CSCI 443: LECTURE 13 STUDENT

Professor David Harrison

OFFICE HOURS

Tuesday	4:00-5:00 PM
Wednesday	12:30-2:30 PM

HOMEWORK 4

Will be handed hopefully Thursday

DATES OF INTEREST

March 21	Homework 4 handed out
March 28	Homework 4 due

BLACKBOARD & GITHUB

Slides up through lecture 12 on blackboard.

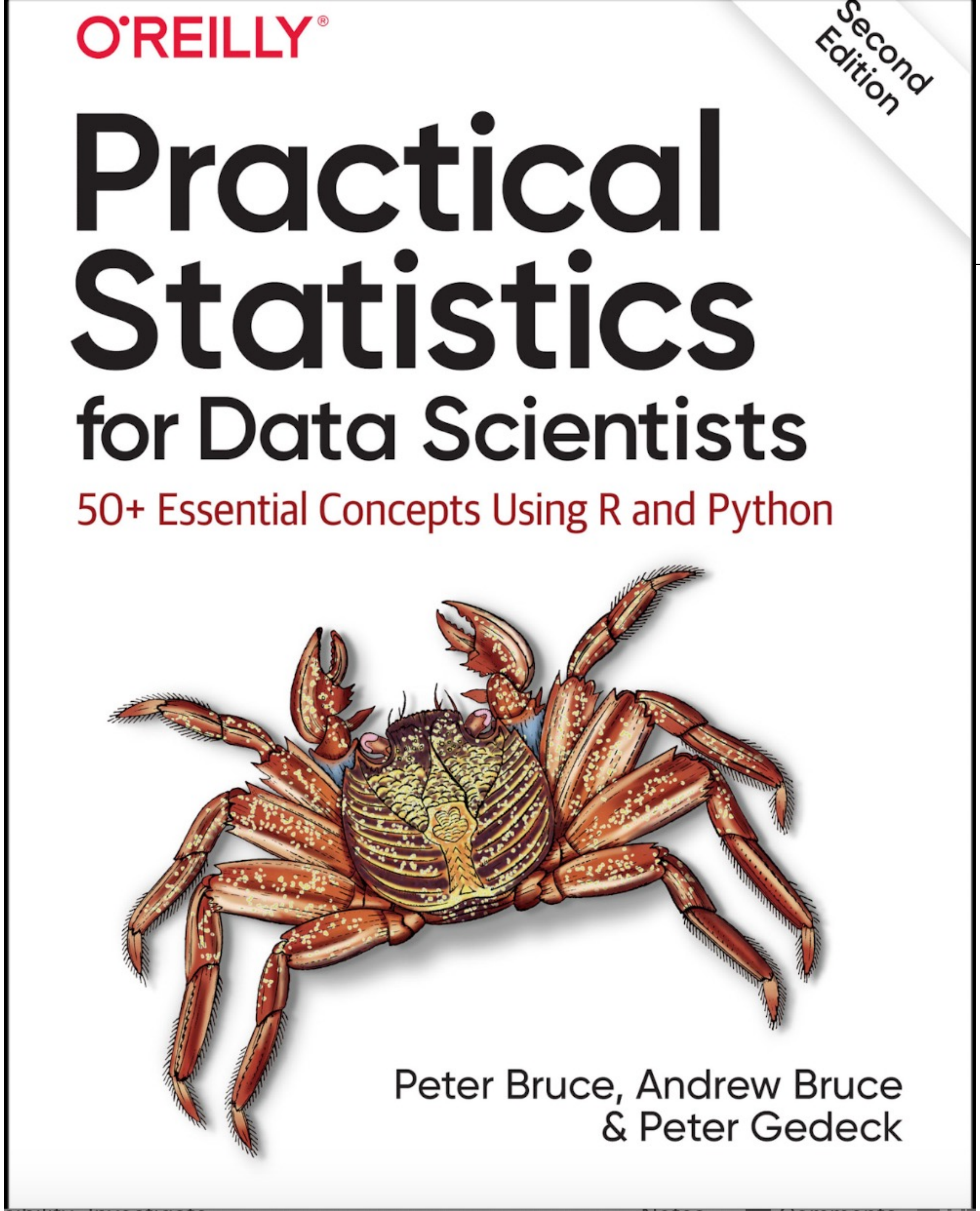
Lecture slides and examples committed to GitHub also up through lecture 12.

The project is at

https://github.com/dosirrah/CSCI443_AdvancedDataScience

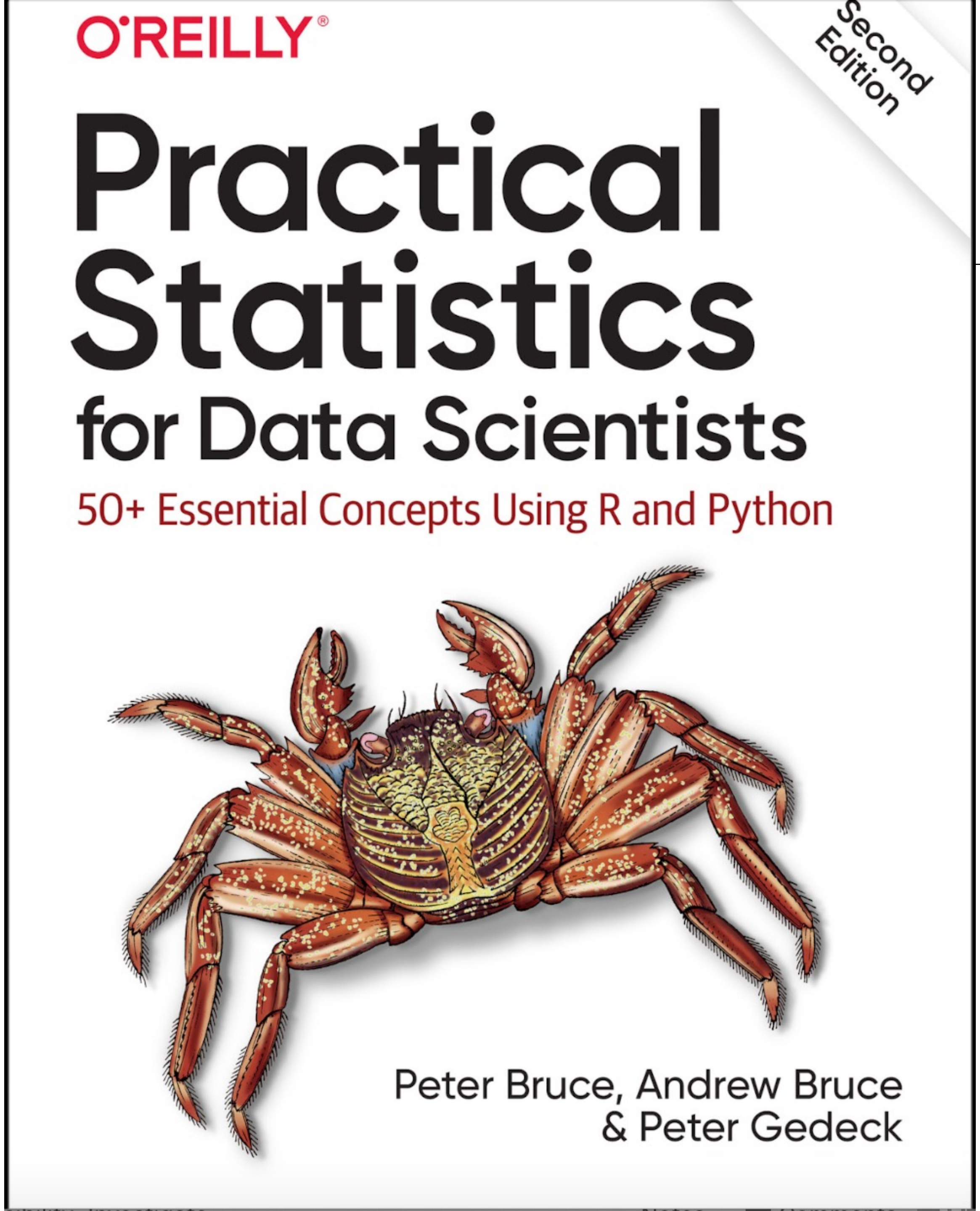
READ ABOUT

- Finishing chapter 2
  - Long-tailed distributions
  - Student t-distribution
  - Binomial distribution
  - Chi-Squared distribution
  - F distribution
  - Poisson distribution
  - Exponential distribution
- Entering chapter 3: experiments, hypothesis testing
  - A/B testing
  - Control groups
  - Null hypotheses



THINGS I WANT TO COVER TODAY

- More on confidence intervals and what they mean.
- Student t distribution

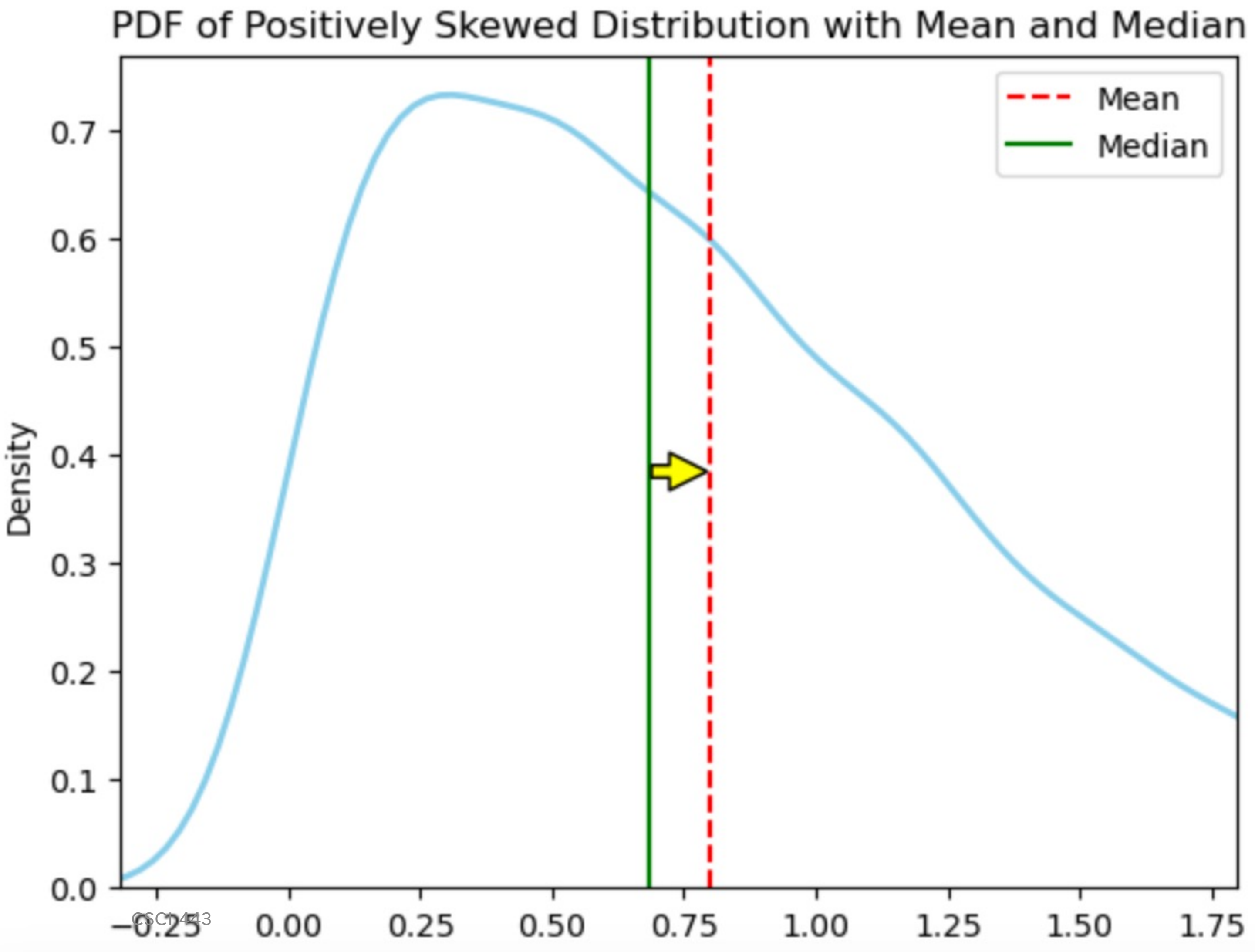


PREVIOUS LECTURE (12): SKEWNESS

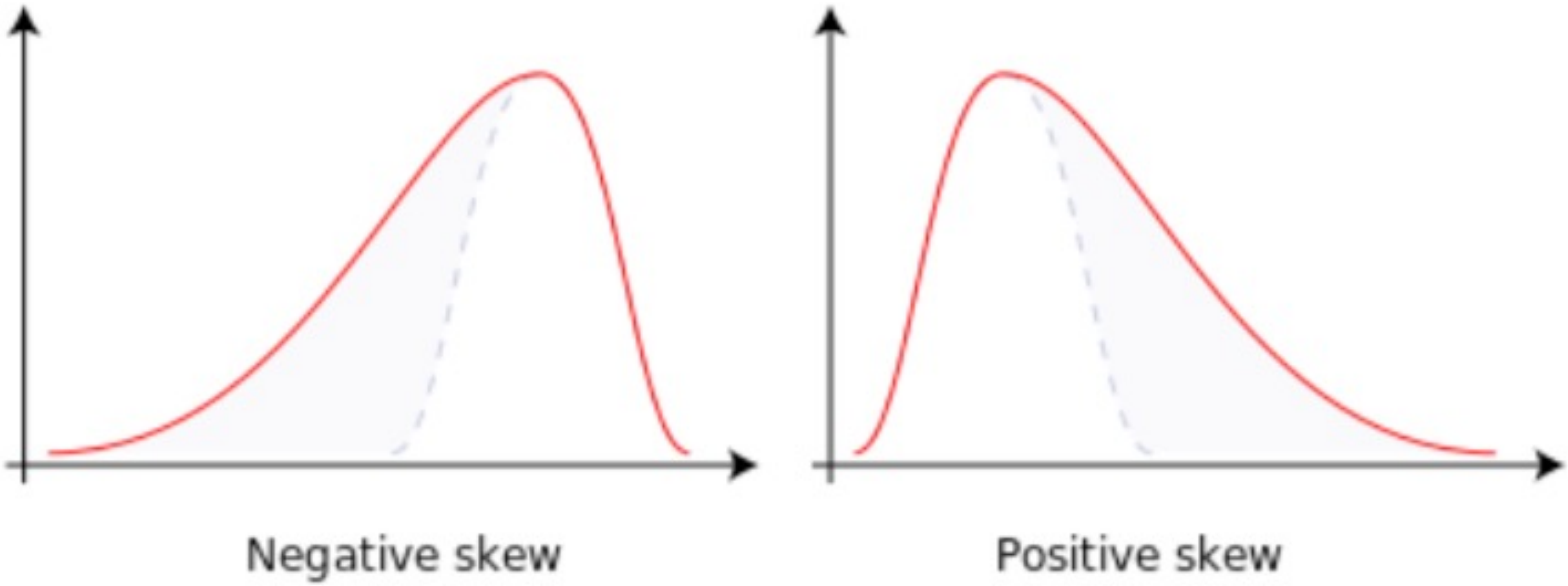
When is a distribution skewed?

Rule of thumb: “When the mean deviates from the median.”

