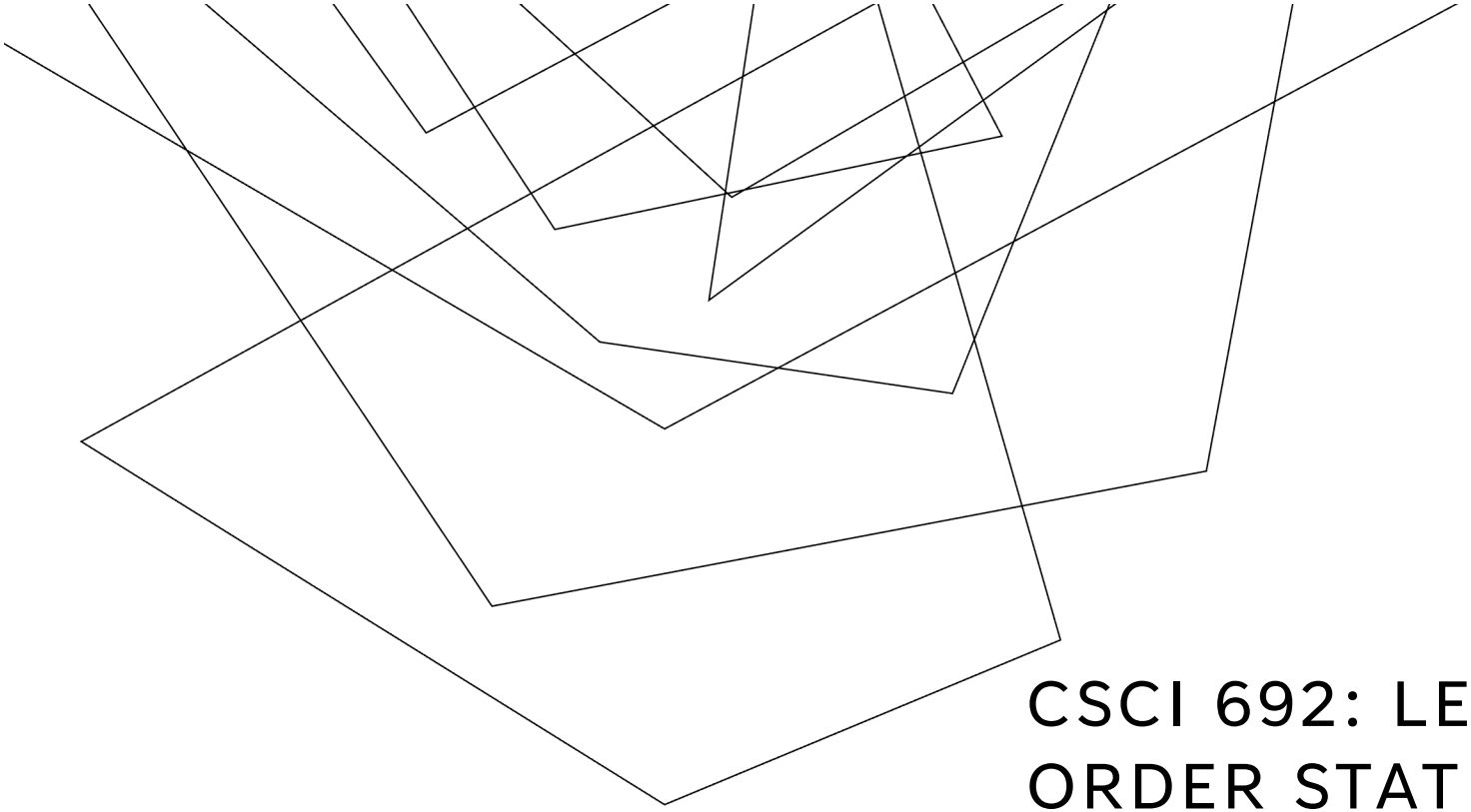


CSCI 692: LECTURE 5 ORDER STATISTICS, PERCENTILES

Professor David Harrison

OFFICE HOURS

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

.

HOMEWORK 2

Will be posted tonight (February 8)

