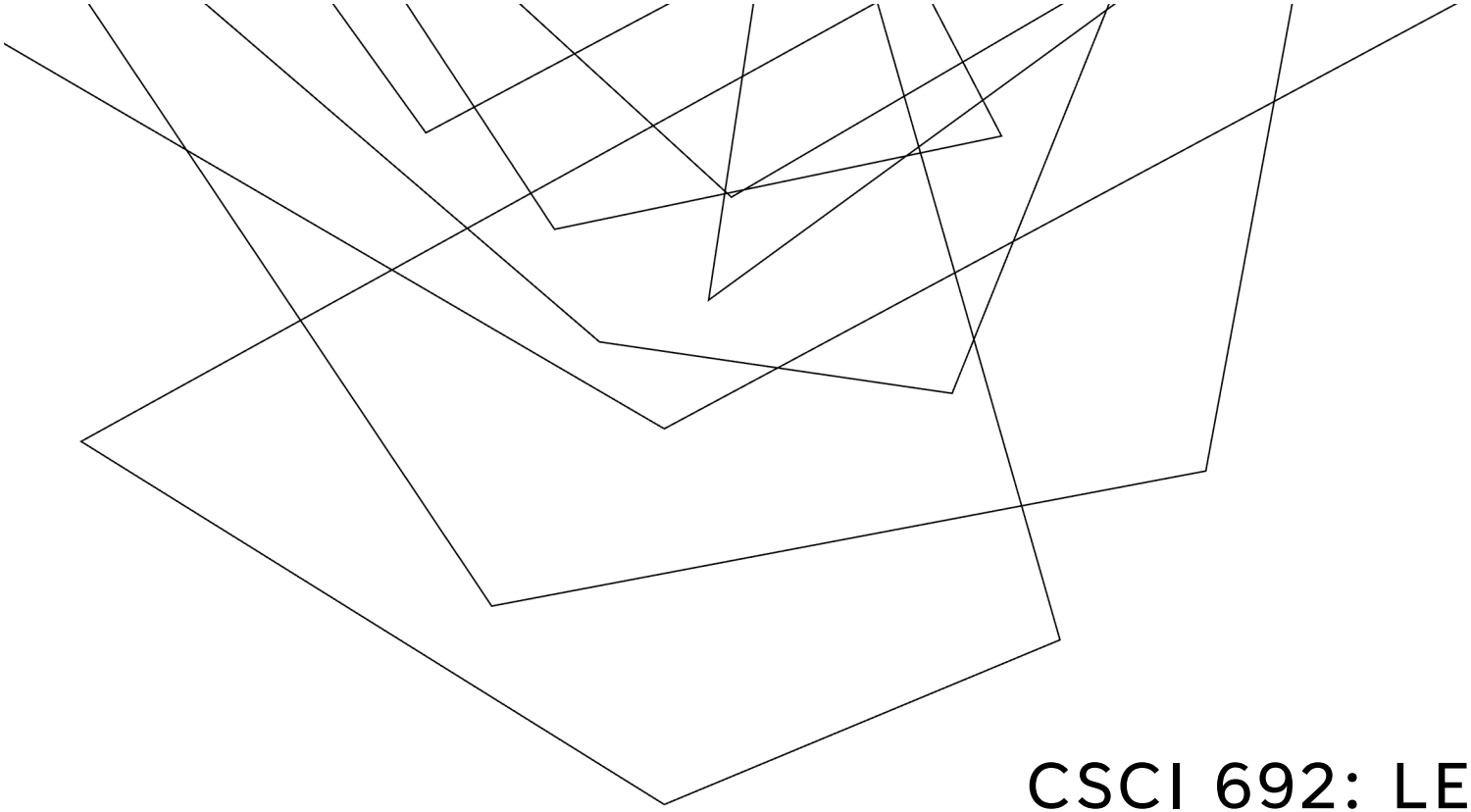


CSCI 692: LECTURE 10 REVIEW

Professor David Harrison

OFFICE HOURS

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

HOMEWORK 3

Was due before class.

NOTE REGARDING EXAMS

Note in homework 2:

Note regarding the midterm and final: The midterm and final will be written, so students will not have access to Databricks or Jupyter or Python. The questions asked on an exam would be computed on small datasets as are used for question 31 in Part 7 and all the problems in Parts 8 and 9. I recommend that you answer the questions in these sections without using Python or a calculator. The problems are not difficult, and doing them by hand may prepare you for answering such questions on the exams.

DATES OF INTEREST

February 8	HW2 handed out
February 15	HW2 due,
February 23	HW3 handed out
February 27	HW3 due
February 27	Review
February 29	Midterm (THIS THURSDAY)
March 4	Progress Reports
March 8	Deadline for Withdrawal
March 9-17	Spring Break

BLACKBOARD & GITHUB

Slides up through lecture 9 on blackboard.