$$\gamma_1 = \frac{E[(X-\mu)^3]}{\sigma^3}$$



FROM LECTURE 12: SKEWNESS



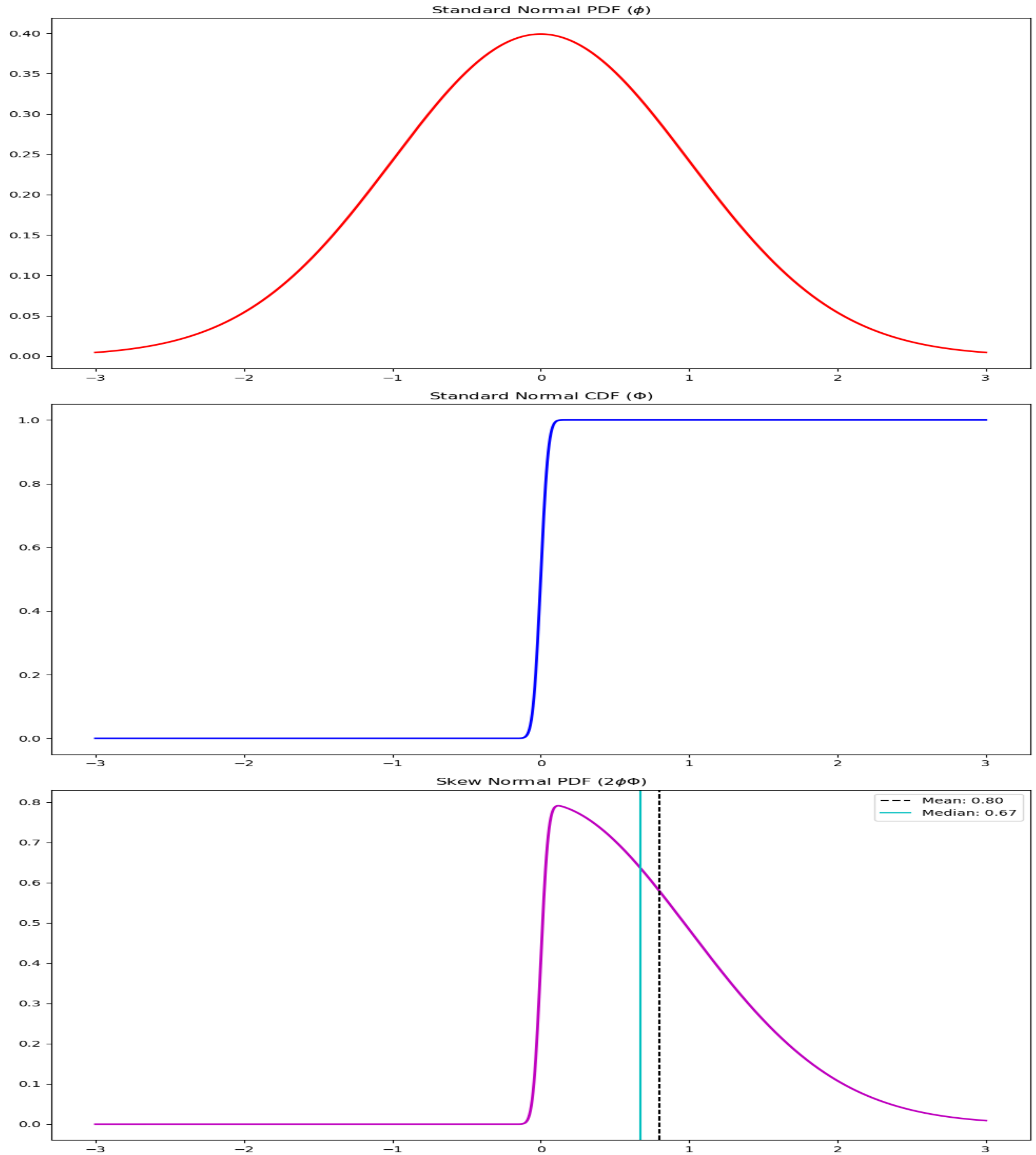
Follow the tail…

FROM LECTURE 12: SKEWNORM

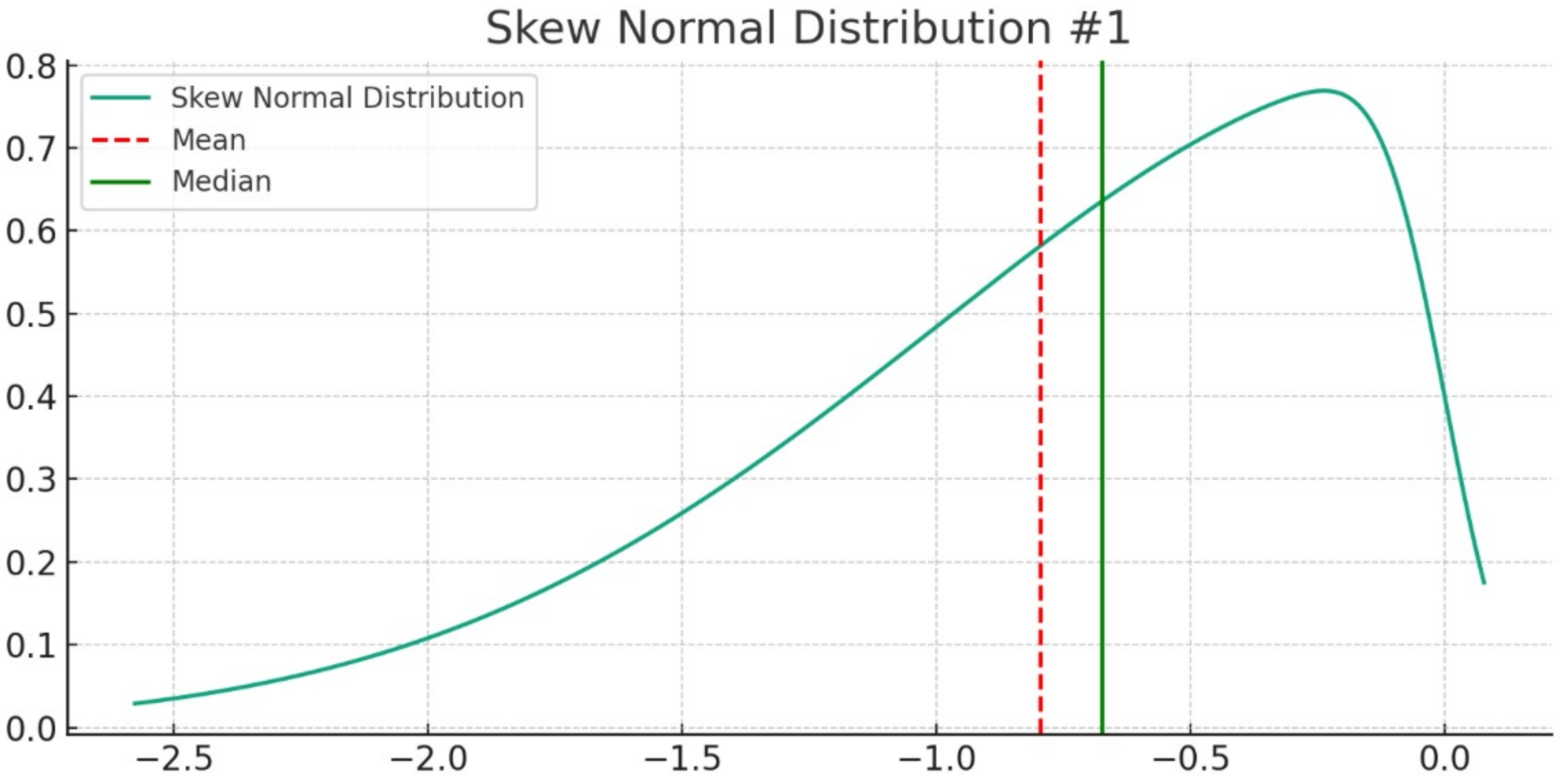
Play with scipy.stats.skewnorm

$$f(x;\alpha) = 2\phi(x)\Phi(\alpha x)$$

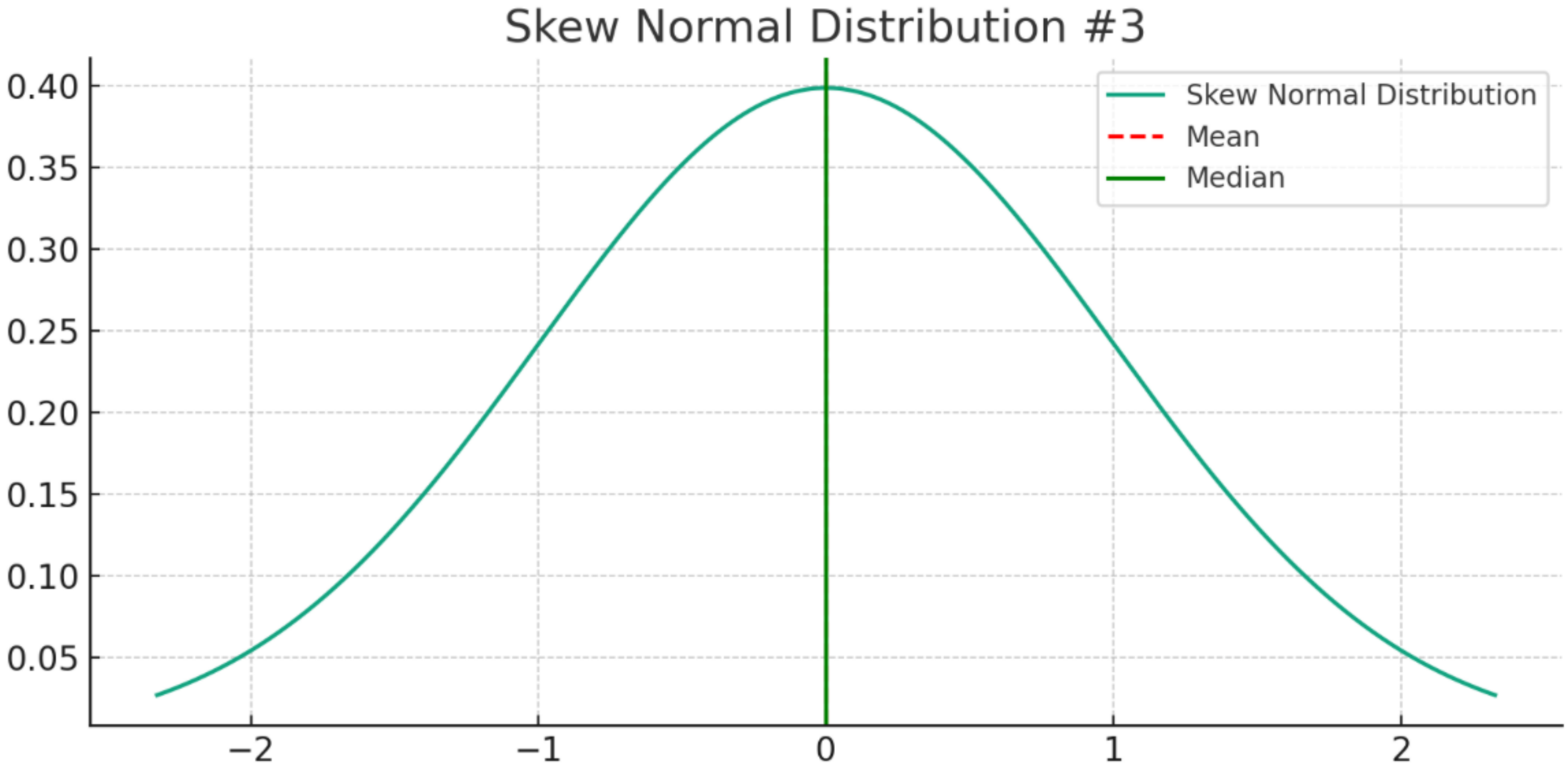
α = 25



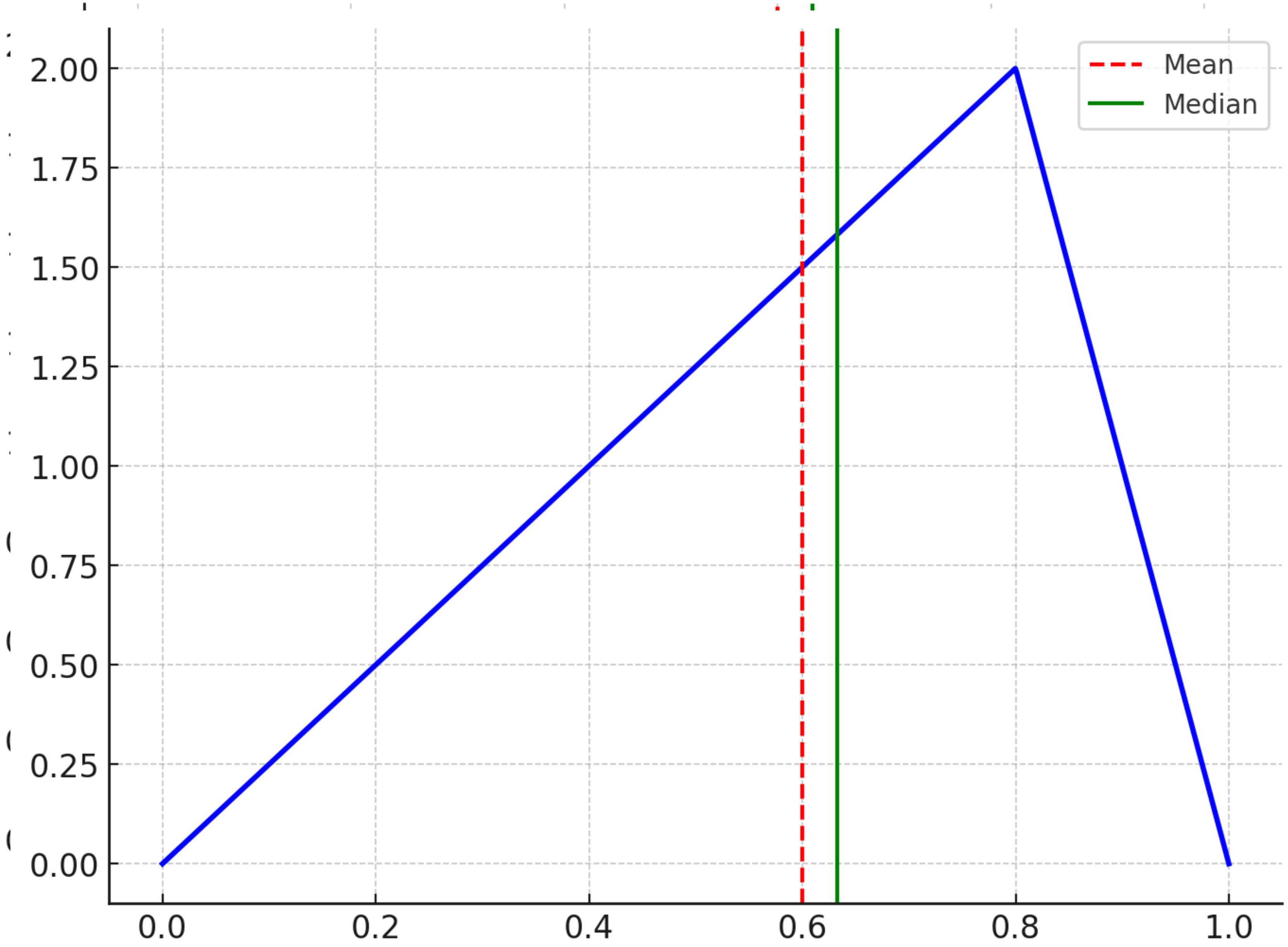
FROM LECTURE 12: WHICH DIRECTION IS THE SKEW? RIGHT

- 

FROM LECTURE 12: WHICH DIRECTION IS THE SKEW? NONE

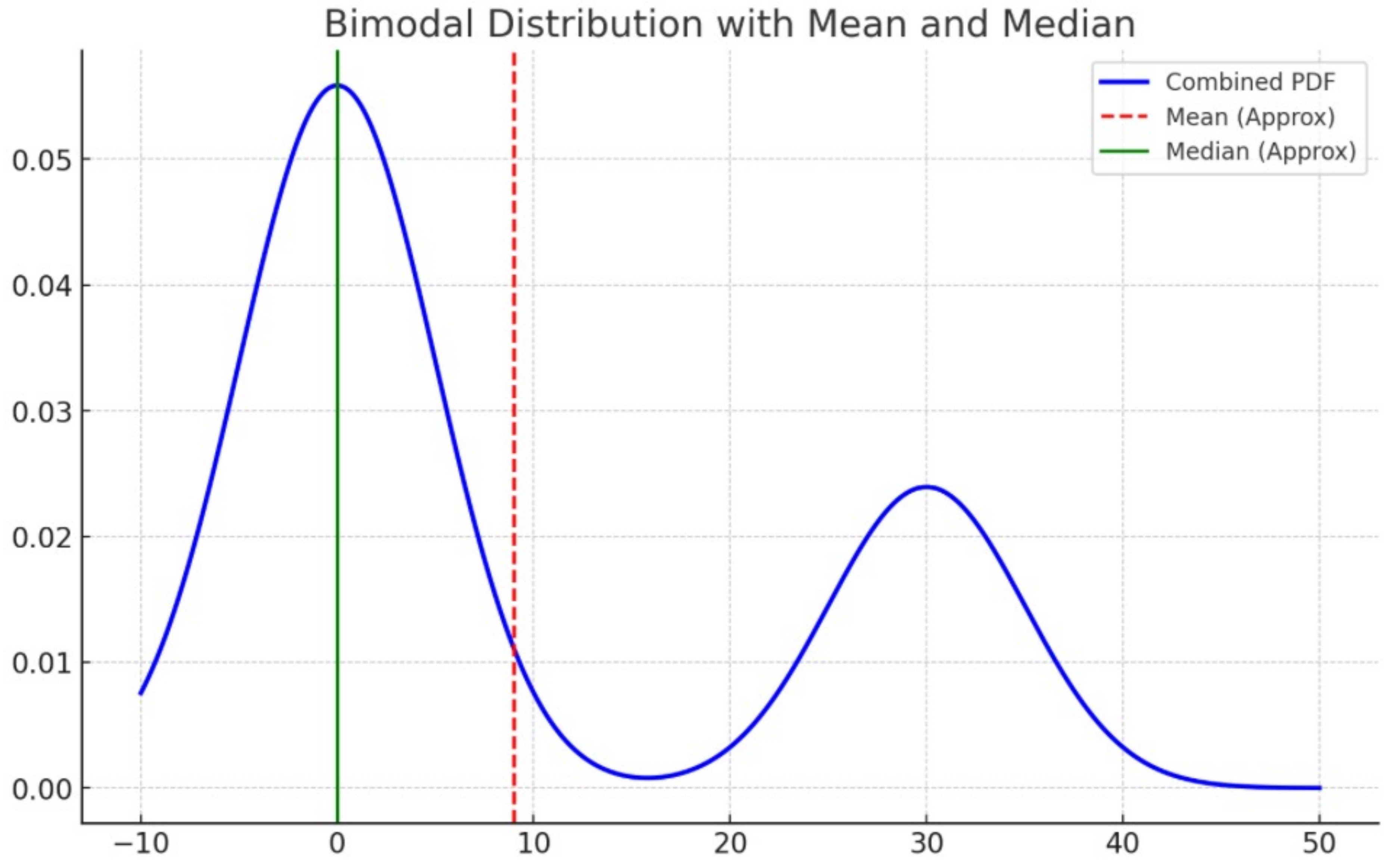


FROM LECTURE 12: WHICH DIRECTION IS THE SKEW? LEFT



FROM LECTURE 12: WHICH DIRECTION IS THE SKEW?

•



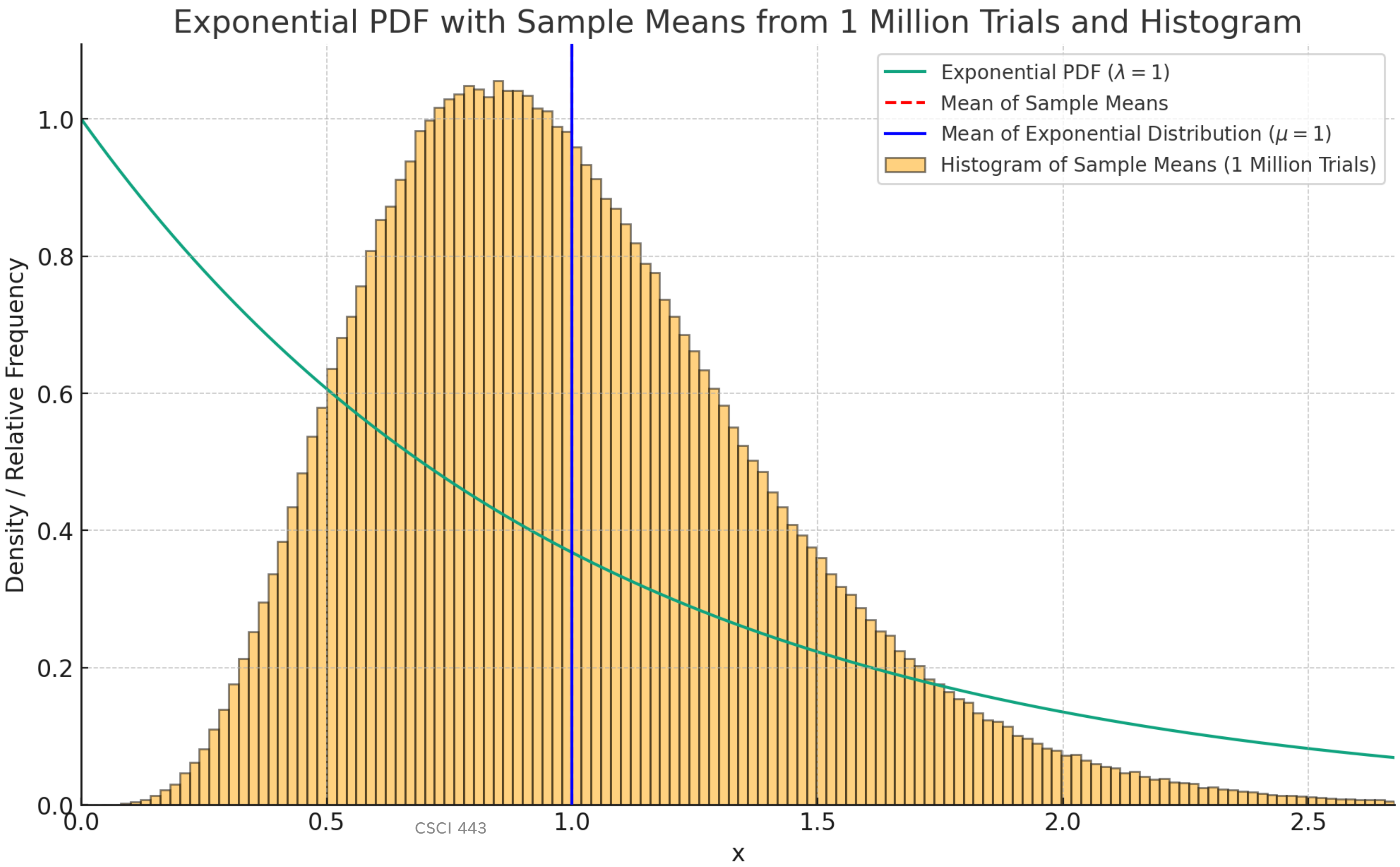
FROM LECTURE 11: SAMPLE MEAN IS ALSO RANDOM

n=6

1 million trials (sample means) Looks kind of like a slightly skewed Gaussian.