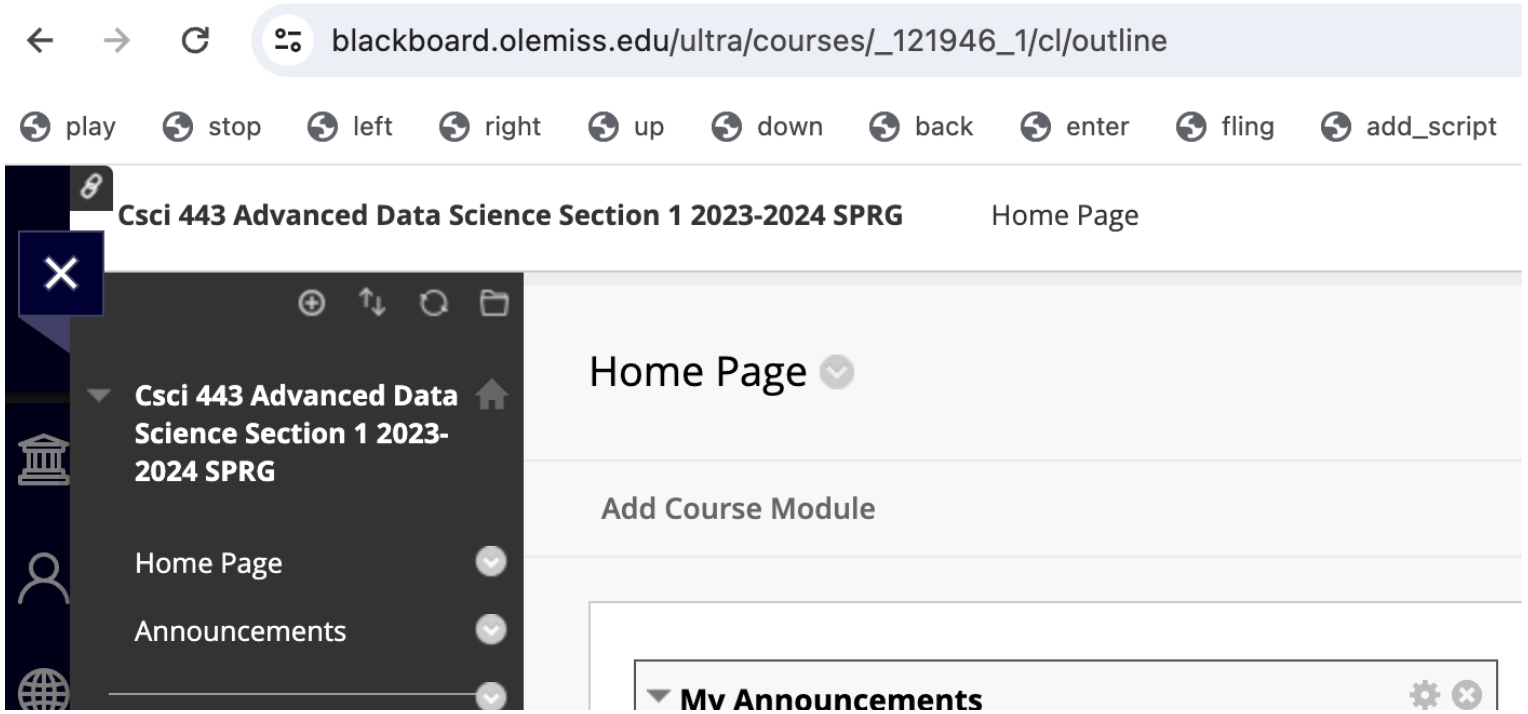
Due February 15

DATES OF INTEREST

February 8	HW2 handed out
February 15	HW2 due, HW3 handed out
February 22	HW3 due
February 27	Review
February 29	Midterm (must be before progress reports)
March 4	Progress Reports
March 8	Deadline for Withdrawal
March 9-17	Spring Break

BLACKBOARD

Slides up through lecture 4 on blackboard.

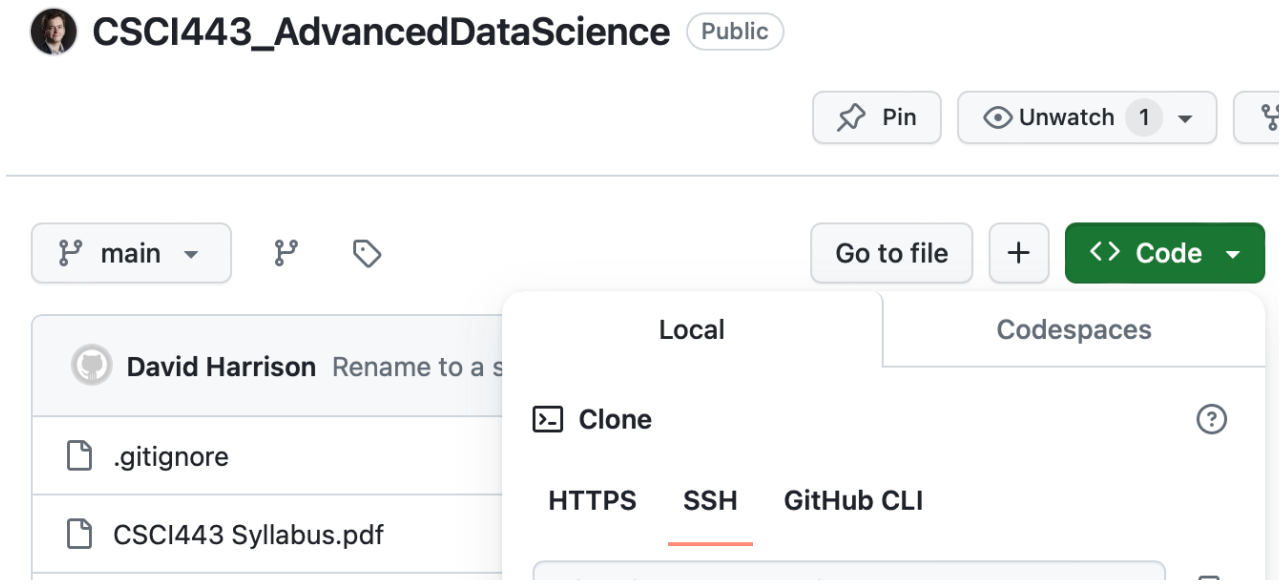


GITHUB

Lecture slides and examples committed to GitHub also up through lecture 4.

