CSCI 692: LECTURE 6
CORRELATION

Professor David Harrison

OFFICE HOURS

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

HOMEWORK 2

Due this Thursday.

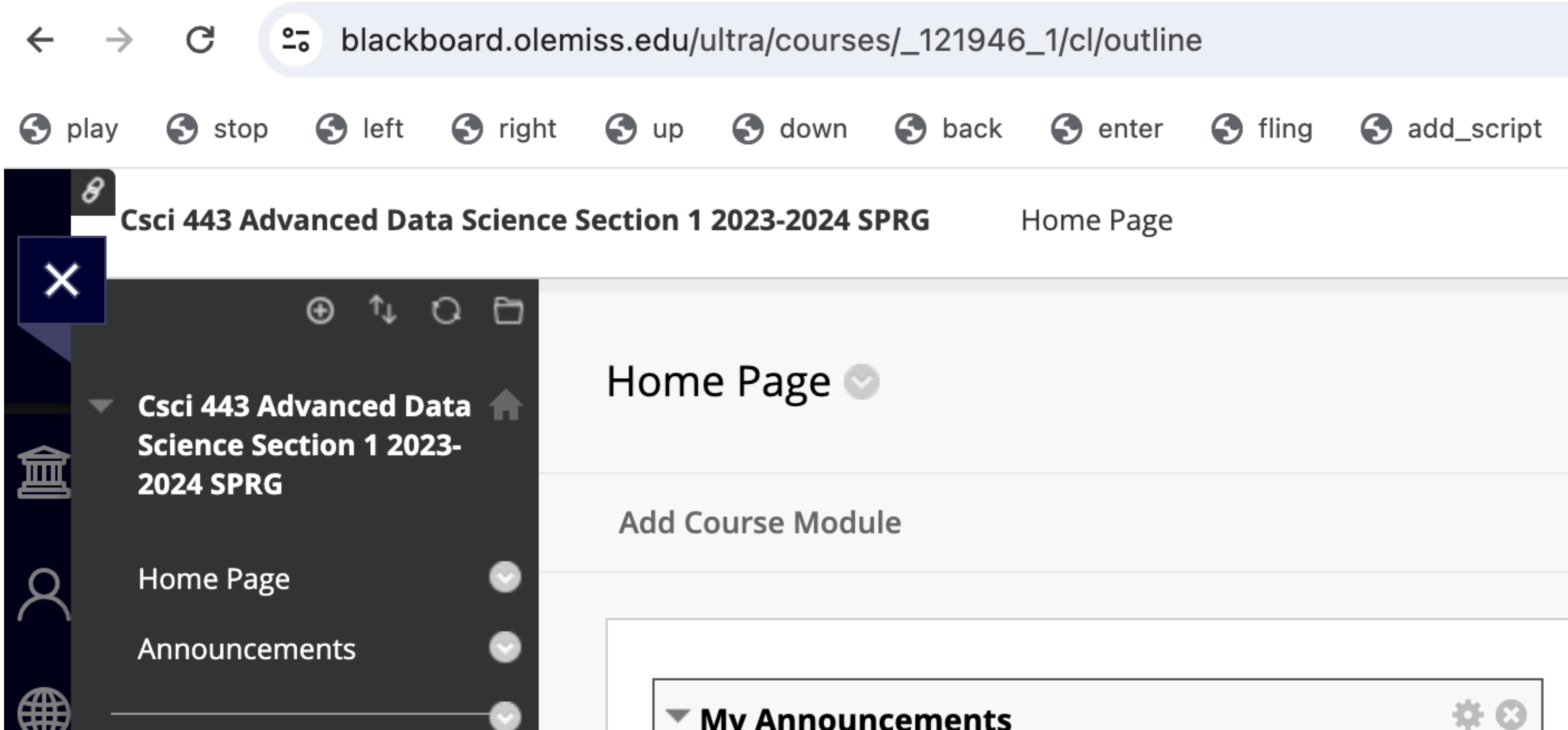
February 15, 11:00 PM

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 5 on blackboard.