With small n in each sample mean, the distribution of sample means may remain skewed.

CLT's effectiveness depends on increasing n.



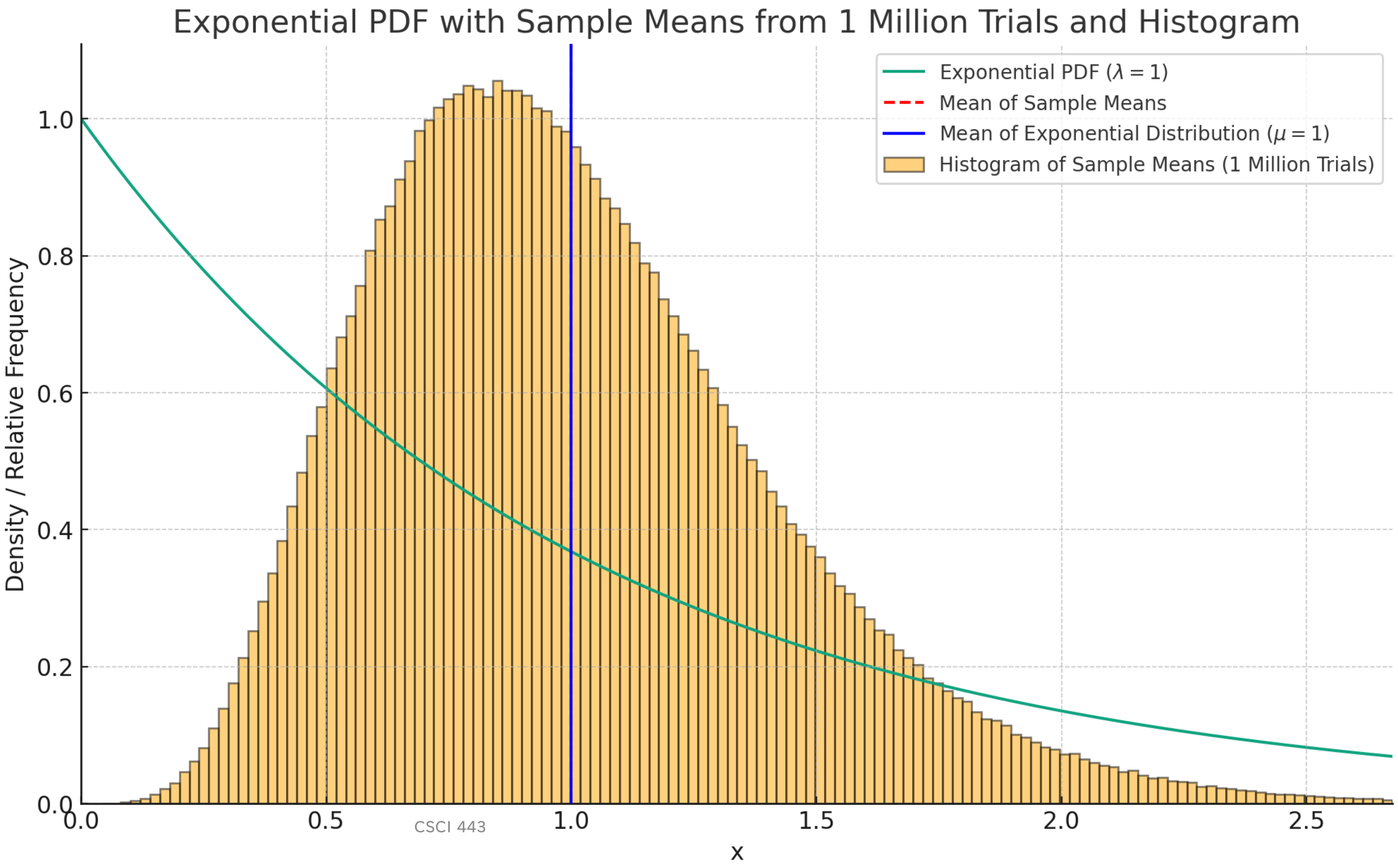
FROM LECTURE 11: SAMPLE MEAN IS ALSO RANDOM

n=6

1 million trials (sample means) Looks kind of like a slightly skewed Gaussian.

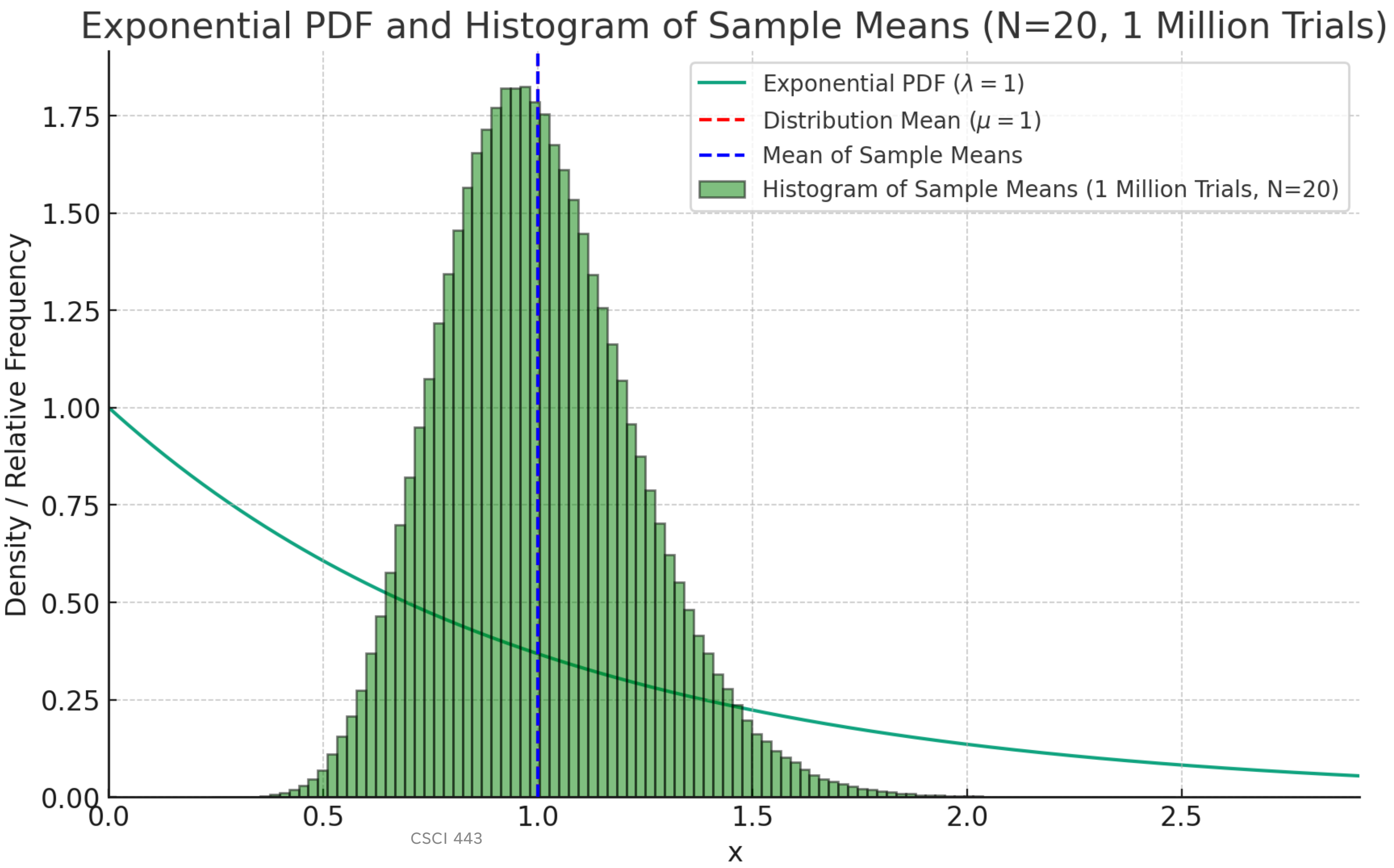
With small n in each sample mean, the distribution of sample means may remain skewed.

CLT's effectiveness depends on increasing n.

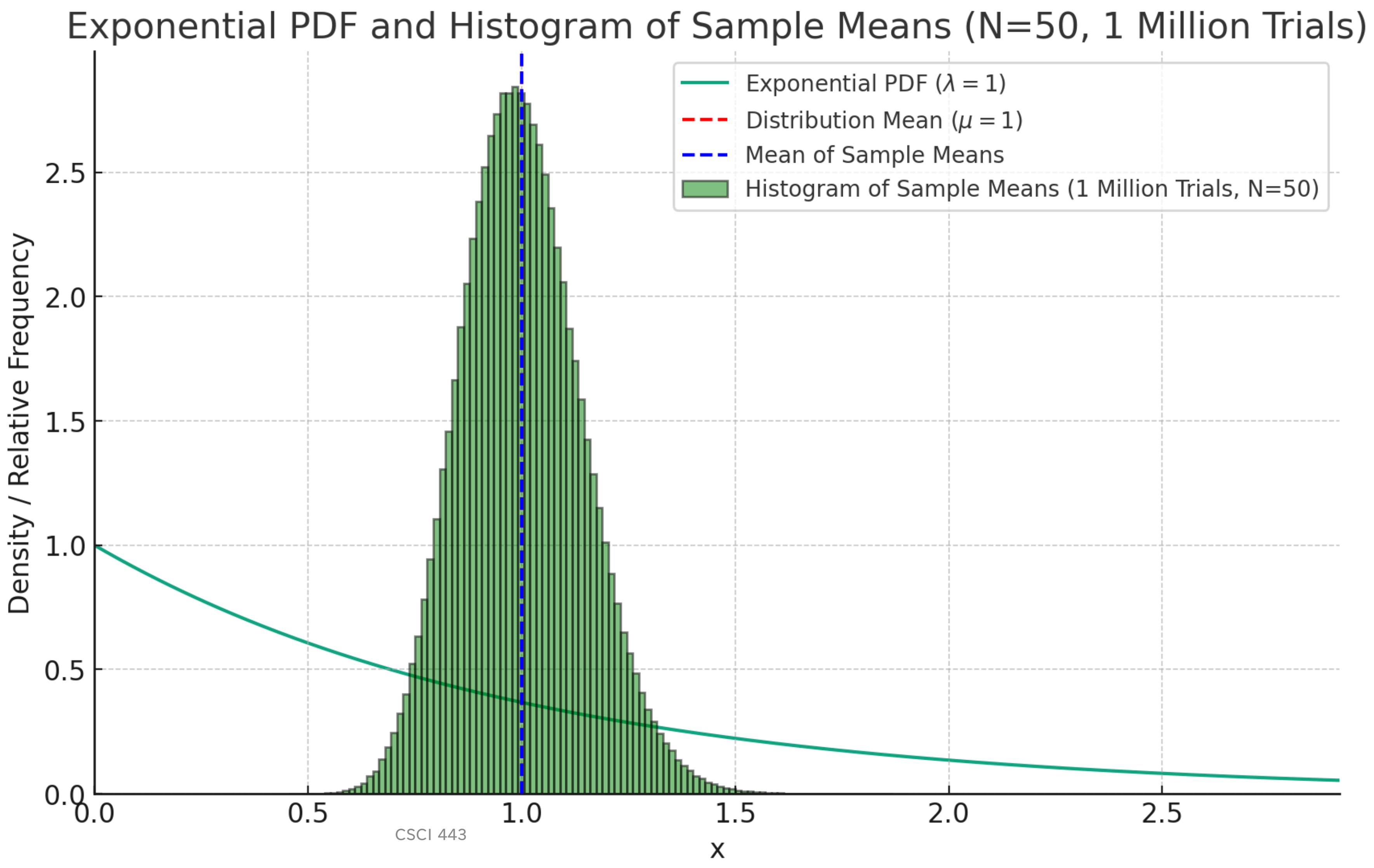


FROM LECTURE 11: SAMPLE MEAN IS ALSO RANDOM BUT N MATTERS

What happens as we increase the number (n) of samples in each sample mean?



FROM LECTURE 11: SAMPLE MEAN IS ALSO RANDOM BUT N MATTERS



What happens as we increase the number (n) of samples in each sample mean?

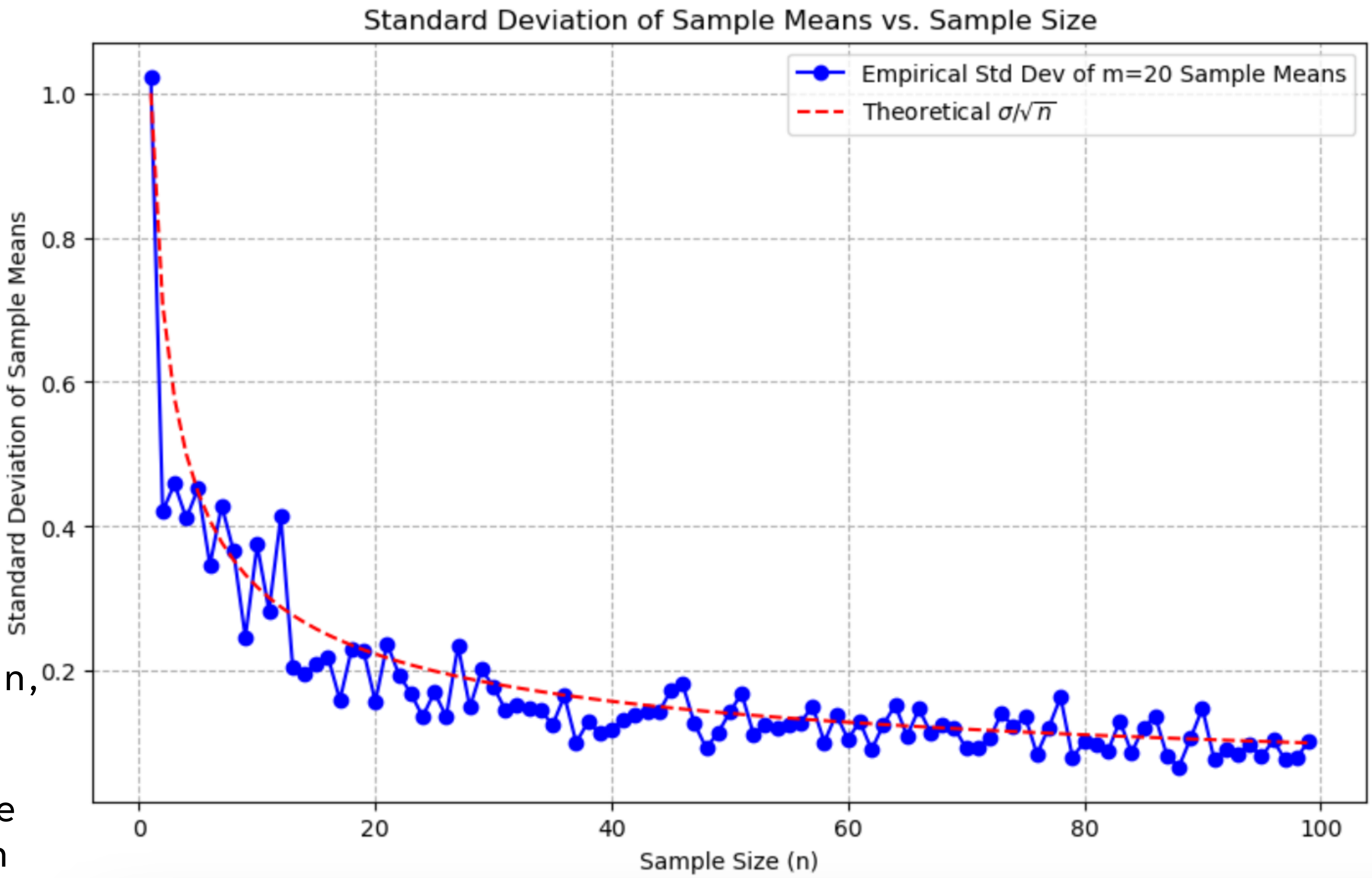
FROM LECTURE 11: STANDARD ERROR AS FUNCTION OF N

Standard deviation of the sampling mean distribution

- Standard Error (SE) decreases with n according to

$$SE = \frac{\sigma}{\sqrt{n}}$$

For moderate to large n, applies across many underlying distributions. Here the underlying distribution is exponential.



WHAT IS A CONFIDENCE INTERVAL?

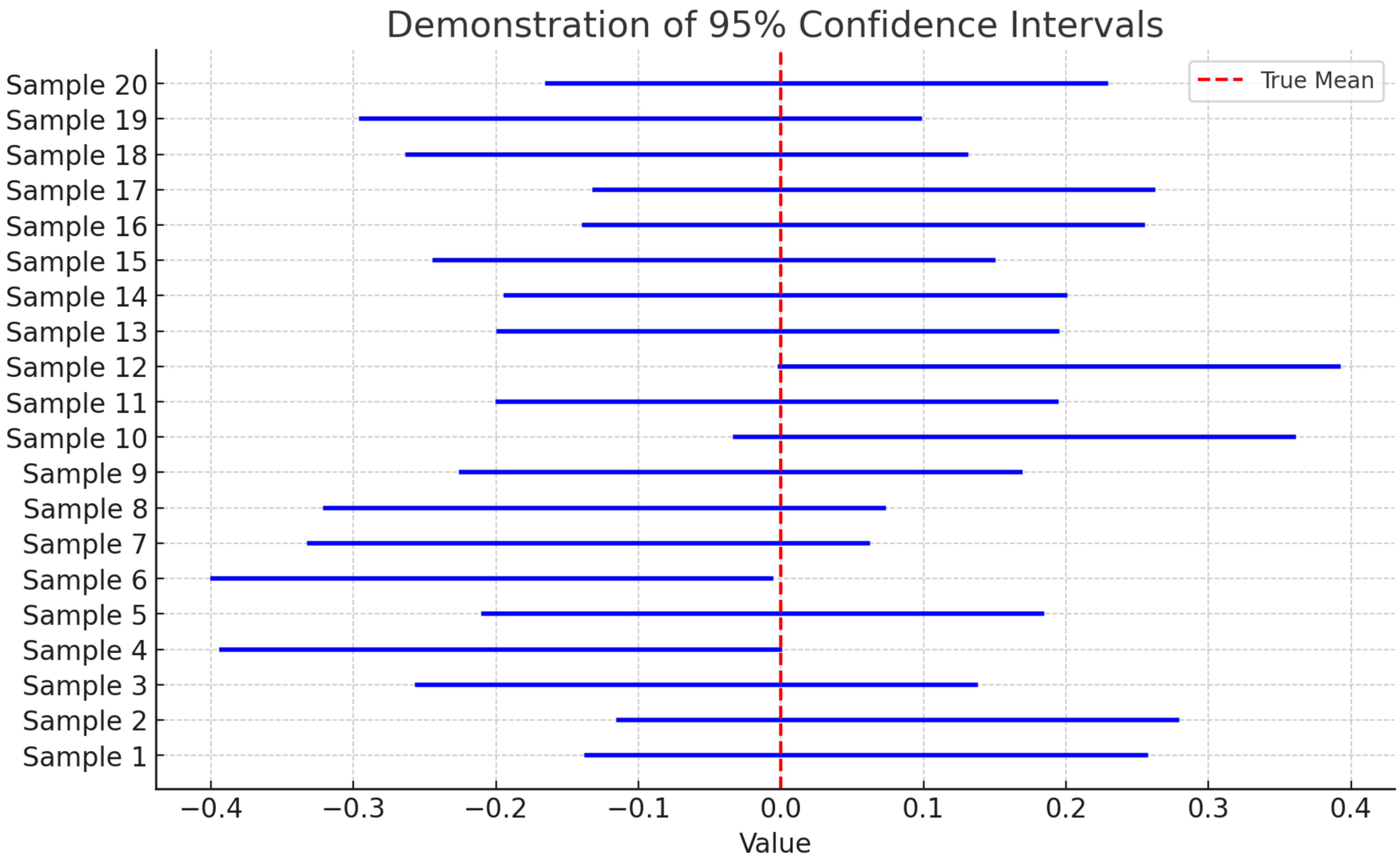
A confidence interval specifies a range around an estimate for which the property being estimated likely resides.

