CSCI 443: LECTURE 11
CONFIDENCE

Professor David Harrison

OFFICE HOURS

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

HOMEWORK 4

Will be handed out after break.

DATES OF INTEREST

March 4	Progress Reports
March 8	Deadline for Withdrawal
March 9-17	Spring Break
March 19	Homework 4 handed out
March 28	Homework 4 due

BLACKBOARD & GITHUB

Slides up through lecture 9 on blackboard.

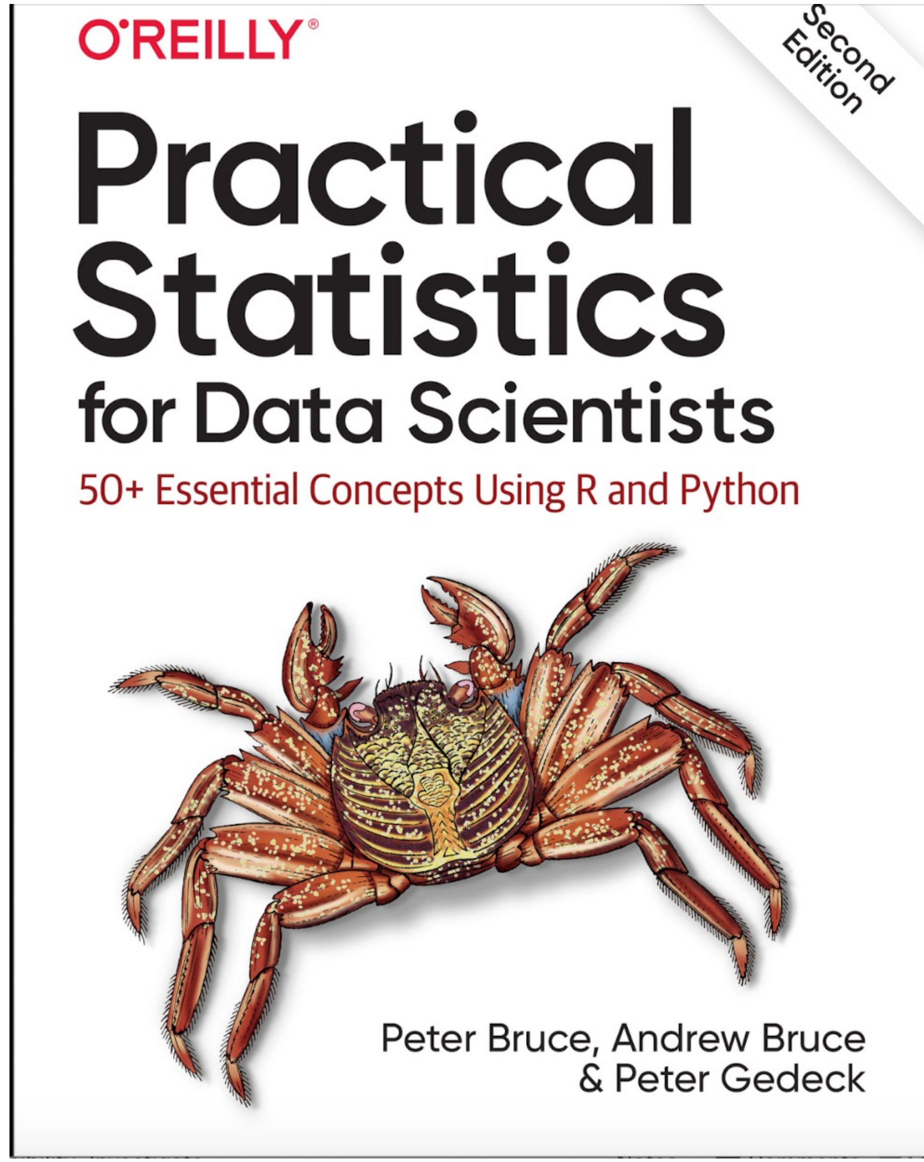
Lecture slides and examples committed to GitHub also up through lecture 9.