The project is at

https://github.com/dosirrah/CSCI443_AdvancedDataScience



THINGS I WANT TO COVER THIS WEEK

- Types of Data
- Definition of a distribution
- Ex: Gaussian
- Visualizing using histograms.
- Samples from a distribution
- Variance
- Standard deviation
- Mean absolute deviation
- Range
- Order statistics
- Percentile (quantile)
- Interquartile Range
- Box plots
- Correlation
- Correlation coefficient
- Correlation matrix
- Scatter plots

THINGS I WANT TO COVER THIS WEEK

- ~~Types of Data~~
- ~~Definition of a distribution~~
- ~~Ex: Gaussian~~
- ~~Visualizing using histograms.~~
- ~~Samples from a distribution~~
- ~~Variance~~
- ~~Standard deviation~~
- ~~Mean absolute deviation~~
- ~~Range~~

- Order statistics
- Percentile (quantile)
- Interquartile Range
- Box plots
- Correlation
- Correlation coefficient
- Correlation matrix
- Scatter plots

THINGS I WANT TO COVER TODAY

- ~~Types of Data~~
- ~~Definition of a distribution~~
- ~~Ex: Gaussian~~
- ~~Visualizing using histograms.~~
- ~~Samples from a distribution~~
- ~~Variance~~
- ~~Standard deviation~~
- ~~Mean absolute deviation~~
- ~~Range~~

- Order statistics
- Percentile (quantile)
- Interquartile Range
- Box plots
- Correlation
- Correlation coefficient
- Correlation matrix
- Scatter plots

Little more

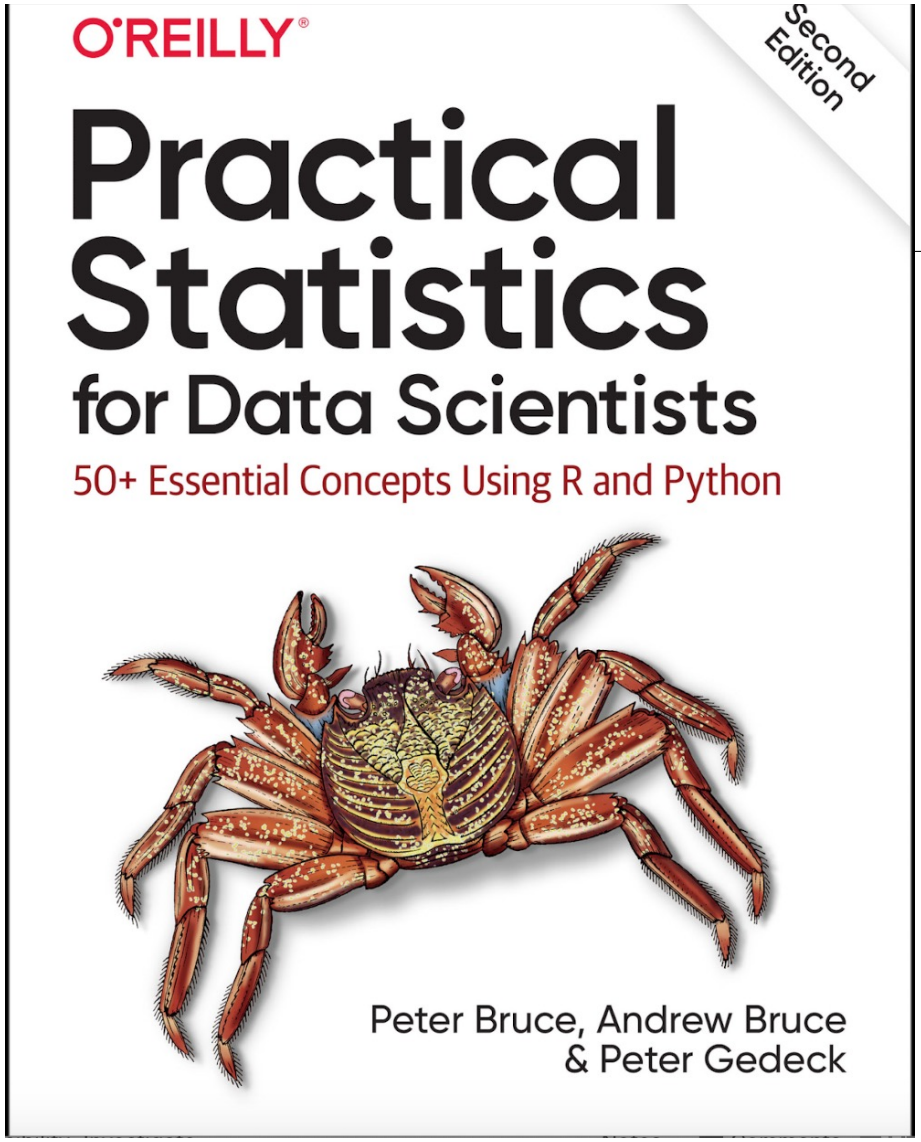
(we didn't get to this)

