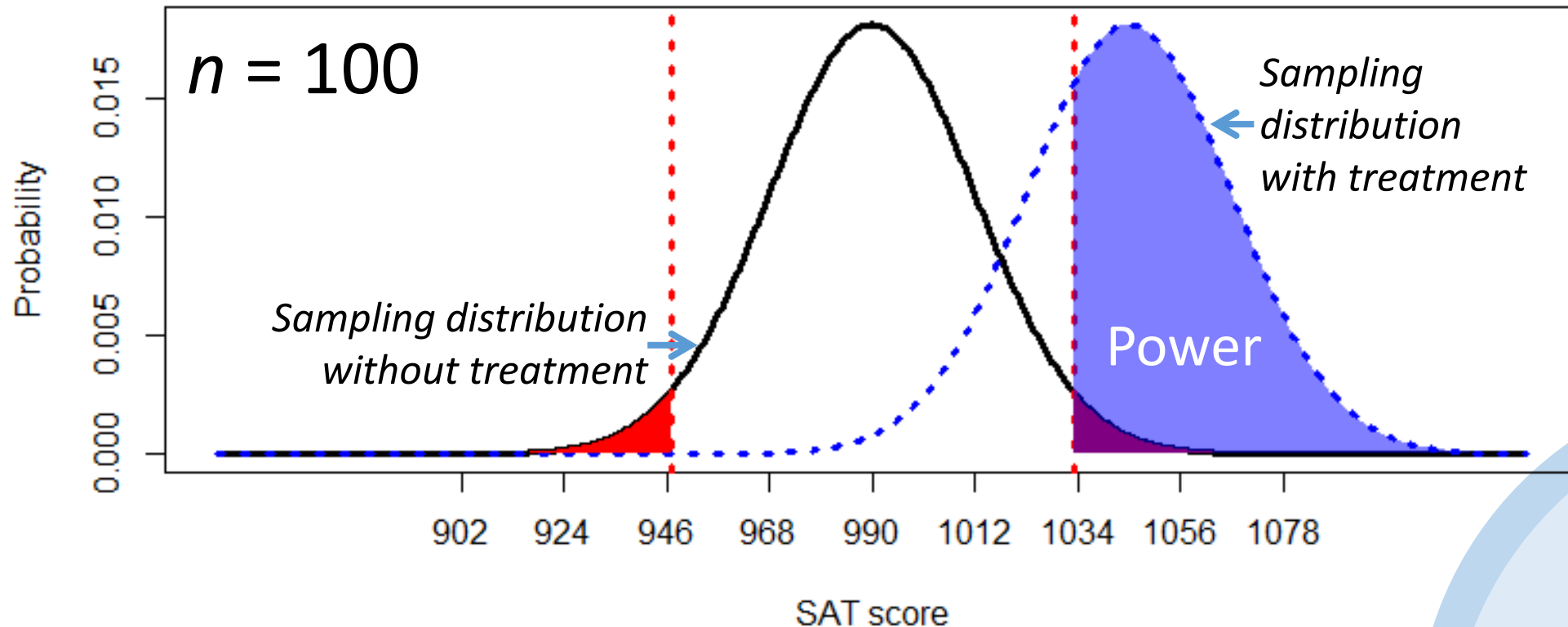
- Probability of rejecting H_0 when there is a treatment effect in reality (i.e., H_1 is true)



Sampling distributions (width of the curves denote standard errors)

Statistical Power: Reality = Real Treatment Effect

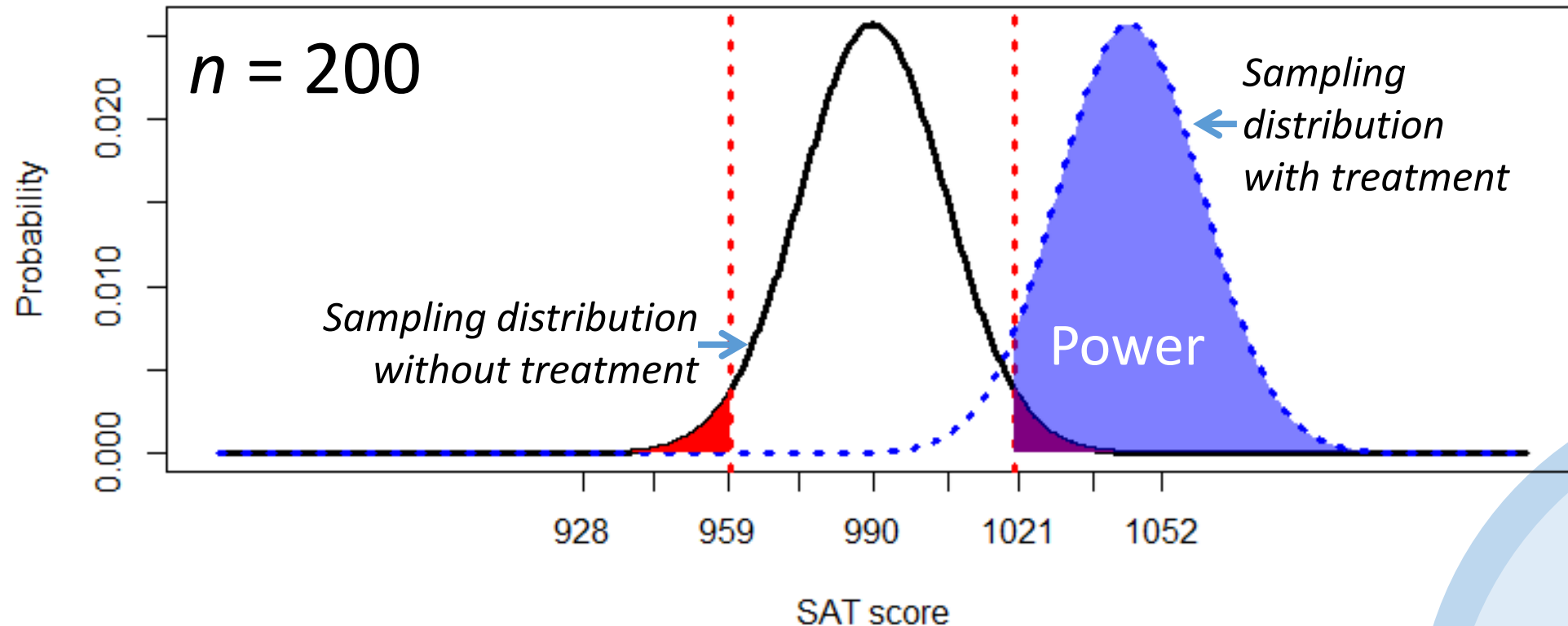
- Probability of rejecting H_0 when there is a treatment effect in reality (i.e., H_1 is true)