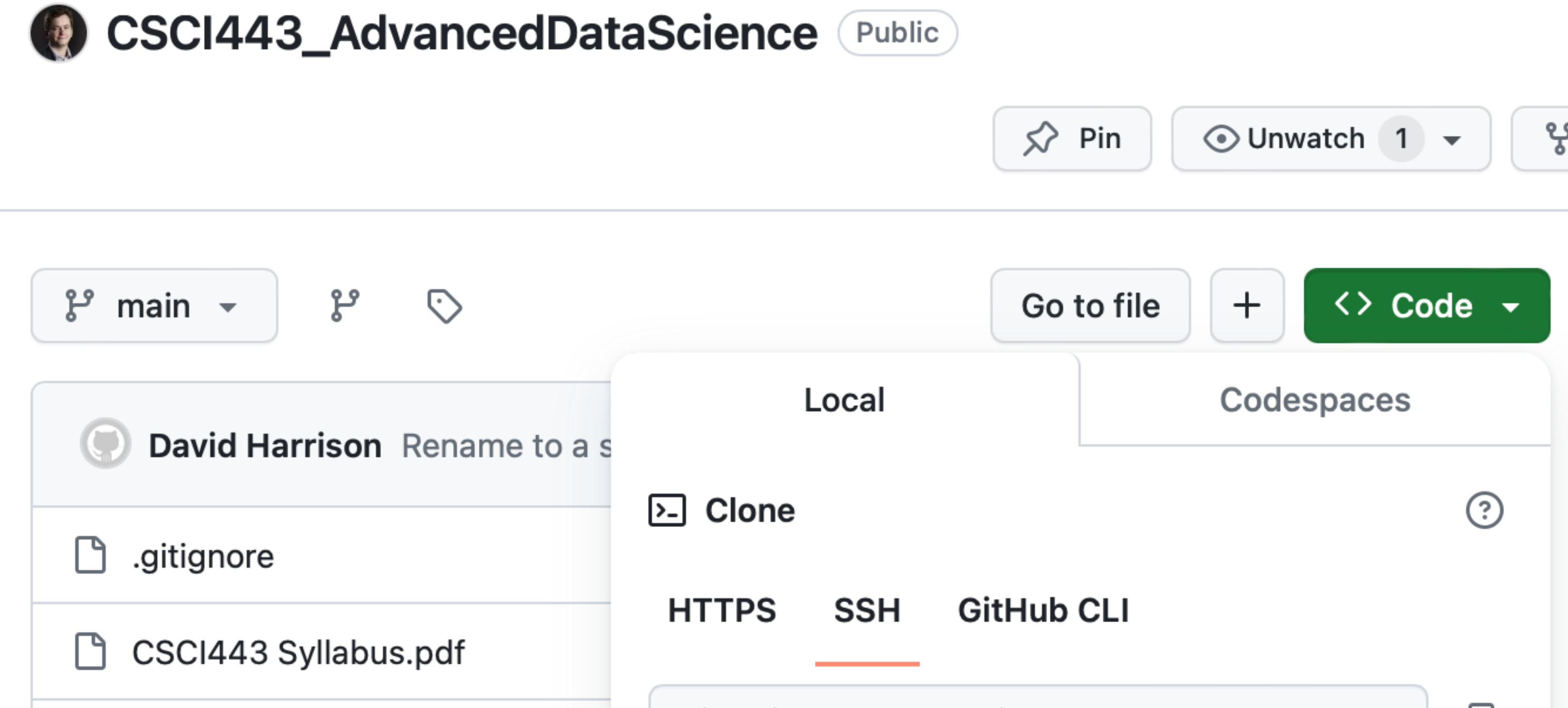


GITHUB

Lecture slides and examples committed to GitHub also up through lecture 5.