READ ABOUT

- Weighted mean
- Weighted median
- Trimmed mean

Now add

- Modes
- Bar charts
- Pie charts



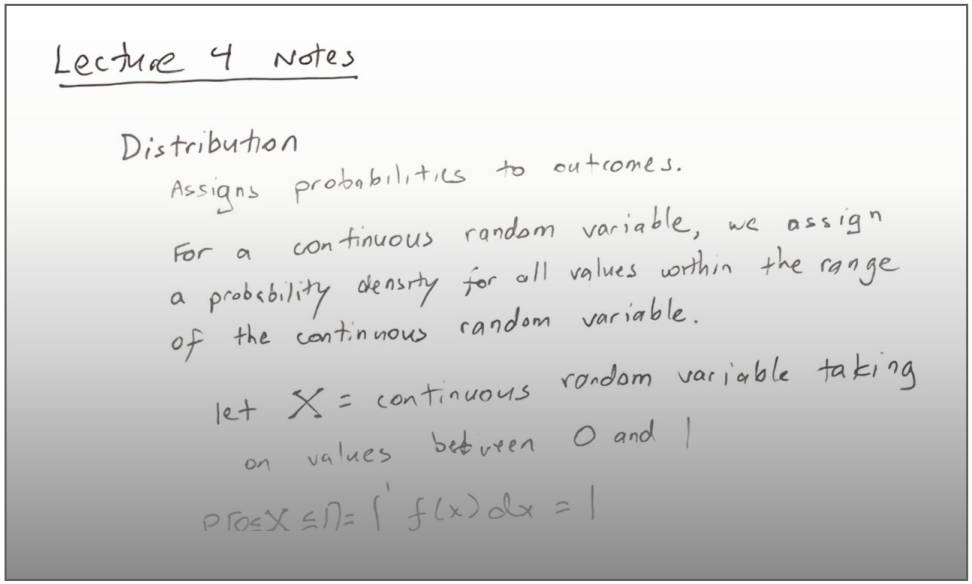
LECUTRE 4 NOTES

I posted lecture notes covering things I wrote on the board in lecture 4.

Ended up with little more detail than given in class specifically on

- Nature vs. Models vs. Samples
- Why histograms work

So, I will review it today. Please look at the notes.


Lecture 4 Notes

Distribution

Assigns probabilities to outcomes.

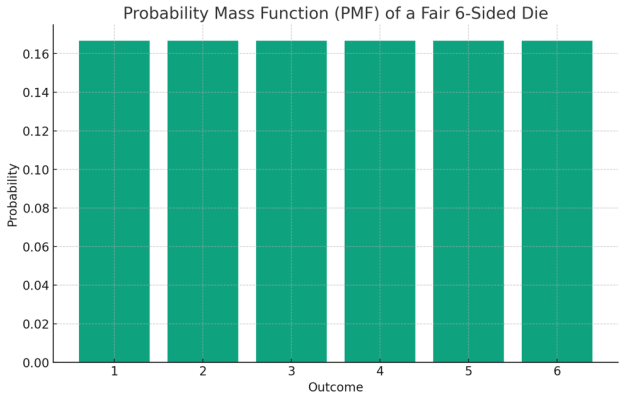
For a continuous random variable, we assign a probability density for all values within the range of the continuous random variable.

let X = continuous random variable taking on values between 0 and 1

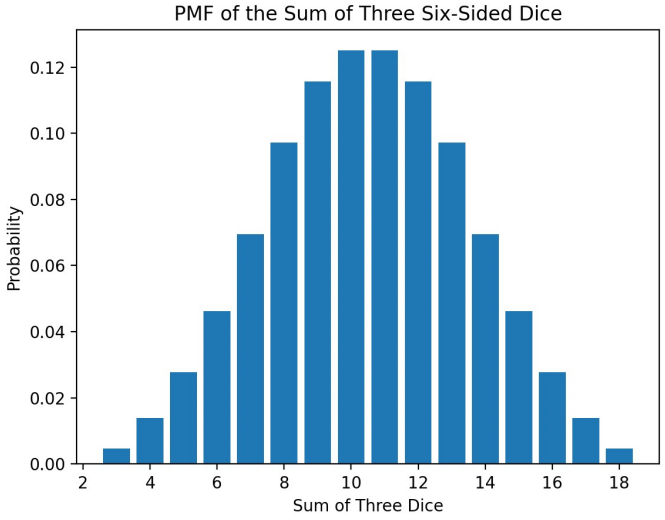
$P(0 \le X \le 1) = \int_0^1 f(x)\,dx = 1$

REVIEW OF LAST LECTURE

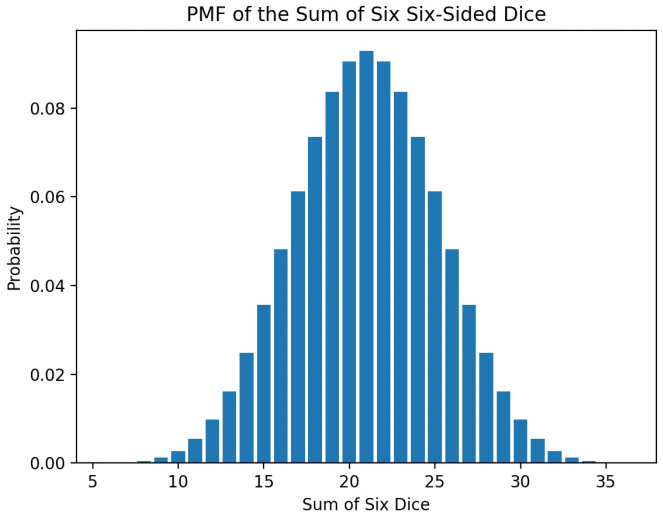
Central Limit Theorem says adding independent random variables tends towards Gaussian.