Slightly more formally: if use the same procedure to generate many p% confidence interval, we could expect to find the true value of the property resides within the interval p% of the time.

WHAT IS A CONFIDENCE INTERVAL?

Ex: many confidence intervals estimating the distribution mean.

We expect 95% will contain the the distribution mean.



HOW DO WE COMPUTE A CONFIDENCE INTERVAL?

When we have many samples, we can assume that the sampling distribution is Gaussian.

Due to the Central Limit theorem this almost always a reasonable assumption.

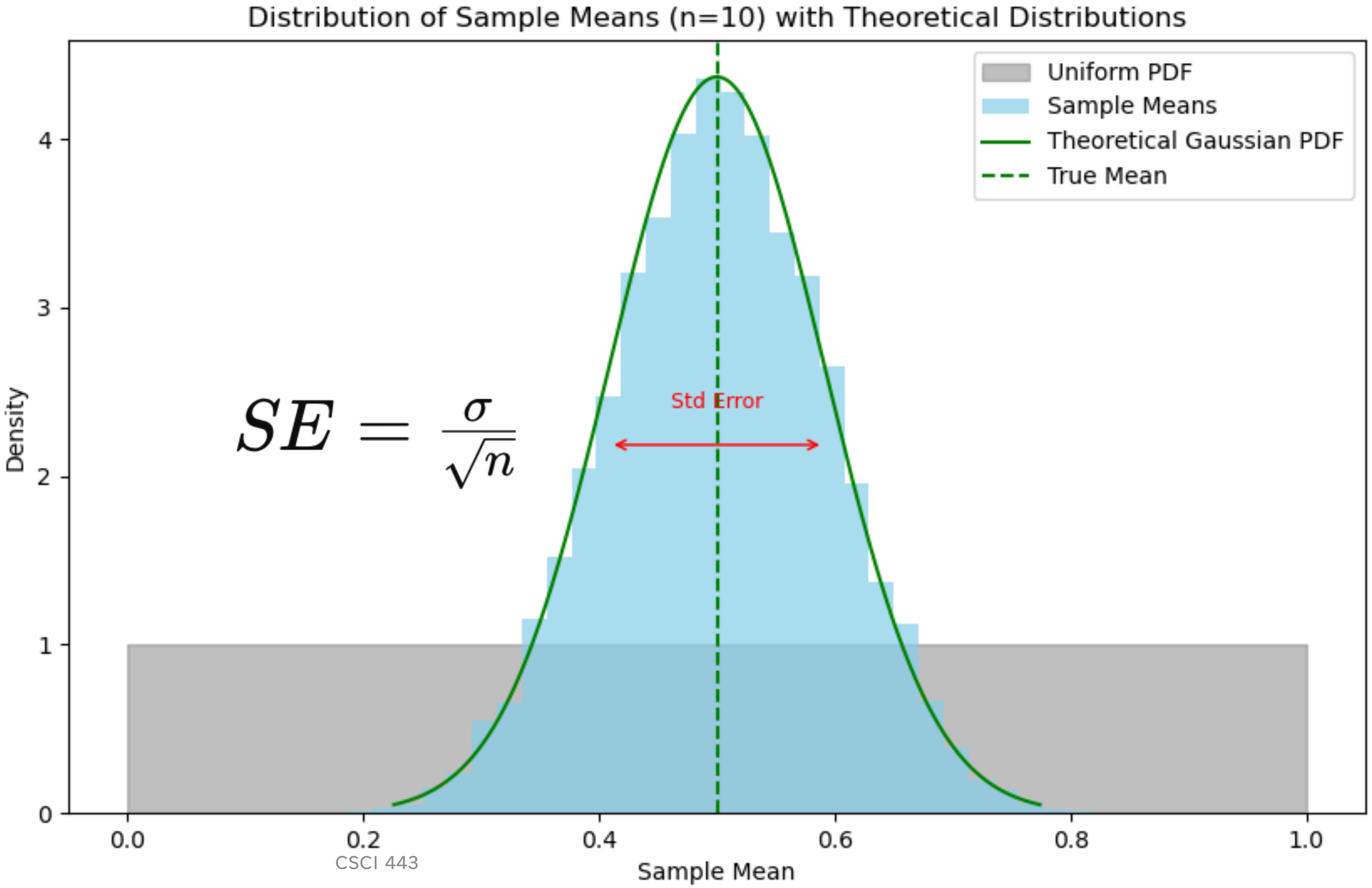
What is enough samples?

FROM LECTURE 11: SAMPLE MEAN DISTRIBUTION OF UNIFORM RV

Let's consider a uniform random variable U[0,1].

For R=10000 sample means created from n=10 samples, we plot a histogram of the sample means.

For a symmetric distribution, the sample mean distribution approaches Gaussian with small n

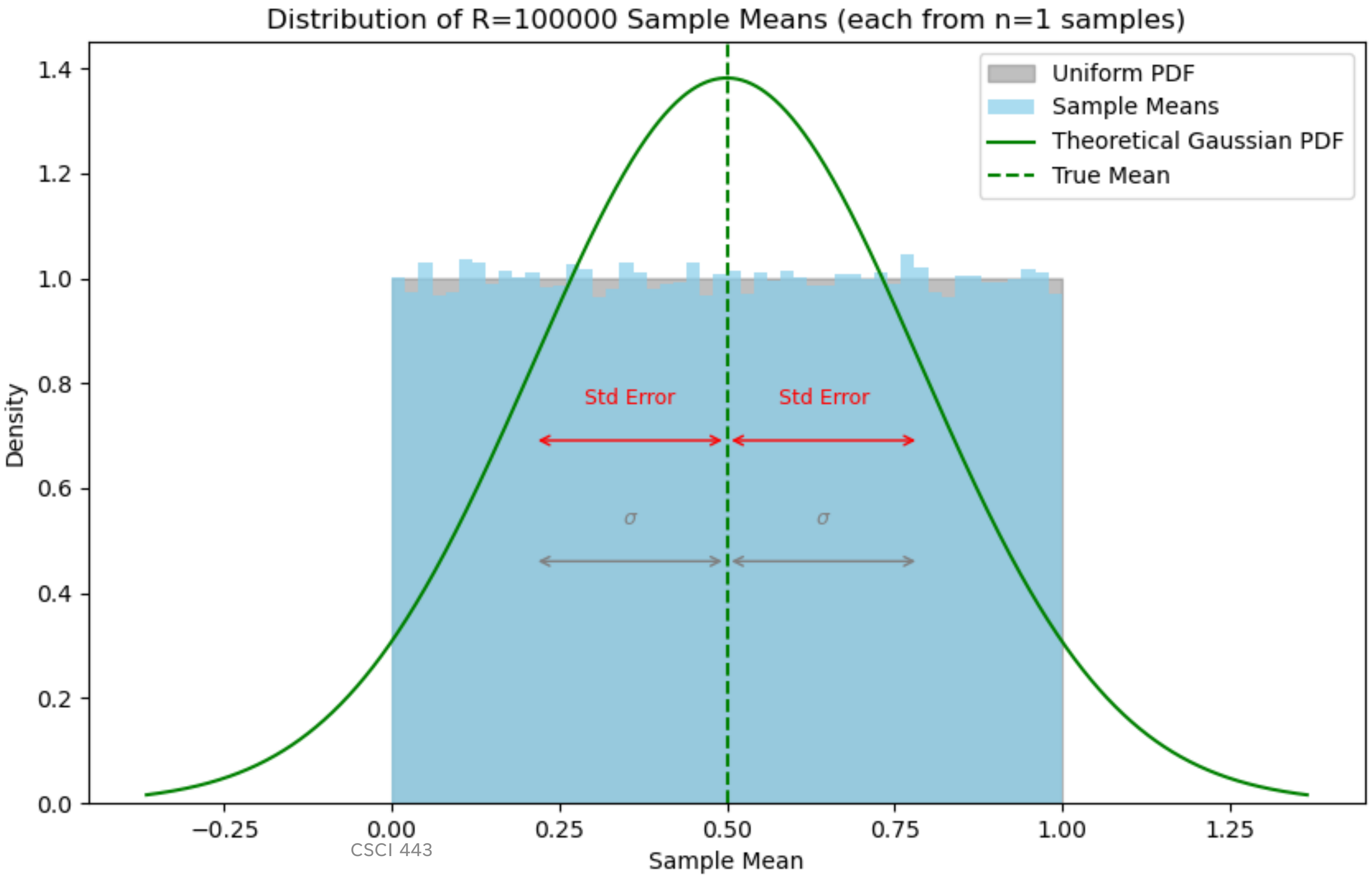


HOW DO WE ESTIMATE SAMPLE MEAN WITH CONFIDENCE FROM A FEW SAMPLES?

Let's consider a uniform random variable U[0,1].

If I get a single sample, the probability that the distribution mean falls within 1σ is approximately 68% if the sampling distribution of the mean is Gaussian.

$$\mathrm{SE} = \frac{\sigma}{\sqrt{n}} = \frac{\sigma}{\sqrt{1}} = \sigma$$

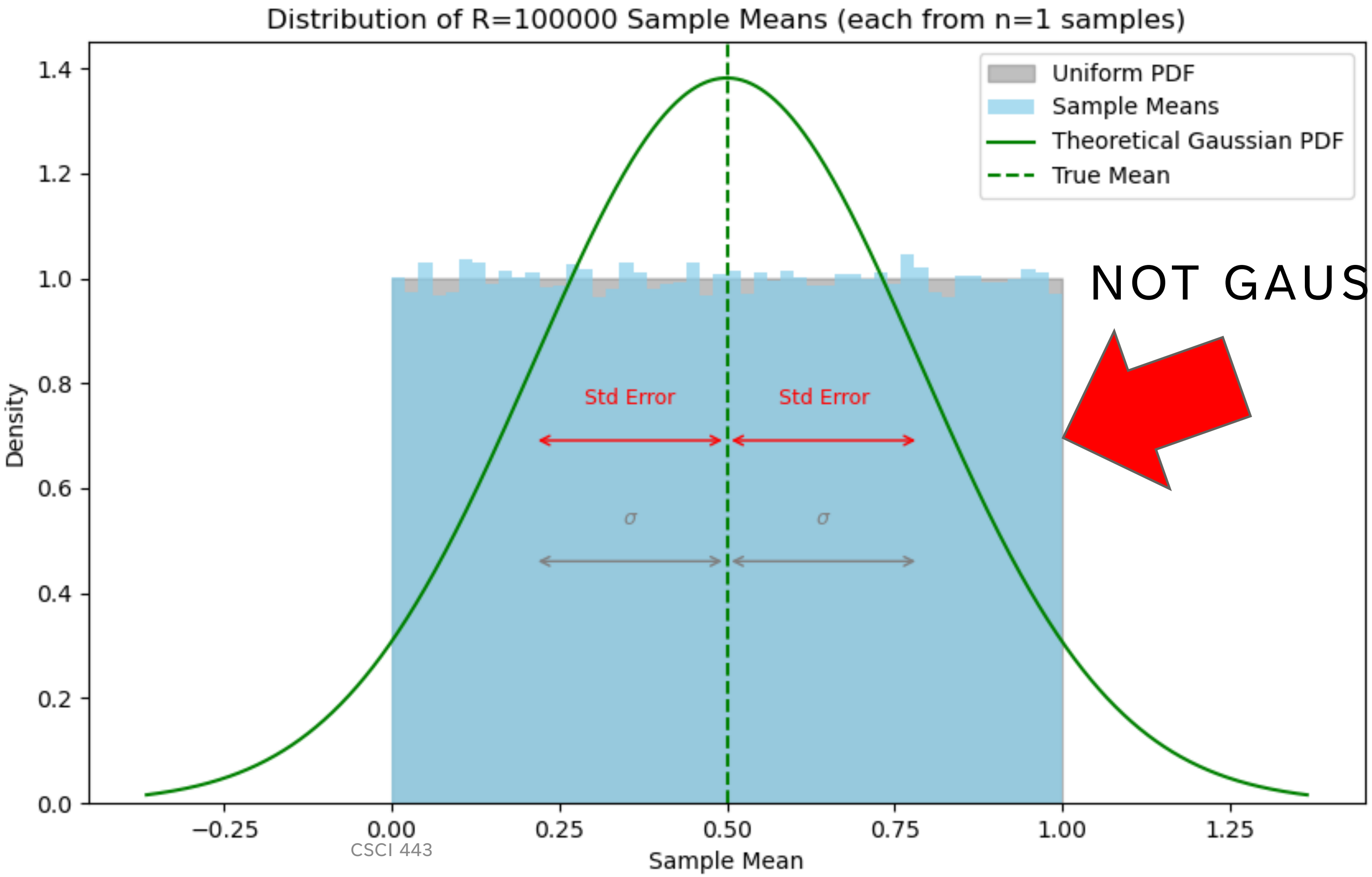


CAN I ESTABLISH A CONFIDENCE INTERVAL ON AN ESTIMATE FROM 1 SAMPLE?

Let’s consider a uniform random variable U[0,1].

Absurd.

1. The sampling distribution is not necessarily Gaussian.
2. No way to estimate the distribution standard deviation from a single sample.



NOT GAUSSIAN!

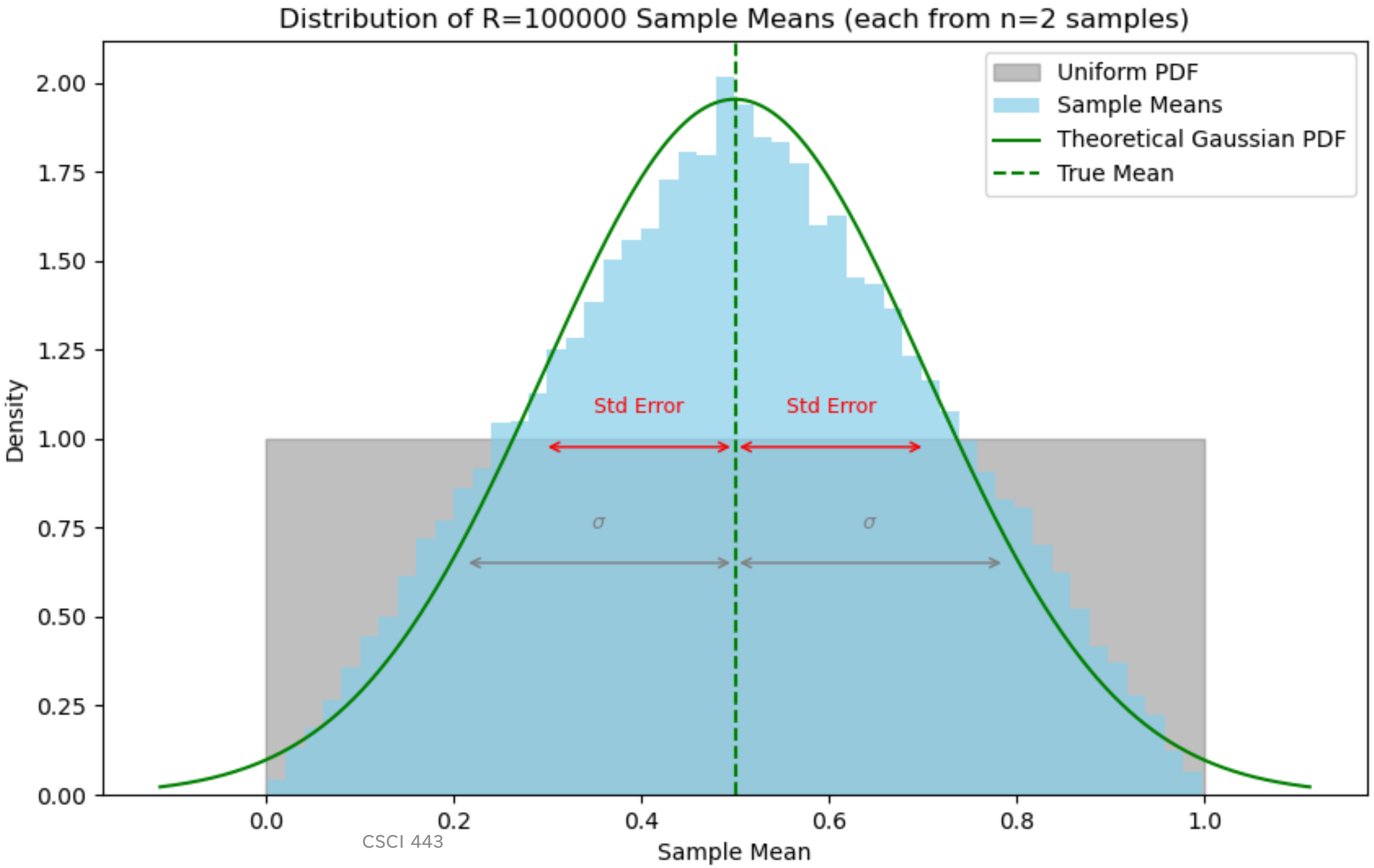
CAN I CREATE AN ESTIMATE OF THE MEAN WITH A CONFIDENCE INTERVAL FROM 2 SAMPLES?

Let's consider a uniform random variable U[0,1].

Extreme example.

1) For n = 2 sampling distribution is STILL NOT Gaussian.