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

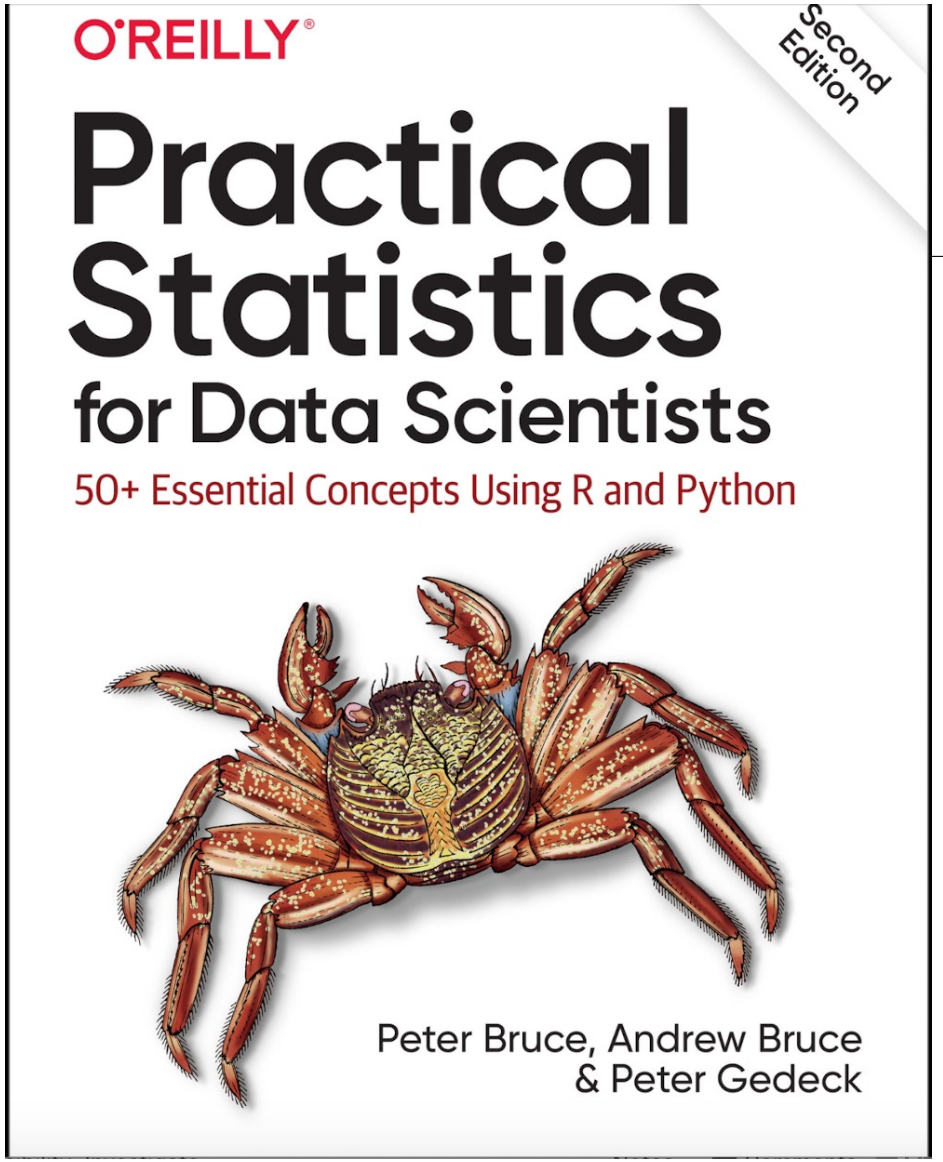
The project is at

https://github.com/dosirrah/CSCI443_AdvancedDataScience

READ ABOUT

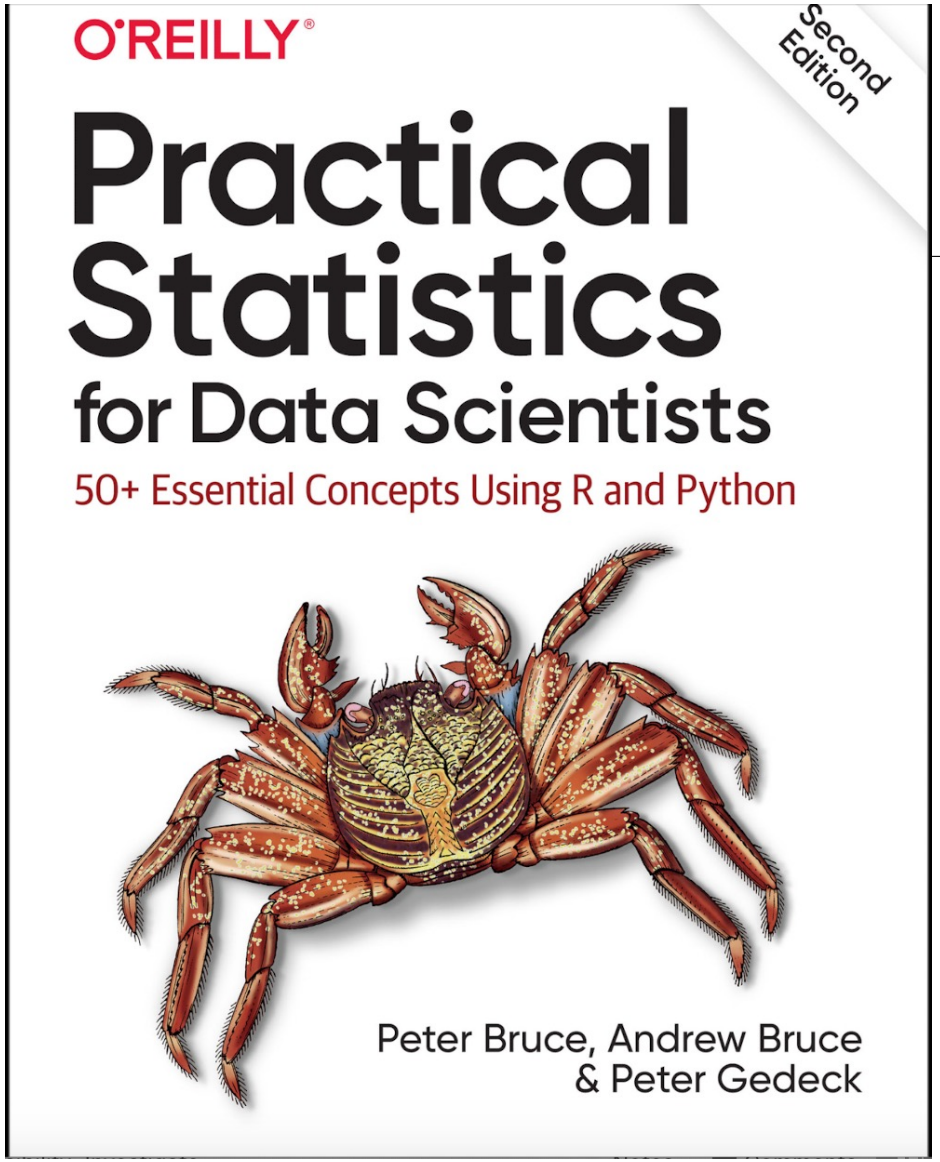
- Bias
  - Examples were already given in class, but book provides good example of selection bias.
- Random selection
  - ways to avoid bias

- Size vs. Quality: When Does Size Matter?