The project is at

https://github.com/dosirrah/CSCI443_AdvancedDataScience



THINGS I WANT TO COVER TODAY

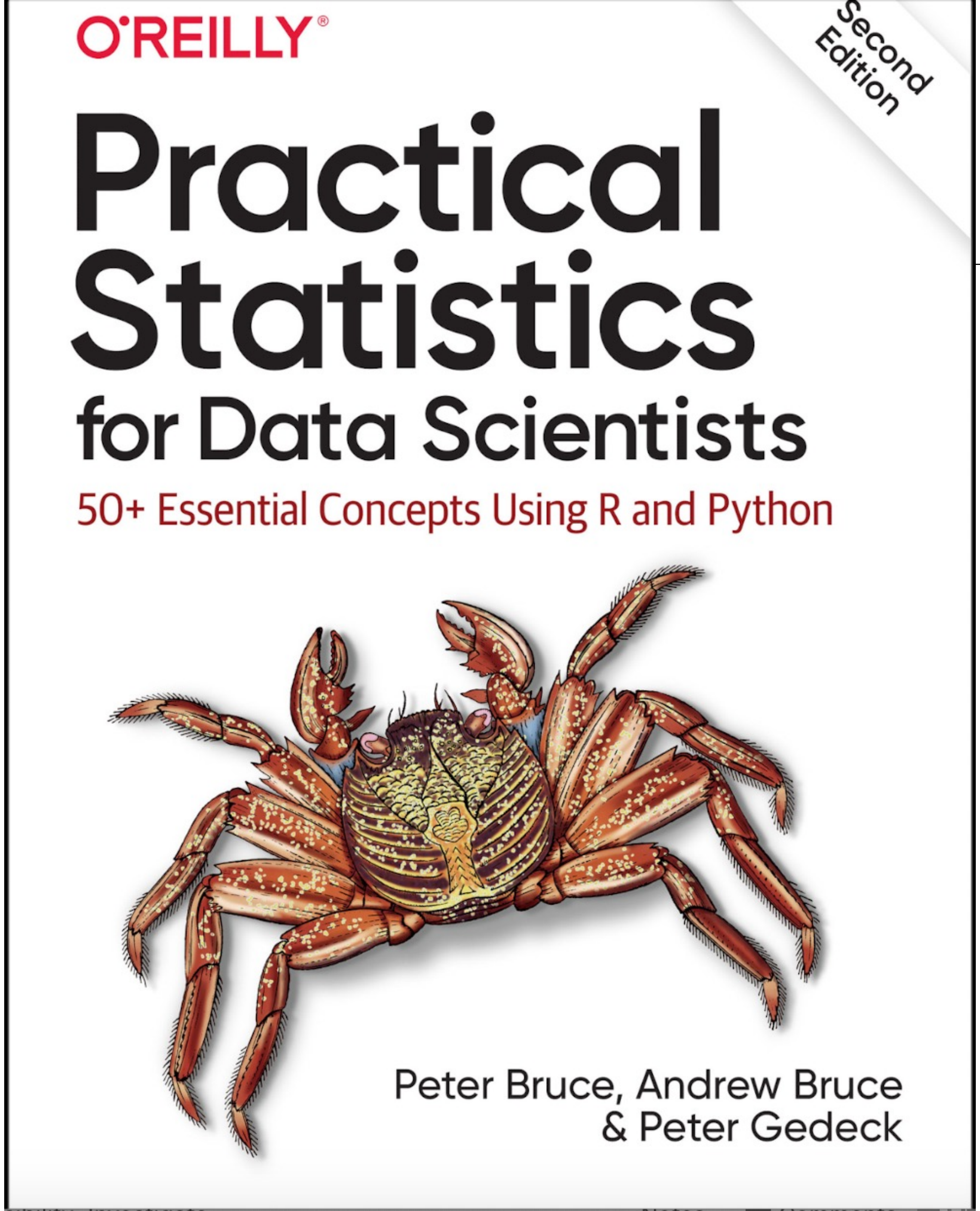
- Box plots
- Rule of thumb for outliers
- Correlation
- Correlation coefficient
- Correlation matrix
- Scatter plots

Little revisit

(we didn't get to this)