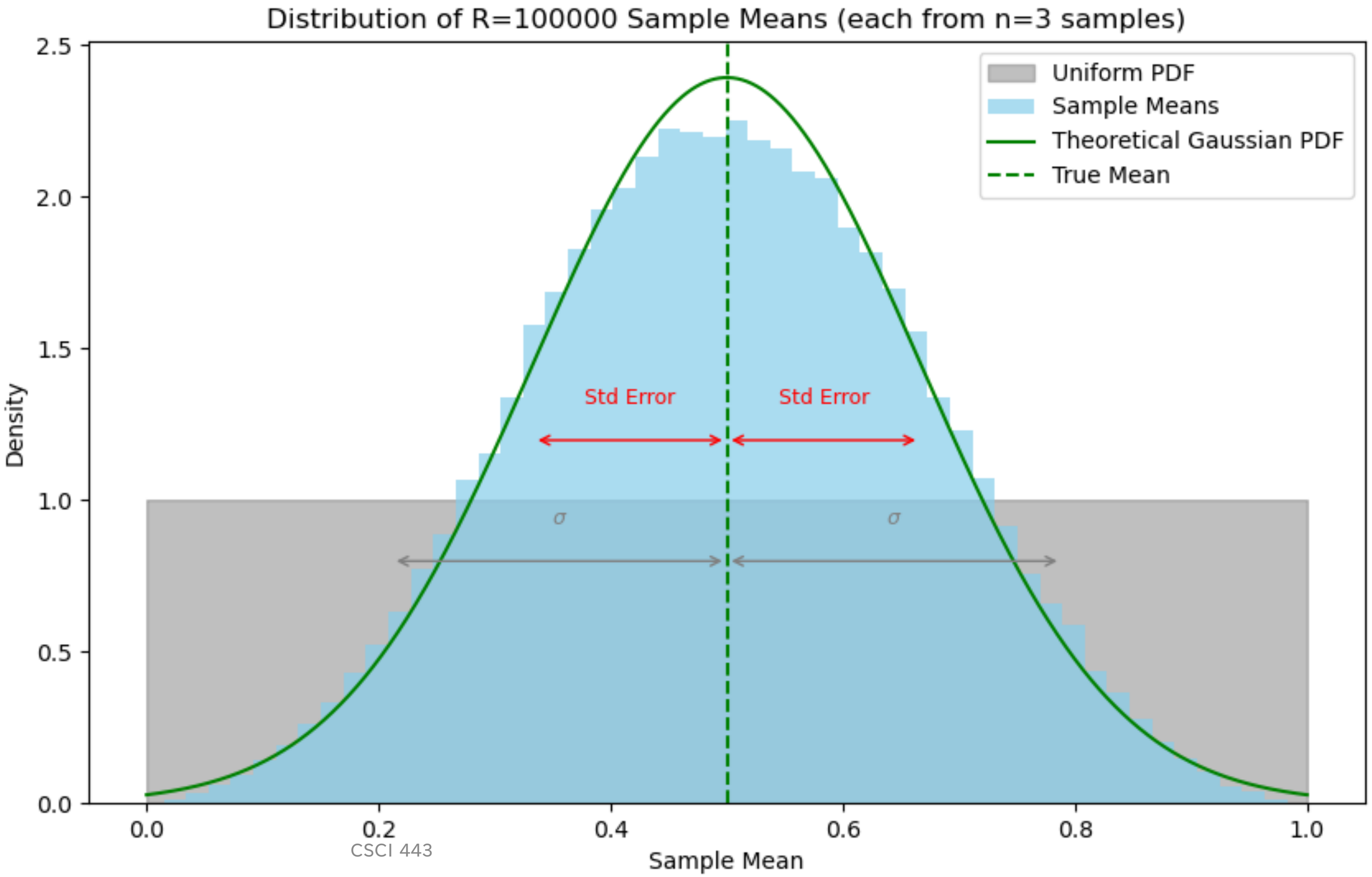
AND 2) CANNOT ACCURATELY ESTIMATE STDDEV.



FROM LECTURE 11: CONFIDENCE INTERVALS USING U[0,1]

Let's consider a uniform random variable U[0,1].

Extreme example. For n = 3, sampling distribution is STILL NOT Gaussian but much closer.



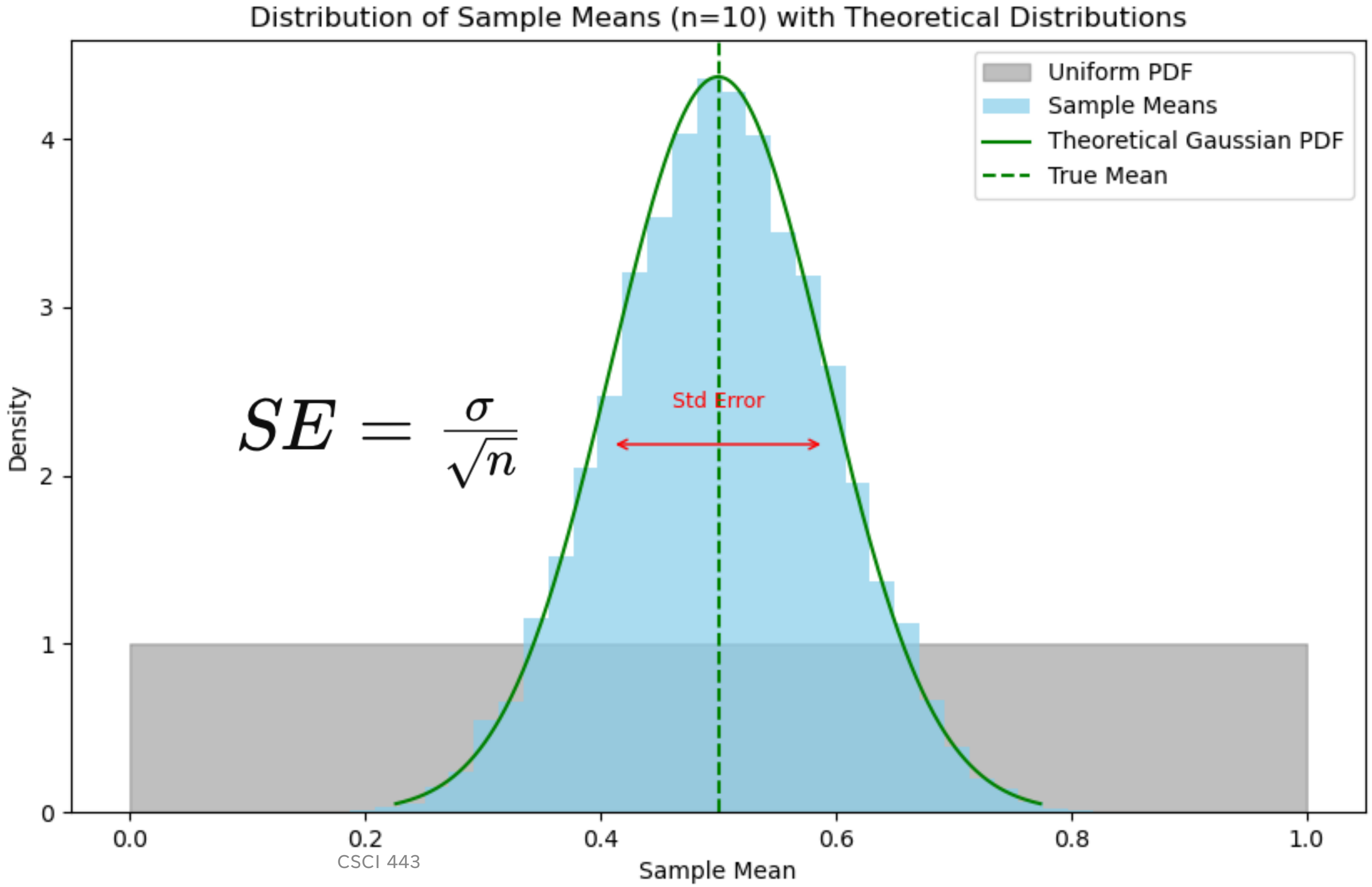
FROM LECTURE 11: SAMPLE MEAN DISTRIBUTION OF UNIFORM RV

Let’s consider a uniform random variable U[0,1].

sample means created from n=10 samples.

At n= 10, the Gaussian assumption holds pretty well for an underlying uniform distribution.

I still don’t know σ.



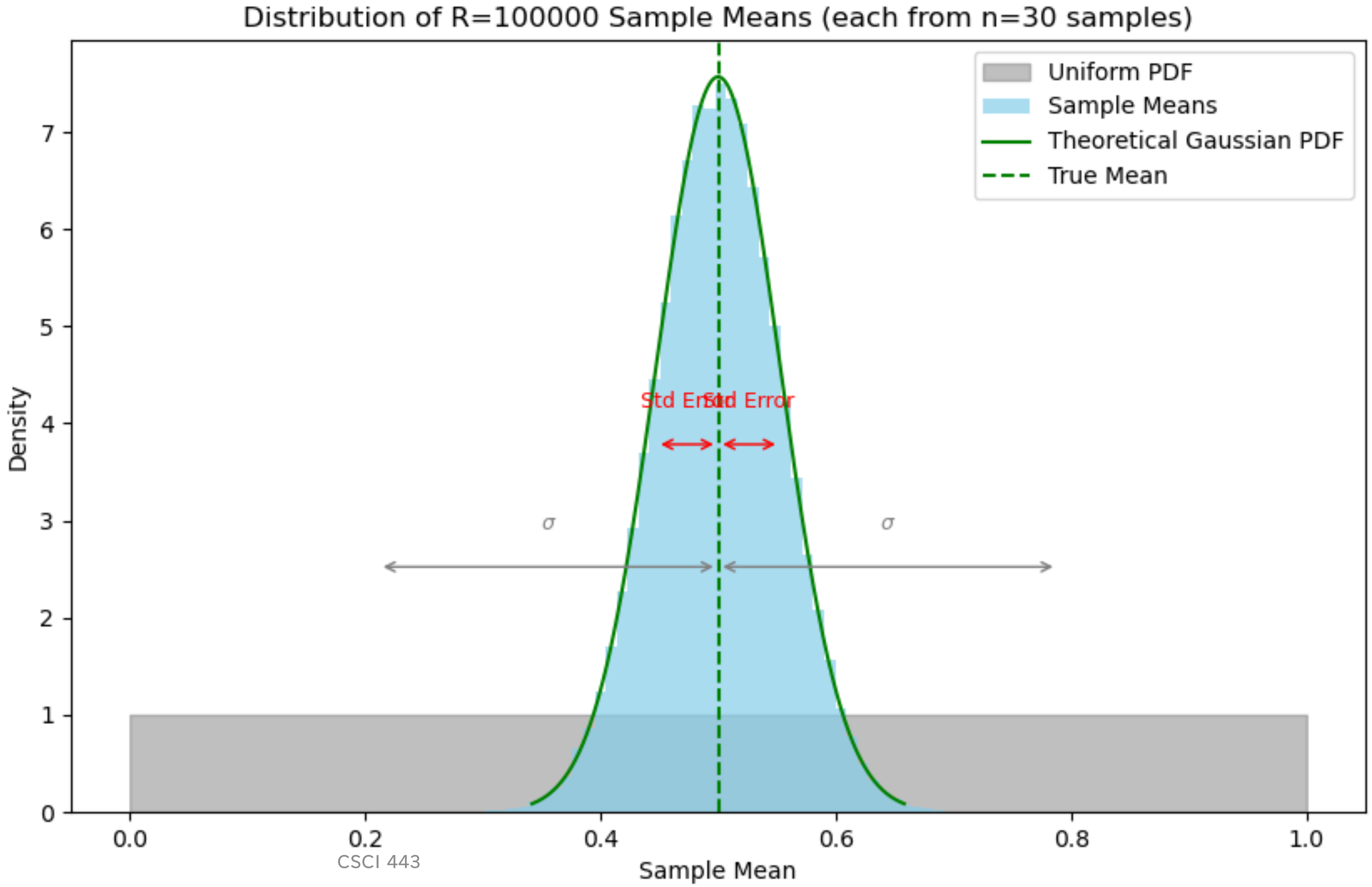
USE SAMPLE STDDEV TO COMPUTE CONFIDENCE INTERVAL?

Let's consider a uniform random variable U[0,1].

Using s_x for σ introduces its own randomness since s_x is estimated from the same random samples.

As a rule of thumb, for n>30, s_x is close enough to σ to use it to compute confidence intervals.

$$SE = \frac{\sigma}{\sqrt{n}} \approx \frac{s_x}{\sqrt{n}}$$



N=30 RULE OF THUMB

For n >= 30, assuming Gaussian sampling distribution for mean.

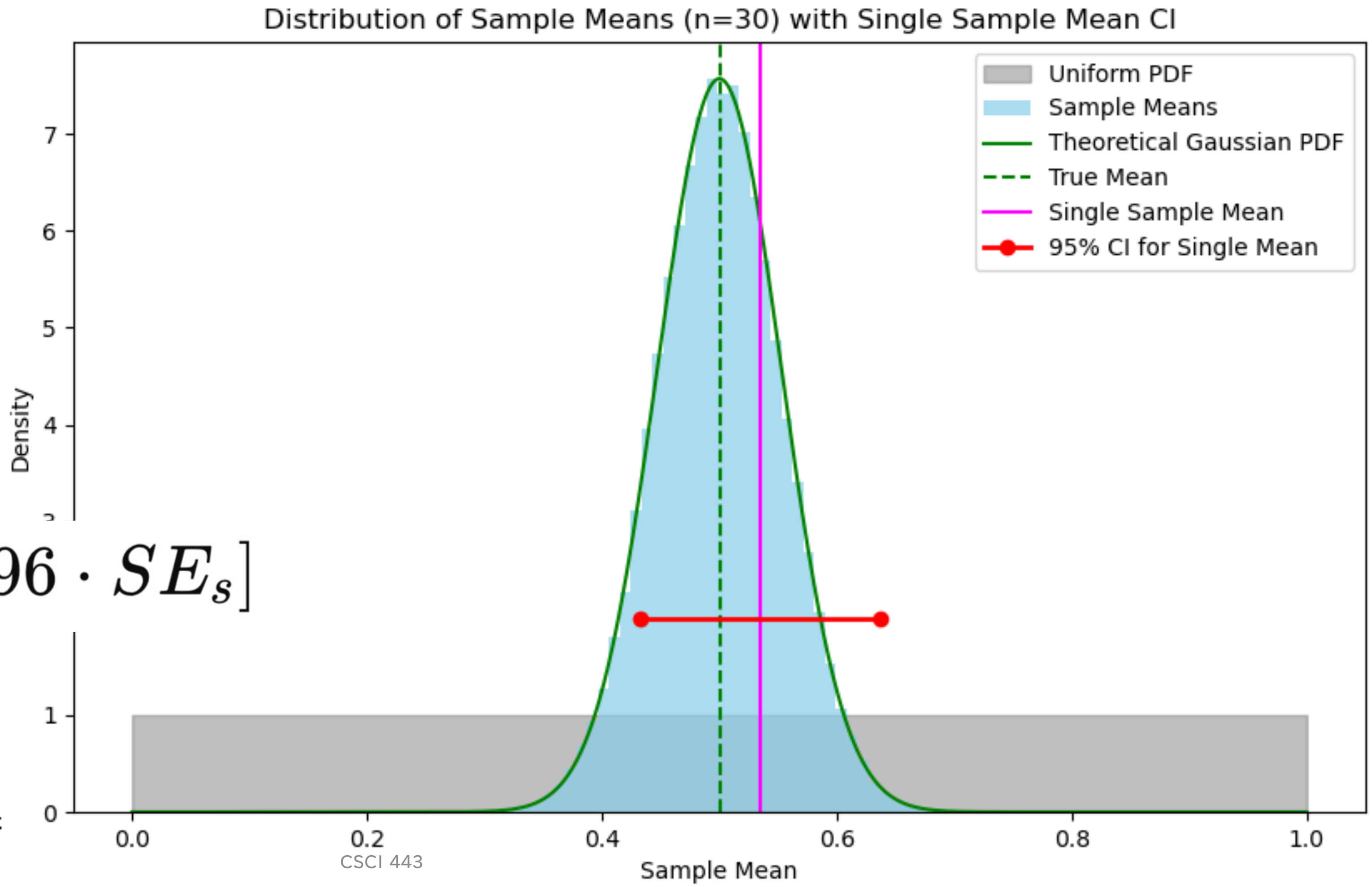
true mean falls within 1.96σ is ≅95%.

95% CI=

$$[\bar{x} - 1.96 \cdot SE_s,\ \bar{x} + 1.96 \cdot SE_s]$$

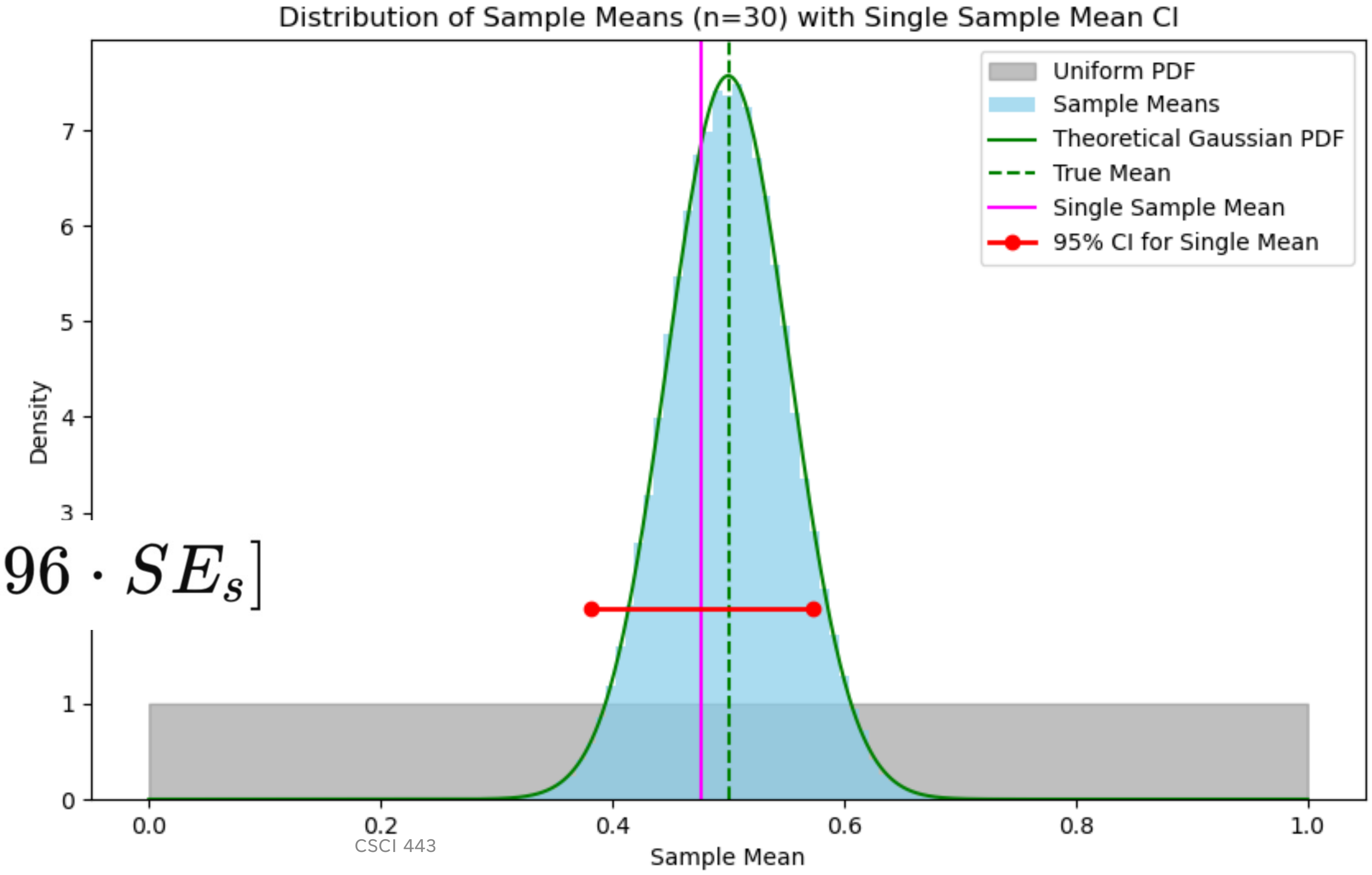
where

$$SE_s = \frac{s_x}{\sqrt{n}} = \frac{s_x}{\sqrt{30}}$$



N=30 RULE OF THUMB

For n >= 30, assuming Gaussian sampling distribution for mean.



