A series of thin, black, intersecting lines of various orientations and lengths, creating a complex, abstract geometric pattern in the upper left portion of the slide.

# CSCI 692: LECTURE 8 EVALUATING GAUSSIAN, BOOSTING, CONFIDENCE INTERVALS

Professor David Harrison



# OFFICE HOURS

Tuesday

4:00-5:00 PM

Wednesday

12:30-2:30 PM



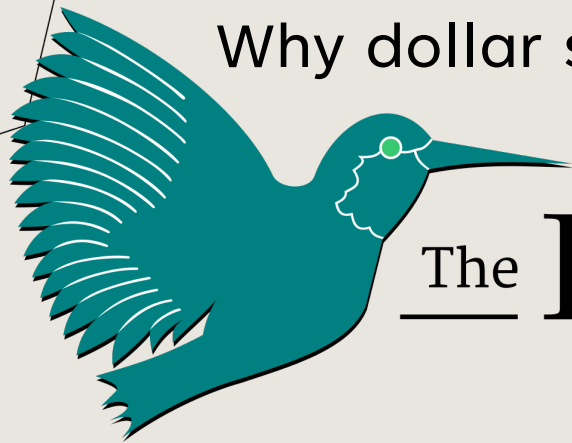
# HOMework 2

Due Today at 11:00 PM.

Today is February 15

`$E$, $X$, $ $ $ $`

Why dollar signs in the homework?



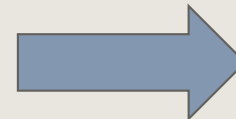
The **L<sup>A</sup>T<sub>E</sub>X** Project

`$E$`



*E*

`$\int_{x=0}^{\infty} f(x) dx$`



$\int_0^{\infty} f(x) dx$



# OUTCOME VS OUTCOME

“In healthcare, the term trial is [...] understood to involve systematic investigations to assess medical interventions’ effectiveness and safety.

In statistics, the term *trial* refers to a single instance of conducting a random experiment [...]

Each individual roll constitutes a trial. [...]

Within statistics, the *outcome* is the result observed from a single trial, exemplified by rolling a die and obtaining a 5. In statistics, an outcome is not a statistical measure like the mean [...]. In a clinical or animal trial, an outcome might refer to a statistical value calculated across a group of patients, such as the mortality rate.”



# DATES OF INTEREST

February 8  
February 15  
February 15 /16  
February 22  
February 27  
February 29  
March 4  
March 8  
March 9-17

HW2 handed out  
HW2 due,  
HW3 handed out  
HW3 due  
Review  
Midterm (must be before progress reports)  
Progress Reports  
Deadline for Withdrawal  
Spring Break

# BLACKBOARD & GITHUB

Slides up through lecture 7 on blackboard.

Lecture slides and examples committed to GitHub also up through lecture 7.

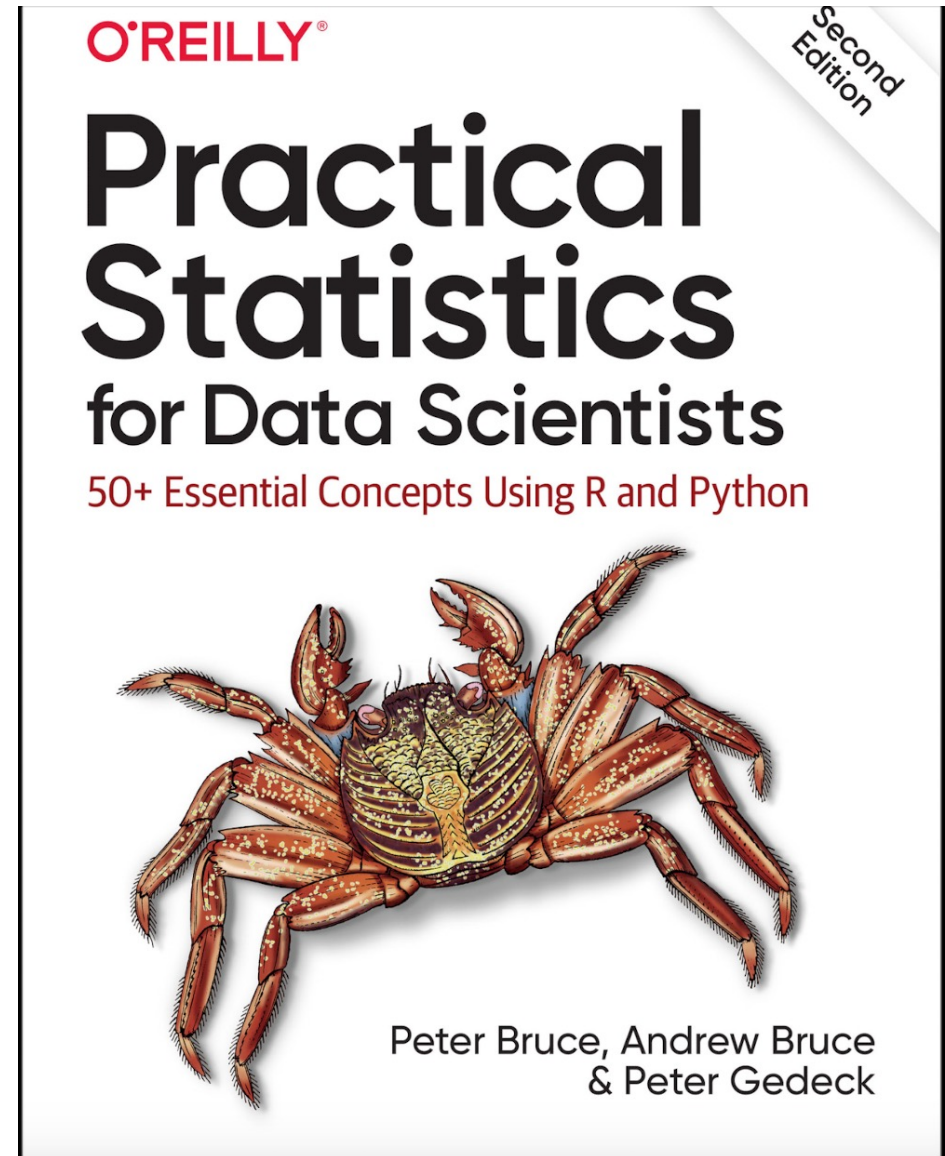
The project is at

[https://github.com/dosirrah/CSCI443\\_AdvancedDataScience](https://github.com/dosirrah/CSCI443_AdvancedDataScience)