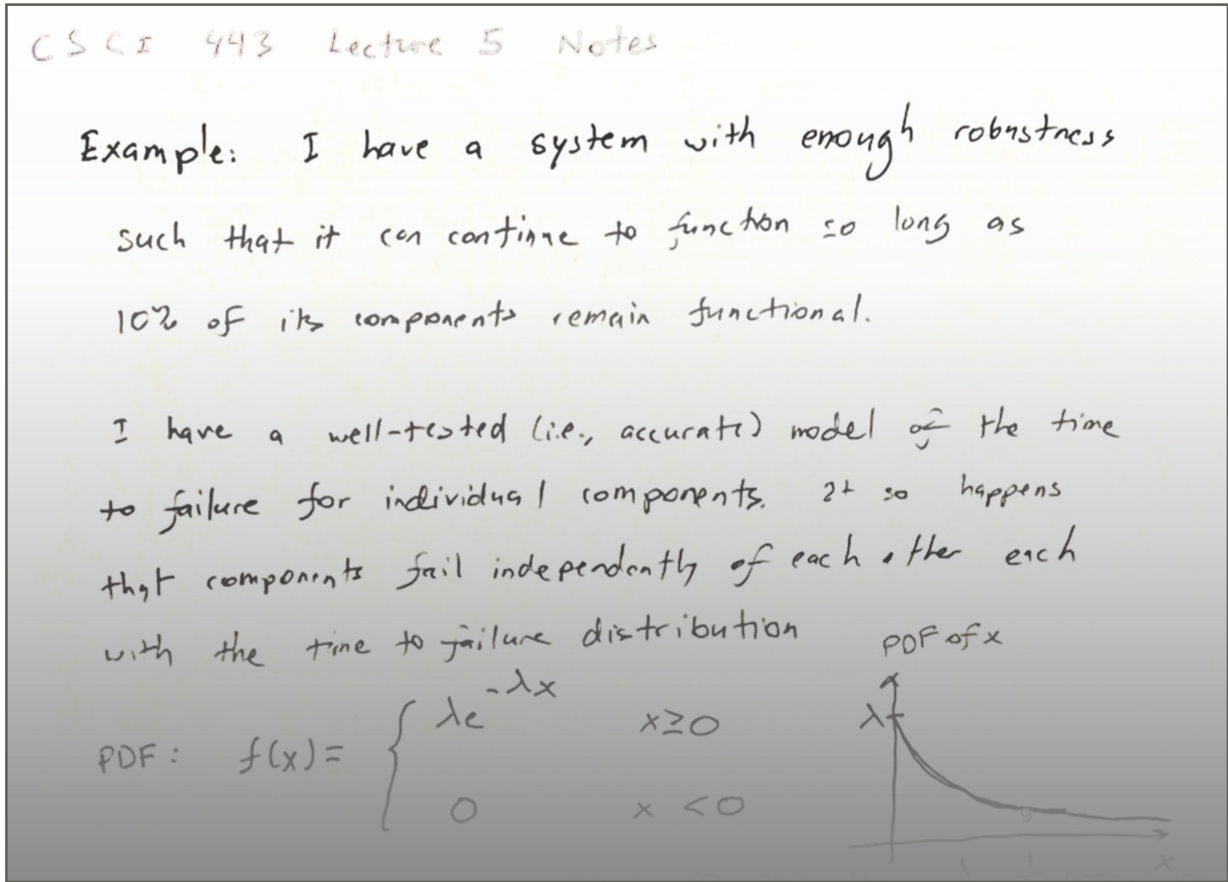
ASKED YOU TO READ ABOUT

- Weighted mean
- Weighted median
- Trimmed mean
- Modes
- Bar charts
- Pie charts



LECTURE 5 NOTES

I posted lecture notes covering what I wrote on the board in lecture 5.


CSCI 443 Lecture 5 Notes

Example: I have a system with enough robustness such that it can continue to function so long as 10% of its components remain functional.

I have a well-tested (i.e., accurate) model of the time to failure for individual components. It so happens that components fail independently of each other each with the time to failure distribution

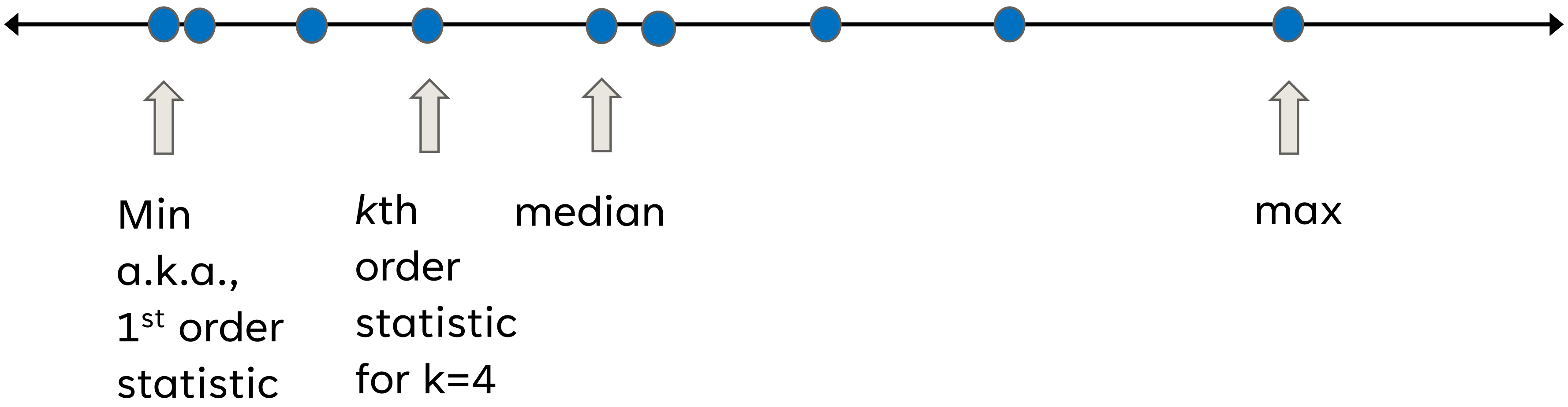
$$\text{PDF:} \quad f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

LAST LECTURE: ORDER STATISTICS

A random variable assign real values to outcomes.