The project is at

https://github.com/dosirrah/CSCI443_AdvancedDataScience

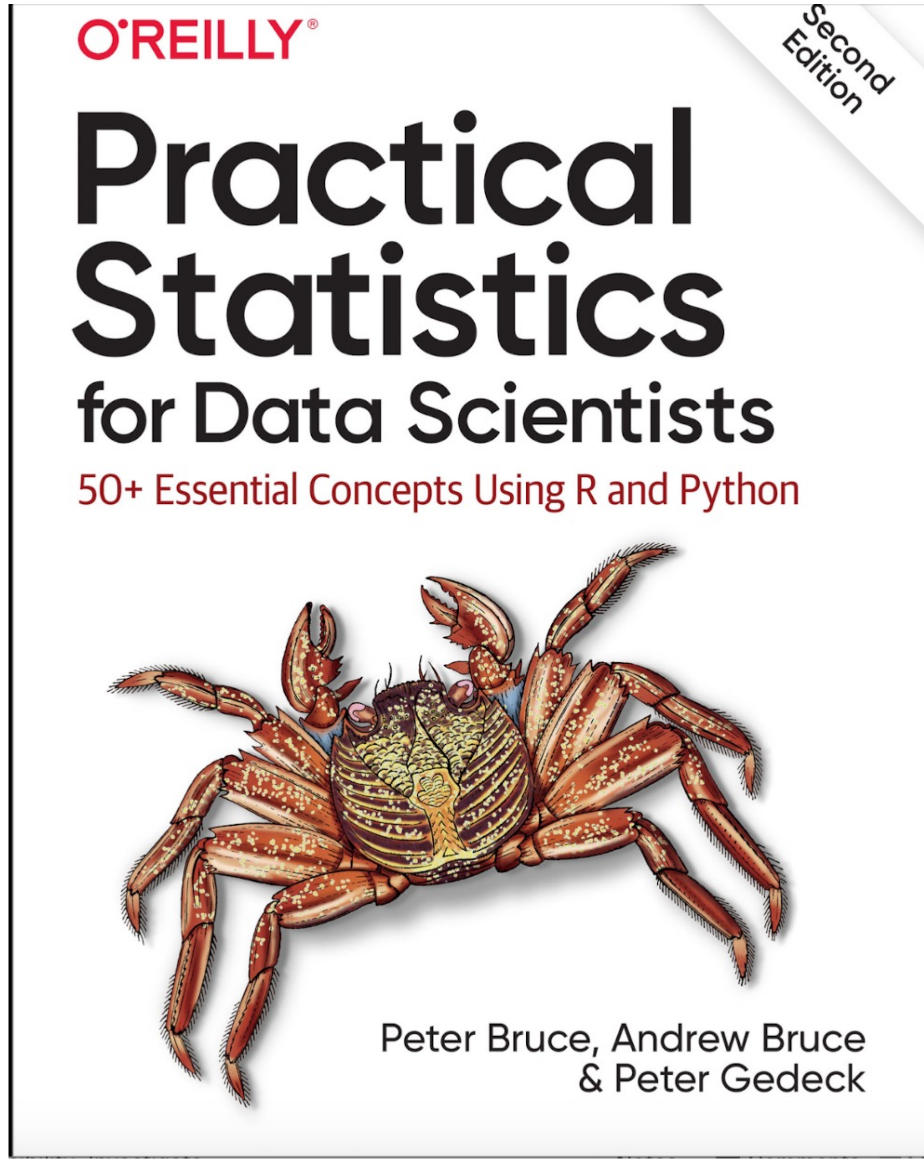
READ ABOUT

- Central Limit Theorem
- Standard Error
- Bootstrap



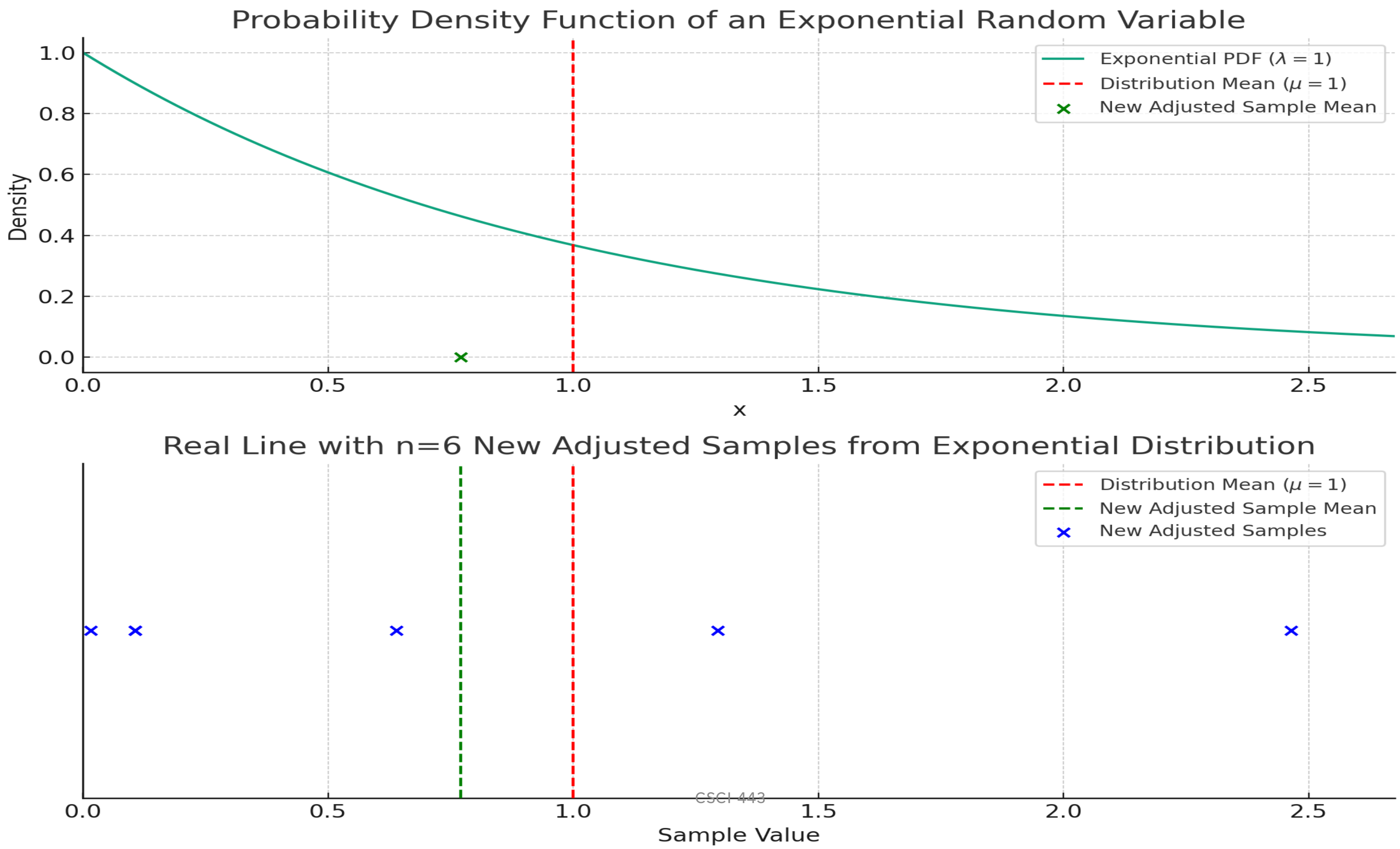
THINGS I WANT TO COVER TODAY

- Discuss exam (to some extent)
- Review Standard Error
- Bootstrap
- Confidence Intervals



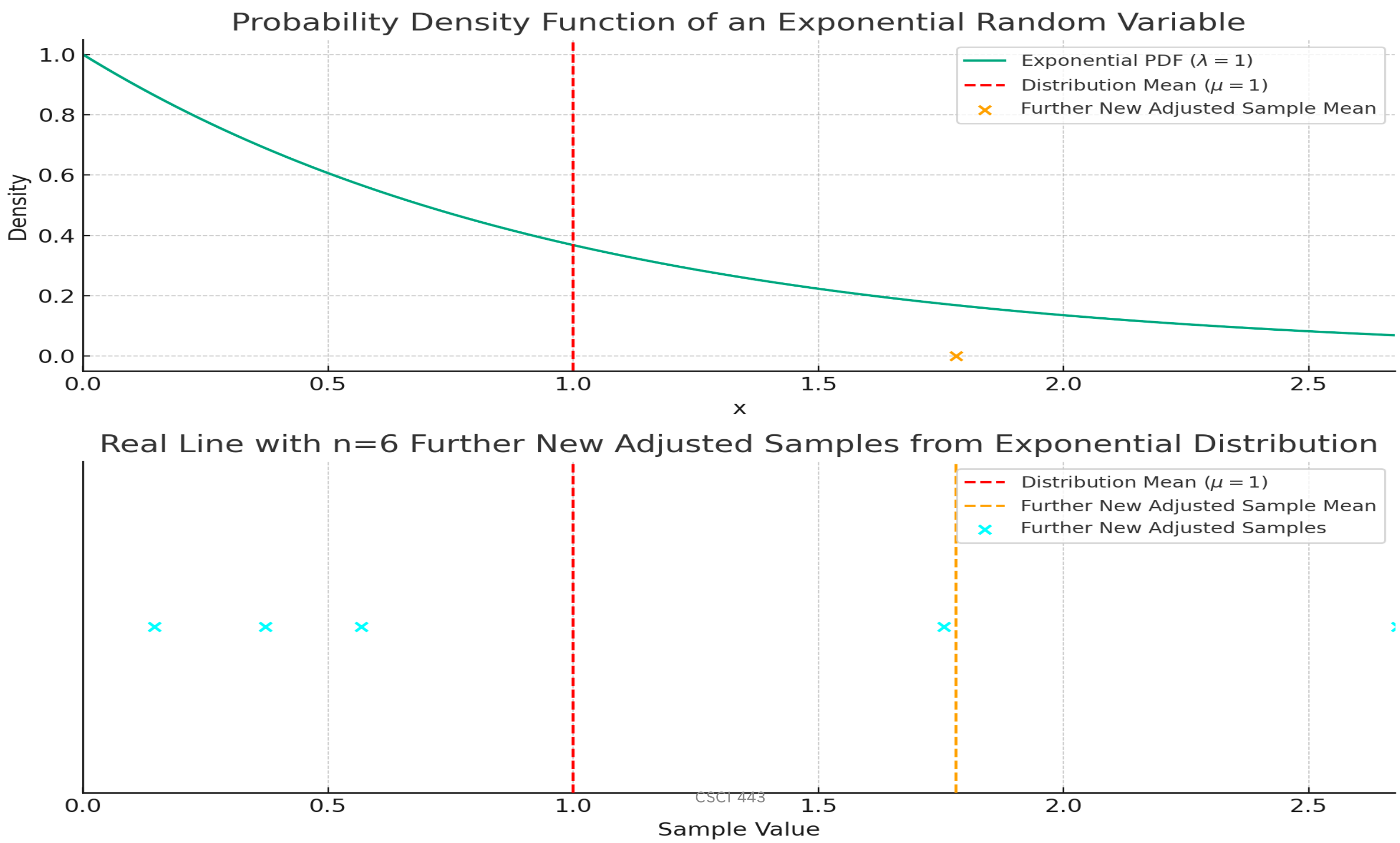
PREVIOUS LECTURE: SAMPLE MEAN IS ALSO RANDOM

Another 6 samples. Sample mean moves.

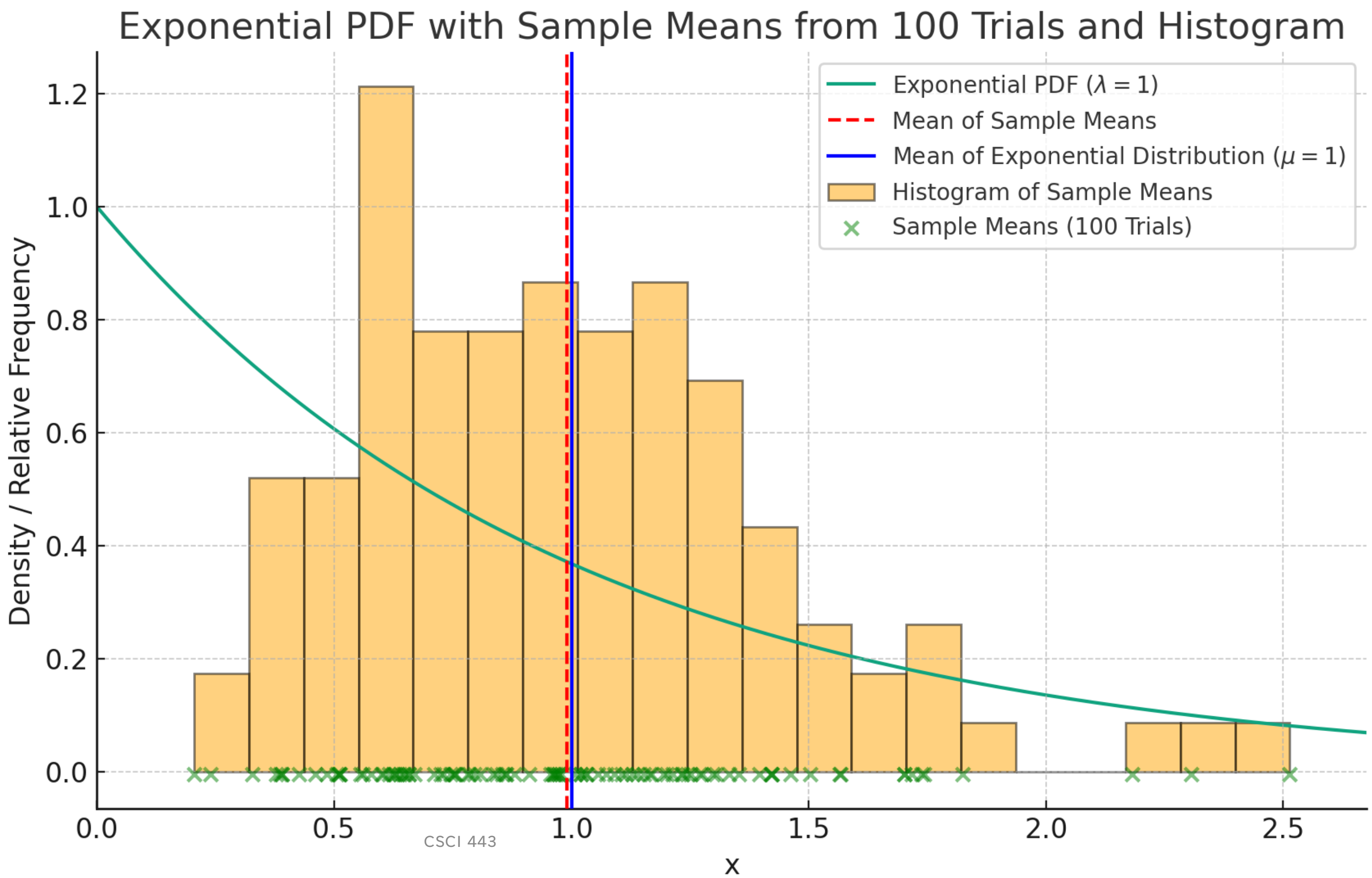


PREVIOUS LECTURE: SAMPLE MEAN IS ALSO RANDOM

Another 6 samples, and we get a different sample mean.