THINGS I WANT TO COVER TODAY

- Issue from last class: Averages of averages
- Review



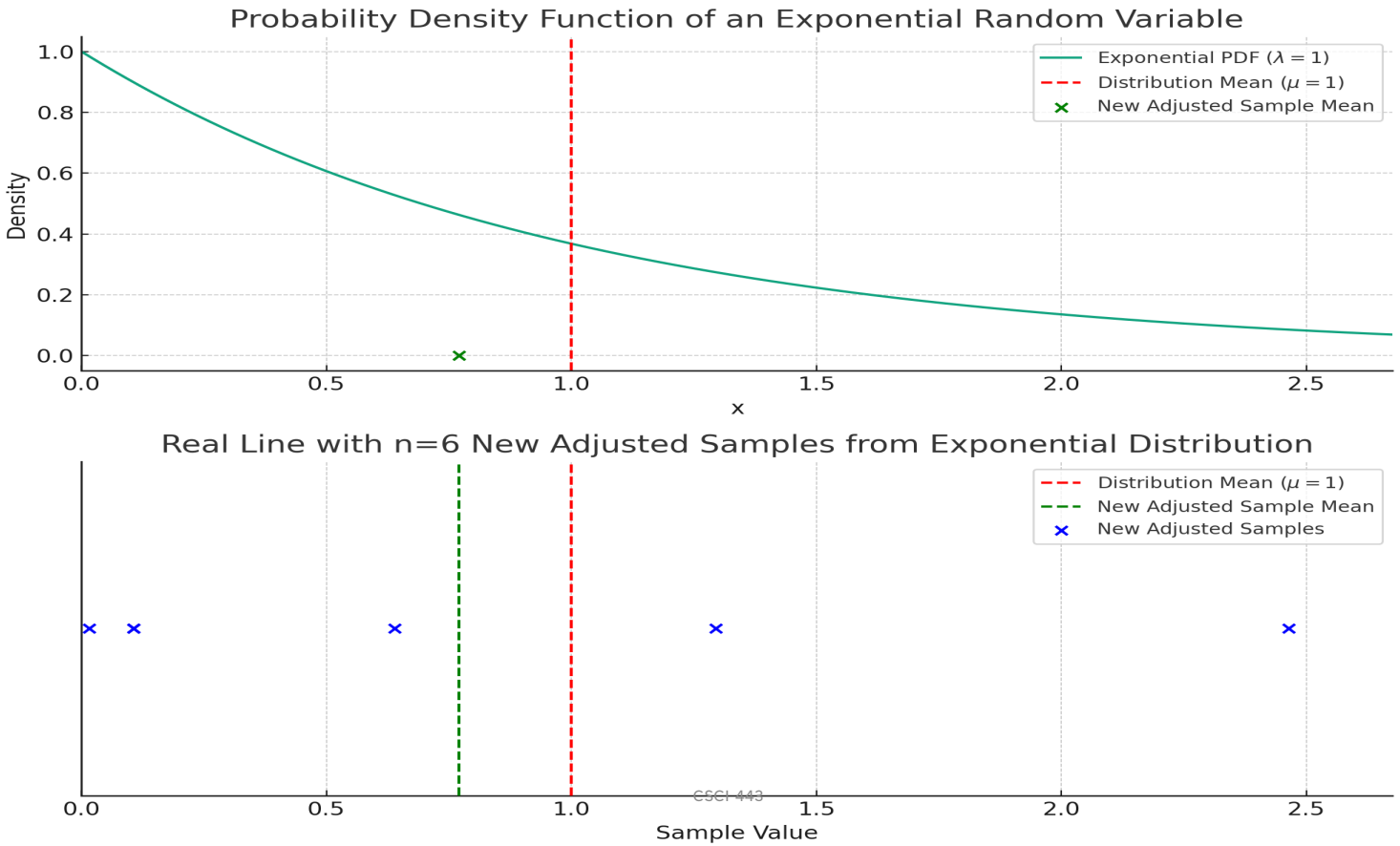
WAYS TO ACCESS DBFS

See CSCI443 Lecture 10 Notes

- Notebook demonstrating multiple