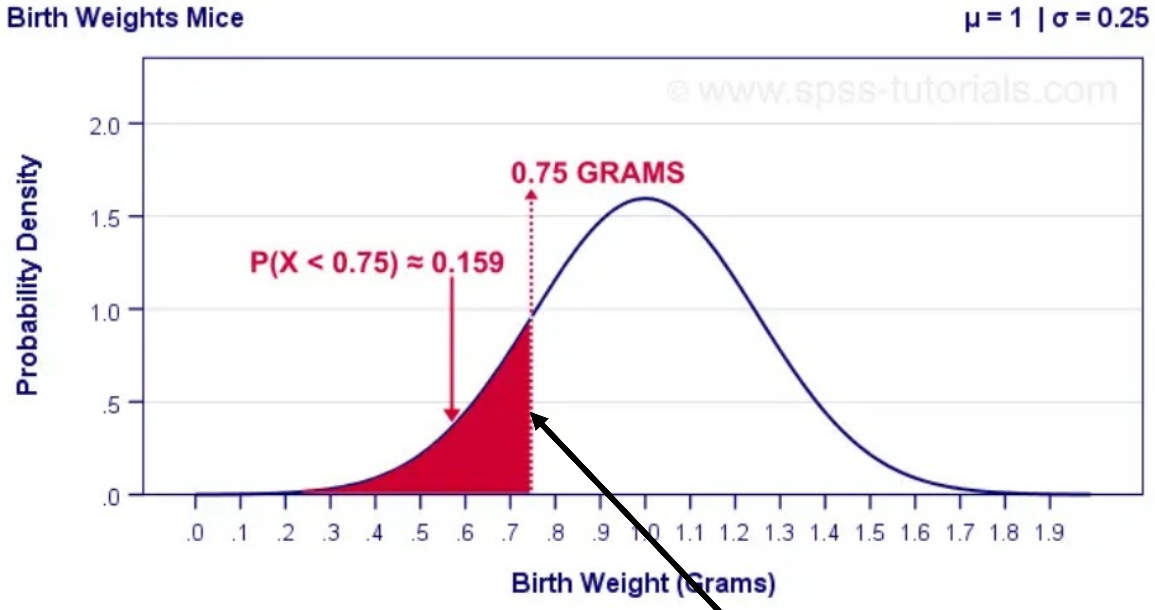
We order outcomes based on this real value.

“Order statistics” are based on this order.



LAST LECTURE: HOW TO COMPUTE PERCENTILE OF A DISTRIBUTION

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



15.9% fall below 0.75 grams.

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

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

LAST LECTURE: HOW TO ESTIMATE PERCENTILE FROM A SAMPLE USING LINEAR INTERPOLATION

A set of samples $\{x_1, x_2, \ldots, x_n\}$ are drawn from the distribution of random variable X. These are then sorted into ascending order $[x_{(1)}, x_{(2)}, \ldots, x_{(n)}]$.

The rank r is given by

$$r = \frac{p}{100}(n+1)$$

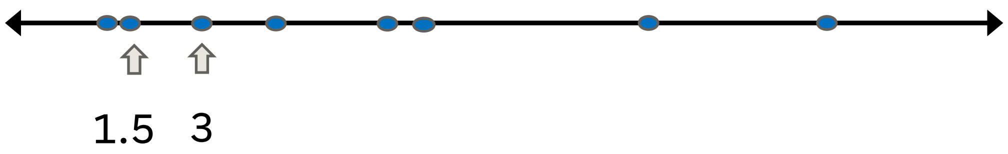
If r is an integer, the p^{th} percentile is $x_{(r)}$.
If r is not an integer, we linearly interpolate.

$$i = \lfloor r \rfloor$$

$$\alpha = r - i$$

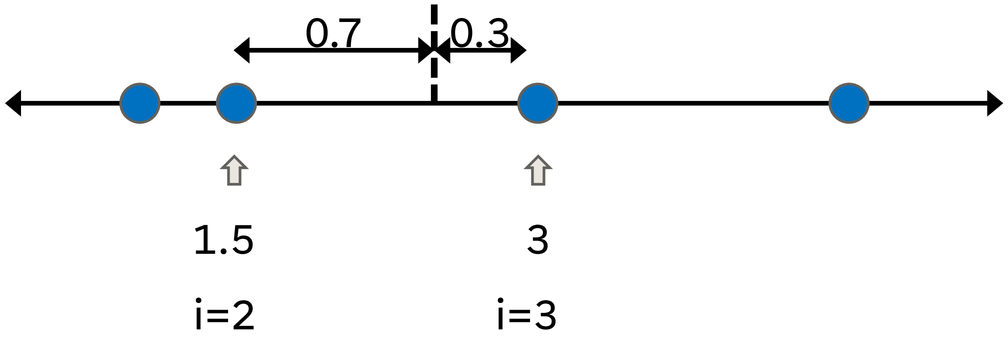
$$p^{th}\text{percentile} = x_{(i)} \cdot (1-\alpha) + x_{(i+1)} \cdot \alpha$$

$$[1, 1.5, 3, 4, 6, 6.6, 12, 14]$$



$$r = \frac{30}{100} \cdot (8+1) = 0.3 * 9 = 2.7$$

$$\alpha = 0.7$$



$$\begin{aligned} 30^{th}\text{percentile} &= x_{(2)} \cdot 0.3 + x_{(3)} \cdot 0.7 \\ &= 1.5 \cdot 0.3 + 3 \cdot 0.7 = \boxed{2.55} \end{aligned}$$