95% CI =

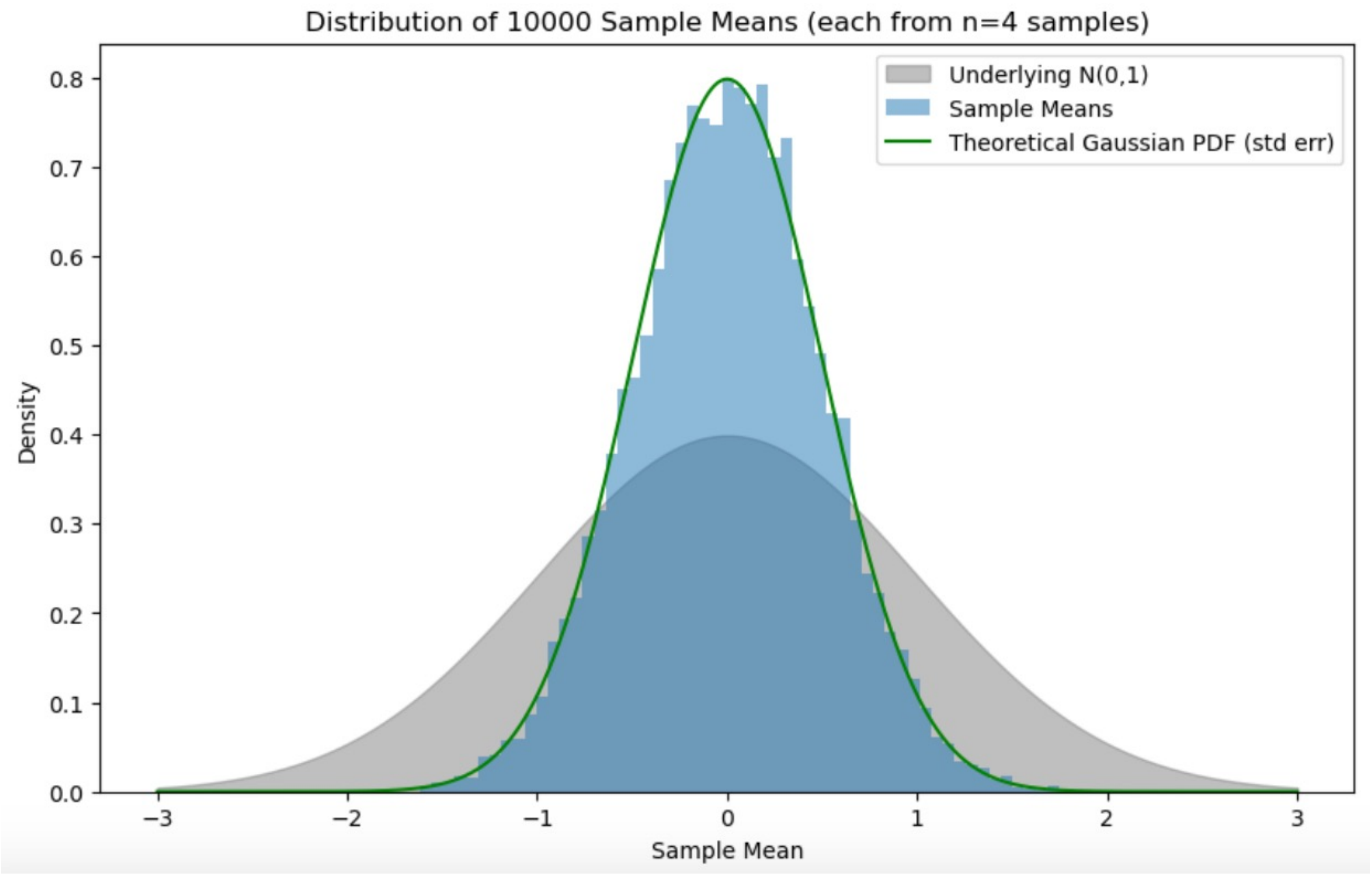
$$[\bar{x} - 1.96 \cdot SE_s,\ \bar{x} + 1.96 \cdot SE_s]$$

where

$$SE_s = \frac{s_x}{\sqrt{n}}$$

WHAT IF UNDERLYING DISTRIBUTION IS GAUSSIAN?

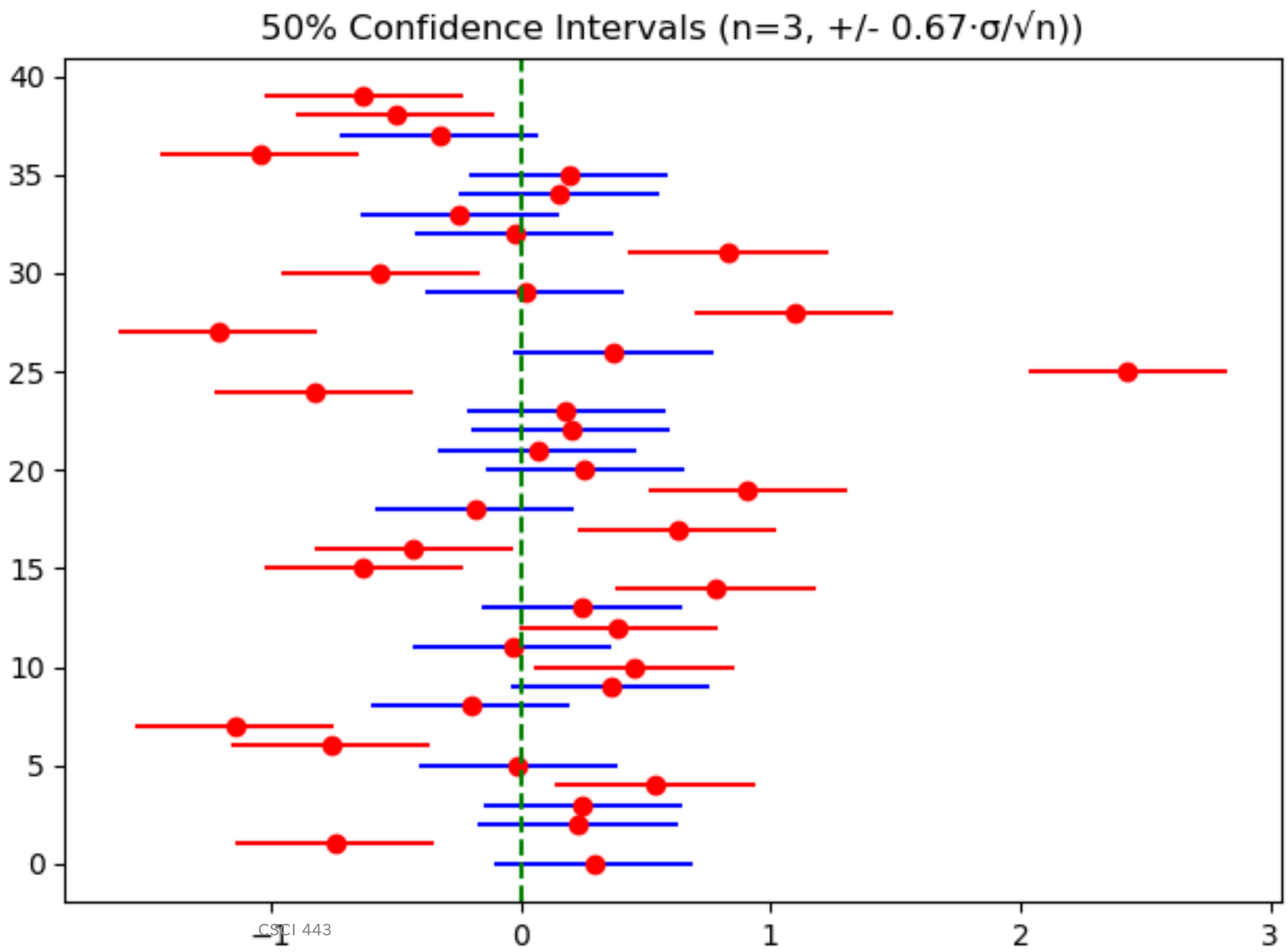
Sampling distribution looks Gaussian for any n≥1.



WHAT IF I KNOW σ?

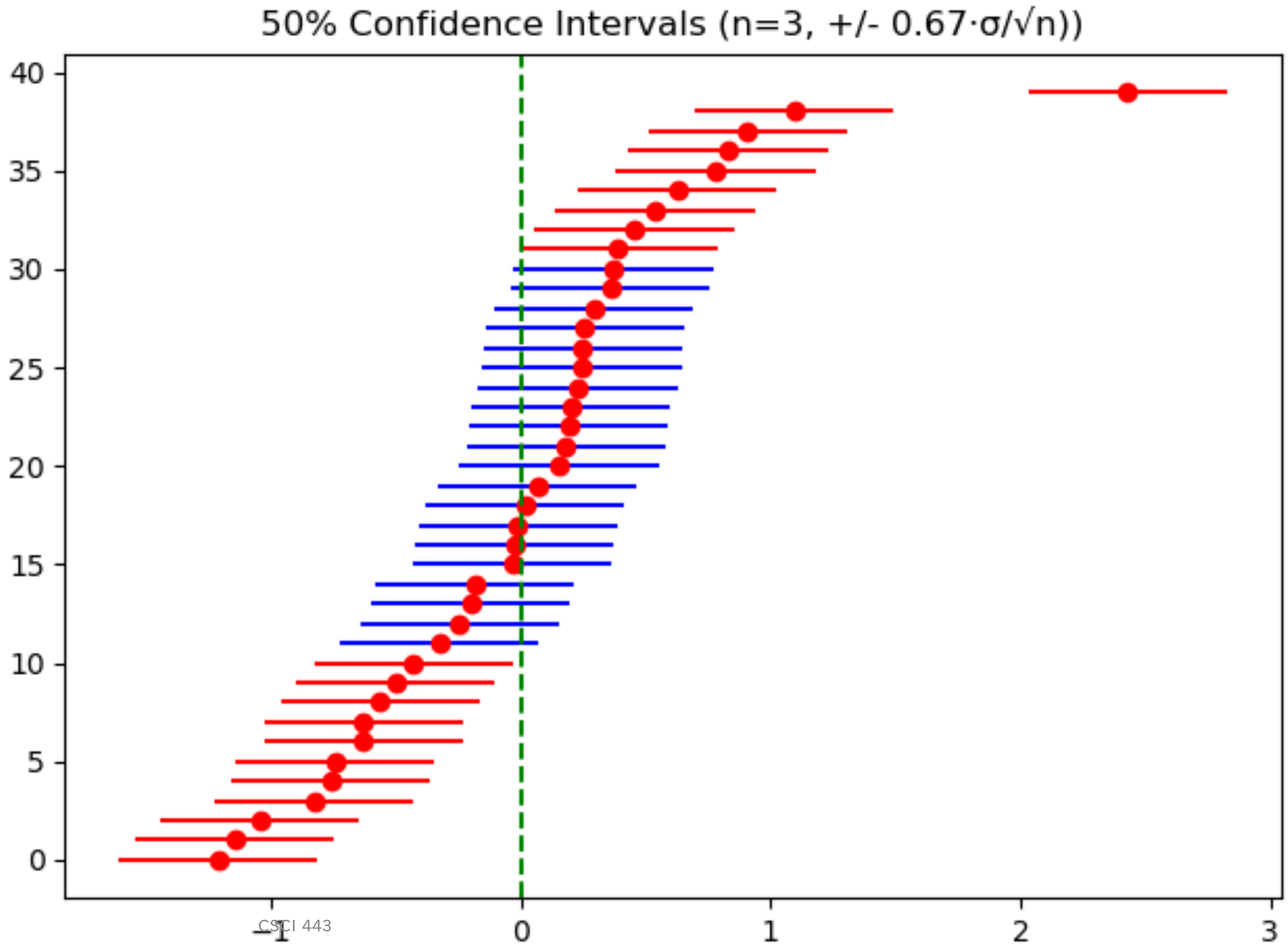
Compute 50% Confidence Interval as

$$\bar{x} \pm 0.67\frac{\sigma}{\sqrt{n}}$$



WHAT IF I KNOW σ?

Order CIs for clarity

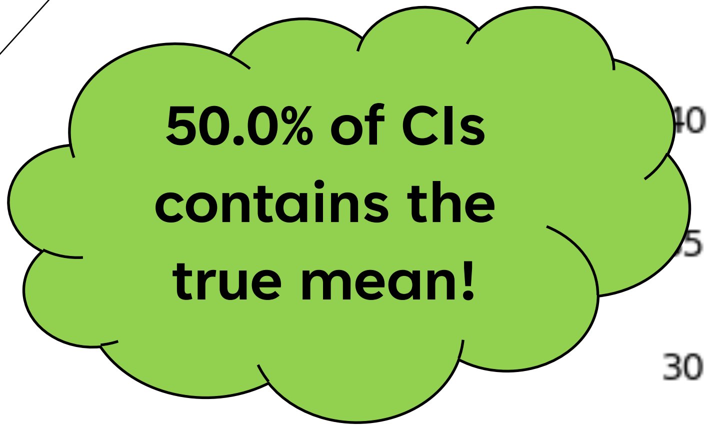


Compute 50% Confidence Interval as

$$\bar{x} \pm 0.67 \frac{\sigma}{\sqrt{n}}$$

Because σ is known, all intervals have equal length.

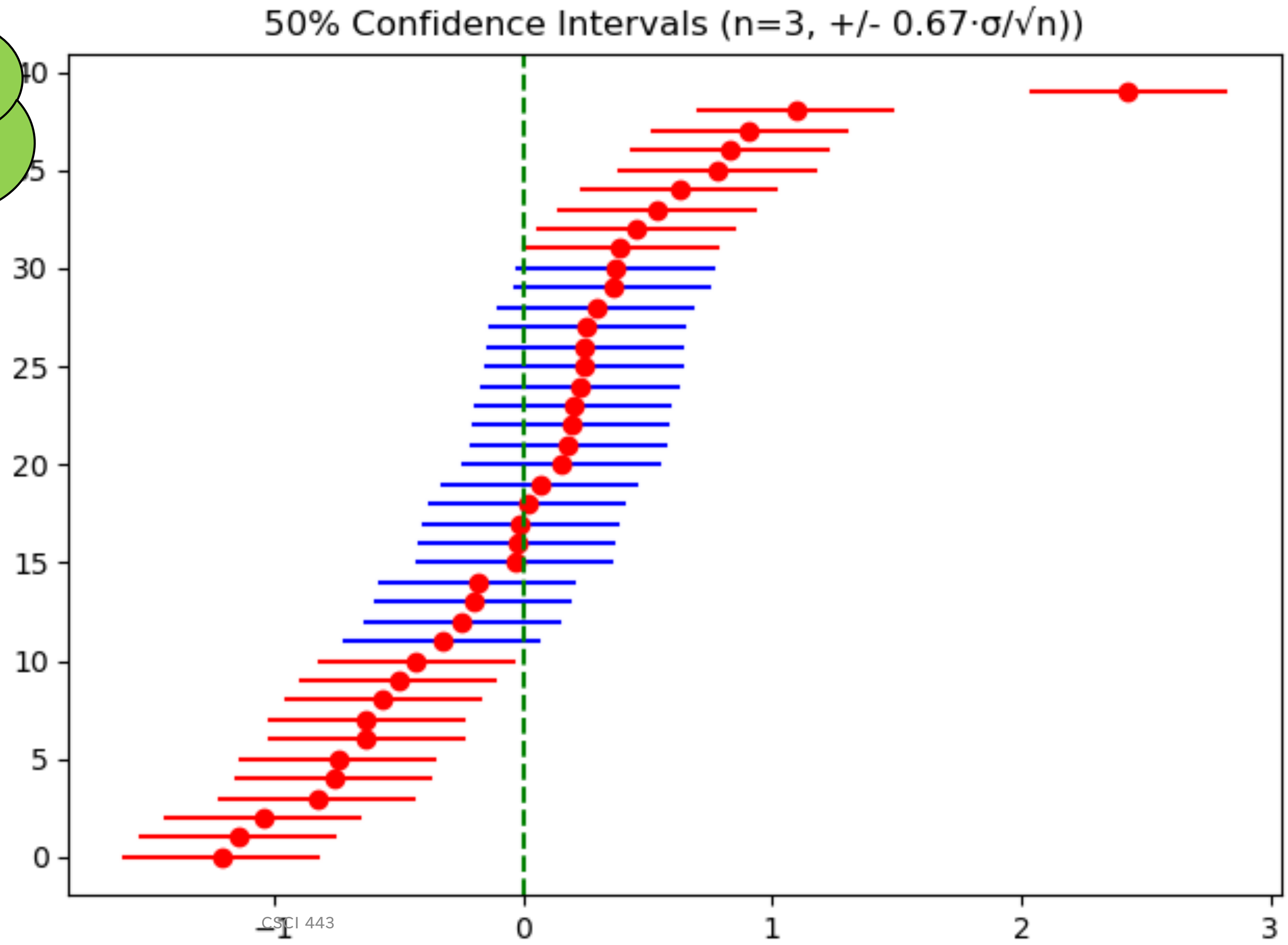
WHAT IF I KNOW σ?



Compute 50% Confidence Interval as

$$\bar{x} \pm 0.67 \frac{\sigma}{\sqrt{n}}$$

Because σ is known, all intervals have equal length.



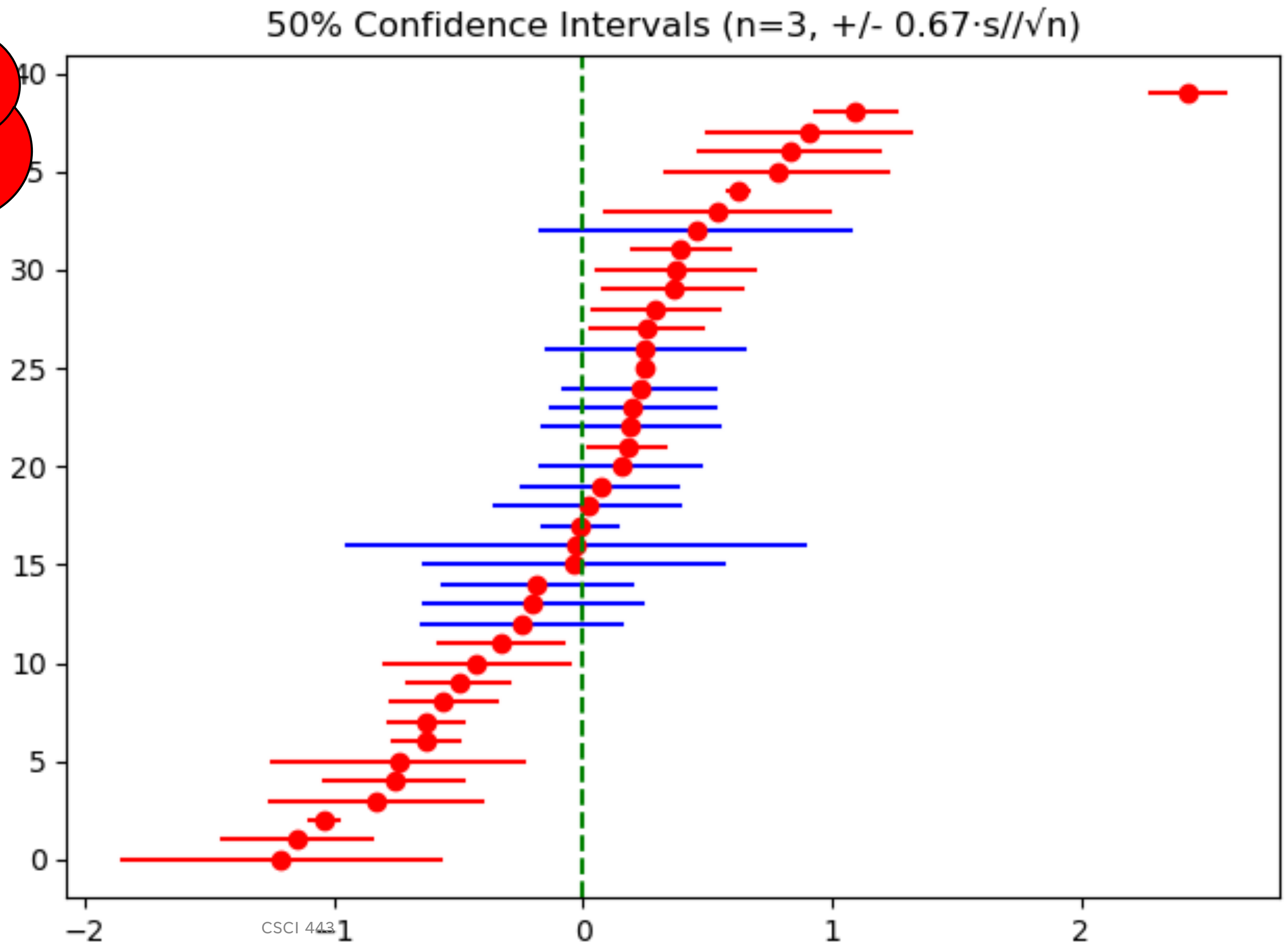
WHAT IF I DON'T KNOW σ AND USE s_x INSTEAD?