## READ ABOUT

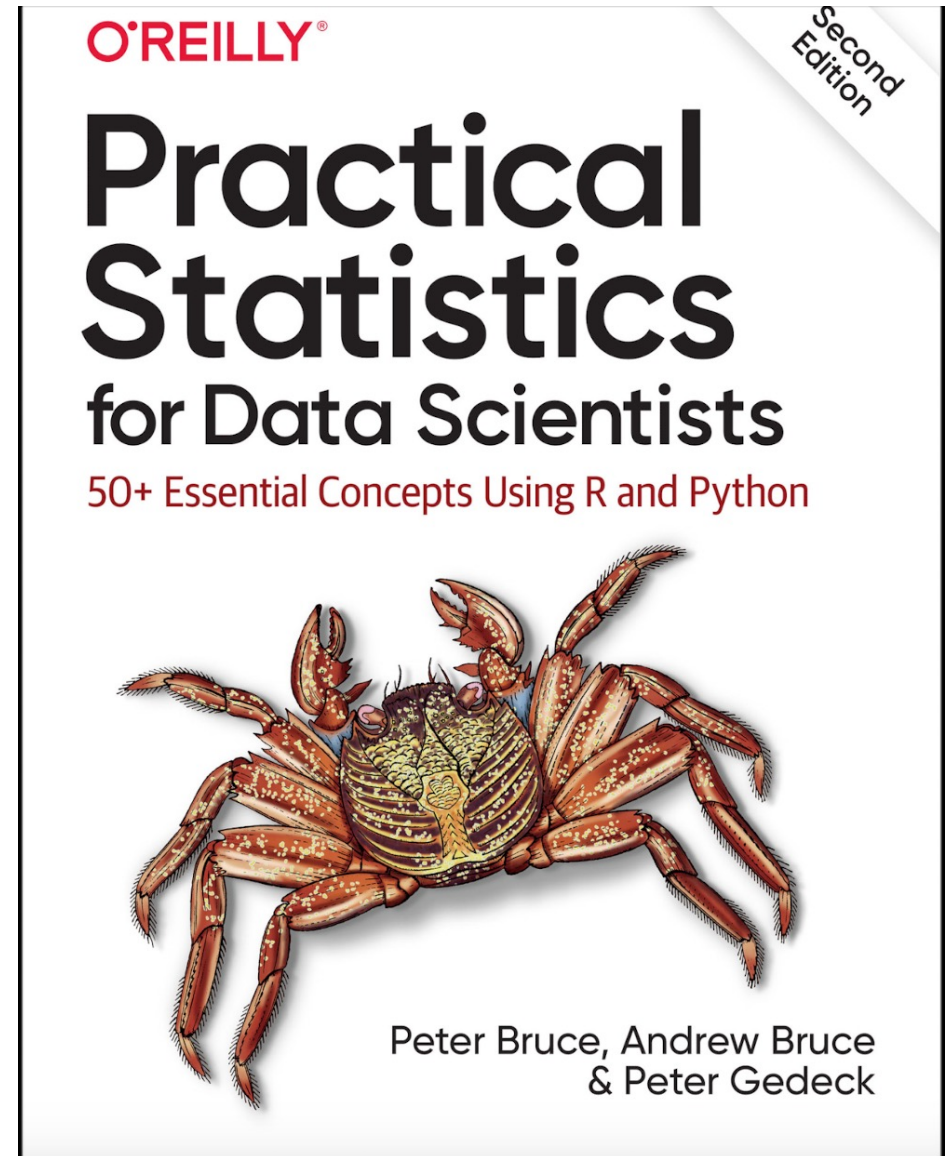
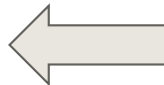
- Bias
  - Examples were already given in class, but book provides good example of selection bias.
- Random selection
  - Examples were already given in class, but book provides good example





ADD ONE MORE TO READ

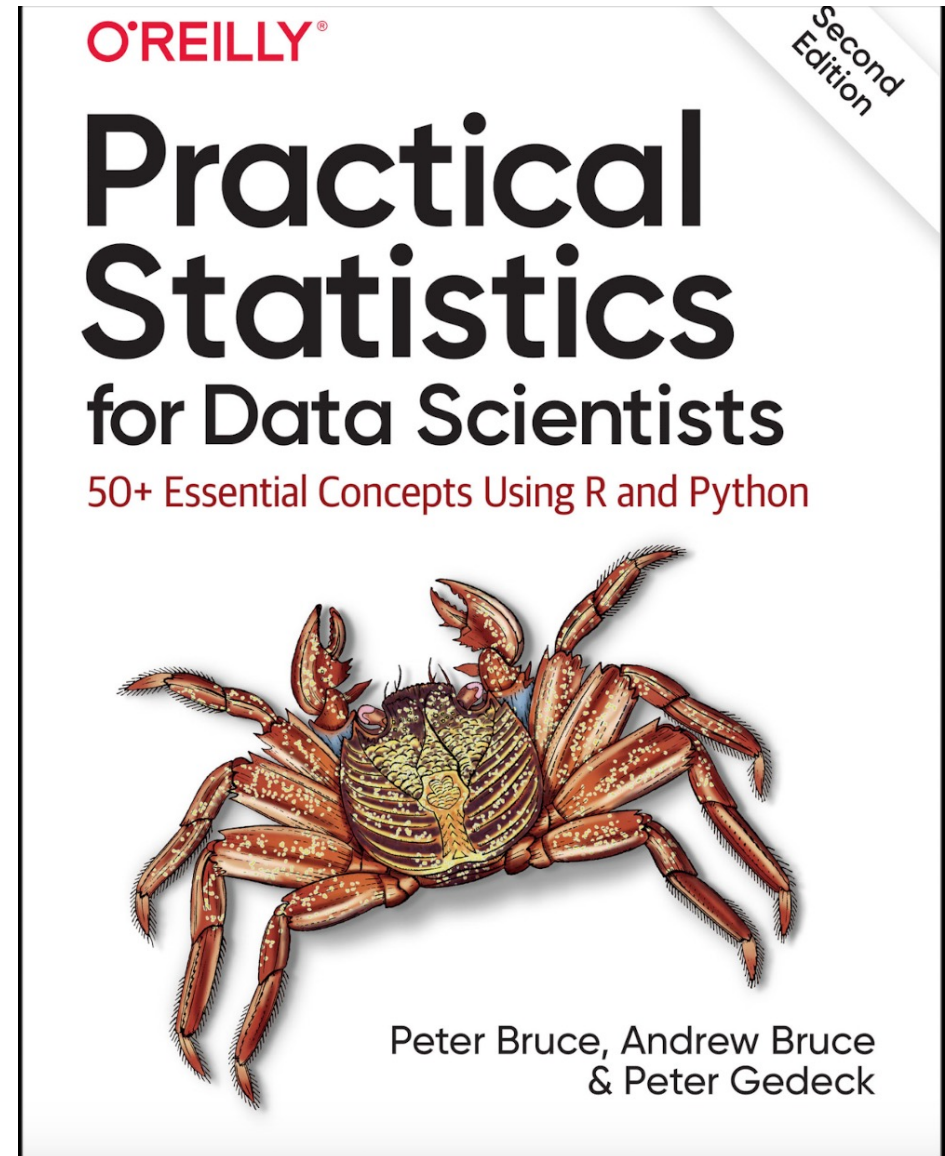
- Weighted mean
- Weighted median
- Trimmed mean
- Modes
- Bar charts
- Pie charts
- **Contour plots**