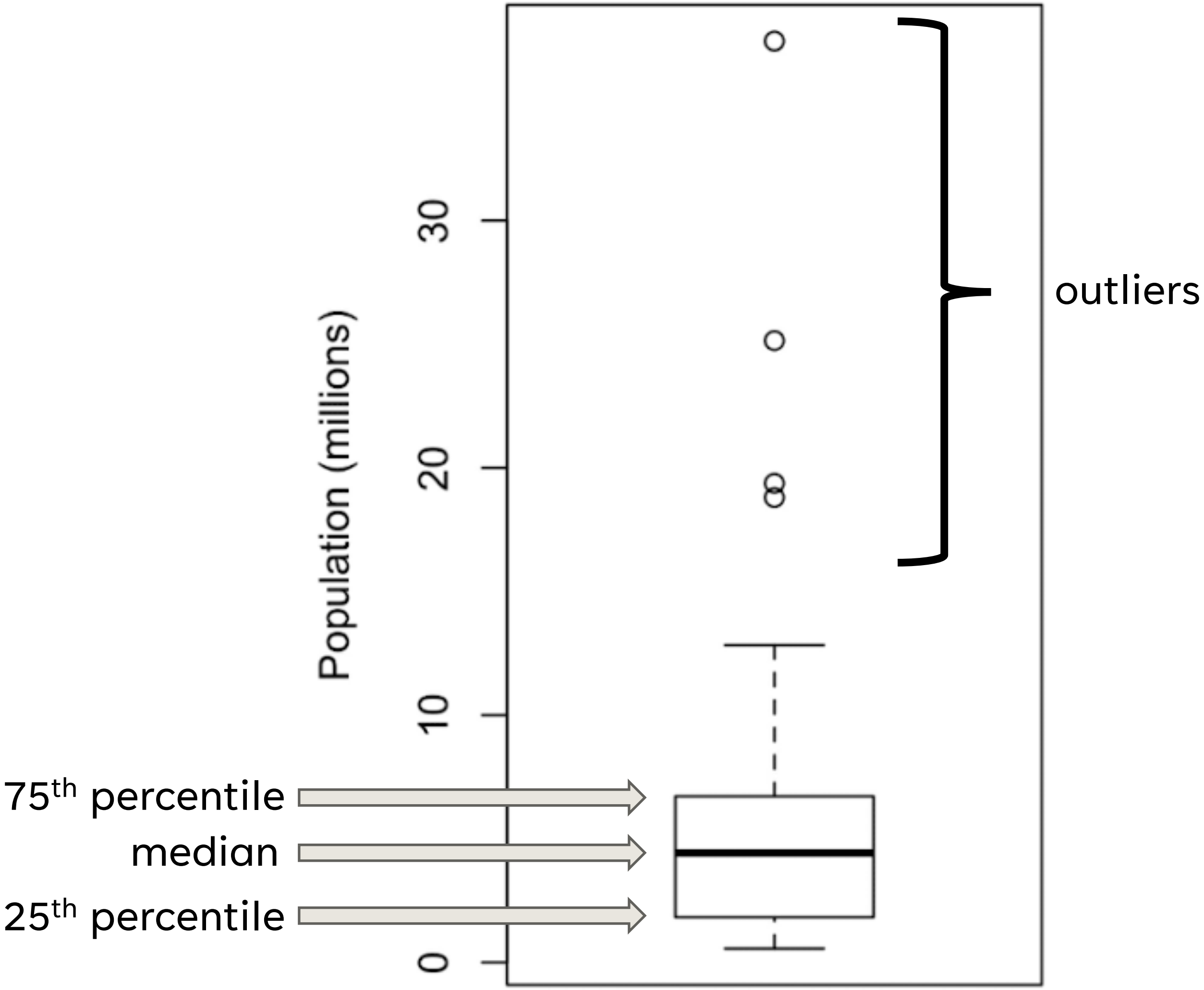
COMMENT ON BOX PLOTS

Box plot uses a box to show the 25th and 75th percentiles.

A line inside the box denotes the median.

Three common methods for whiskers:

- Min to max (highly affected by outliers)
- Smallest and largest falling within 1.5 IQR.
- 10th to 90th percentile