Only 35% of Cis now contain the true means!

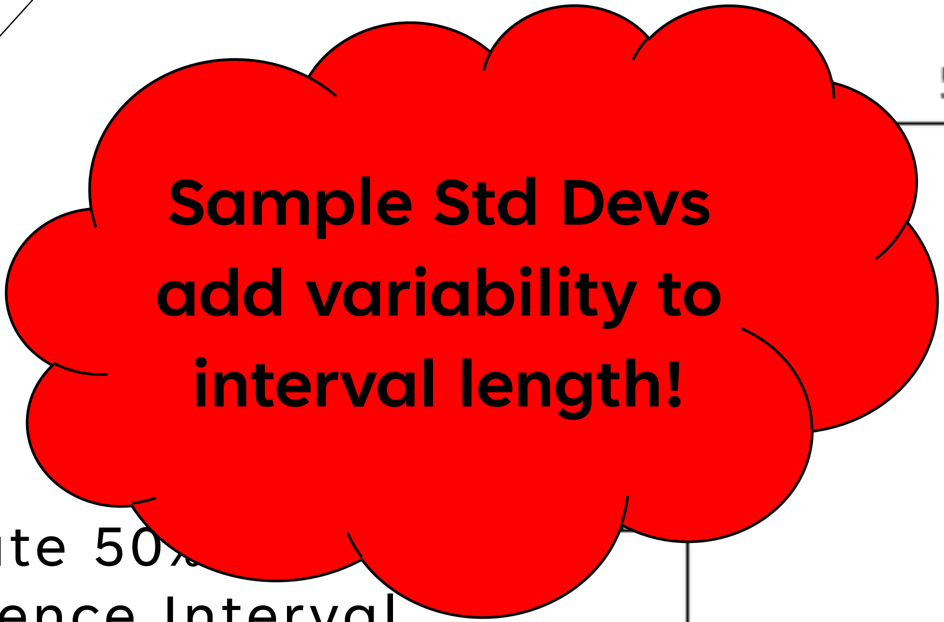
Compute 50% Confidence Interval as

$$\bar{x} \pm 0.67 \frac{\sigma}{\sqrt{n}}$$

$$\bar{x} \pm 0.67 \frac{s_x}{\sqrt{n}}$$



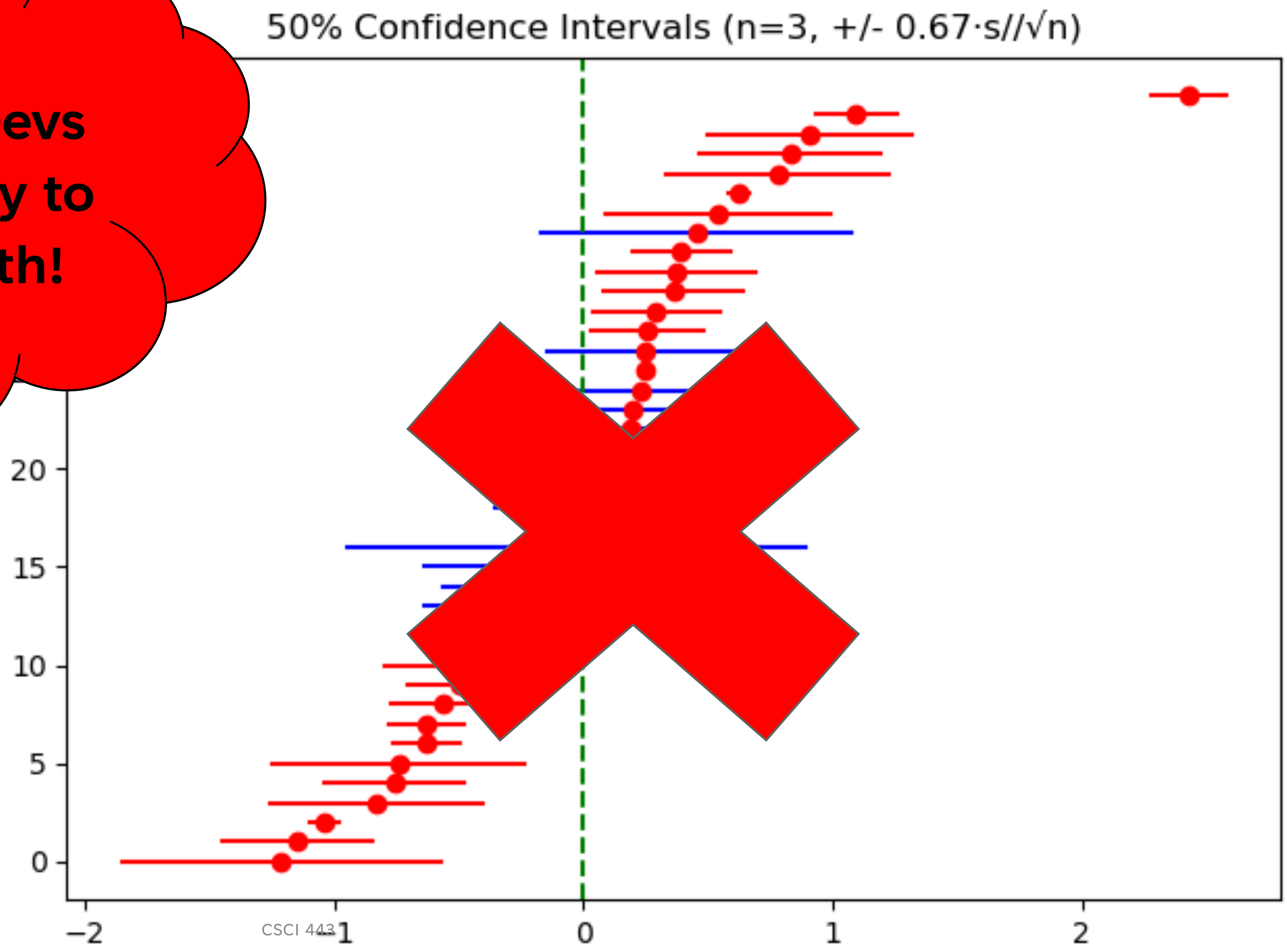
WHAT IF I DON'T KNOW σ AND USE s_x INSTEAD?



Compute 50% Confidence Interval as

$$\bar{x} \pm 0.67 \frac{\sigma}{\sqrt{n}}$$

$$\bar{x} \pm 0.67 \frac{s_x}{\sqrt{n}}$$

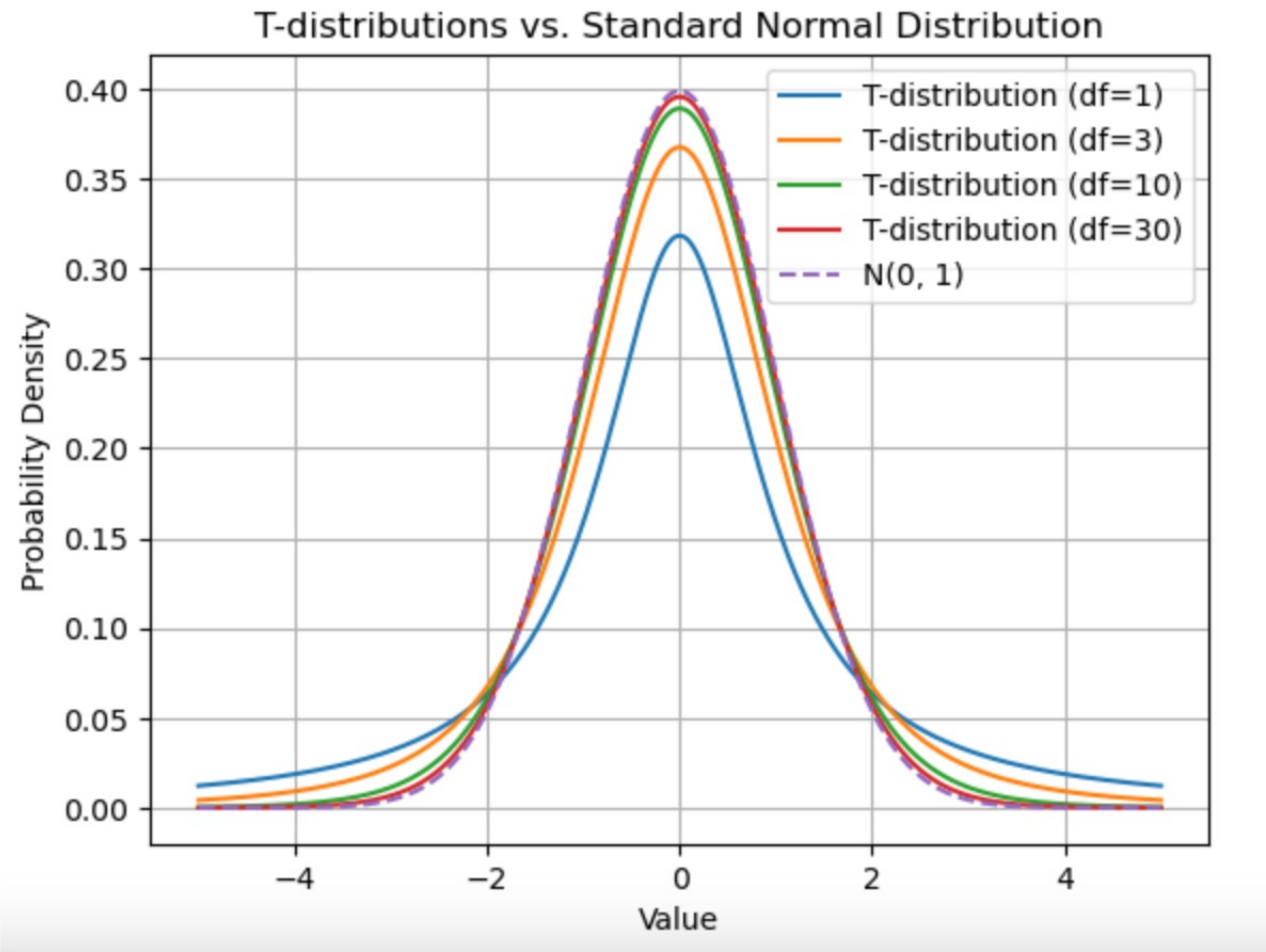


HOW DO I ADJUST CONFIDENCE INTERVALS TO USE s_x?

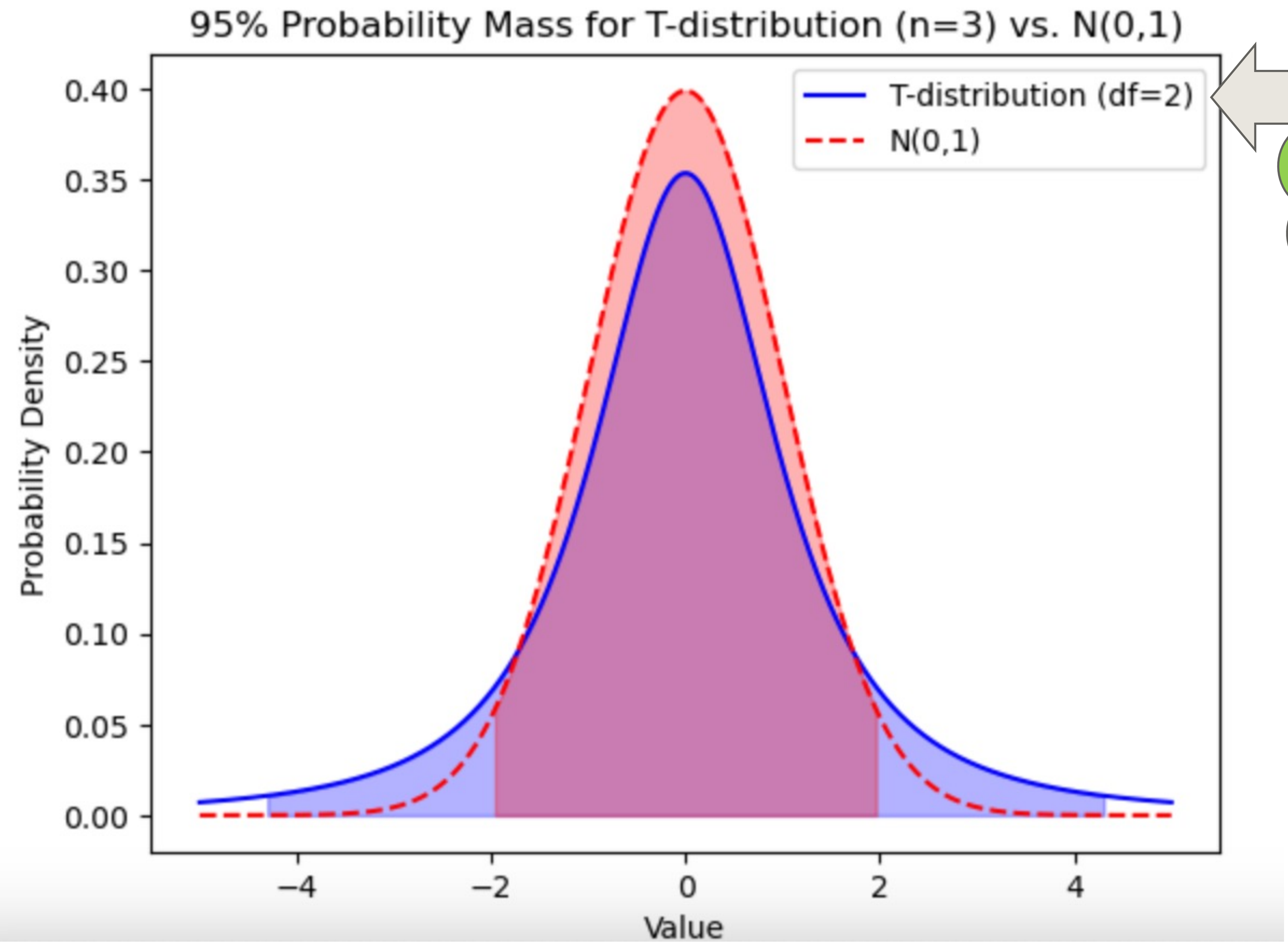
We need a distribution that takes into account the additional randomness introduced by estimating the standard error.

$$\bar{x} \pm 0.67 \frac{s_x}{\sqrt{n}}$$

STUDENT'S T-DISTRIBUTION



STUDENT'S T-DISTRIBUTION



df → degrees of freedom df=n-1=3-1=2

I commented in class that I wasn't sure that I had set degrees of freedom correctly.

It is correct.

LET'S INTRODUCE A NOTATION THAT ALLOWS US TO TALK ABOUT Z-SCORES FOR PARTICULAR CONFIDENCE LEVELS

The 0.67 refers to the critical Z score for a 50% confidence interval. We could say z_{50}=0.67, z_{90}=1.96, z_{99} = 2.58. Let's denote z_p to mean the z-score needed to a *p* confidence level.

$$\bar{x} \pm z_{50} \frac{s_x}{\sqrt{n}} = \bar{x} \pm 0.67 \frac{s_x}{\sqrt{n}}$$

$$\bar{x} \pm z_{95} \frac{s_x}{\sqrt{n}} = \bar{x} \pm 1.96 \frac{s_x}{\sqrt{n}}$$

NOW AN ADJUSTMENT FOR THE VARIABILITY INTRODUCED BY s_x

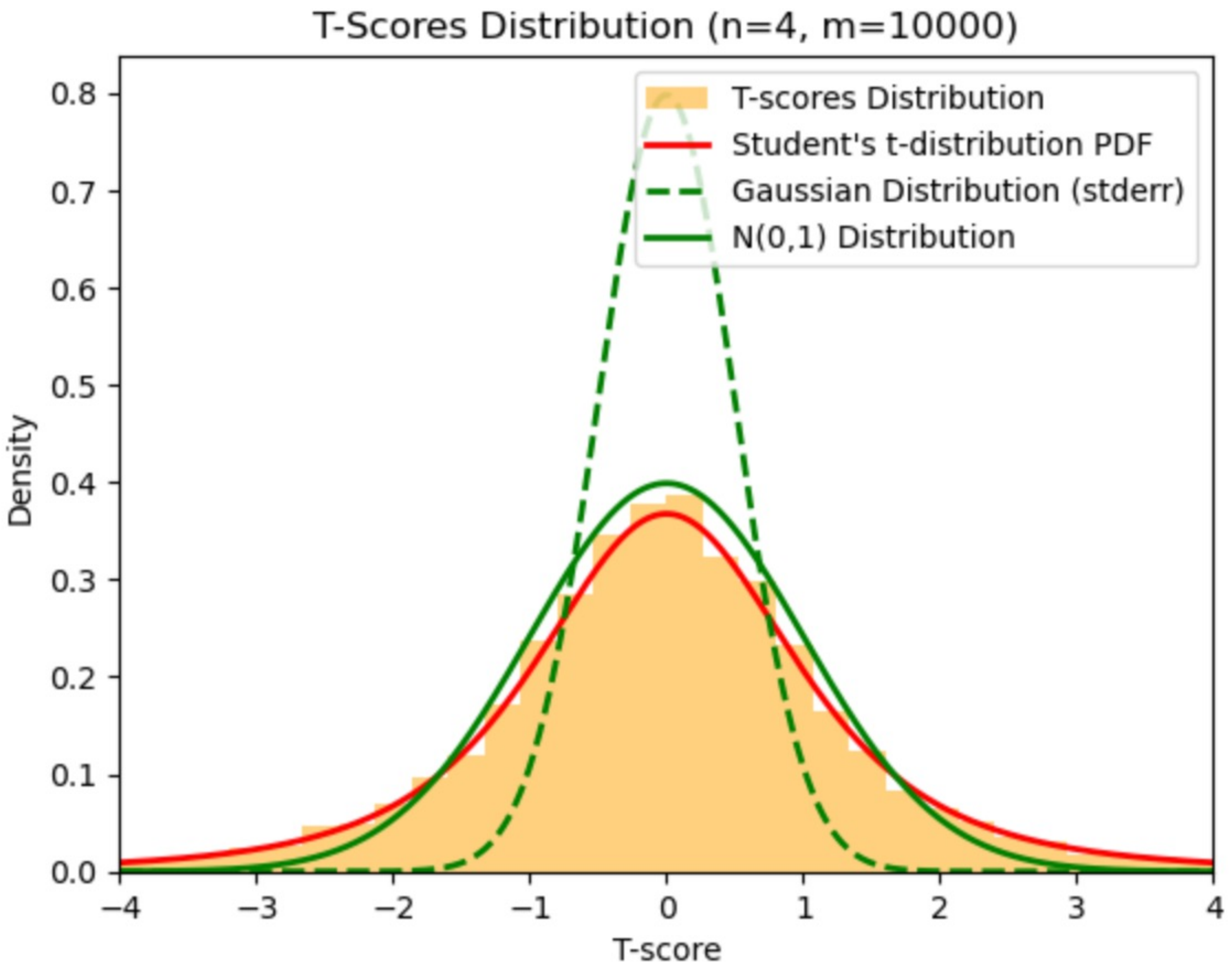
Instead of z-score, we now use a t-score. But t-score is a function of n. So, for n = 3,

$$\bar{x} \pm t_{95} \frac{s_x}{\sqrt{n}} = \bar{x} \pm 4.30 \frac{s_x}{\sqrt{n}} = \bar{x} \pm 4.30 \frac{s_x}{\sqrt{3}}$$

$$\bar{x} \pm z_{95} \frac{s_x}{\sqrt{n}} = \bar{x} \pm 1.96 \frac{s_x}{\sqrt{n}}$$