ways to access DBFS via a DataFrame from Databricks Notebooks

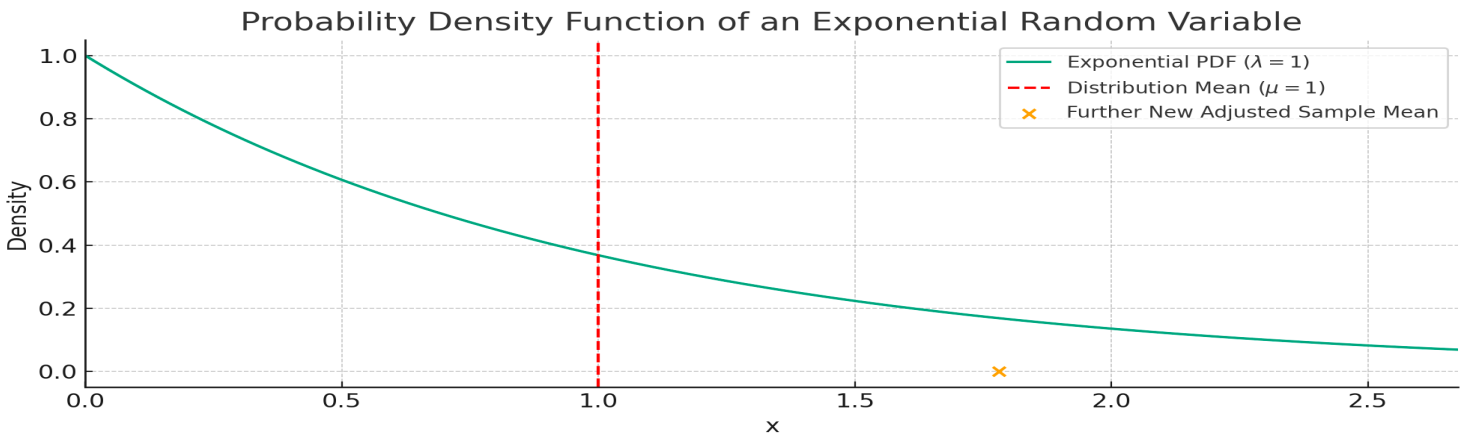
SAMPLE MEAN IS ALSO RANDOM

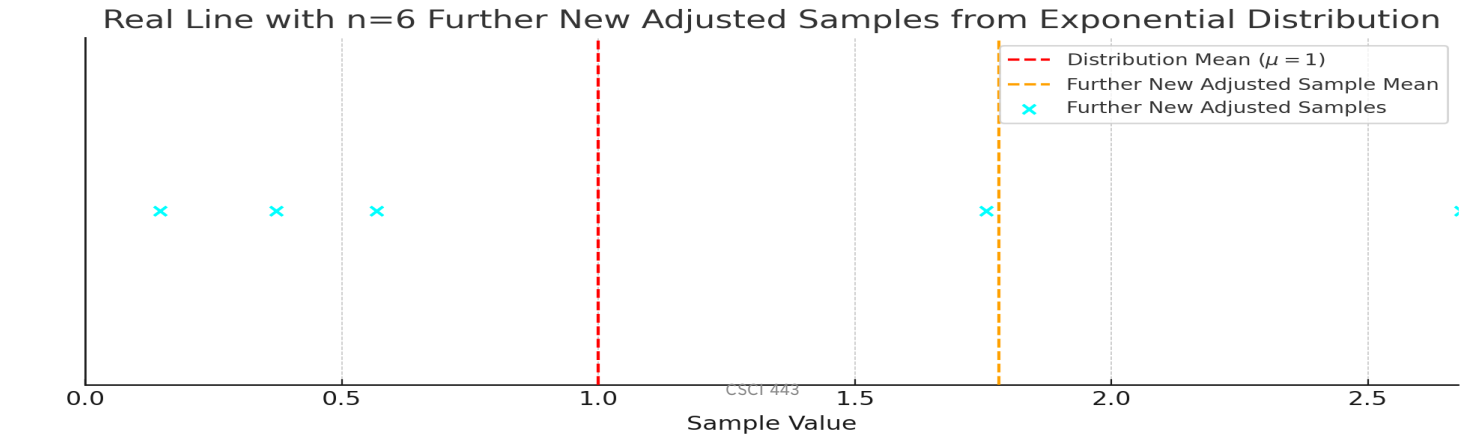
Another 6 samples. Sample mean moves.