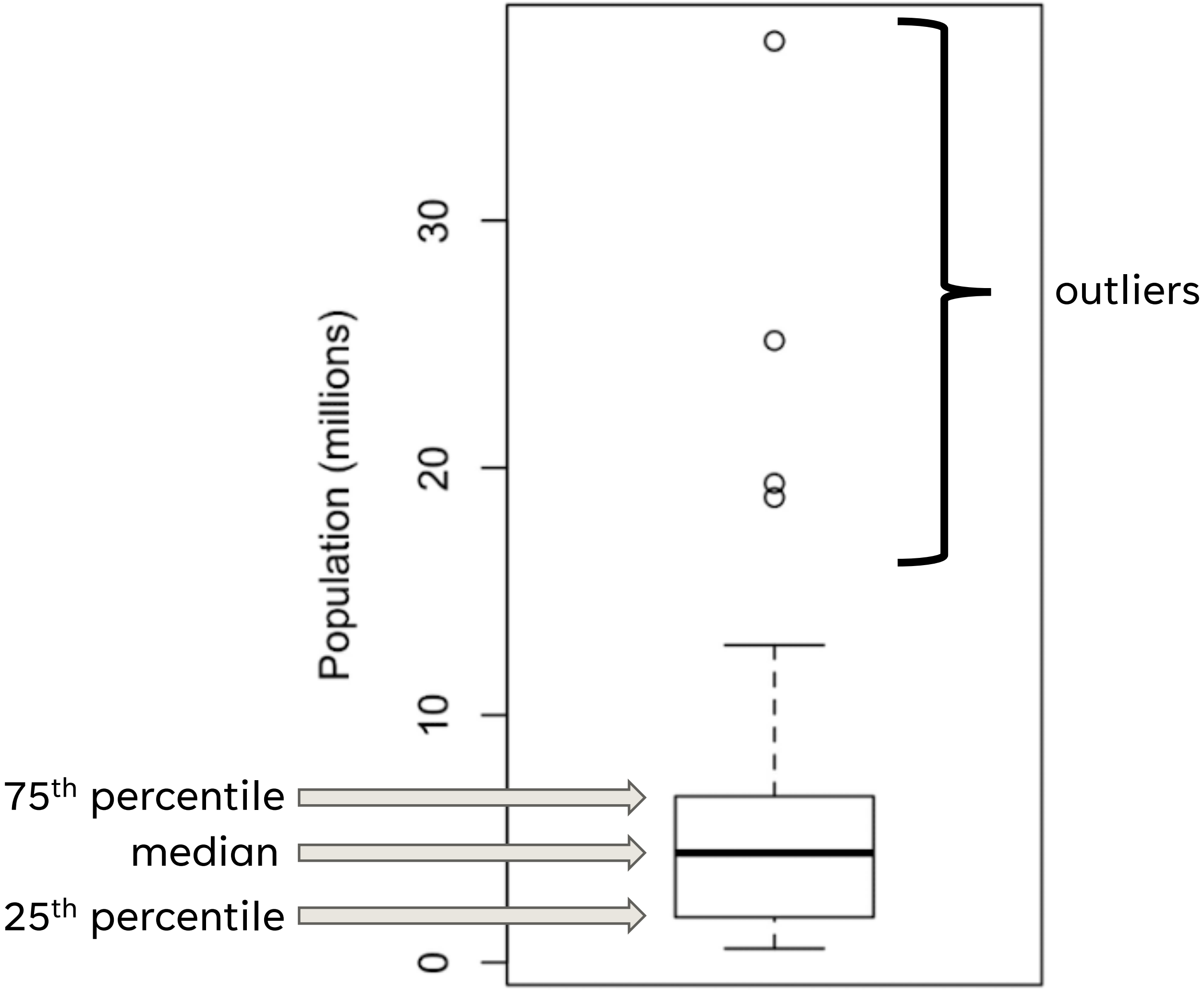
RULE OF THUMB FOR OUTLIERS

Common method is to declare an outlier when

> 3^{rd} quartile + 1.5 IQR

or

< 1^{st} quartile - 1.5 IQR



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

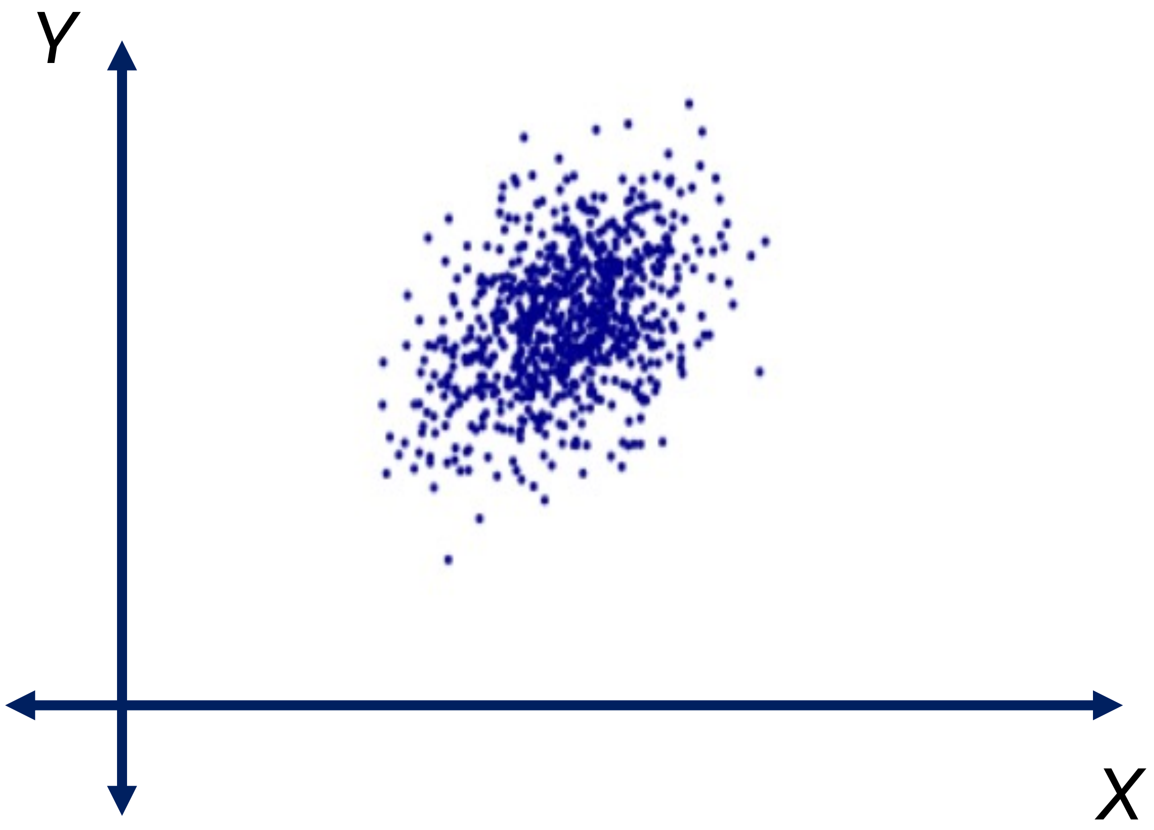
A plot in which the x-axis is the value of one variable, and the y-axis the value of another.