PREVIOUS LECTURE: SAMPLE MEAN IS ALSO RANDOM



n=6 samples in each sample mean.

m=100 trials (sample means)

Hmm…

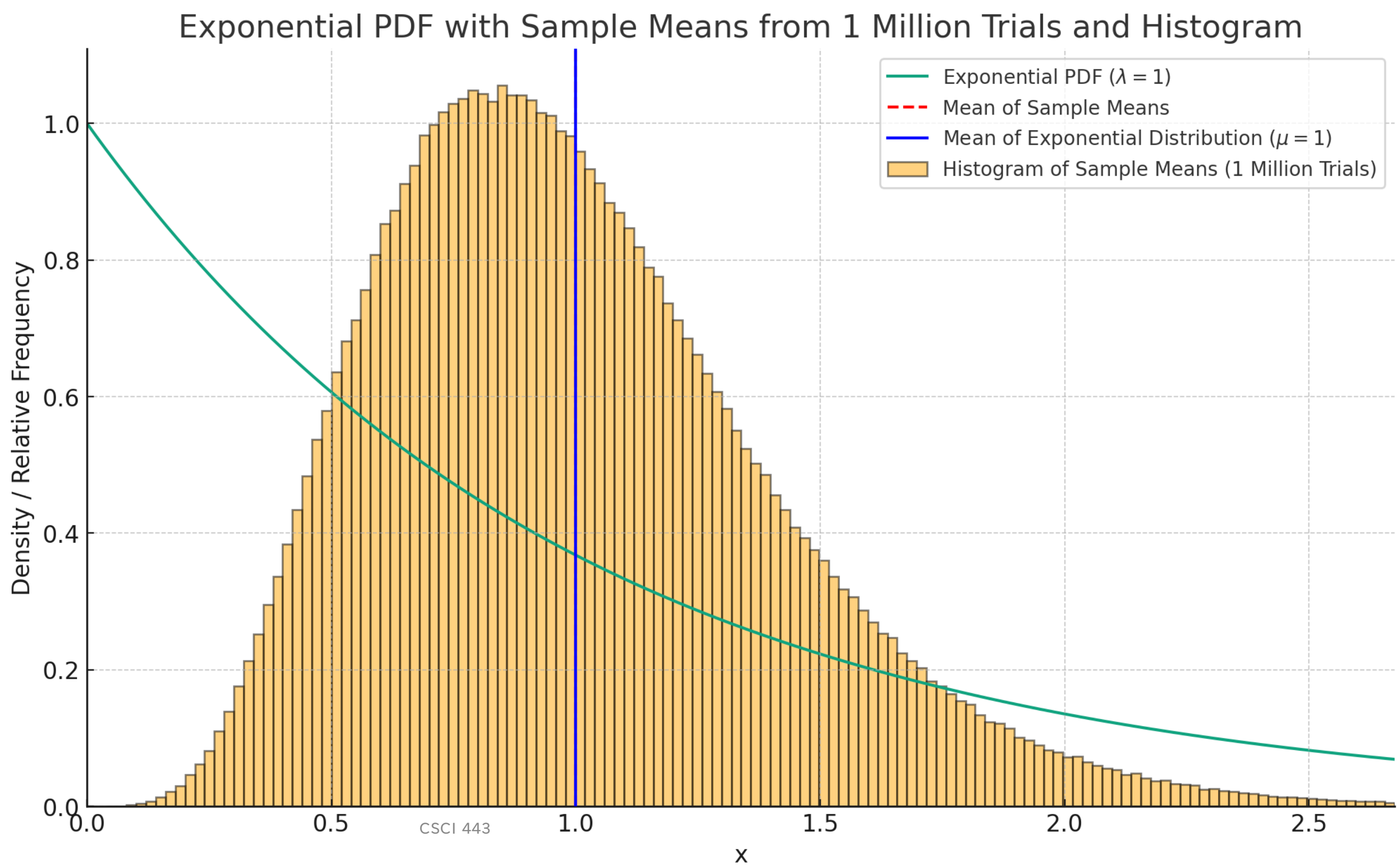
SAMPLE MEAN IS ALSO RANDOM

n=6

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

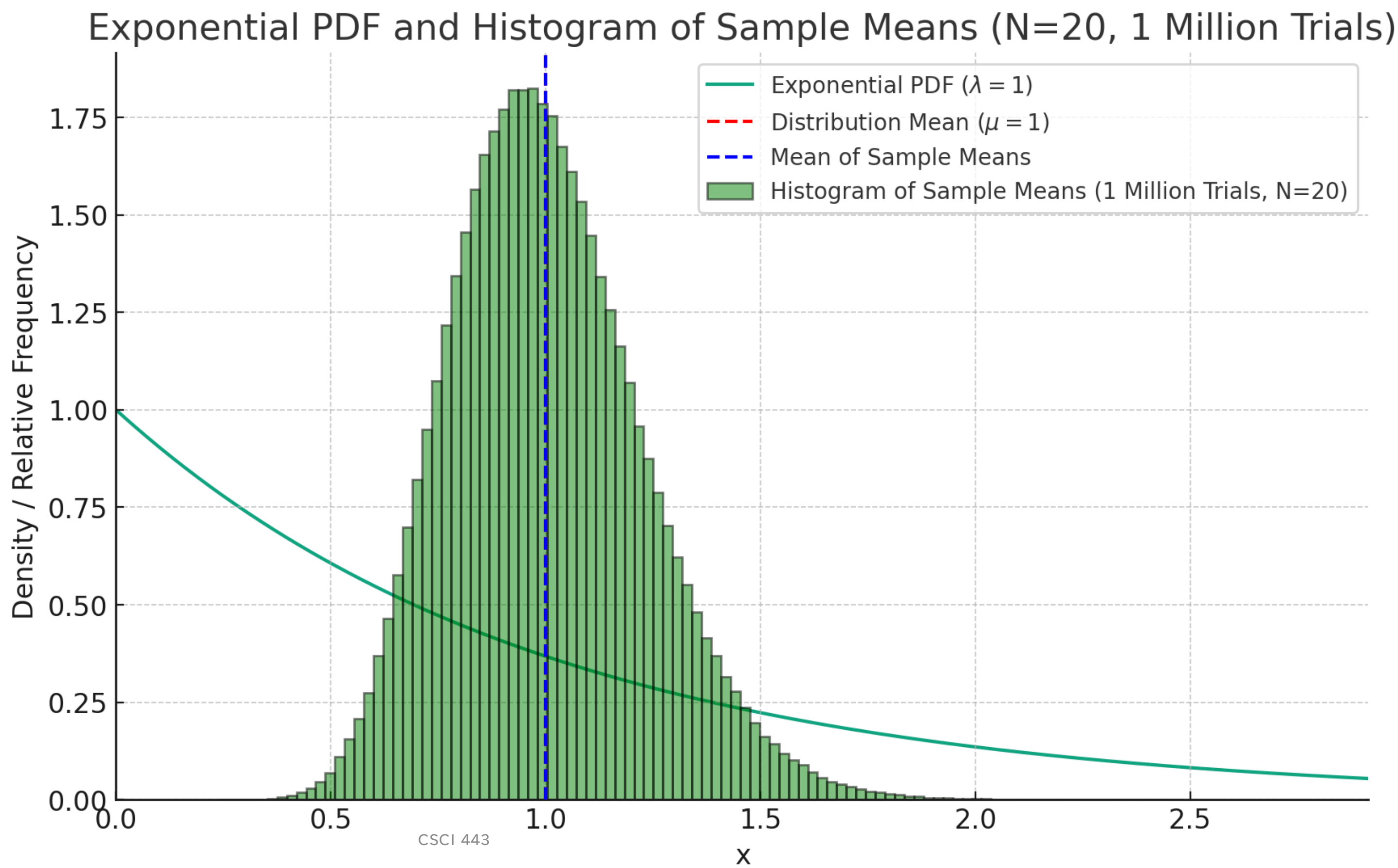
With small n in each sample mean, the distribution of sample means may remain skewed.

CLT's effectiveness depends on increasing *n*.