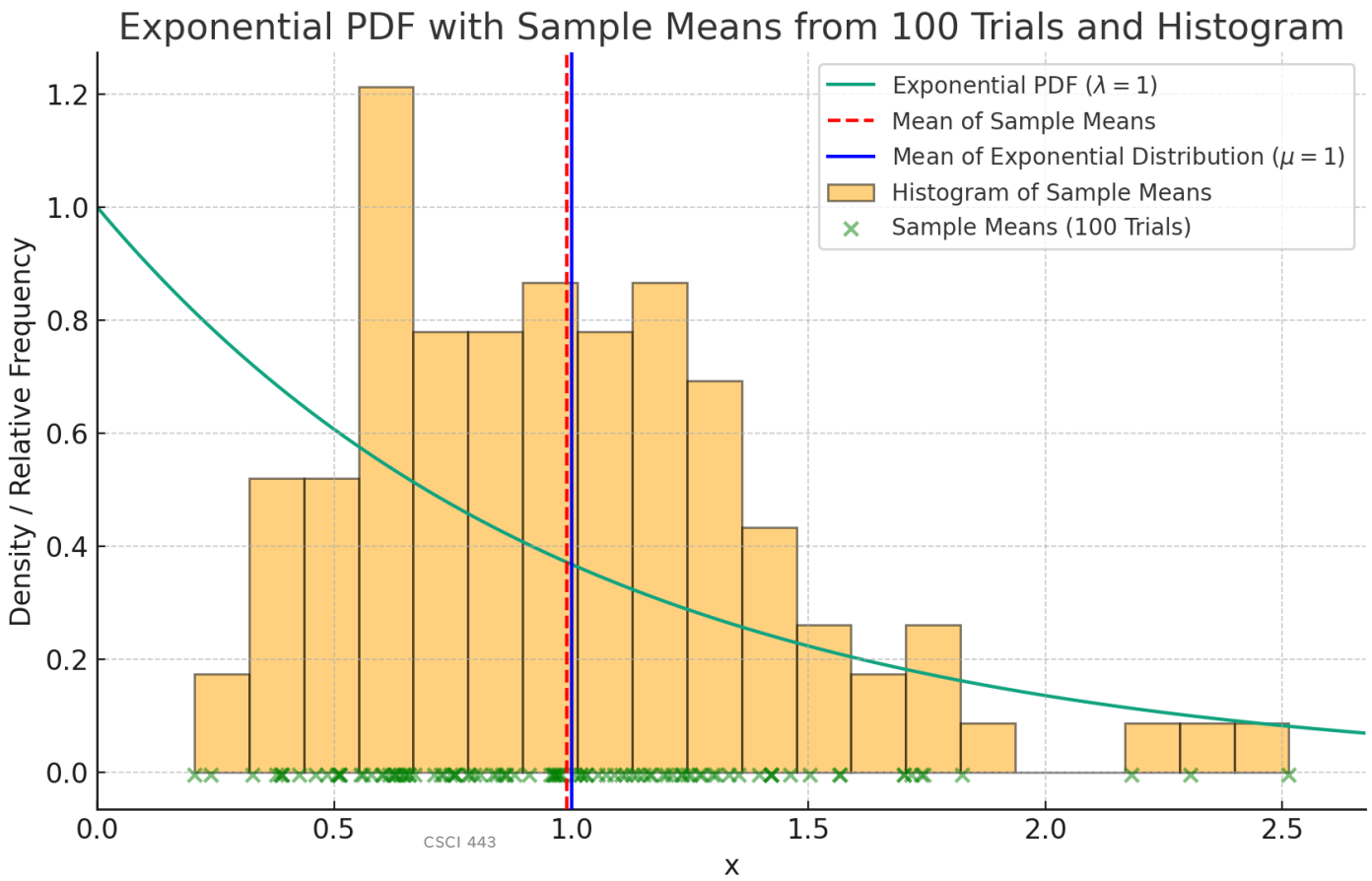
SAMPLE MEAN IS ALSO RANDOM

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





SAMPLE MEAN IS ALSO RANDOM



n=6 samples in each sample mean.

100 trials (sample means)

Hmm...