PMF of 1 die



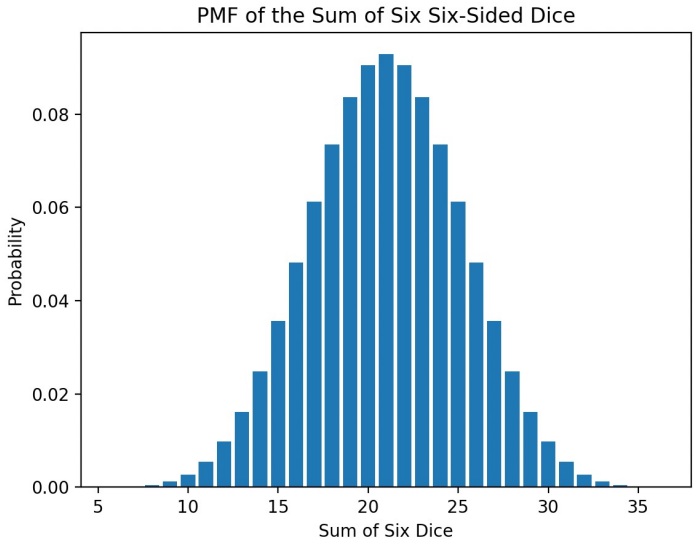
PMF of sum of 3 dice

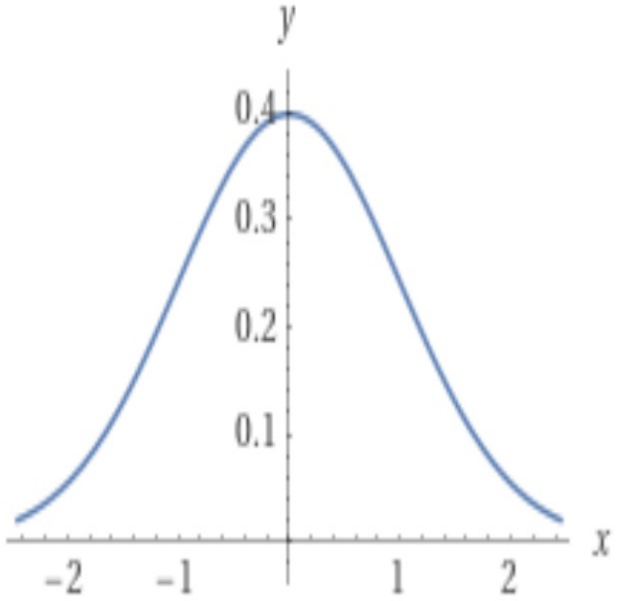


PMF of sum of 6 dice

REVIEW OF LAST LECTURE: SUM OF DICE VERSUS GAUSSIAN

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$





REVIEW OF LAST LECTURE: STANDARD DEVIATION AND GAUSSIAN DISTRIBUTION

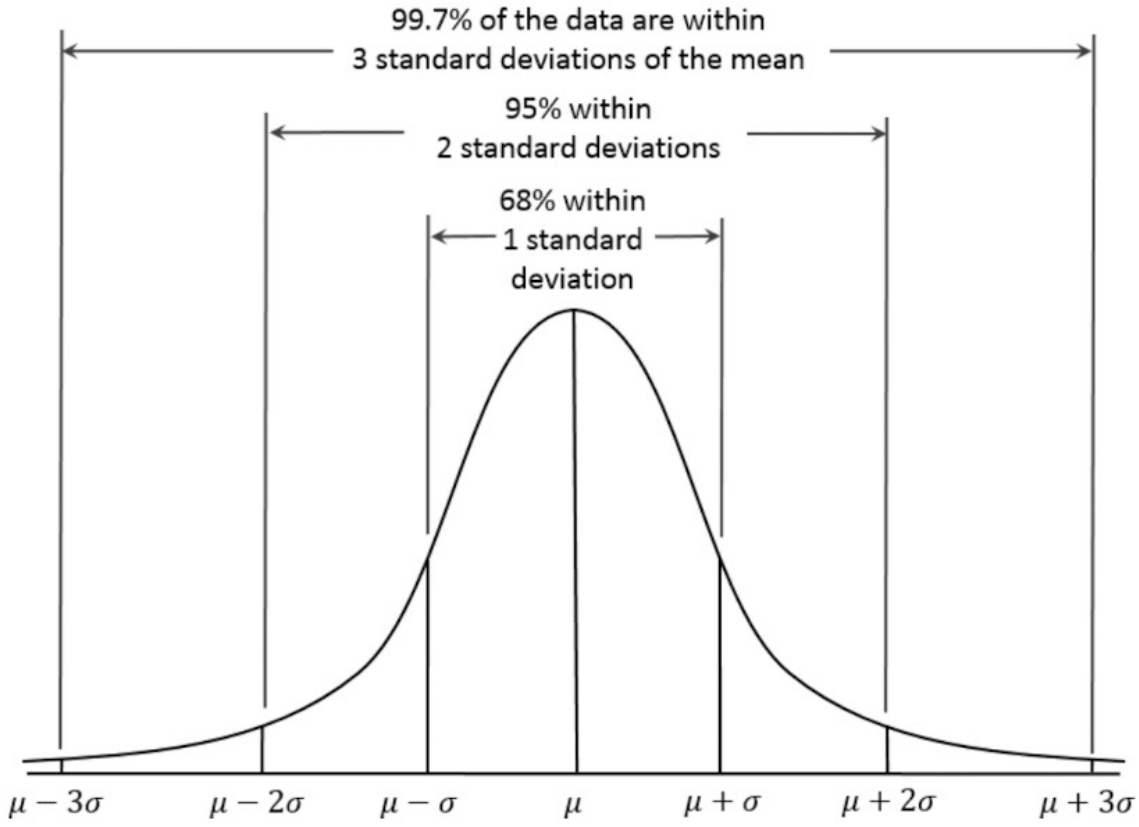


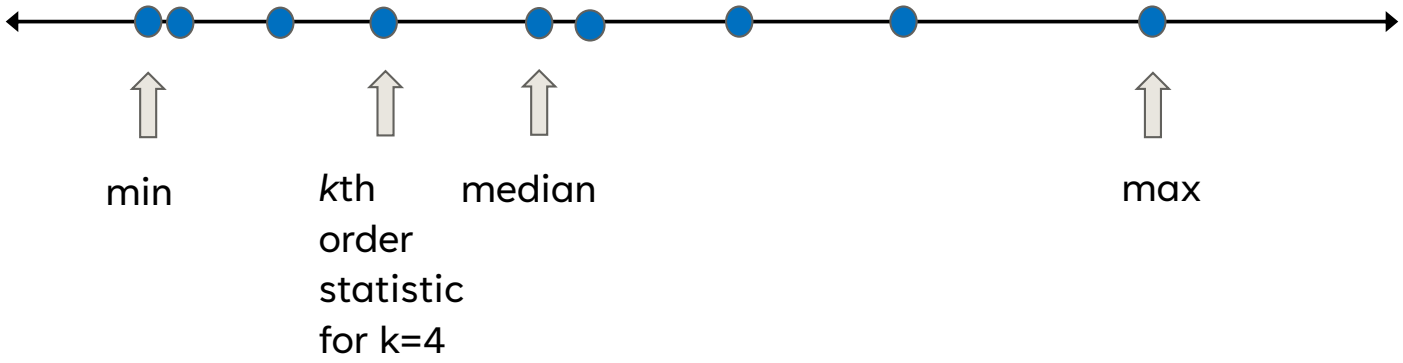