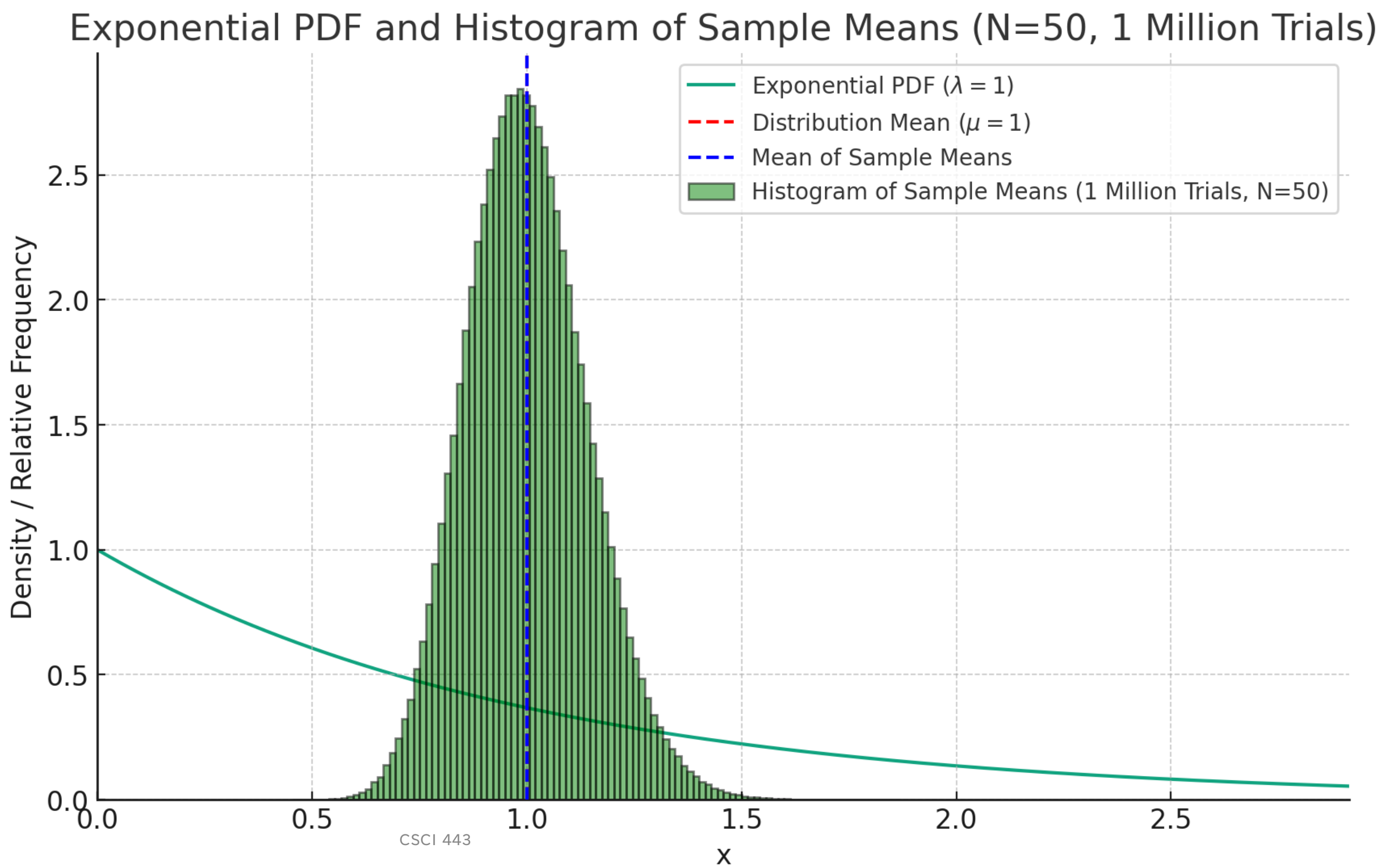


SAMPLE MEAN IS ALSO RANDOM BUT N MATTERS

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



SAMPLE MEAN IS ALSO RANDOM BUT N MATTERS

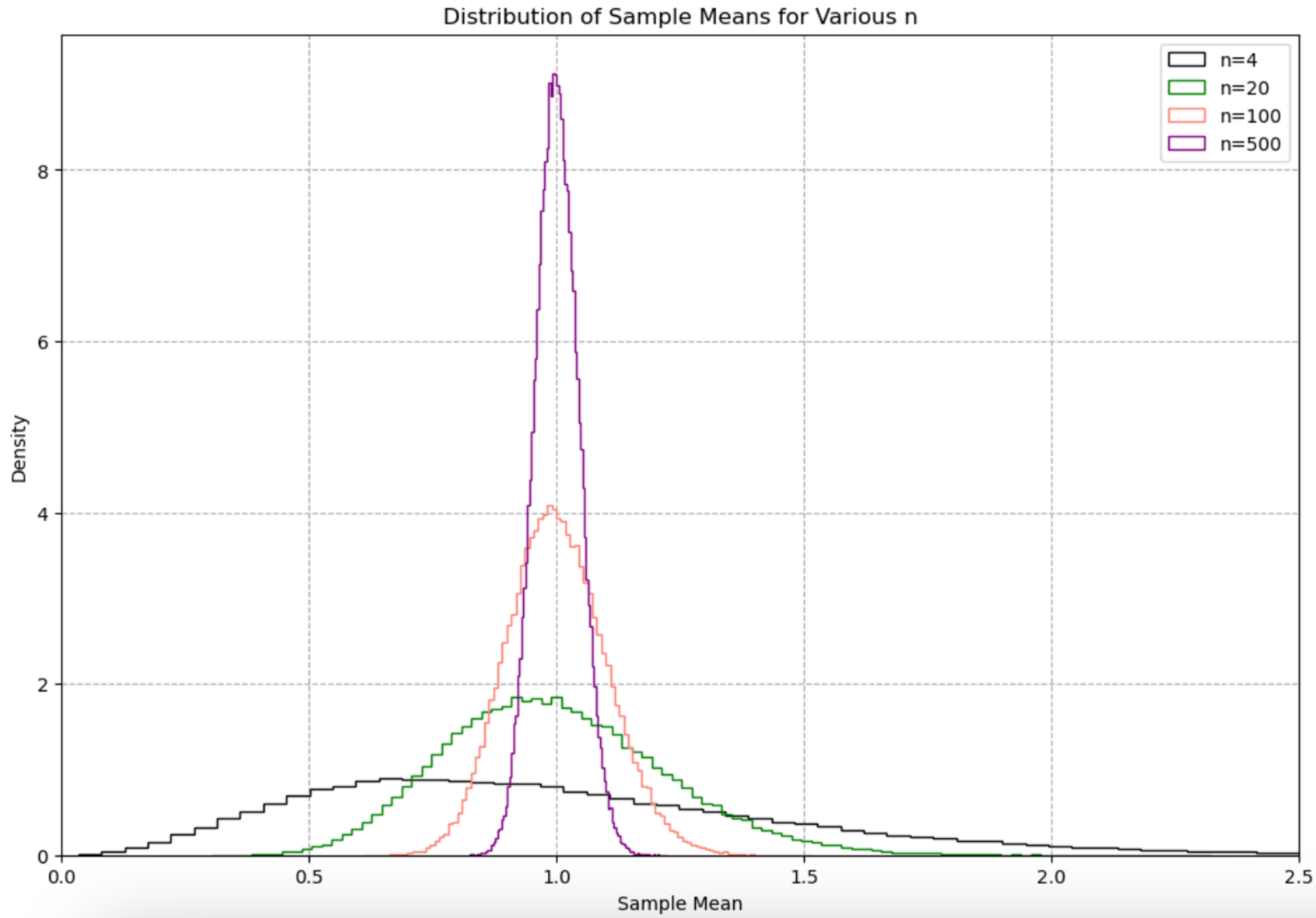


What happens as we increase the number (n) of samples in each sample mean?

SAMPLE MEAN IS ALSO RANDOM BUT N MATTERS

Several sampling mean distributions as we increase n.

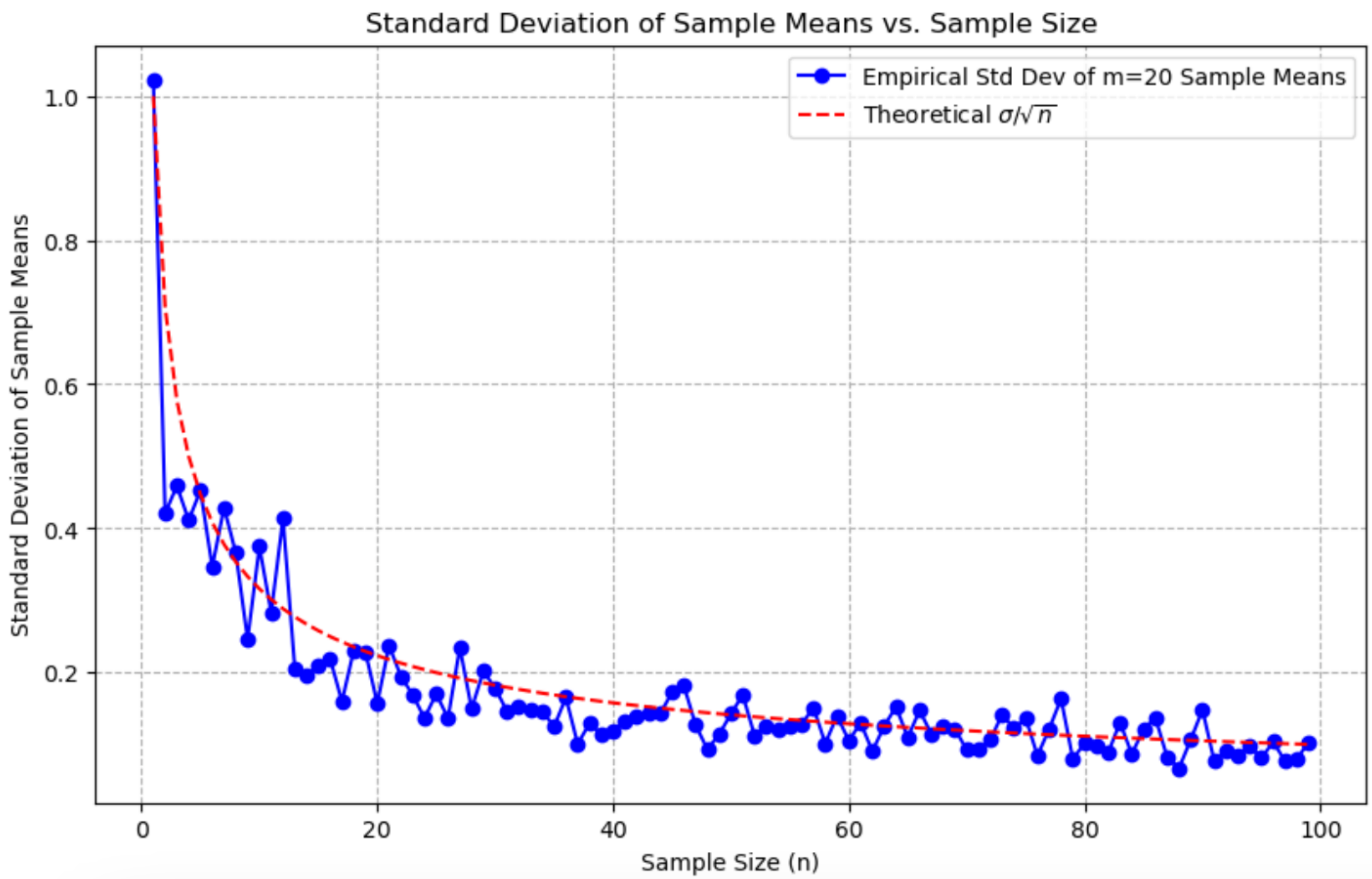
The standard deviation of the sample mean distribution is called the standard error