## SKIP PARTS OF CHAPTER 1

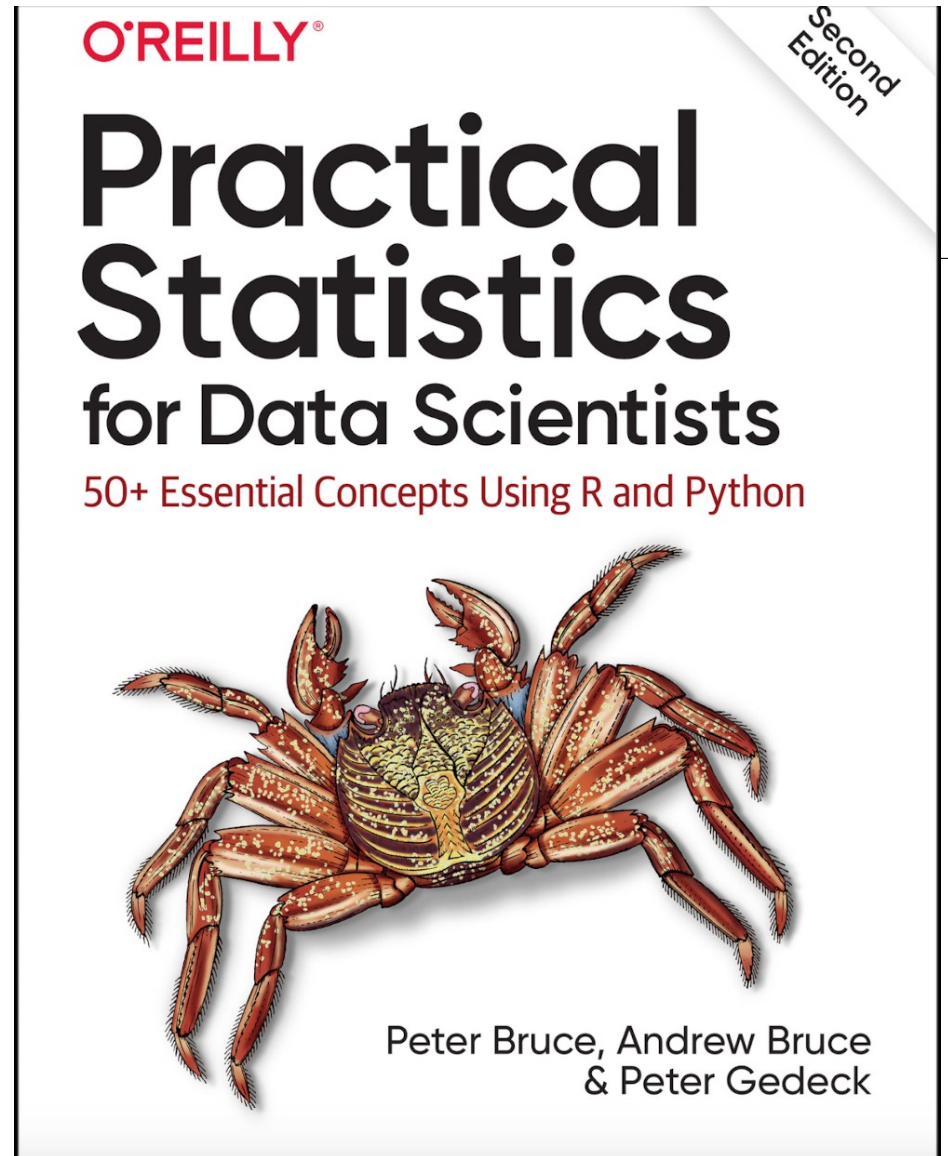
We will not cover in this class  
the following:

- Hexagonal binning
- Violin Plots
- Contingency Tables



## THINGS I WANT TO COVER TODAY

- Chapter 2
  - Distribution vs. Sample vs. Population
  - How to evaluate Gaussian (without a computer)





## PREVIOUS LECTURE: CORRELATION

### KEY TERMS FOR CORRELATION

#### ***Correlation coefficient***

A metric that measures the extent to which numeric variables are associated with one another (ranges from  $-1$  to  $+1$ ).

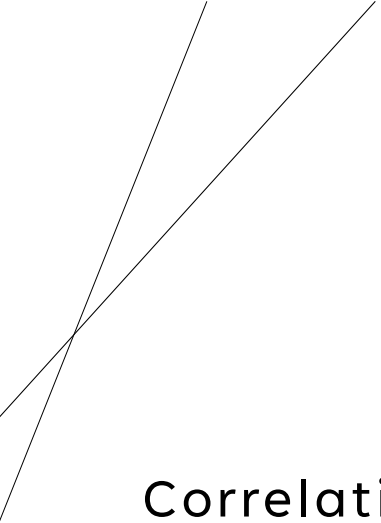
#### ***Correlation matrix***

A table where the variables are shown on both rows and columns, and the cell values are the correlations between the variables.

#### ***Scatterplot***

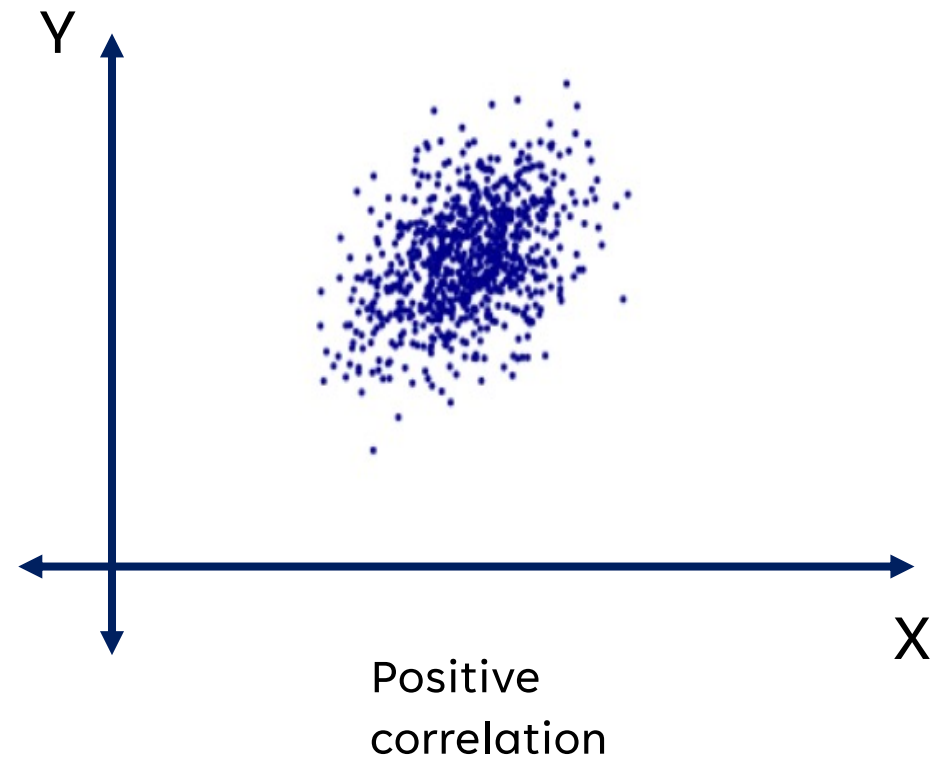
A plot in which the x-axis is the value of one variable, and the y-axis the value of another.