Figure 2-10. Normal curve

ORDER STATISTICS

A random variable assign real values to outcomes.

We order outcomes based on this real value.

“Order statistics” are based on this order.

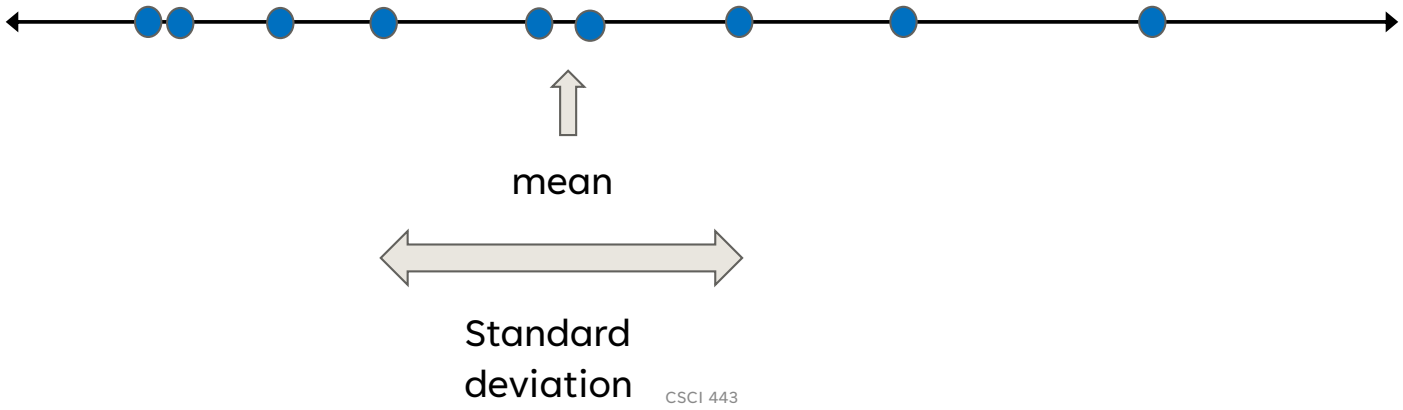


NOT ORDER STATISTICS

An order statistic is the value of a specific sample that arises from arranging the samples in ascending (or descending order).