STANDARD ERROR AS FUNCTION OF N

Sampling mean distribution, i.e., Standard Error (SE) decreases with n according to

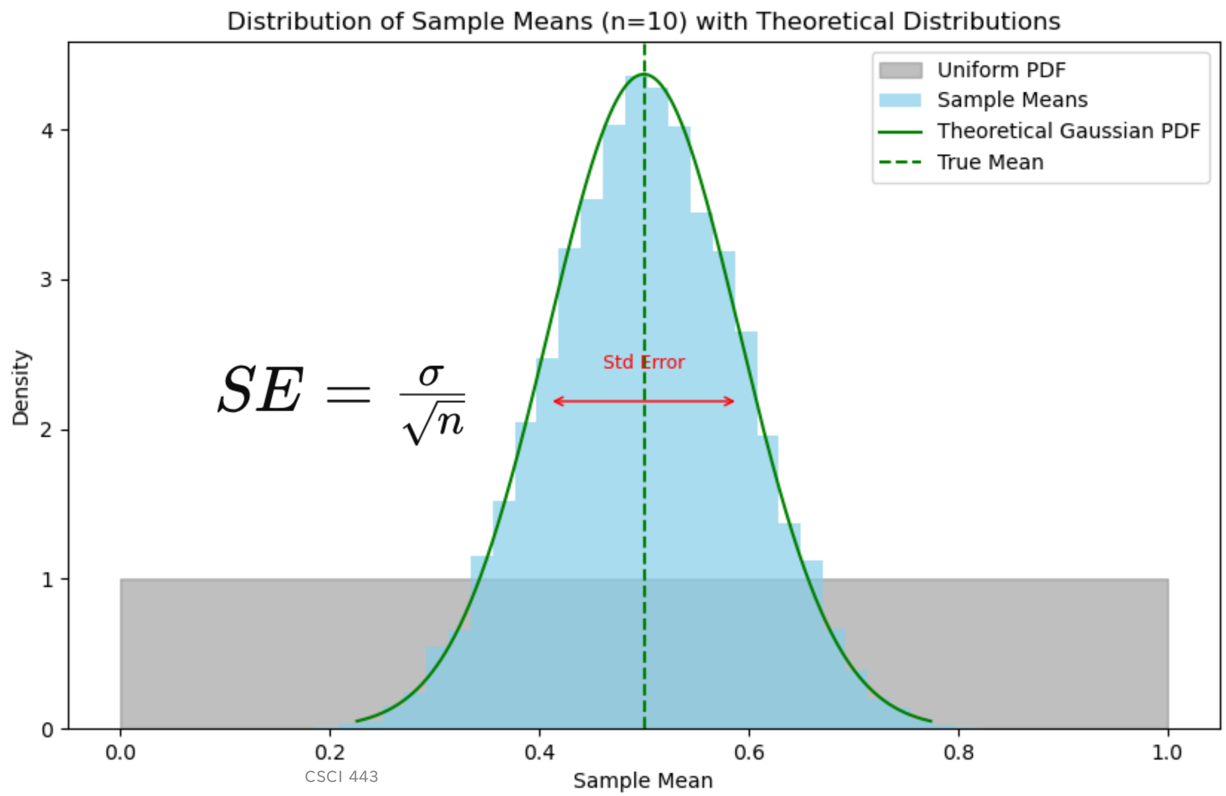
$$SE = \frac{\sigma}{\sqrt{n}}$$



SAMPLE MEAN DISTRIBUTION OF UNIFORM RV

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

For R=10000 sample means created from n=5 samples, we plot a histogram of the sample means.

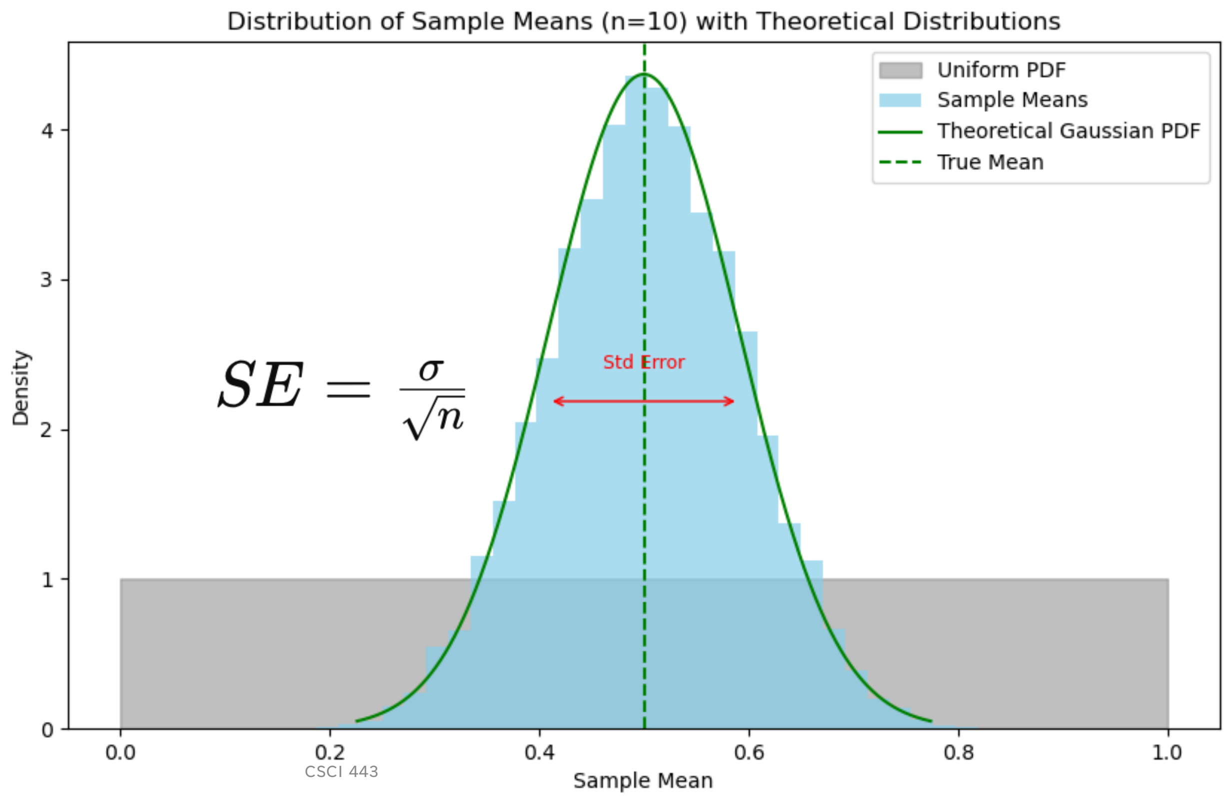


SAMPLE MEAN DISTRIBUTION OF UNIFORM RV

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

For R=10000 sample means created from n=5 samples, we plot a histogram of the sample means.