Sampling distributions (width of the curves denote standard errors)

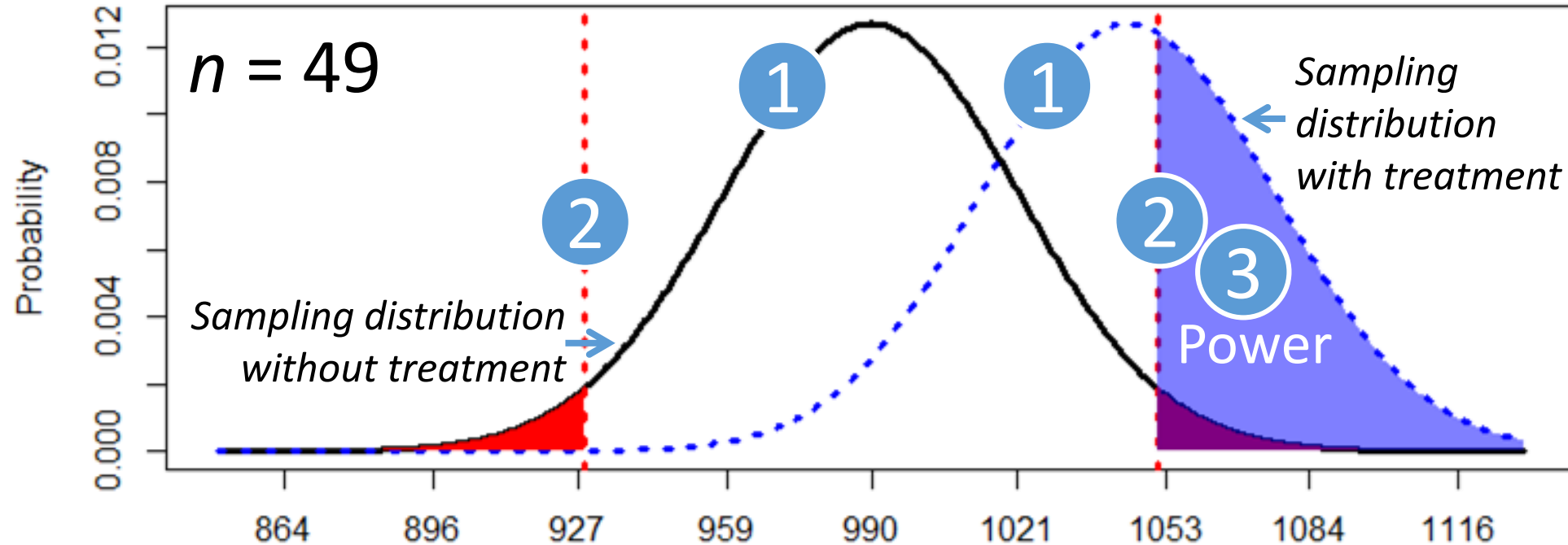
Statistical Power: Reality = Real Treatment Effect

- Probability of rejecting H_0 when there is a treatment effect in reality (i.e., H_1 is true)



Sampling distributions (width of the curves denote standard errors)

Calculating Statistical Power: Overview



Step 1:

Find the standard error or the sampling distributions

Step 2:

Find the cutoff points in raw scores (w.r.t H_0 distribution)

Step 3:

Find the probability of getting a sample mean more extreme than the cutoff scores under H_1 (w.r.t. H_1 distribution)

Sampling distributions (width of the curves denote standard errors)