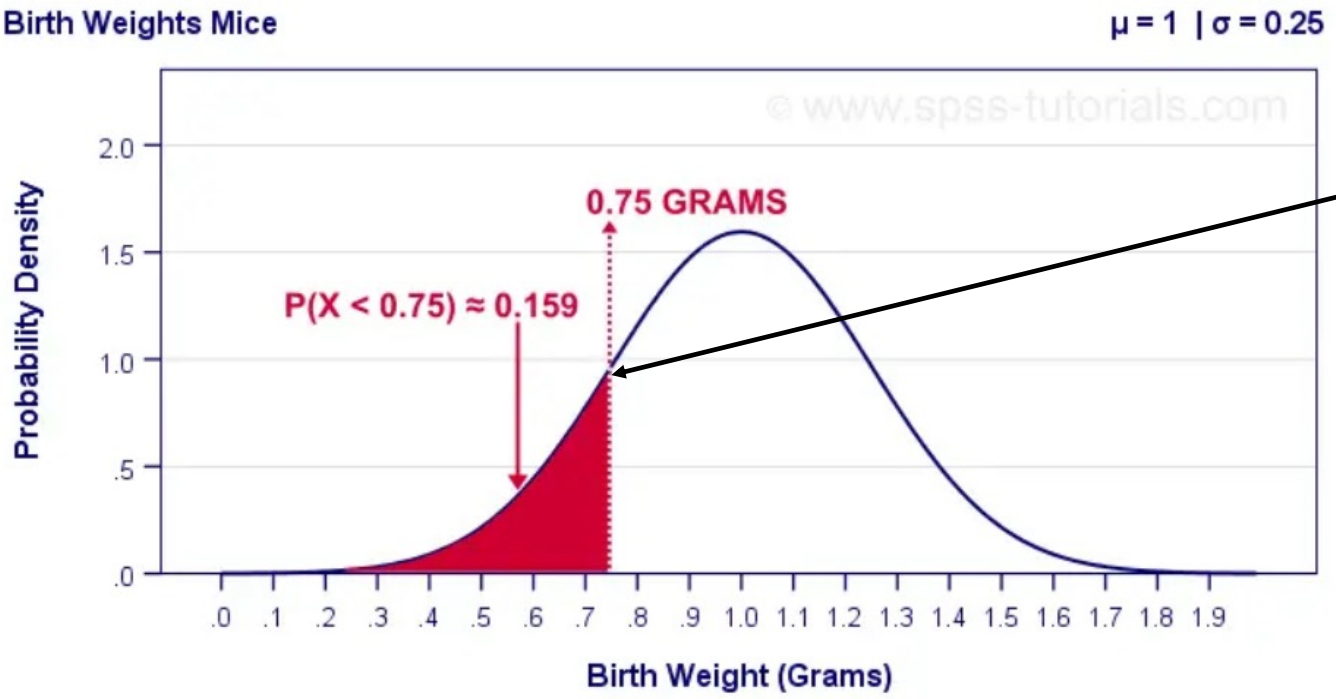
- Special case. Median with even number of samples is average of two sample values, but still often called an order statistic

Range, mean, mean absolute deviation, standard deviation are NOT order statistics.



HOW TO COMPUTE PERCENTILE OF A DISTRIBUTION

Computing p^{th} percentile ($p\%$), find the x in which $p\%$ of the probability mass below x.



p=15.9%,
x=0.75

15.9% the distribution falls below 0.75 grams.

HOW TO COMPUTE PERCENTILE OF A DISTRIBUTION

Computing p^{th} percentile ($p\%$), find the x in which $p\%$ of the probability mass below x.

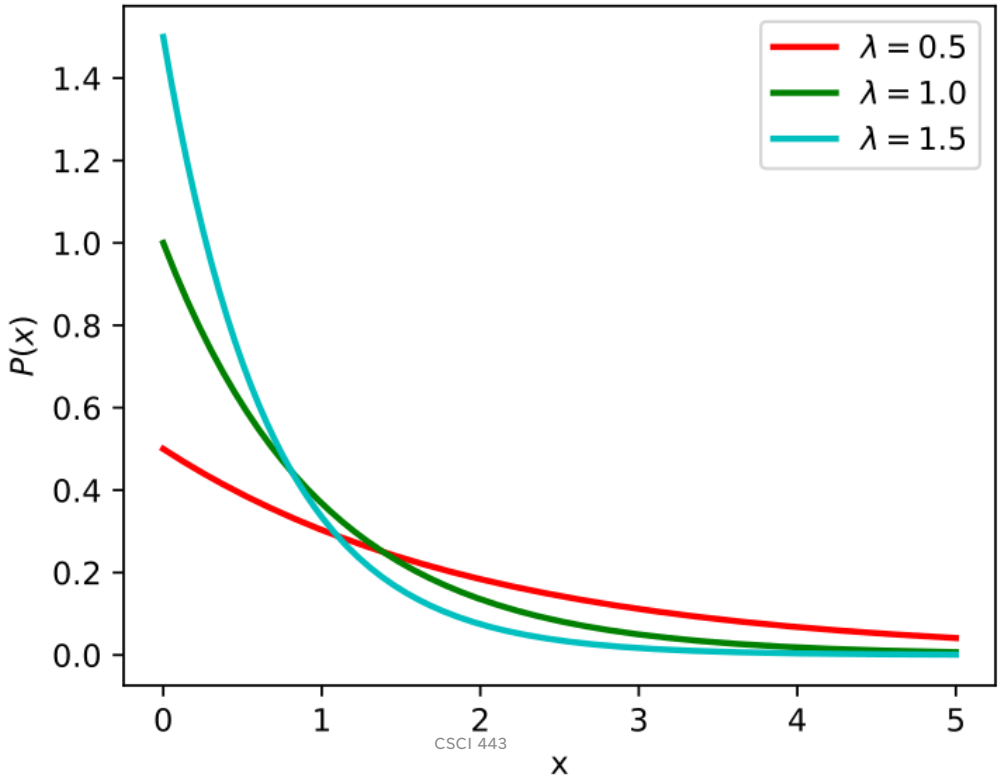
.

$$P[X < a] = \int_{-\infty}^{x=a} f(x)dx = p$$

Solve for a such that the integral equals p where p is the desired percentile expressed as a fraction in [0,1].