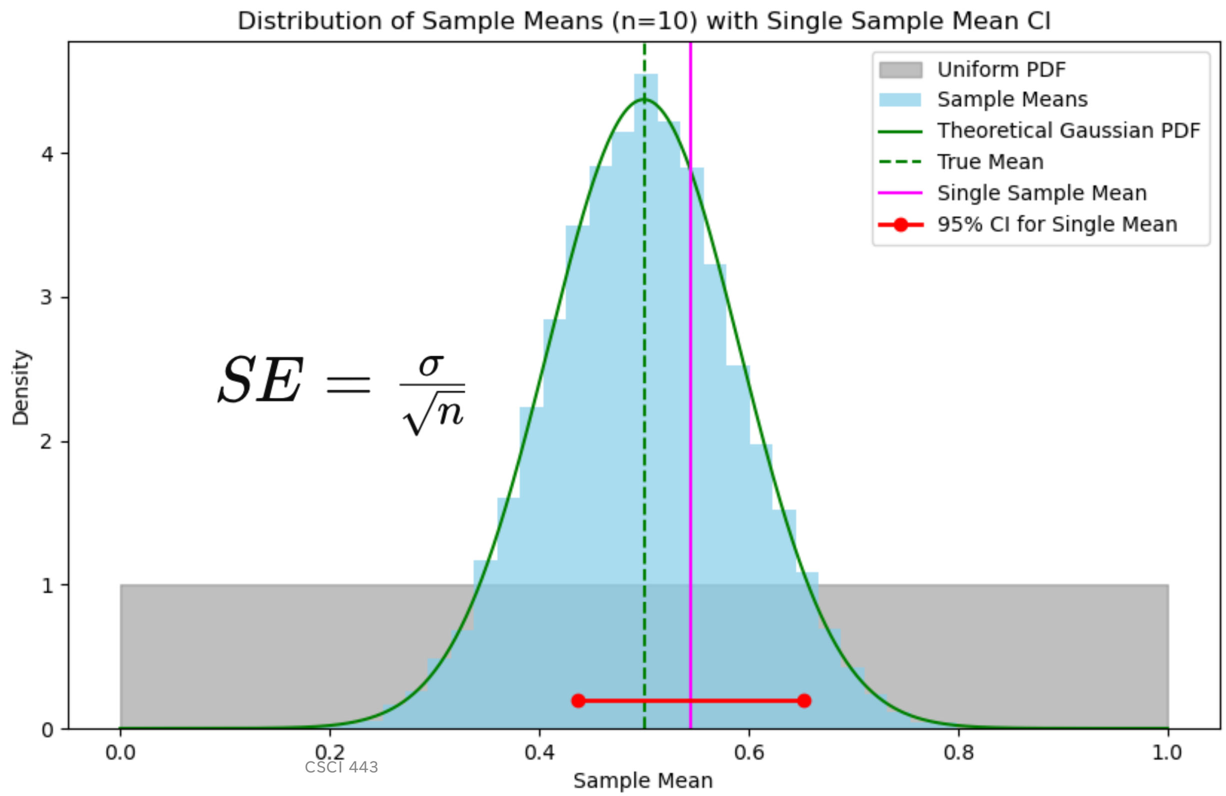
For a symmetric distribution, the sample mean distribution approaches Gaussian with small n



CONFIDENCE INTERVALS USING U[0,1]

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

95% Confidence interval of the sample mean = "if I generate many confidence intervals the same way, approximately 95% will include the true mean"

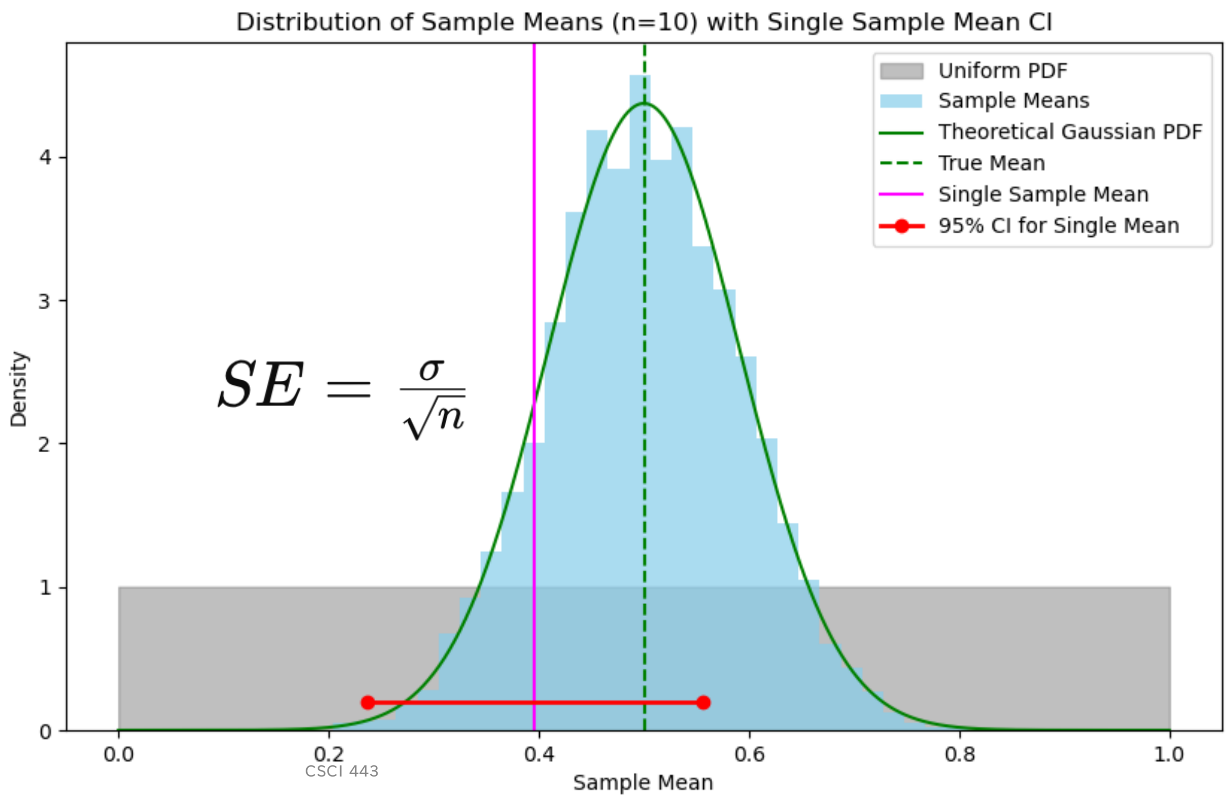


$$SE = \frac{\sigma}{\sqrt{n}}$$

CONFIDENCE INTERVALS USING U[0,1]

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

95% Confidence interval = "the probability that the sample mean will fall within the given range is approximately 95% each time I generate this sample mean"

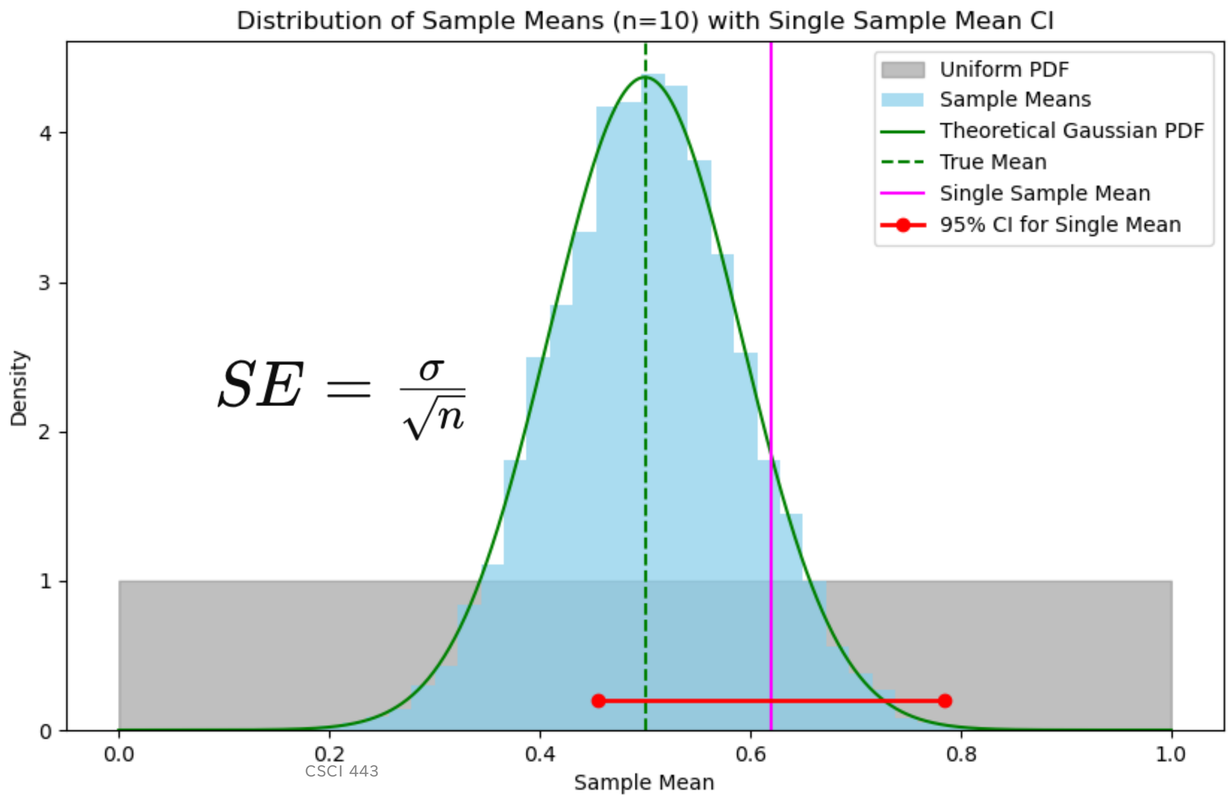


$$SE = \frac{\sigma}{\sqrt{n}}$$

CONFIDENCE INTERVALS USING U[0,1]

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

95% Confidence interval = "the probability that the sample mean will fall within the given range is approximately 95% each time I generate this sample mean"