Calculating Statistical Power

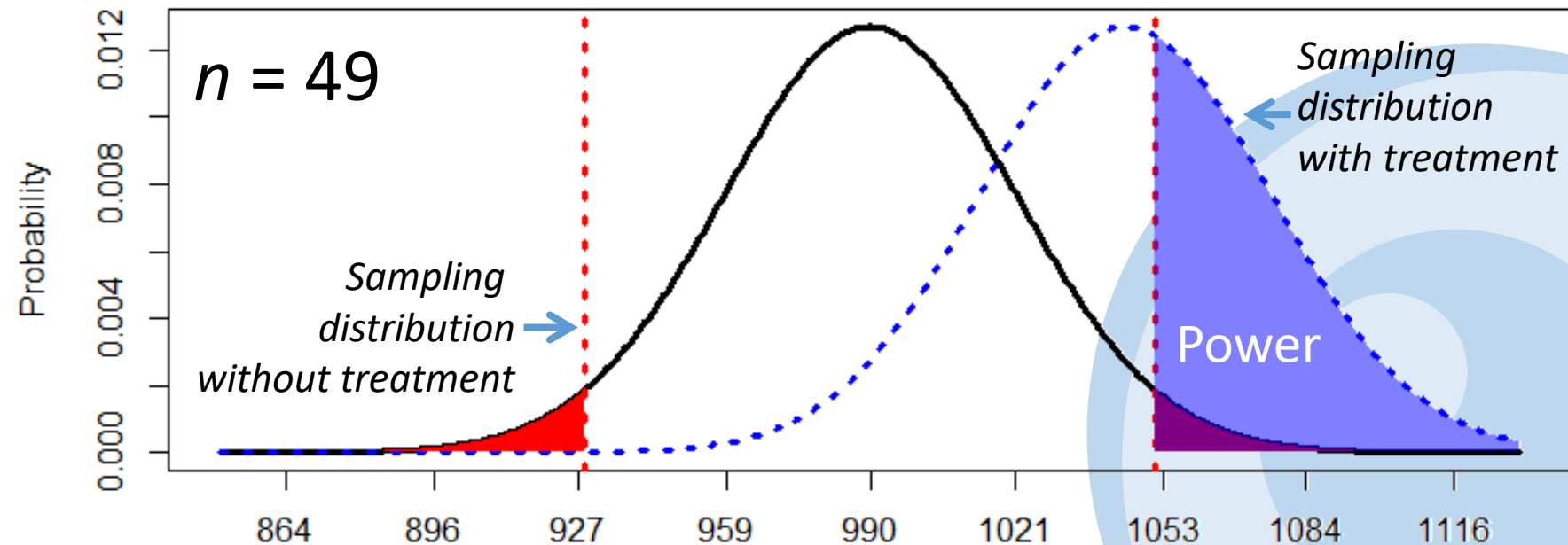
Information Needed:

- Baseline population mean
- Baseline population standard deviation
- Sample mean with treatment
(usually an estimation before you conduct the study)
- Sample size
- α level

Step 1:

Find the standard error of the sampling distributions

se =



Calculating Statistical Power

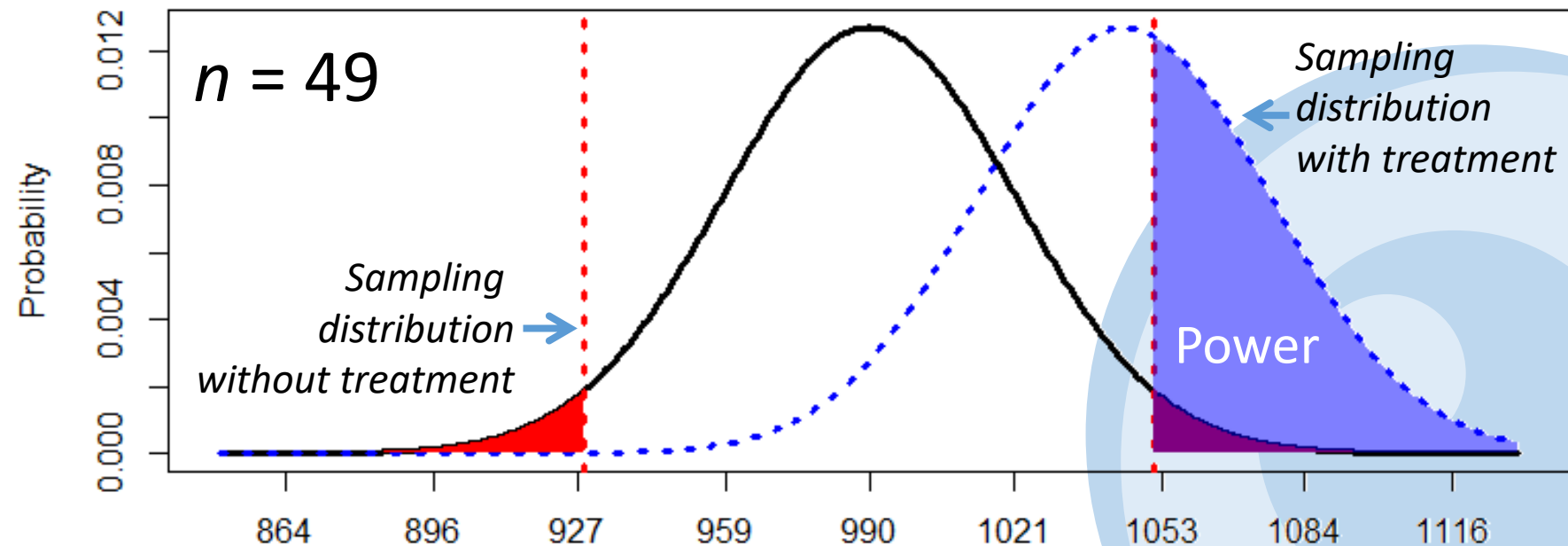
Information Needed:

- Baseline population mean (990)
- Baseline population standard deviation (220)
- Sample mean with treatment (1045)
(usually an estimation before you conduct the study)
- Sample size (49)
- α level (0.05)

Step 1:

Find the standard error of the sampling distributions

se =



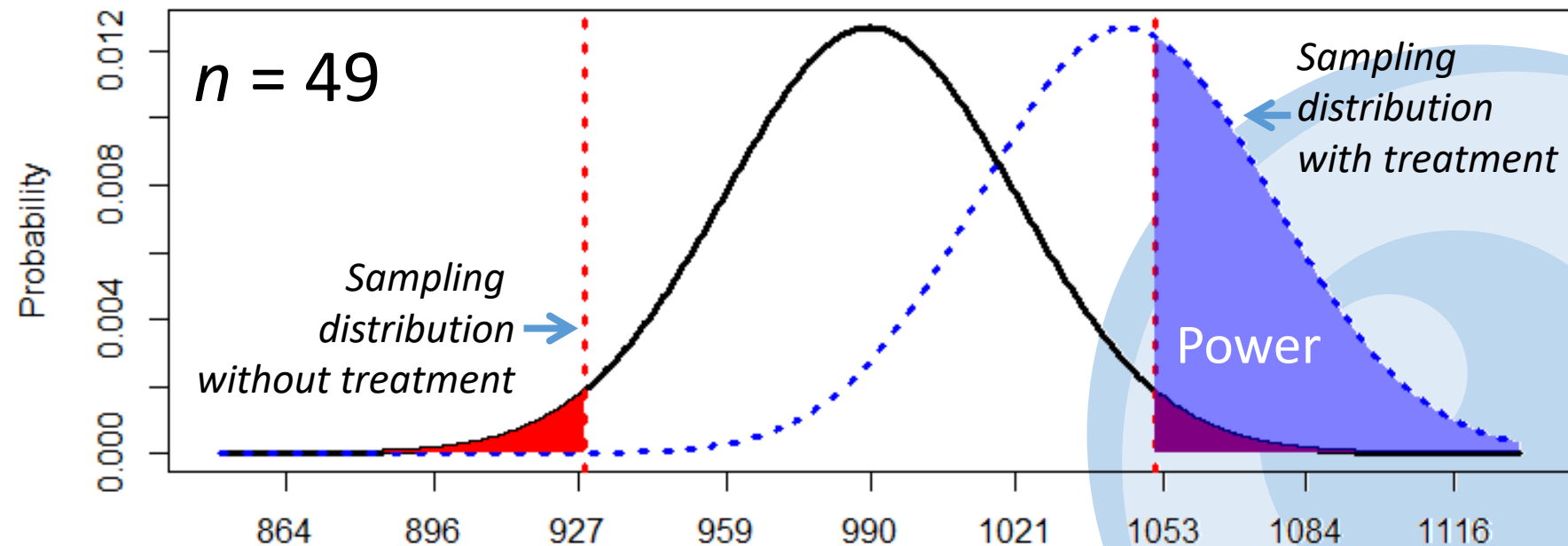
Calculating Statistical Power

Information Needed:

- Baseline population mean (990)
- Baseline population standard deviation (220)
- Sample mean with treatment (1045)
(usually an estimation before you conduct the study)
- Sample size (49)
- α level (0.05)
- S.E.

Practical Research Tips:

ALWAYS sketch the sampling distributions



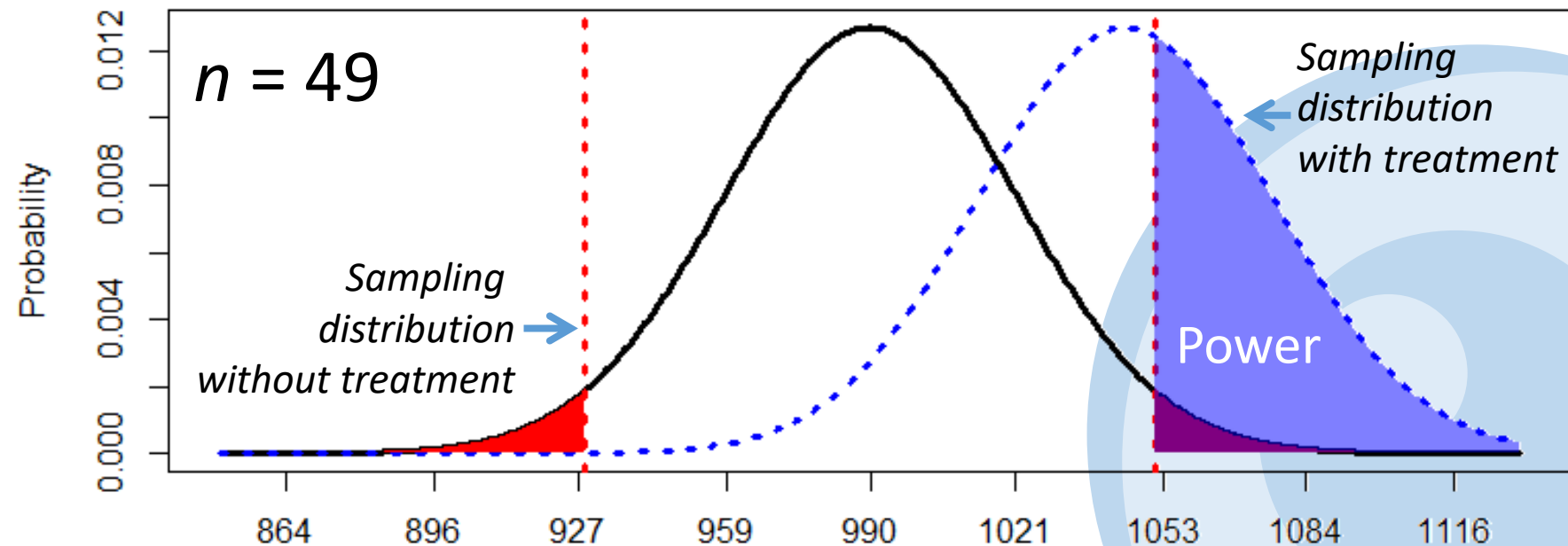
Calculating Statistical Power

Information Needed:

- Baseline population mean (990)
- Baseline population standard deviation (220)
- Sample mean with treatment (1045)
(usually an estimation before you conduct the study)
- Sample size (49)
- α level (0.05)
- S.E.

Step 2:

Find the cutoff points in raw scores
cutoff =



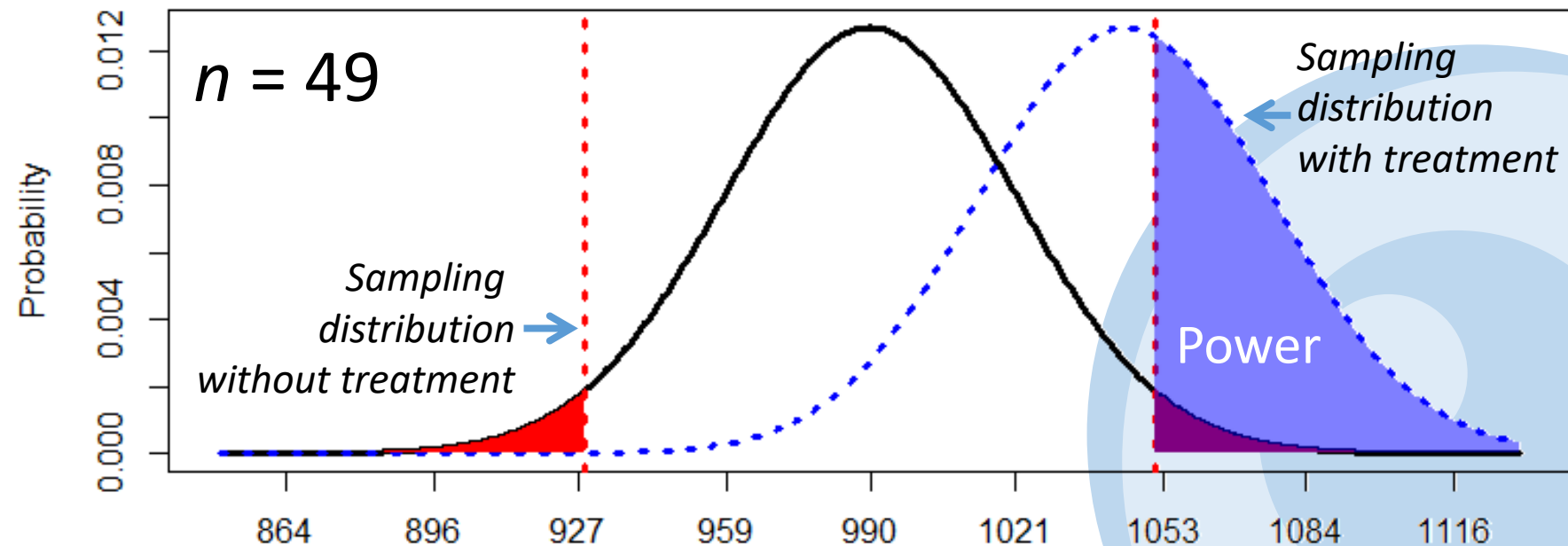
Calculating Statistical Power

Information Needed:

- Baseline population mean (990)
- Baseline population standard deviation (220)
- Sample mean with treatment (1045)
(usually an estimation before you conduct the study)
- Sample size (49)
- α level (0.05)
- S.E.