EXAMPLE: EXPONENTIAL DISTRIBUTION

Probability density function of an exponential distribution



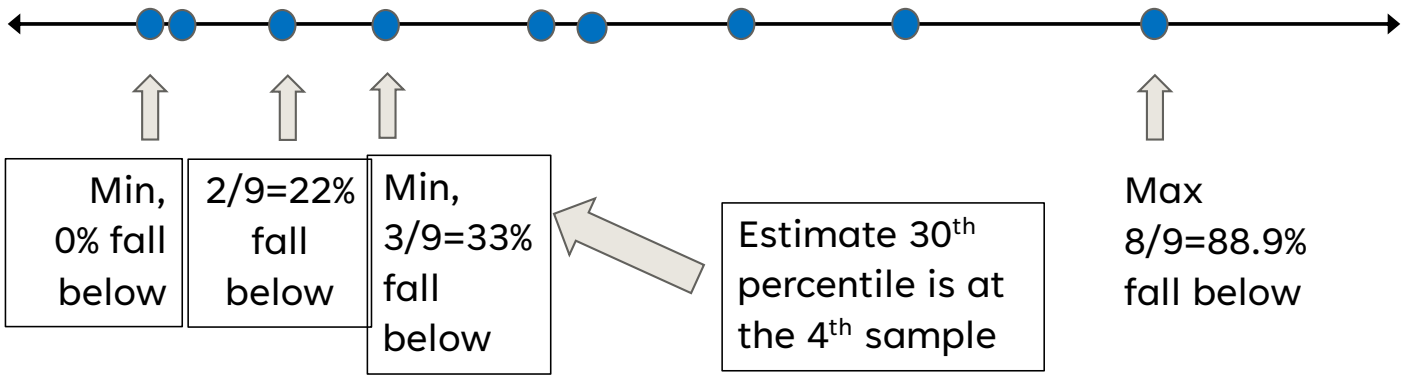
ESTIMATE PERCENTILE FROM A SAMPLE

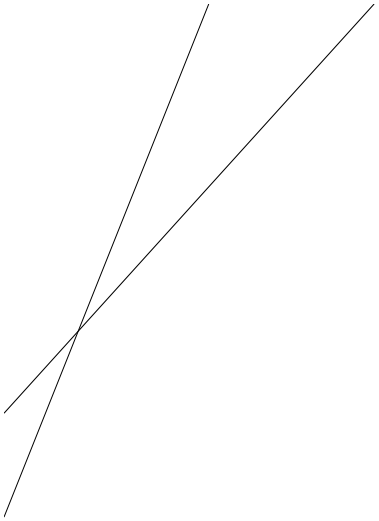
If looking for the pth percentile,

1. Order the samples on the real line
2. Count in ascending order until p% of the samples fall below your sample. This sample estimates your pth percentile for the underlying distribution.

- the estimate is likely slightly high especially with small sample sizes.

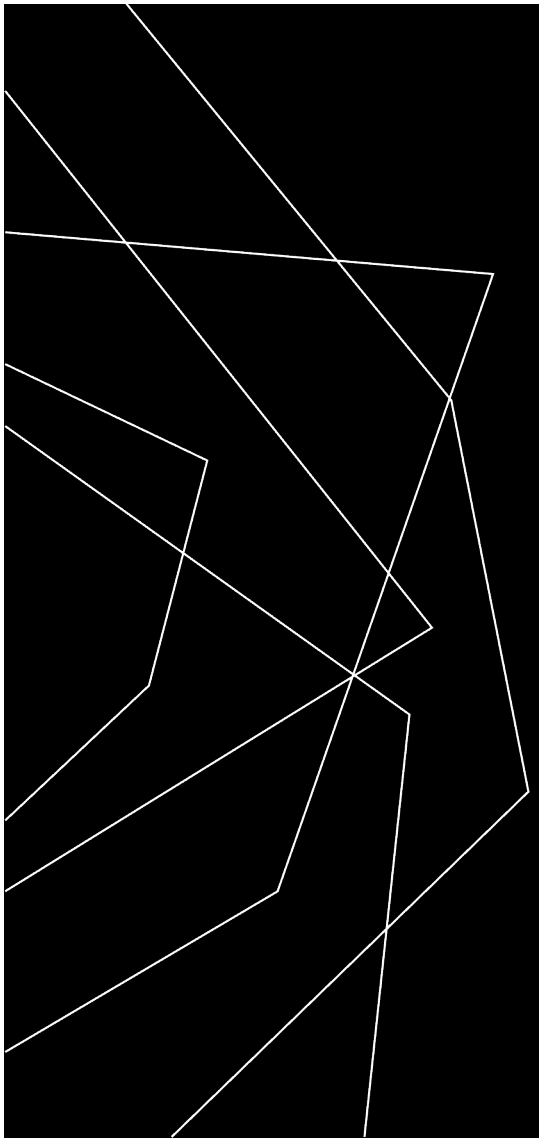
Example looking for 30th percentile.





MOST COMMON SOLUTION IS TO LINEARLY INTERPOLATE

See lecture notes



THANK YOU

David Harrison

Harrison@cs.olemiss.edu