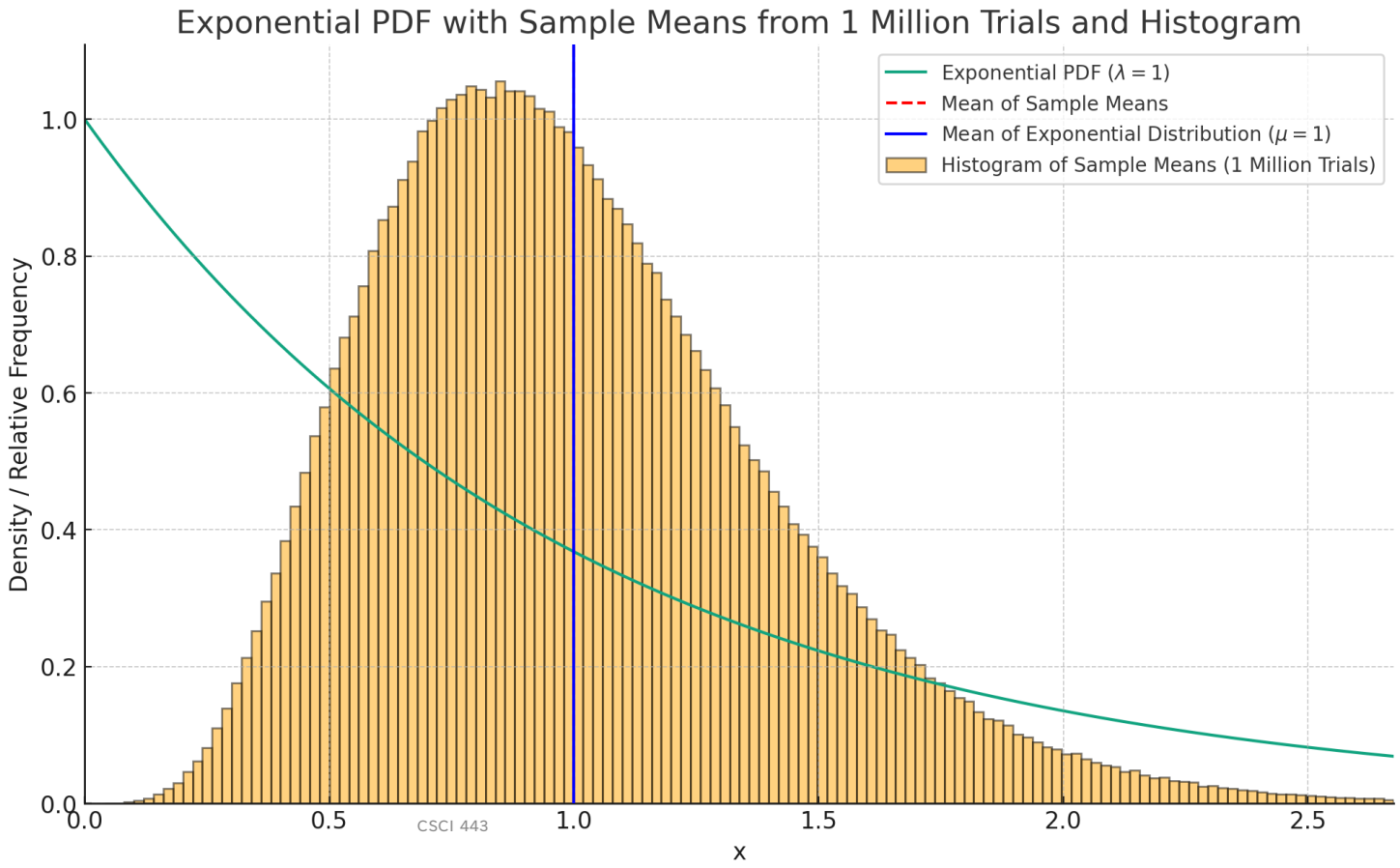
SAMPLE MEAN IS ALSO RANDOM

n=6

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

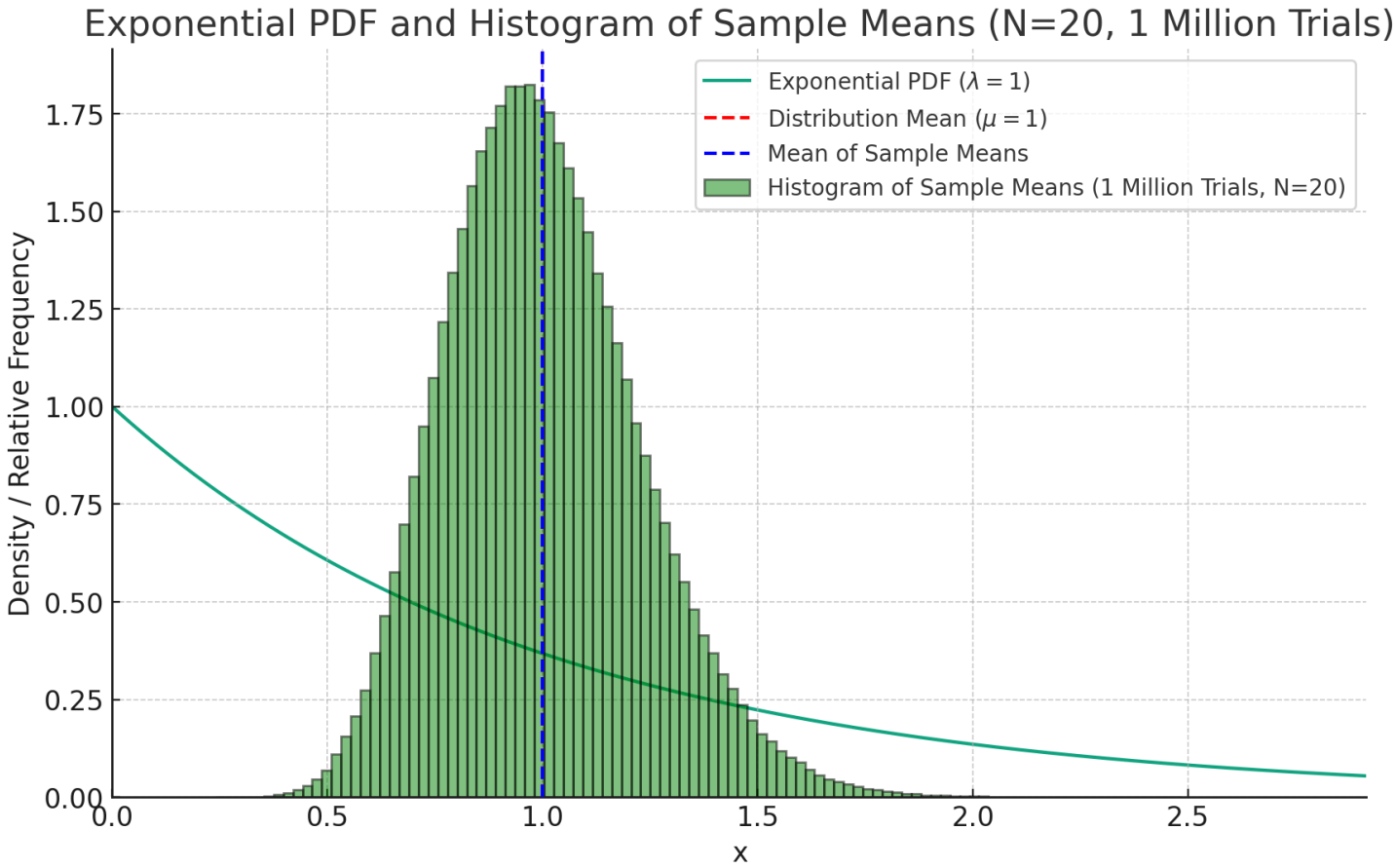
With small n in each sample mean, the distribution of sample means may remain skewed.

CLT's effectiveness depends on increasing n.