## PREVIOUS LECTURE: CORRELATION

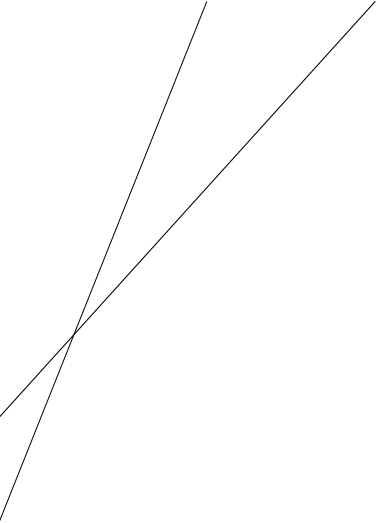


Correlation between two random variables means they tend to move together.

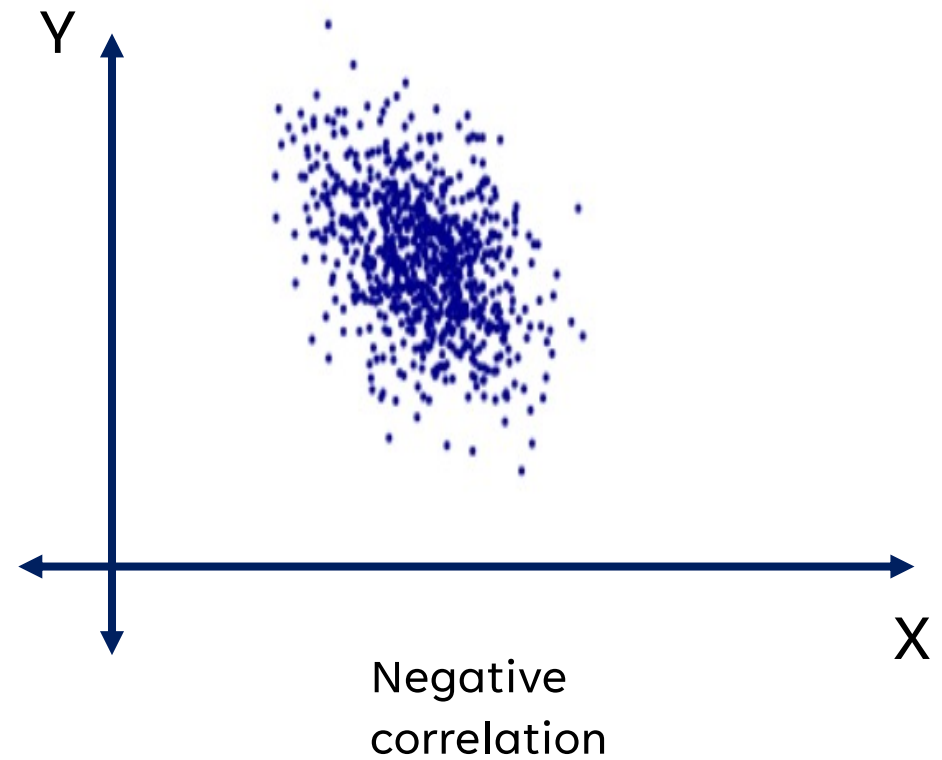
When one increases the other does, and vice versa.



## PREVIOUS LECTURE: CORRELATION



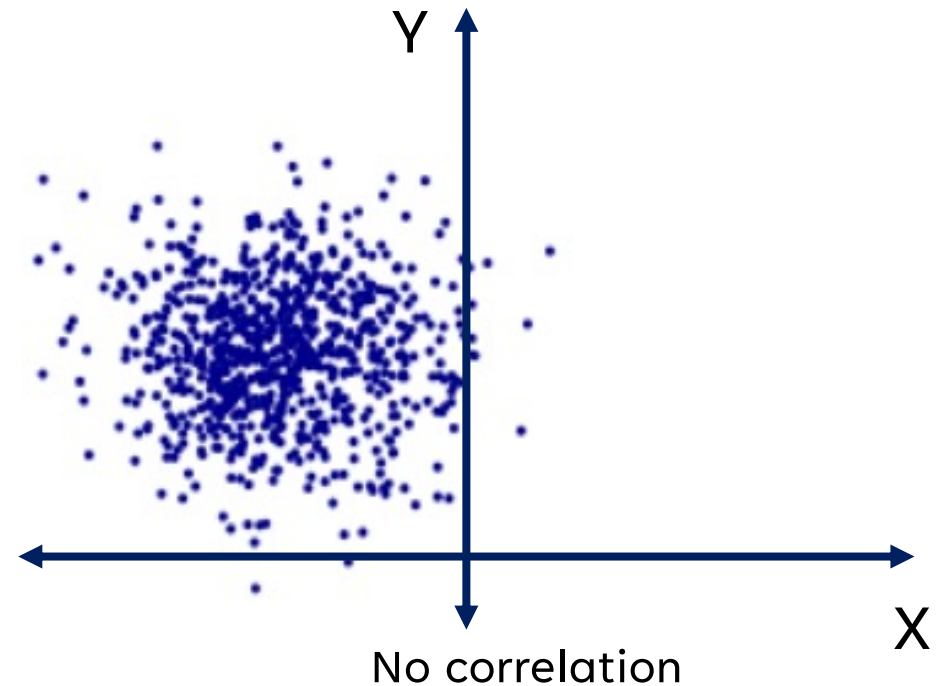
Negative correlation means they tend to move opposite to one another.



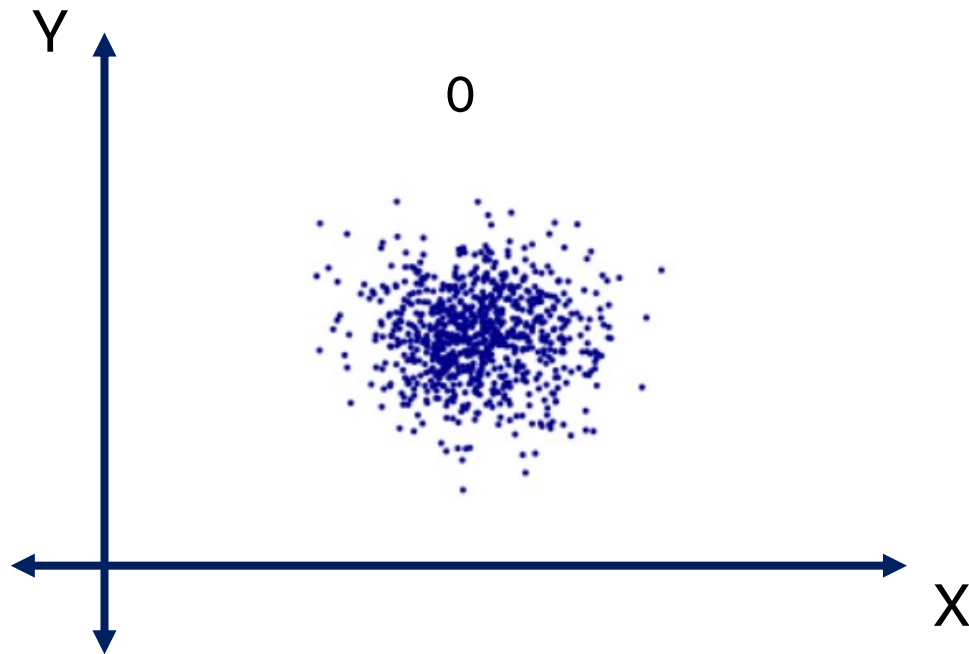
## PREVIOUS LECTURE: CORRELATION

There is no correlation if they do not move together.