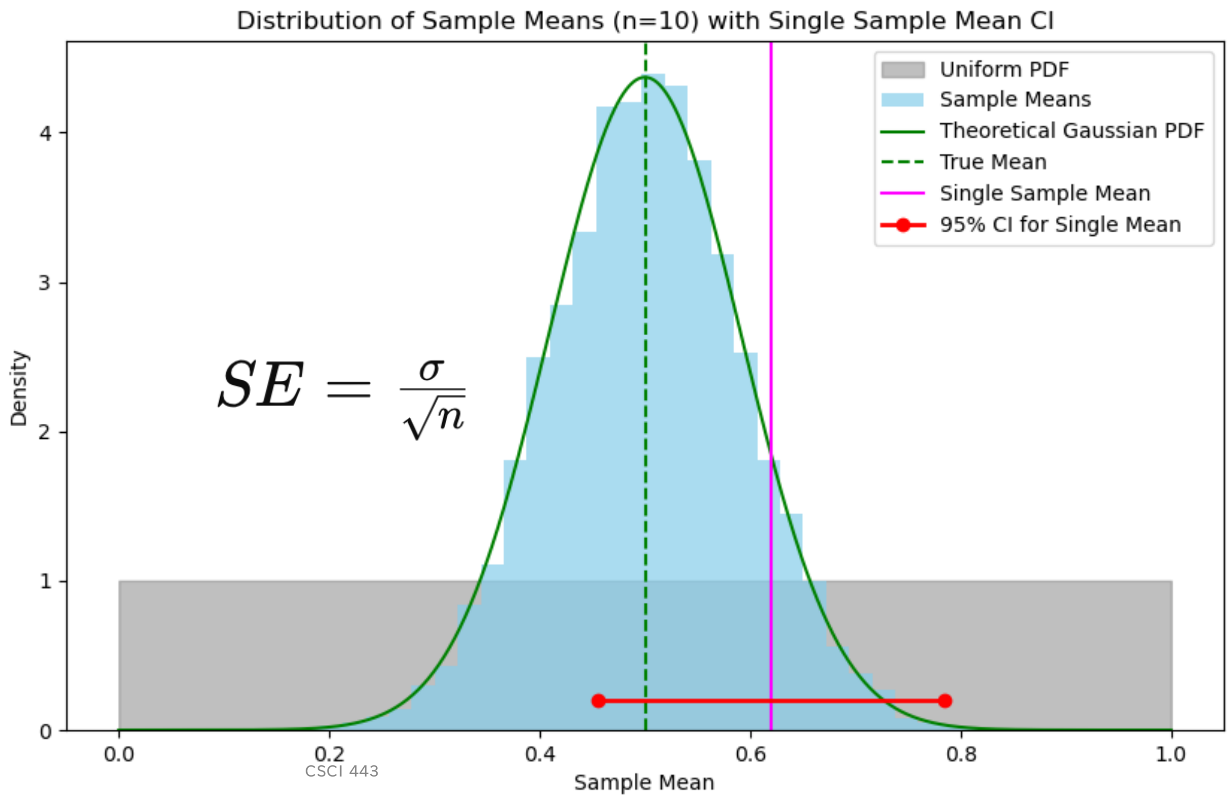


$$SE = \frac{\sigma}{\sqrt{n}}$$

CONFIDENCE INTERVALS USING U[0,1]

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

HERE TALK ABOUT HOW TO COMPUTE A CONFIDENCE INTERVAL.



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

BOOTSTRAPPING

I estimated the shape of the sampling distribution by creating m sample means.

For each sample mean I generated n samples.

I thus needed m*n samples to generate a single plot.

What if I only have n samples but I want to see the distribution of the sample mean? If n is not TOO small, I can create m sample means by drawing n samples with replacement from N original samples

```
Original Sample:
[4 7 6 3 3 7 4 0 6 0 7 2]

Bootstrap Samples:
Bootstrap Sample 0: [6 7 6 7 2 3]
Bootstrap Sample 1: [4 7 7 2 7 0]
Bootstrap Sample 2: [4 7 7 0 2 3]
Bootstrap Sample 3: [3 0 6 4 6 7]
Bootstrap Sample 4: [7 3 2 0 7 2]
Bootstrap Sample 5: [3 6 3 4 0 7]
Bootstrap Sample 6: [7 3 3 4 7 7]
Bootstrap Sample 7: [4 7 7 0 7 3]
Bootstrap Sample 8: [3 7 2 0 3 3]
Bootstrap Sample 9: [3 6 4 7 3 6]
```

BOOTSTRAPPING

We can do the same thing for an exponential distribution. We have n samples from an exponential distribution. Let's estimate the shape of the sampling distribution from only these n samples.

```
Original Sample:
[2.50558974 0.43413249 0.34428447 0.85090025 2.98173239]

Bootstrap Samples:
Bootstrap Sample 0: [2.50558974 2.50558974 0.43413249 2.98173239 2.50558974]
Bootstrap Sample 1: [0.43413249 0.85090025 0.43413249 2.98173239 2.98173239]
Bootstrap Sample 2: [0.34428447 2.50558974 2.98173239 0.34428447 2.98173239]
Bootstrap Sample 3: [2.50558974 0.43413249 2.50558974 0.34428447 2.50558974]
Bootstrap Sample 4: [2.50558974 0.43413249 0.34428447 2.50558974 0.85090025]
Bootstrap Sample 5: [0.34428447 2.98173239 0.43413249 2.98173239 0.85090025]
Bootstrap Sample 6: [0.34428447 0.85090025 0.34428447 0.85090025 0.85090025]
Bootstrap Sample 7: [0.34428447 0.85090025 0.43413249 2.98173239 0.43413249]
Bootstrap Sample 8: [0.43413249 0.85090025 0.43413249 2.50558974 0.85090025]
Bootstrap Sample 9: [2.98173239 0.34428447 0.85090025 0.43413249 2.50558974]
```

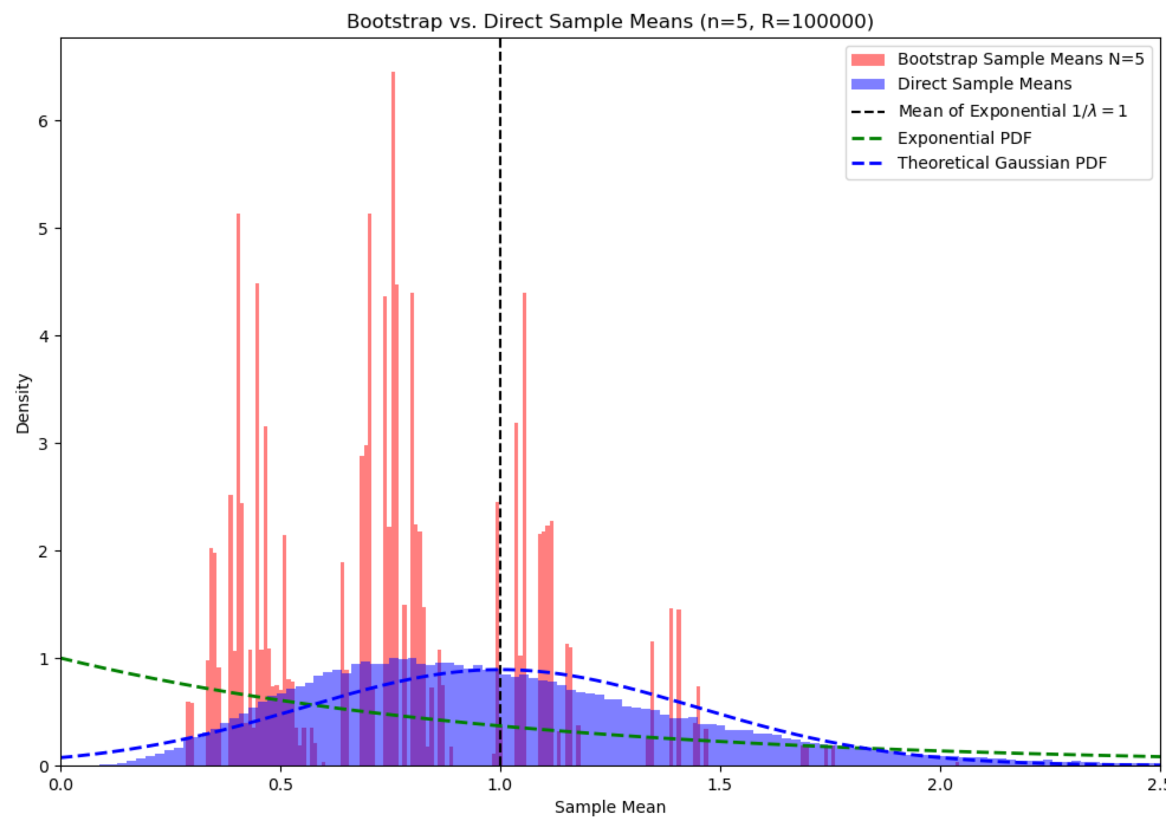
BOOTSTRAPPING TO FIND SAMPLE MEAN DISTRIBUTION N=5

Let n = number of samples in each sample mean = 5

Let R = number of sample means = 100000

Let N = size of original sample that we are bootstrapping = 5

With an original distribution of 5 sample means, bootstrapping does not work well.