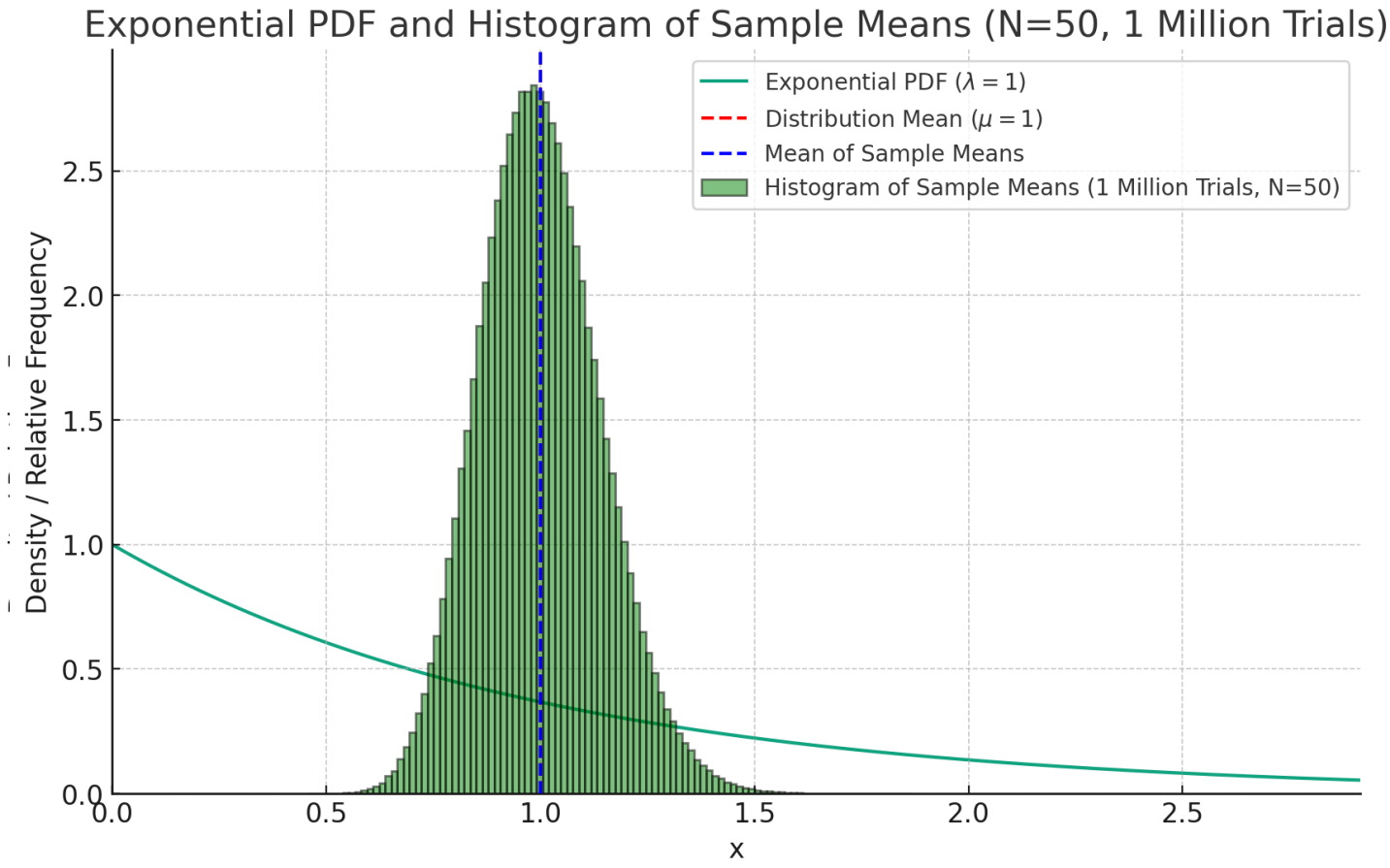
SAMPLE MEAN IS ALSO RANDOM BUT N MATTERS

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



SAMPLE MEAN IS ALSO RANDOM BUT N MATTERS

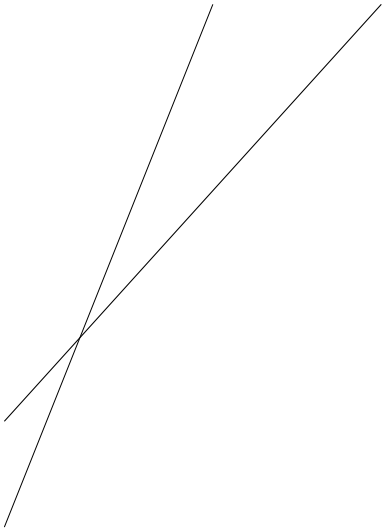
What happens as we increase the number (n) of samples in each sample mean?



THEN I SAID SOMETHING...

”It is strictly better to compute the mean over larger n than to divide n into batches and take the mean of means.”

- But I didn’t explain why...
- See CSCI443 Lecture 10 Notes Part II



MIDTERM

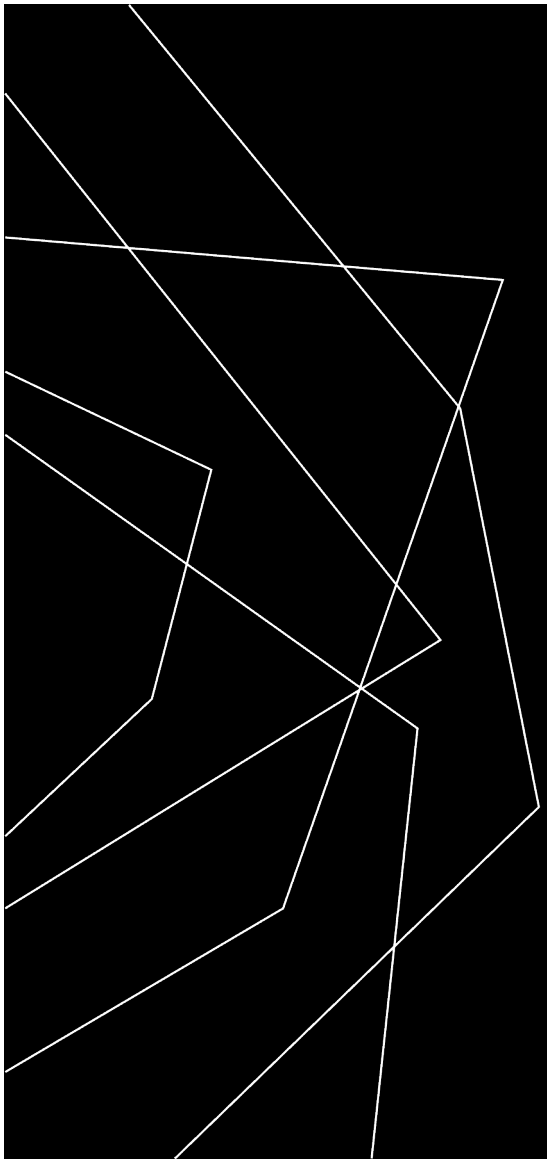
- Covers similar problems to those on homeworks 1, 2, and 3.
- Except those that require a computer to comptue.

TOPICS COVERED IN HW2

- Types of data
- Random Experiments, Outcomes, Sample Spaces
- Random Variables
- Events
- Distributions and Samples
- Range, Means, Medians, Trimmed Means, Percentiles
- Effects of outliers

TOPICS COVERED IN HW3

- Bias
- Variance, Covariance, Correlation
- Population vs. Sample Statistics
- Z-Scores
- Phi, Erf, Gaussians
- Sampling Distributions



THANK YOU

David Harrison

Harrison@cs.olemiss.edu