On a Cartesian plane this appears as NO tilt to the scatter of samples.

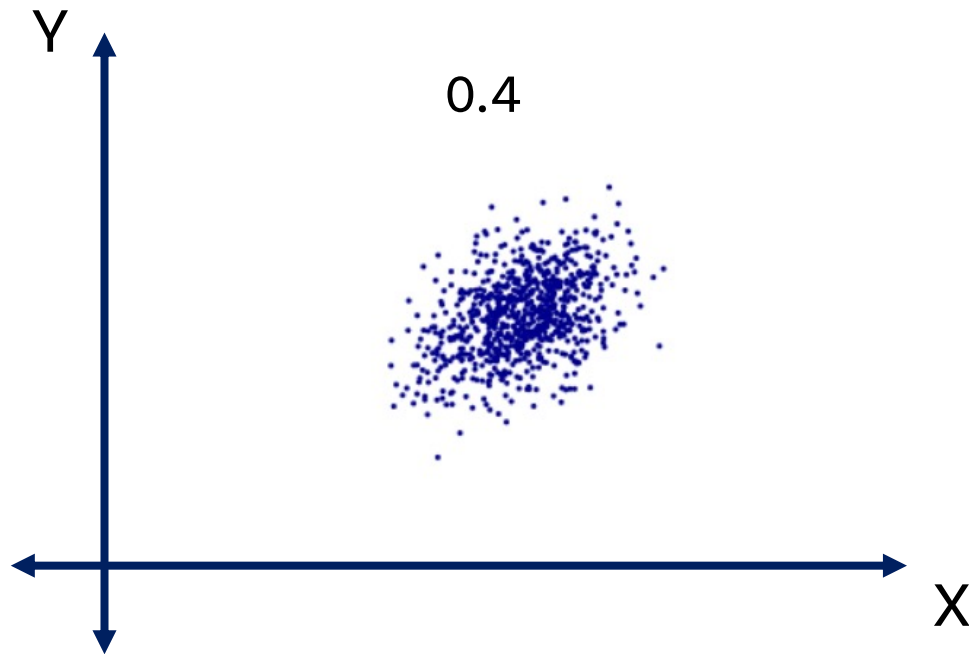


## PREVIOUS LECTURE: PEARSON CORRELATION





## PREVIOUS LECTURE: PEARSON CORRELATION



# COVARIANCE

Let  $X$  and  $Y$  be two random variables, covariance is

$$\text{cov}(X, Y) = E[(X - \mu_x)(Y - \mu_Y)]$$

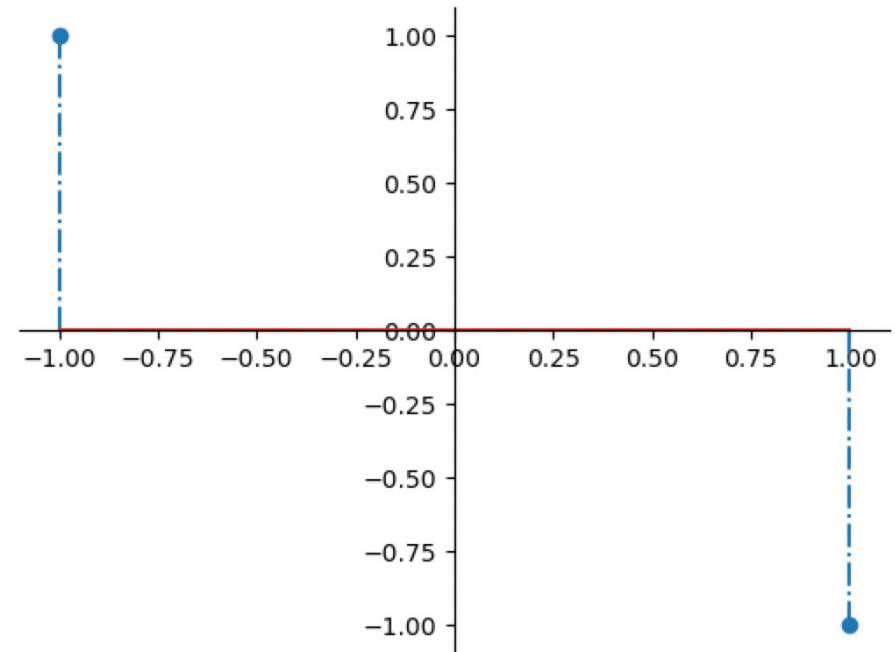
If  $\mu_x = 0$  and  $\mu_y = 0$  then this simplifies to

$$\text{cov}(X, Y) = E[XY]$$

$$E[XY] = \frac{1}{n} \sum_{i=1}^n x_i y_i$$

$$S = [(-1, 1), (1, -1)]$$

$$E[XY] = \frac{1}{2}[(-1 \cdot 1) + (1 \cdot -1)] = -1$$



REMINDER! Specifically chose mean = 0 to simplify equation.

## POPULATION COVARIANCE

Let  $X$  and  $Y$  be two random variables, covariance is

$$\text{cov}(X, Y) = E[(X - \mu_x)(Y - \mu_y)]$$

If  $\mu_x = 0$  and  $\mu_y = 0$

$$\text{cov}(X, Y) = E[XY]$$

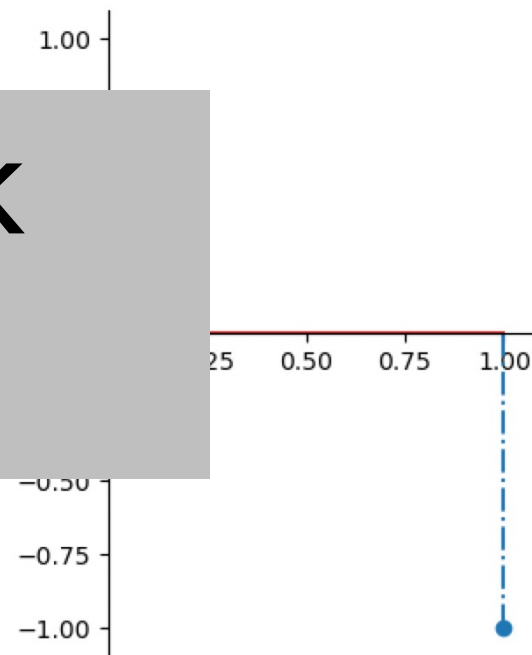
$$E[XY]$$

$$S = [(-1, 1), (1, -1)]$$

$$E[XY] = \frac{1}{2}[(-1 \cdot 1) + (1 \cdot -1)] = -1$$

REMINDER! Specifically chose mean = 0 to simplify equation.

I'll come back to this.