Step 2:

Find the cutoff points in raw scores
cutoff =



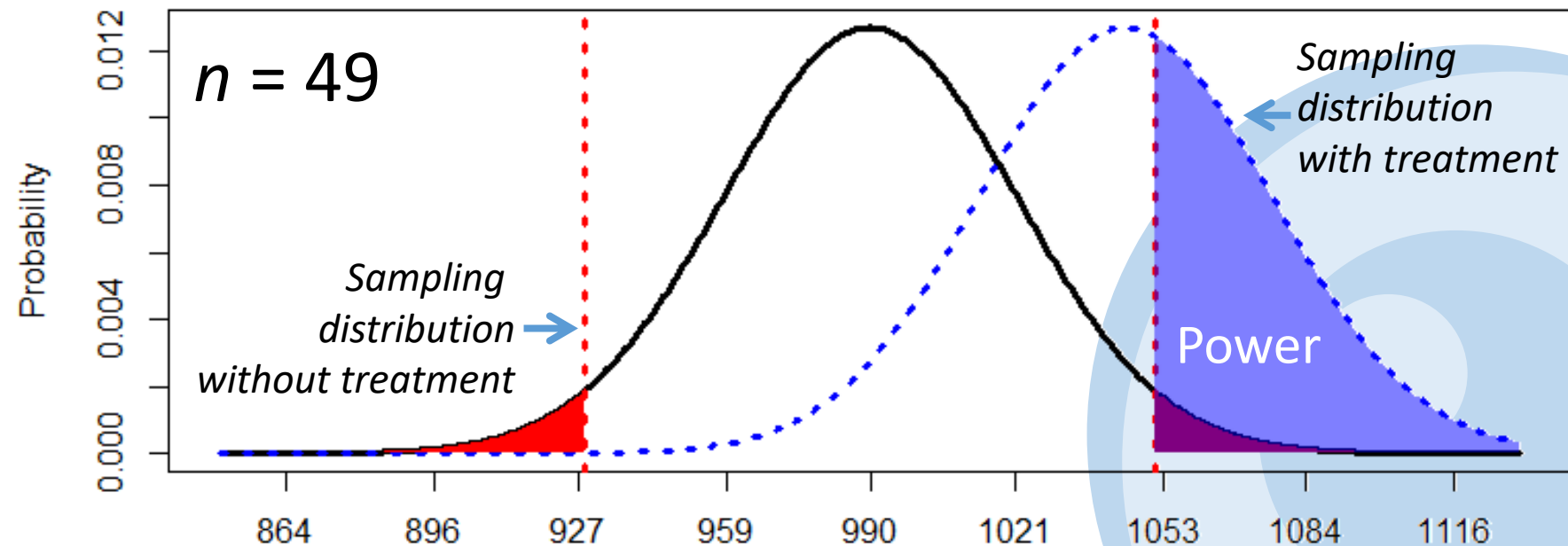
Calculating Statistical Power

Information Needed:

- Baseline population mean (990)
- Baseline population standard deviation (220)
- Sample mean with treatment (1045)
(usually an estimation before you conduct the study)
- Sample size (49)
- α level (0.05)
- S.E.
- Cut-off (raw)

Step 3:

Find the probability of getting a sample mean more extreme than the cutoff scores under H_1
power =



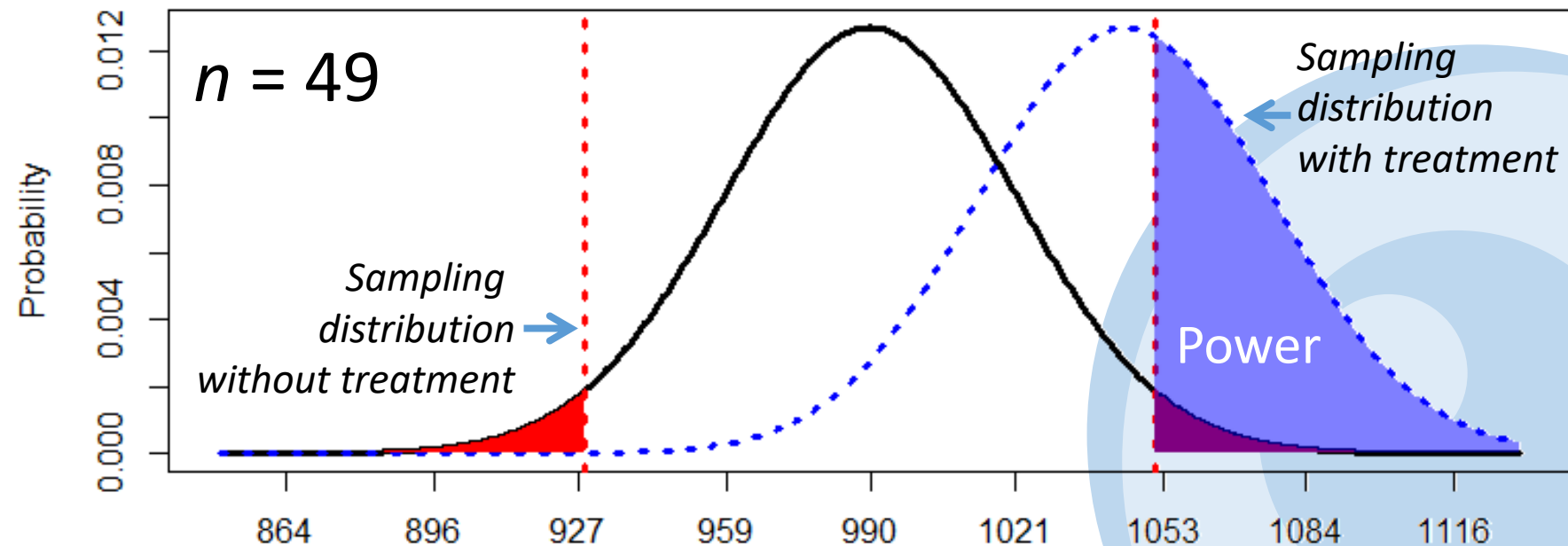
Calculating Statistical Power

Information Needed:

- Baseline population mean (990)
- Baseline population standard deviation (220)
- **Sample mean with treatment (1045)**
(usually an estimation before you conduct the study)
- Sample size (49)
- α level (0.05)
- S.E.
- Cut-off (raw)

Step 3:

Find the probability of getting a sample mean more extreme than the cutoff scores under H_1
power =



Class Challenge

- Design a "power" function that takes in:
 - Sample mean with treatment
 - Baseline population mean
 - Baseline population standard deviation
 - Sample size
 - α level
- And returns an array (sample_size, power)

This function only works when:
 $\text{treatment_mean} > \text{baseline_pop_mean}$
You can modify it to make it universal

```
power = function(treatment_mean,
baseline_pop_mean, pop_sd, sample_size,
alpha = 0.05){
  # Calculate the standard error
  se = NA

  # Find the cutoff points in raw scores
  cutoff = NA

  # Find the probability of
  # getting a sample mean
  # more extreme than the cutoff scores
  # under H1
  stat_power = NA

  # save sample_size and power to an array
  answer = NA

  # return the array
  return (answer)
}
```

Test your code:

```
> power(1045, 990, 220, 49, 0.05)
```

For-loop

- A for() loop lets you perform the same computations for a specific number of times

"i" is a counter,
it takes the next
value in the array
after each round

"1:10" denotes
an array from
1 to 10

```
for(i in 1:10){  
  #print out the number i  
  print(i)  
}
```

Some
computations to
perform

Output:

```
[1] 1  
[1] 2  
[1] 3  
...  
[1] 10
```

Ending the for()
loop

Statistical Power: Why do we care?

- You want to estimate the probability of getting a significant result *before* you perform the study
 - If there is little chance to show a treatment effect (usually $> 80\%$ power)
→ You don't invest the money, time and other resources
- You want to estimate the number of participants you need to get a pre-specified power (e.g., 80%)
 - Data collection is expensive:
 - Time (investigators, researchers, participants)
 - \$ (researchers, participants)
 - Effects on participants
(ineffective clinical trial, loss of other treatment opportunities)

A use case for a for() loop?

Class Challenge

- Design a `for()` loop that:
 - Assigns `i` to a number from 5 to 200
 - Assumes that baseline population mean, treatment mean, population standard deviation, and `alpha` do not change
 - Calculates the power with a sample size of `i`
- Prints the sample size and the corresponding power

If we want to achieve an 80% power, how many participants do we need?

Set up the parameters for the `power()` function:

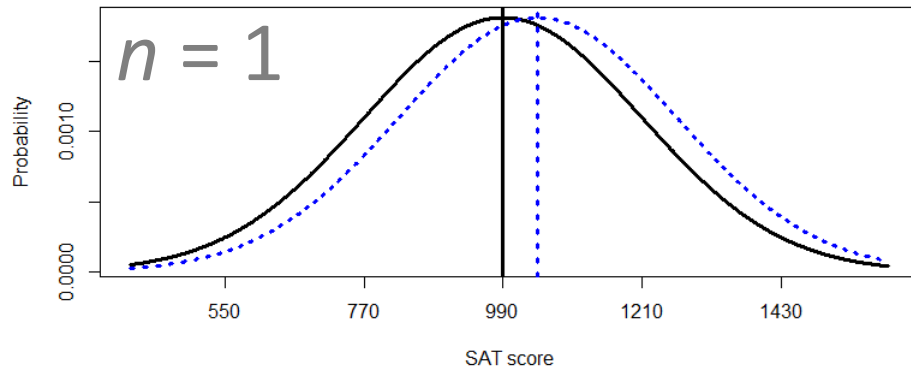
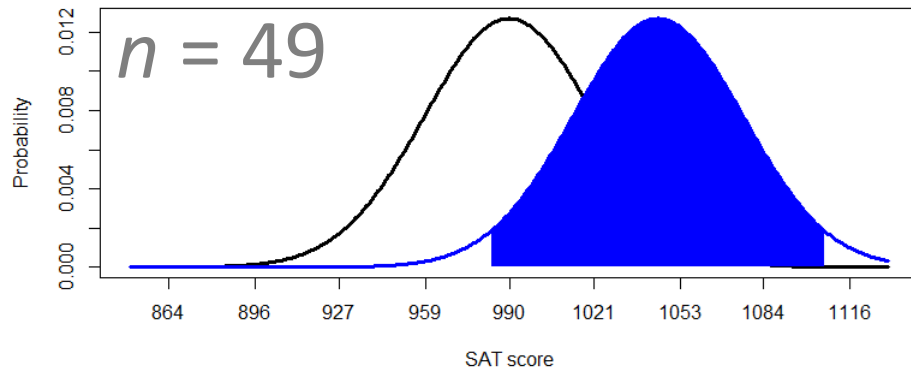
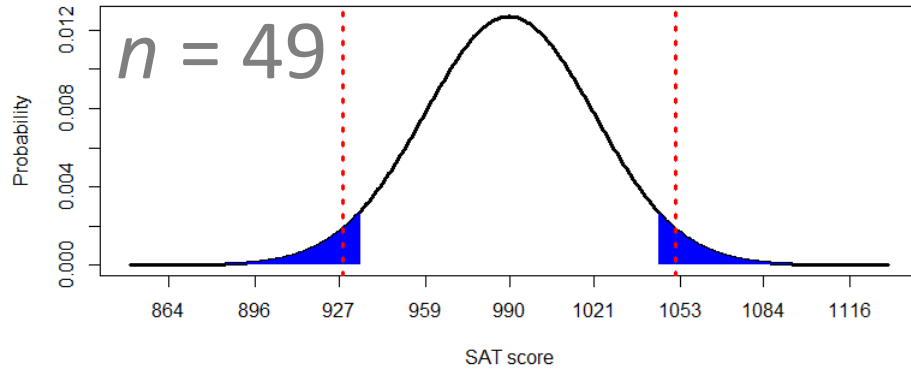
```
treatment_mean = NA
baseline_pop_mean = NA
pop_sd = NA
alpha = NA
```