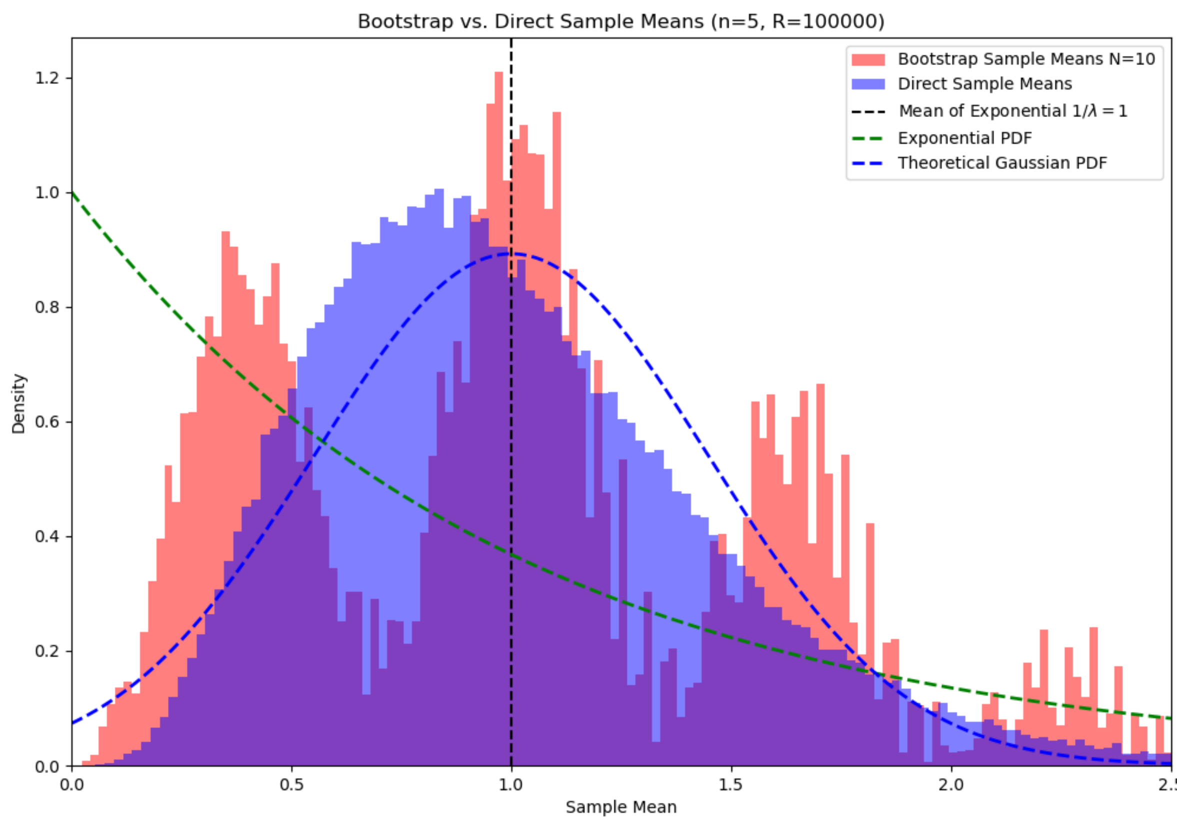
BOOTSTRAPPING TO FIND SAMPLE MEAN DISTRIBUTION N=10

Let n = number of samples in each sample mean = 5

Let R = number of sample means = 100000

Let N = size of original sample that we are bootstrapping = 10

N=10 still doesn't work well.



BOOTSTRAPPING TO FIND SAMPLE MEAN DISTRIBUTION N=20

Let n = number of samples in each sample mean = 5

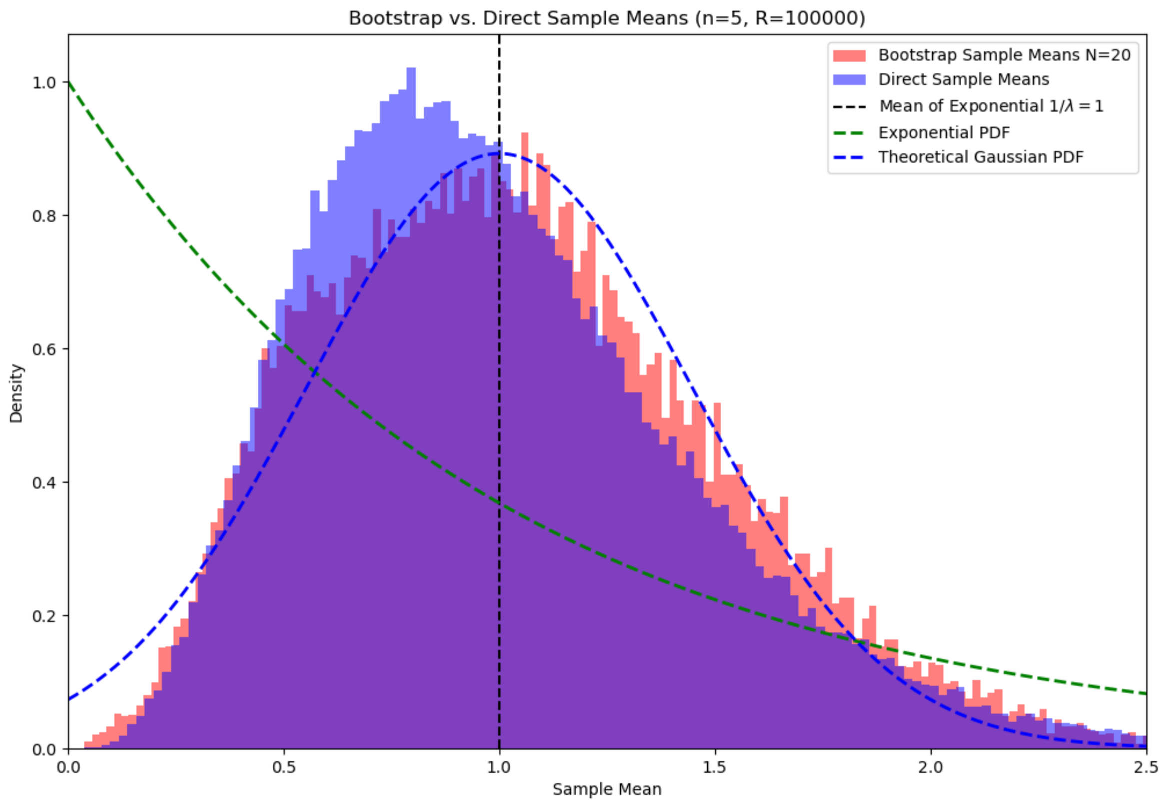
Let R = number of sample means = 100000

Let N = size of original sample that we are bootstrapping = 20

N=20 closer.

Direct sampling required R*n samples = 500,000

Bootstrapping required 20.



BOOTSTRAPPING TO FIND SAMPLE MEAN DISTRIBUTION N=500

Let n = number of samples in each sample mean = 5

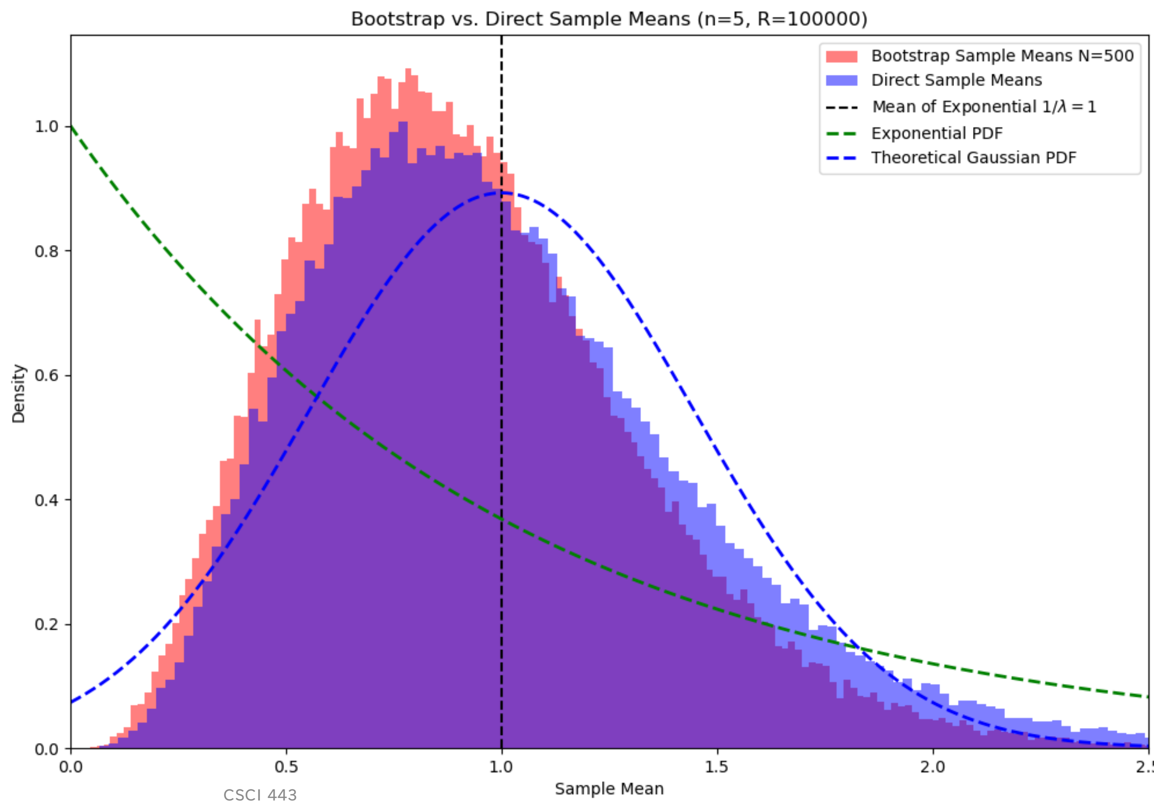
Let R = number of sample means = 100000

Let N = size of original sample that we are bootstrapping = 500

N=500 very close.

Direct sampling required R*n samples = 500,000

Bootstrapping required 500.



THANK YOU

David Harrison

Harrison@cs.olemiss.edu