# COVARIANCE

Let  $X$  and  $Y$  be two random variables, covariance is

$$\text{cov}(X, Y) = E[(X - \mu_x)(Y - \mu_Y)]$$

If  $\mu_x = 0$  and  $\mu_y = 0$  then this simplifies to

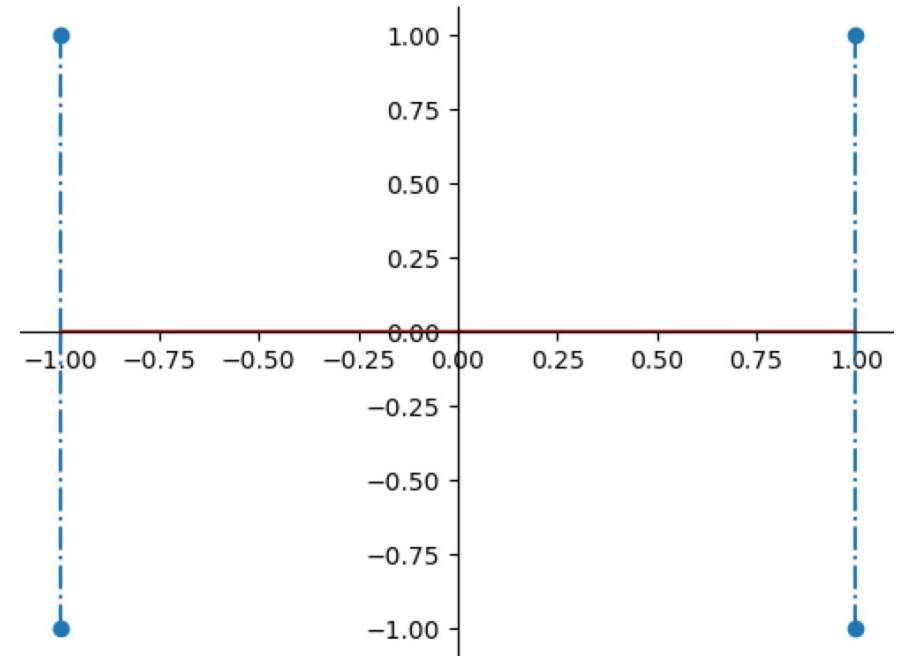
$$\text{cov}(X, Y) = E[XY]$$

$$E[XY] = \frac{1}{n} \sum_{i=1}^n x_i y_i$$

$$S = [(-1, 1), (-1, -1), (1, 1), (1, -1)]$$

$$E[XY] = \frac{1}{4}[(-1 \cdot 1) + (-1 \cdot -1) + (1 \cdot 1) + (1 \cdot -1)] = 0$$

REMINDER! Specifically chose mean = 0 to simplify equation.



## POPULATION COVARIANCE

Let  $X$  and  $Y$  be two random variables, covariance is

$$\text{cov}(X, Y) = E[(X - \mu_x)(Y - \mu_y)]$$

If  $\mu_x = 0$  and  $\mu_y = 0$

$$\text{cov}(X, Y) = E[XY]$$

$$E[XY]$$

$$S = [(-1, 1), (-1, -1), (1, 1), (1, -1)]$$

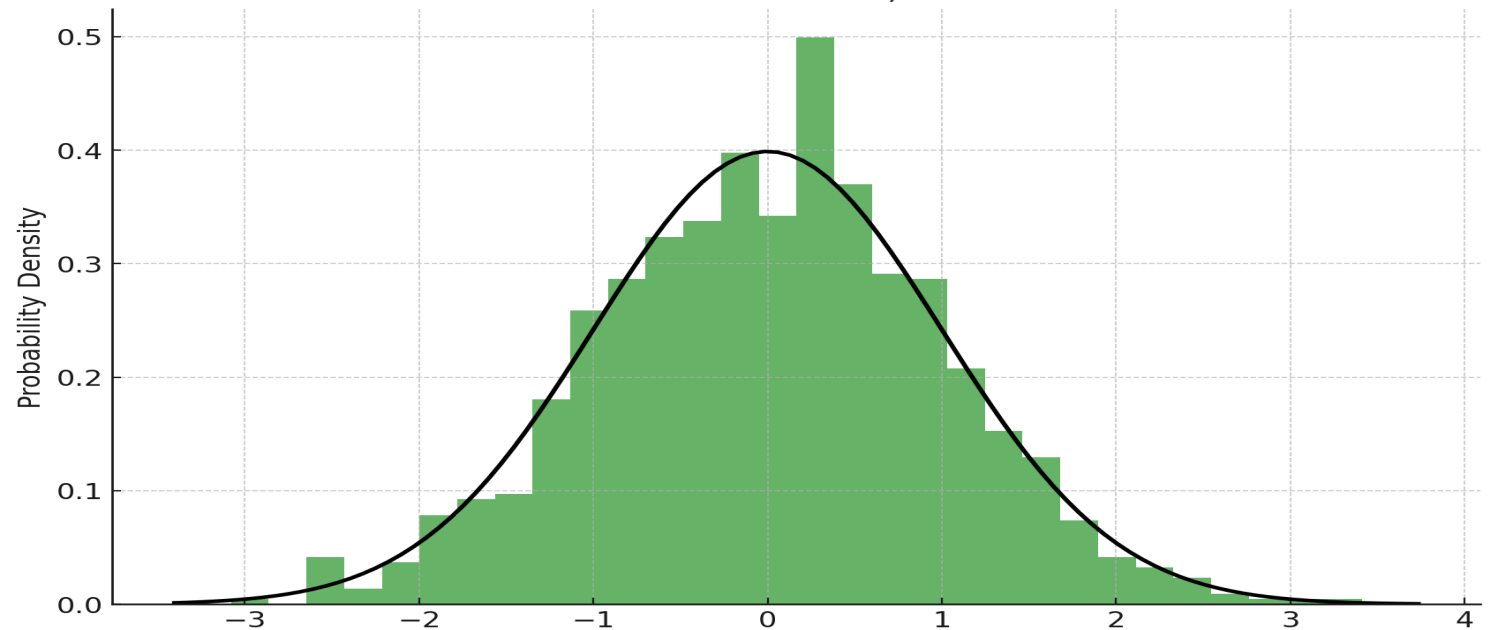
$$E[XY] = \frac{1}{4}[(-1 \cdot 1) + (-1 \cdot -1) + (1 \cdot 1) + (1 \cdot -1)] = 0$$

REMINDER! Specifically chose mean = 0 to simplify equation.

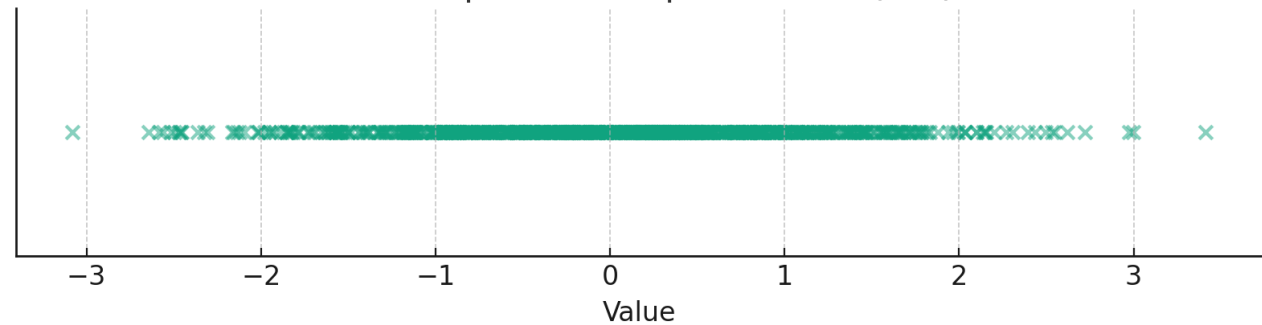
I'll come back to this.