The `for()` loop:

```
for(NA in seq(from = NA, to = NA, by = NA)){
  # Calculate the power for
  # that given i as sample size
  stat_power = NA

  # Print the sample size and
  # the estimated power
  print(paste0("Sample size: ", i, ",
Power = ", stat_power[2]))
}
```

Wrap up!



Hypothesis testing

Test if the treatment effect is reliable
(finding p-value)

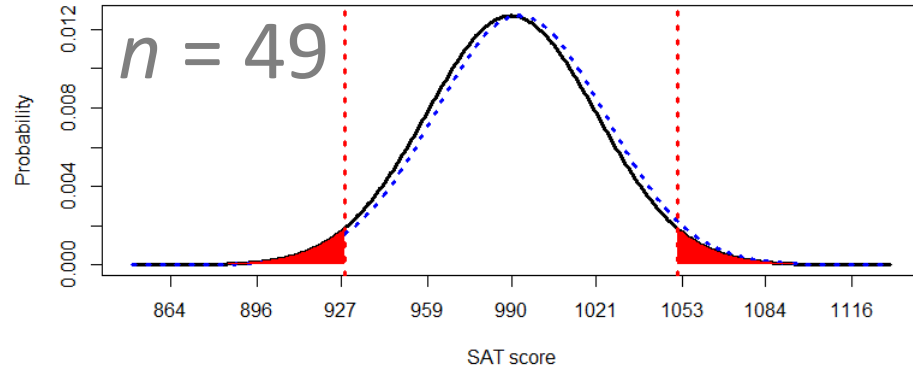
Confidence Interval

Reliability of the treatment effect
if the experiment is repeated under the same
condition, [95%] of the time we will get a
sample mean in between $[x1, x2]$

Effect Size (on population distributions)

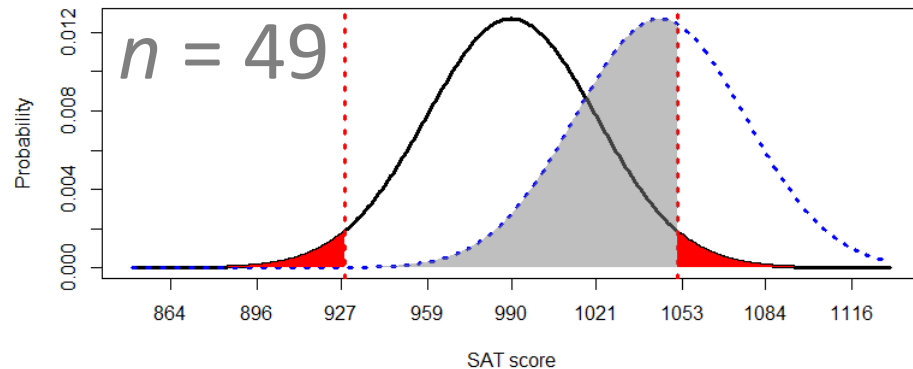
On average, how big is the change with
treatment on an individual [in standard
deviations]

Wrap up!



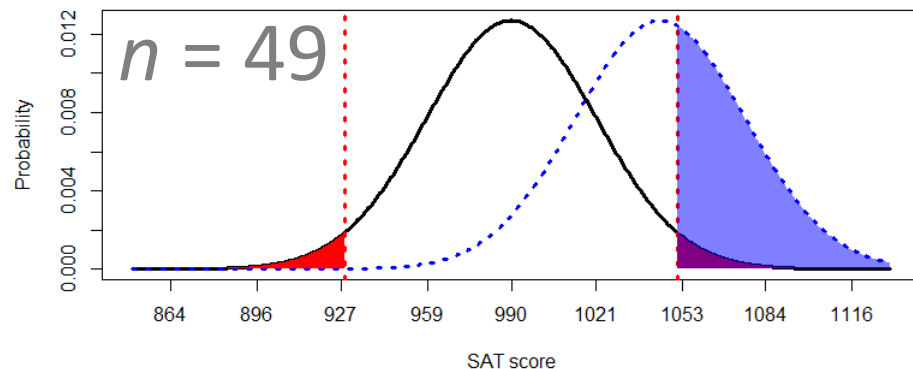
Type I error

The probability of rejecting H_0 when there is no treatment effect



Type II error

The probability of not rejecting H_0 when there is a treatment effect



Power

The probability of rejecting H_0 when there is a treatment effect

→ Estimation of sample size needed to show a significant treatment effect