Include the CI using z-scores for comparison

STUDENT'S T-DISTRIBUTION



STUDENT'S T-DISTRIBUTION

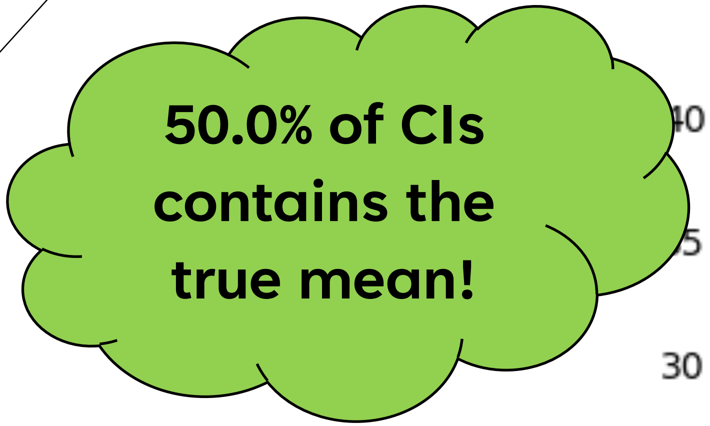
$$f(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}}$$

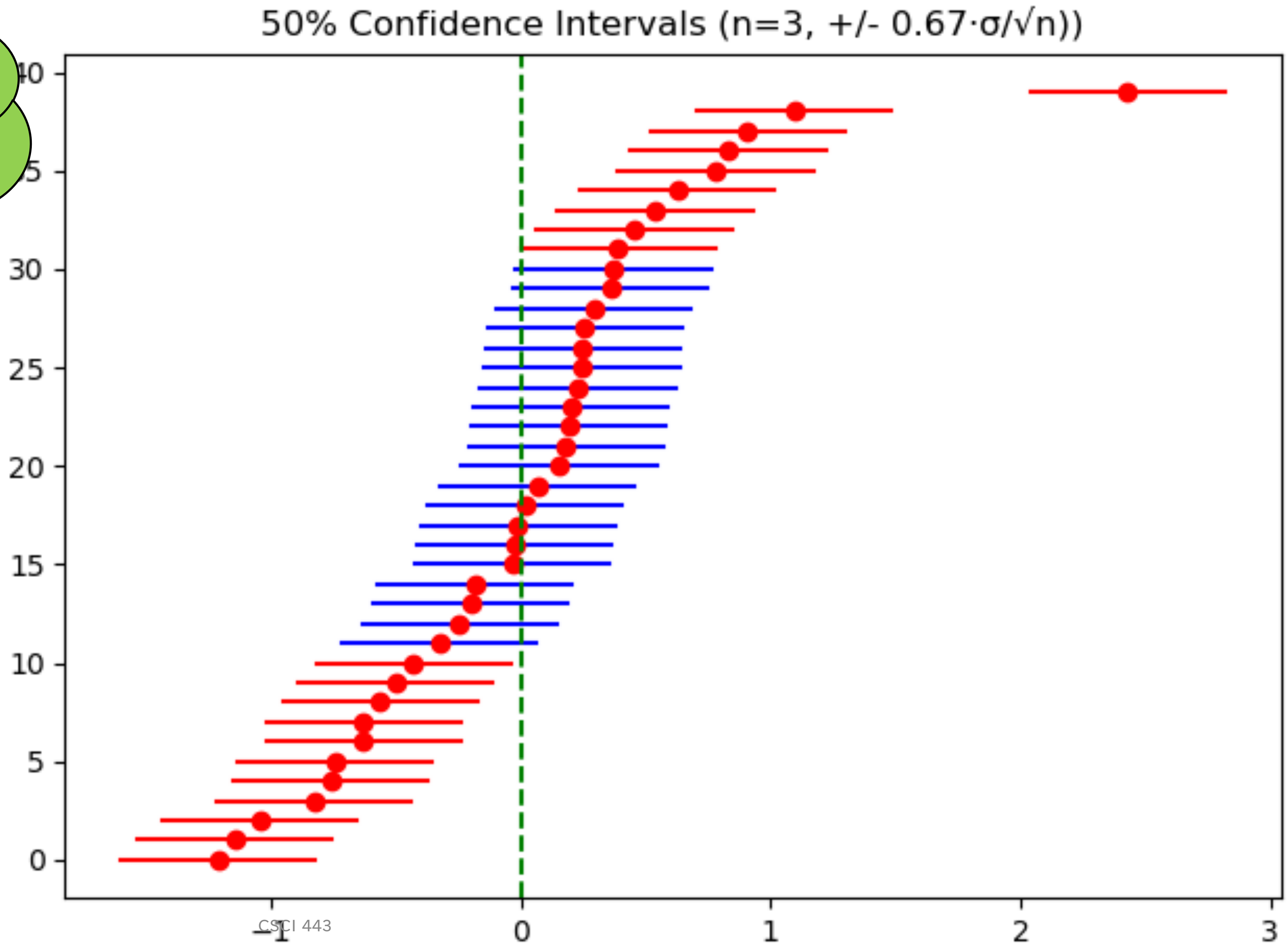
Where *n* is the number of samples.

We won't use this function directly. As with Gaussian PDF, we use a table or a computer to evaluate.

•

I WE KNOW σ, BUT WE USUALLY DON'T.





Compute 50% Confidence Interval as

$$\bar{x} \pm 0.67 \frac{\sigma}{\sqrt{n}}$$

Because σ is known, all intervals have equal length.

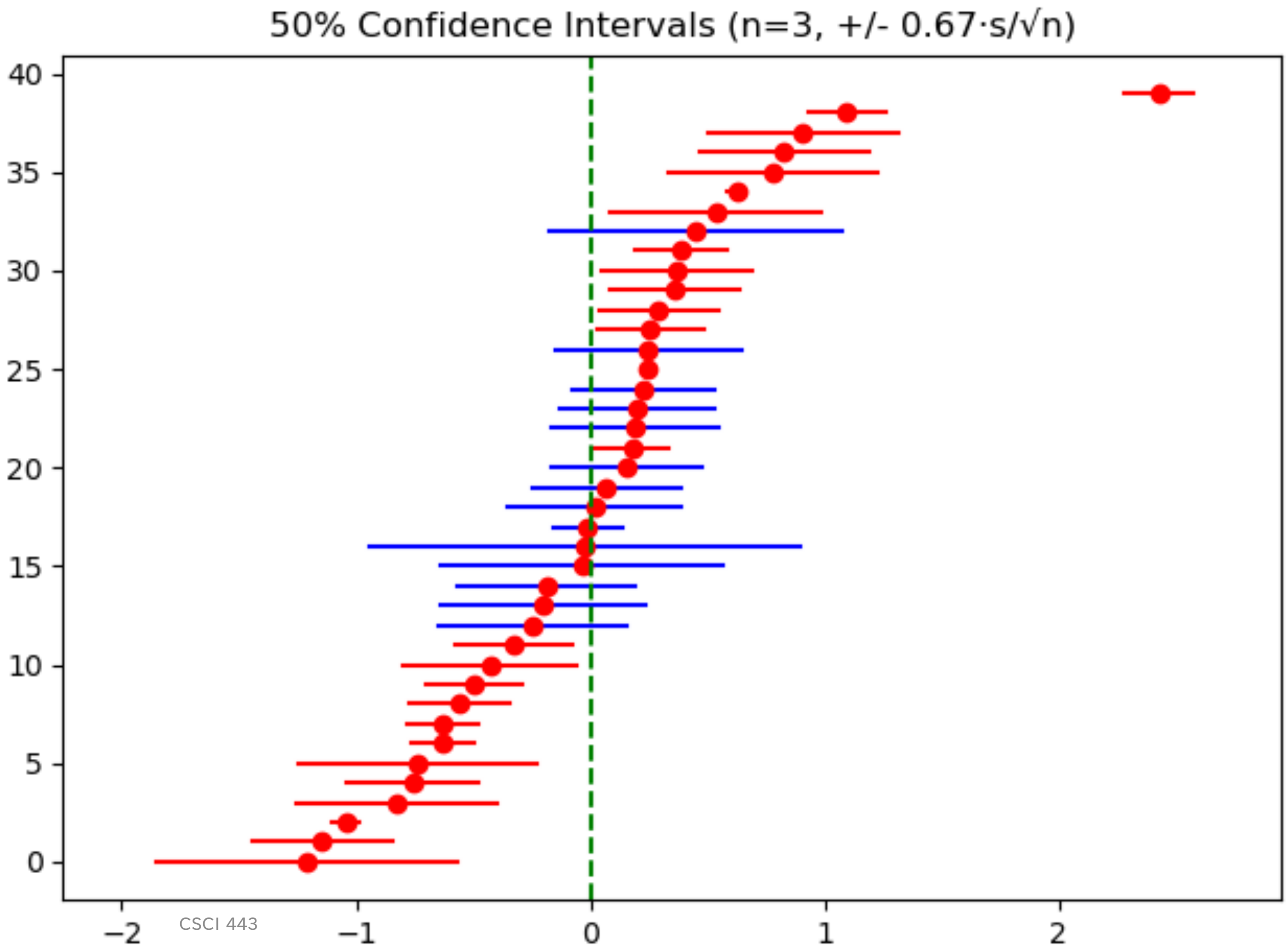
BUT IF WE USE s_x…

Only 35% of Cis now contain the true means!

Compute 50% Confidence Interval for n=3 as

$$\bar{x} \pm 0.67 \frac{\sigma}{\sqrt{n}}$$

$$\bar{x} \pm 0.67 \frac{s_x}{\sqrt{n}}$$



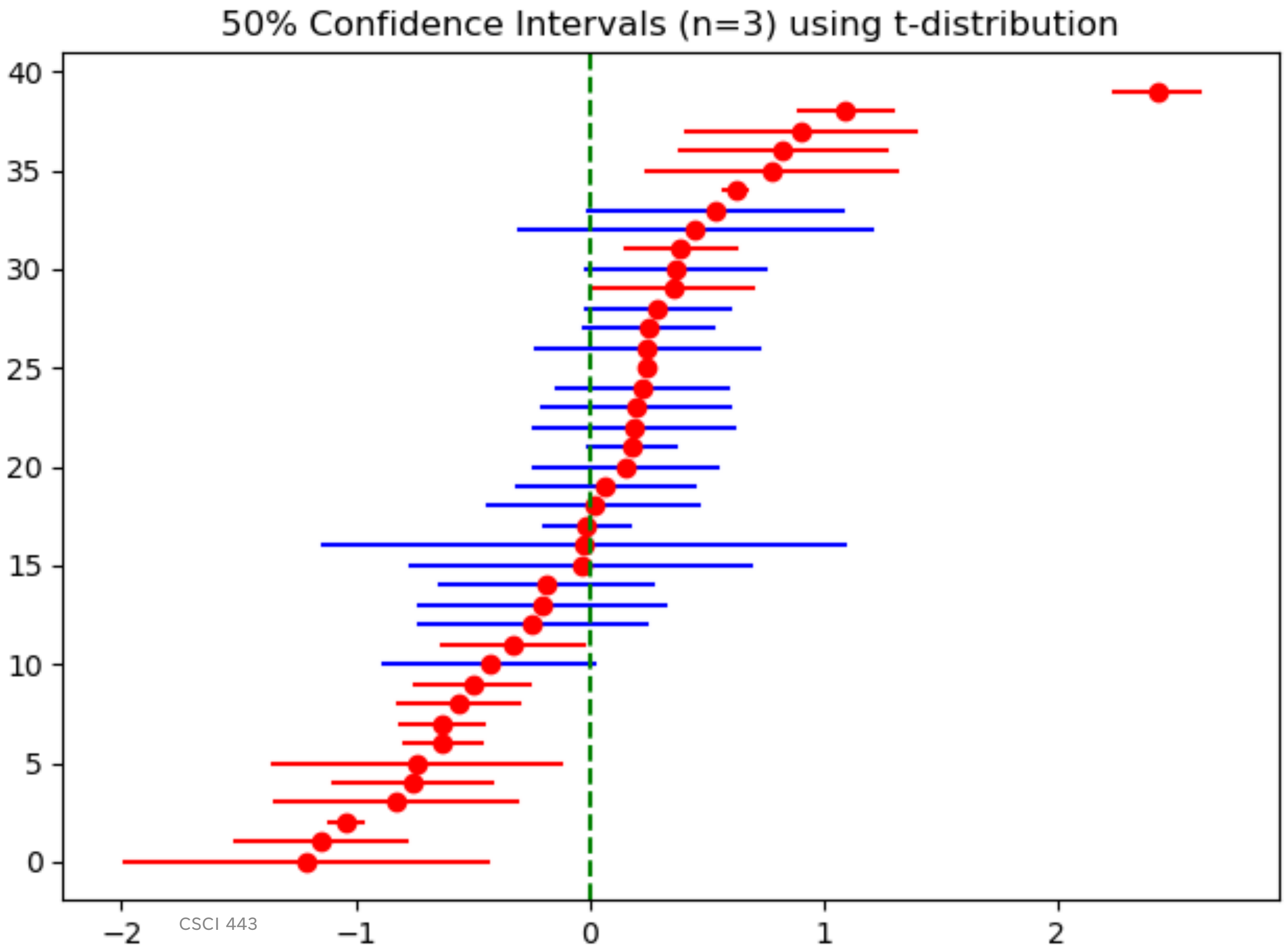
BUT IF WE USE T-SCORES...

50.0% of CIs contain the true mean!

Compute 50% Confidence Interval for n=3 as

$$\bar{x} \pm t_{50} \frac{s_x}{\sqrt{n}}$$

$$\bar{x} \pm 0.82 \frac{s_x}{\sqrt{3}}$$



COMBINE WITH N=30 RULE OF THUMB

Compute α Confidence Interval (CI) for n samples

$$\bar{x} \pm t_p \frac{s_x}{\sqrt{n}}$$ if n < 30

$$\bar{x} \pm Z_p \frac{s_x}{\sqrt{n}}$$ if n ≥ 30

COMPUTING GAUSSIAN CONFIDENCE INTERVALS IN PYTHON

Gaussian confidence interval

$$\bar{x} \pm Z_p \frac{s_x}{\sqrt{n}} \qquad \text{if } n \geq 30$$

p is the confidence level. *p* can either be expressed as a percentage or as a probability in [0,1]. Z_{95} refers to the 95% confidence level.

$$Z_p = \Phi^{-1}\left(1 - \frac{1 - \frac{p}{100}}{2}\right)$$

Φ^{-1} is the inverse Gaussian Cumulative Distribution Function (CDF).

The inverse CDF is also called the *Percent Point Function (PPF).*

COMPUTING GAUSSIAN CONFIDENCE INTERVALS IN PYTHON (2)

Gaussian confidence interval

$$Z_p = \Phi^{-1}\left(1 - \frac{1 - \frac{p}{100}}{2}\right)$$

Φ^{-1} is the inverse Gaussian Cumulative Distribution Function (CDF), a.k.a., PPF.

In Python

```
from scipy.stats import norm
```

```
...
p = confidence_level / 100
z_p = norm.ppf((1 + p) / 2)
```

COMPUTING GAUSSIAN CONFIDENCE INTERVALS IN PYTHON (3)

Gaussian confidence interval

```
def compute_gaussian_confidence_interval(samples, confidence_level):
    sample_mean = np.mean(samples)                       # (1)
    n = len(samples)                                     # (2)
    sample_std = np.std(samples, ddof=1)                 # (3)
    stderr = sample_std  / sqrt(n)                       # (4)

    p = confidence_level / 100
    z_p = norm.ppf((1 + p) / 2)

    lower_bound = sample_mean - z_p * stderr
    upper_bound = sample_mean + z_p * stderr
    return sample_mean, lower_bound, upper_bound
```

COMPUTING CONFIDENCE INTERVALS IN PYTHON (3)

Gaussian confidence interval

```
def compute_gaussian_confidence_interval(samples, confidence_level):
    sample_mean = np.mean(samples)                     # (1)
    n = len(samples)                                   # (2)
    sample_std = np.std(samples, ddof=1)               # (3)
    stderr = sample_std  / sqrt(n)                     # (4)

    p = confidence_level / 100
    z_p = norm.ppf((1 + p) / 2)

    lower_bound = sample_mean - z_p * stderr
    upper_bound = sample_mean + z_p * stderr
    return sample_mean, lower_bound, upper_bound
```

COMPUTING t-DISTRIBUTION CONFIDENCE INTERVALS IN PYTHON (1)

t confidence interval

$$\bar{x} \pm t_p \frac{s_x}{\sqrt{n}}$$

We compute t_p in Python with

```
p = confidence_level / 100
df = n - 1  # Degrees of freedom for t-distribution, n-1
t_p = t.ppf((1 + p) / 2, df)
```

COMPUTING *t*-DISTRIBUTION CONFIDENCE INTERVALS IN PYTHON (2)

t confidence interval

```
def compute_t_confidence_interval(samples, confidence_level):
    sample_mean = np.mean(samples)                       # (1)
    n = len(samples)                                     # (2)
    sample_std = np.std(samples, ddof=1)                 # (3)
    stderr = sample_std  / sqrt(n)                       # (4)

    # Calculate critical t-value for {confidence_level}% CI using t-distribution
    p = confidence_level / 100
    df = n - 1  # Degrees of freedom for t-distribution, n-1
    t_p = t.ppf((1 + p) / 2, df)

    lower_bound = sample_mean - t_p * stderr
    upper_bound = sample_mean + t_p * stderr
    return sample_mean, lower_bound, upper_bound
```

THANK YOU

David Harrison

Harrison@cs.olemiss.edu