## DISTRIBUTION VS. SAMPLE

Fit results:  $\mu = 0.00$ ,  $\text{std} = 1.00$



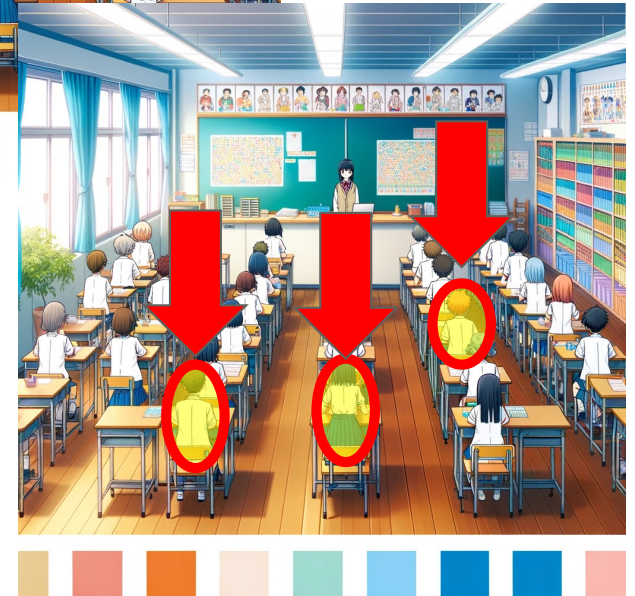
Scatter plot of samples from  $N(0,1)$



- Random variable obeys a distribution.
- Draw samples of a random variable.
- We did this in HW2.

## POPULATION VS. SAMPLE

- Population = my samples include all instances.
  - **All people in a class**
  - **All voters on election day**
- Sample
  - **Subset of people in the class.**
  - **Ex: poll a few voters.**





## WHY SAMPLE?

- Too expensive or time-consuming to talk to everyone in the population.
- Or when considering natural phenomena
  - **Alpha decay (U-238  $\rightarrow$  Th-234)**
    - Time series which goes on forever
  - **Matter across the universe**
    - Beyond our ability to count
- There are cases when we can never compute a statistic over every member

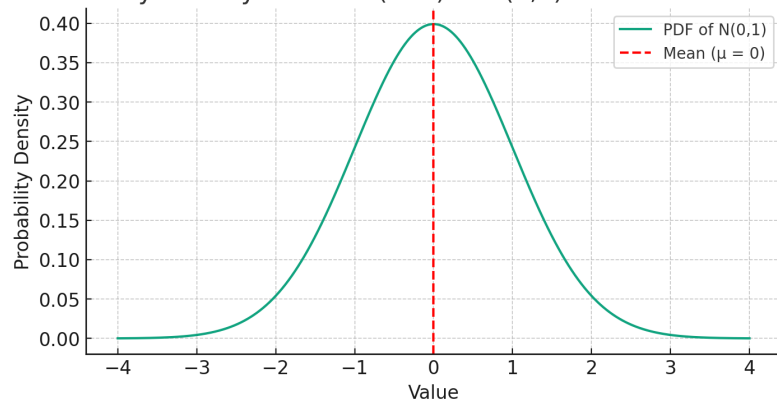


## DISTRIBUTION MEAN

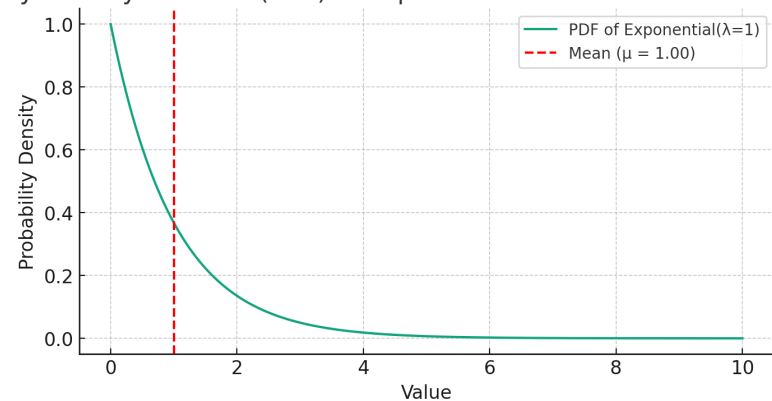
- Distribution mean ( $\mu$ ) can be determined by integrating the pdf.

$$\mu = E[X] = \int_{-\infty}^{\infty} x \cdot f(x) dx$$

Probability Density Function (PDF) of N(0,1) with Mean Denotation



Probability Density Function (PDF) of Exponential Distribution with Mean Denotation



## DISTRIBUTION MEAN VS. POPULATION MEAN VS. SAMPLE MEAN

- Distribution mean ( $\mu$ ) can be determined by integrating the pdf.

$$\mu = E[X] = \int_{-\infty}^{\infty} x \cdot f(x) dx$$

- Doesn't work if we don't know the probability density function
- Population mean ( $\mu$ ) where population is size  $N$

$$\mu = \frac{\sum_{i=1}^N X_i}{N}$$

$X_i$  = the  $i$ th member of the population.

- Sample mean ( $\bar{x}$ ) with sample size  $n$

$x_i$  = the  $i$ th sample in the sample set.

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$



## LAW OF LARGE NUMBERS

- We can use a sample to estimate a population mean or distribution mean.
- Why? Law of Large Numbers.
- Wikipedia says:

In **probability theory**, the **law of large numbers (LLN)** is a **mathematical theorem** that states that the **average** of the results obtained from a large number of independent and identical random samples converges to the true value, if it exists.<sup>[1]</sup> More formally, the LLN states that given a sample of independent and identically distributed values, the **sample mean** converges to the true **mean**.

$$\lim_{n \rightarrow \infty} \frac{1}{n} (X_1 + X_2 + \dots + X_n) = \mu$$



## DISTRIBUTION VS. POPULATION VS. SAMPLE VARIANCE

Distribution variance

$$E[(X - \mu)^2] \quad \sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx$$

Population variance

$$\sigma^2 = \frac{\sum_{i=1}^N (X_i - \mu)^2}{N}$$

Sample variance

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

# DISTRIBUTION VS. POPULATION VS. SAMPLE VARIANCE


Distribution variance

$$E[(X - \mu)^2] \quad \sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx$$

Population variance

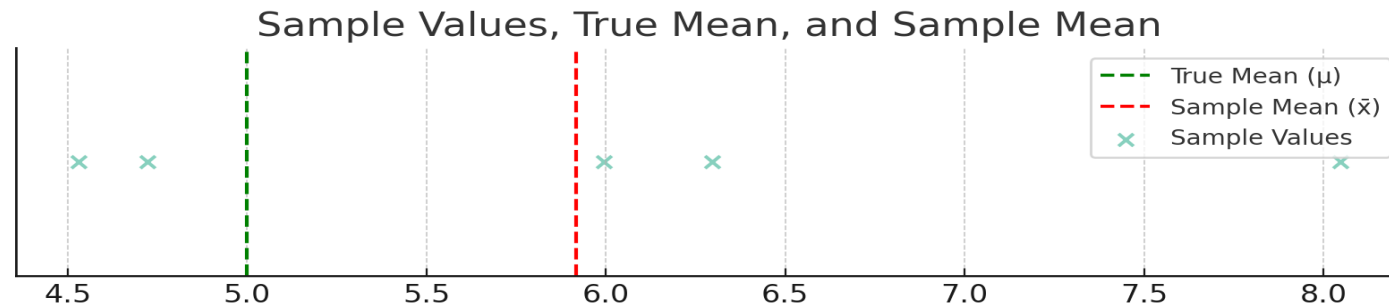
$$\sigma^2 = \frac{\sum_{i=1}^N (X_i - \mu)^2}{N}$$

Population variance

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$


# WHAT HAPPENS WITH N RATHER THAN N-1 IN THE DENOMINATOR?

Distribution variance



true mean ( $\mu$ ) = 5, standard deviation ( $\sigma$ ) = 2

Samples = [5.99, 4.72, 6.30, 8.05, 4.53]

Sample mean = 5.92

$$s_{\text{incorrect}}^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n} = \frac{\sum_{i=1}^5 (x_i - 5.918)^2}{5} = 1.605$$

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1} = \frac{\sum_{i=1}^5 (x_i - 5.918)^2}{5 - 1} = 2.006$$



## POPULATION VS. SAMPLE STANDARD DEVIATION

Standard deviation is the square root of the variance.

$$\sigma = \sqrt{\int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx}$$

Population standard deviation

$$\sigma = \sqrt{\frac{\sum_{i=1}^N (X_i - \mu)^2}{N}}$$

Sample standard deviation

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$$



## POPULATION VS. SAMPLE COVARIANCE

Distribution Covariance

$$\text{Cov}(X, Y) = \iint (x - \mu_X)(y - \mu_Y) f(x, y) dx dy$$

Population Covariance

$$\sigma_{XY} = \frac{1}{N} \sum_{i=1}^N (X_i - \mu_X)(Y_i - \mu_Y)$$

Sample Covariance

$$s_{XY} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$



# COVARIANCE

Let  $X$  and  $Y$  be two random variables, covariance is

$$\text{cov}(X, Y) = E[(X - \mu_x)(Y - \mu_Y)]$$

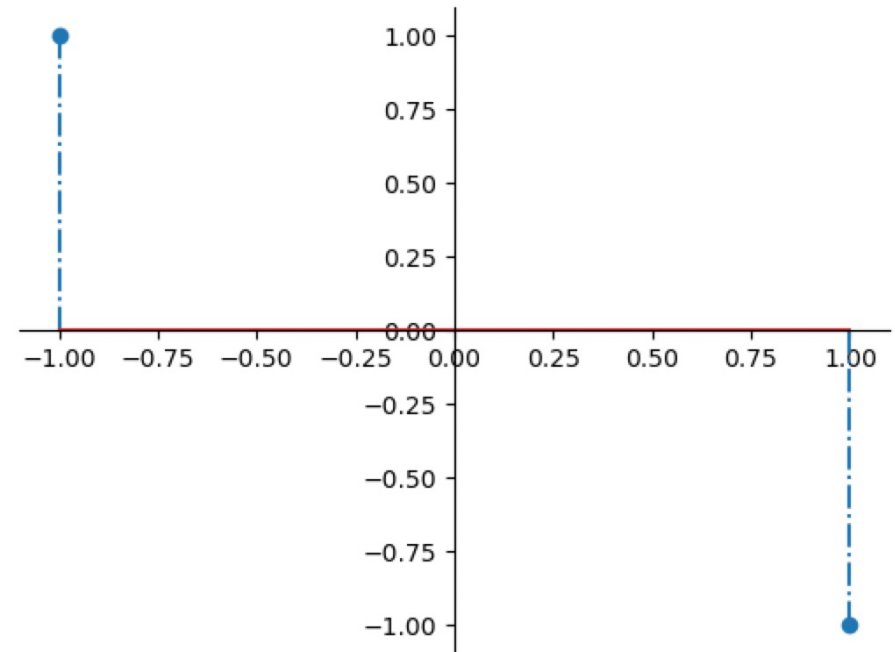
If  $\mu_x = 0$  and  $\mu_y = 0$  then this simplifies to

$$\text{cov}(X, Y) = E[XY]$$

$$E[XY] = \frac{1}{n} \sum_{i=1}^n x_i y_i$$

$$S = [(-1, 1), (1, -1)]$$

$$E[XY] = \frac{1}{2}[(-1 \cdot 1) + (1 \cdot -1)] = -1$$



Am I assuming population covariance or sample covariance?

# COVARIANCE

Let  $X$  and  $Y$  be two random variables, covariance is

$$\text{cov}(X, Y) = E[(X - \mu_x)(Y - \mu_Y)]$$

If  $\mu_x = 0$  and  $\mu_y = 0$  then this simplifies to

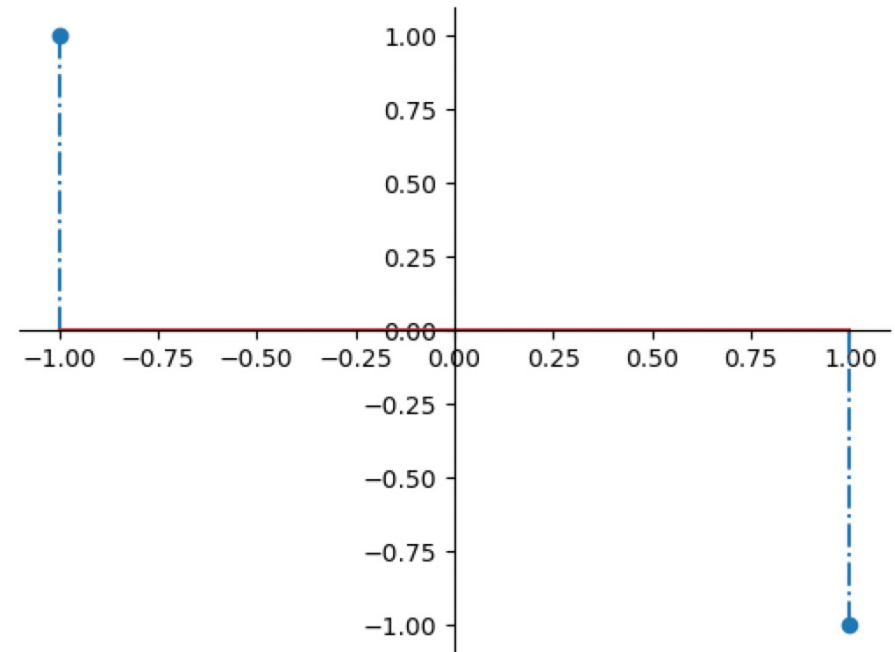
$$\text{cov}(X, Y) = E[XY]$$

$$E[XY] = \frac{1}{n} \sum_{i=1}^n x_i y_i$$

$$S = [(-1, 1), (1, -1)]$$

$$E[XY] = \frac{1}{2}[(-1 \cdot 1) + (1 \cdot -1)] = -1$$

This is an example of population covariance





# THANK YOU

David Harrison

[Harrison@cs.olemiss.edu](mailto:Harrison@cs.olemiss.edu)