SCATTER PLOT

There are two random variables *X* and *Y*.

Each sample has an *x* and *y* value.

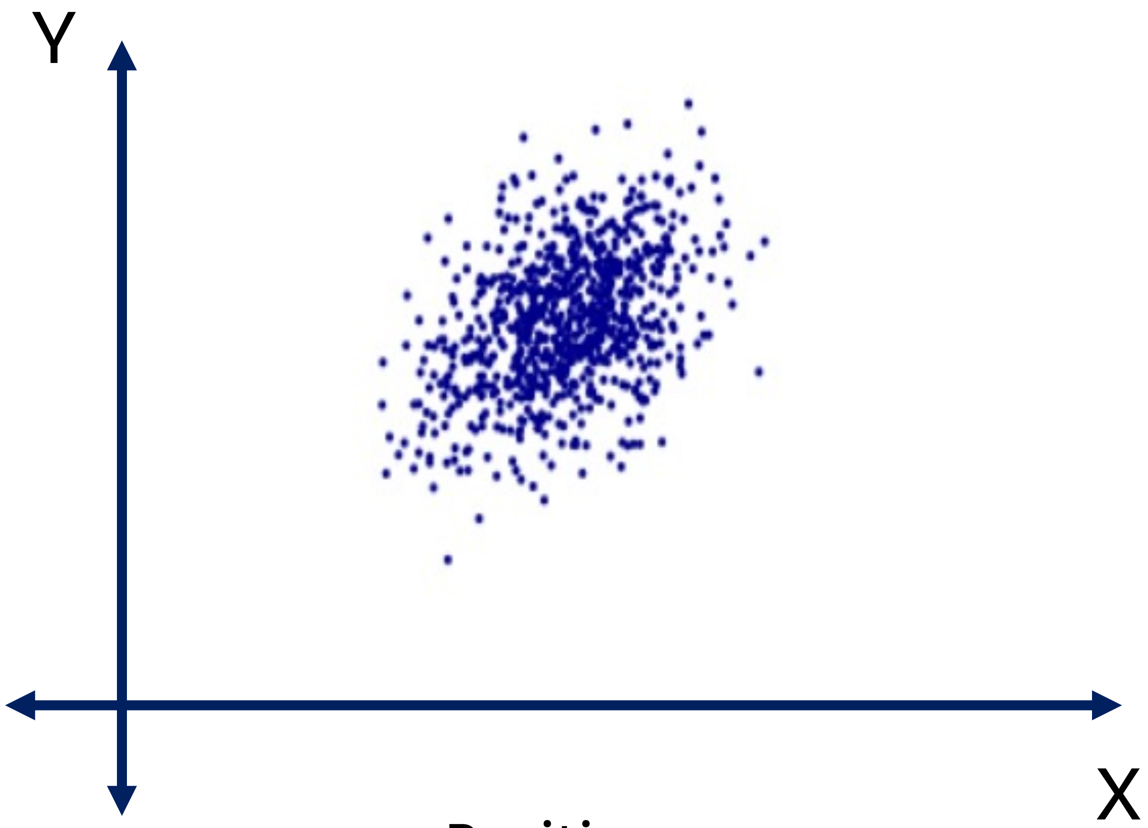
We plot each sample with its x and y values as a point at the coordinate (x,y) on a cartesian plane.



CORRELATION

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

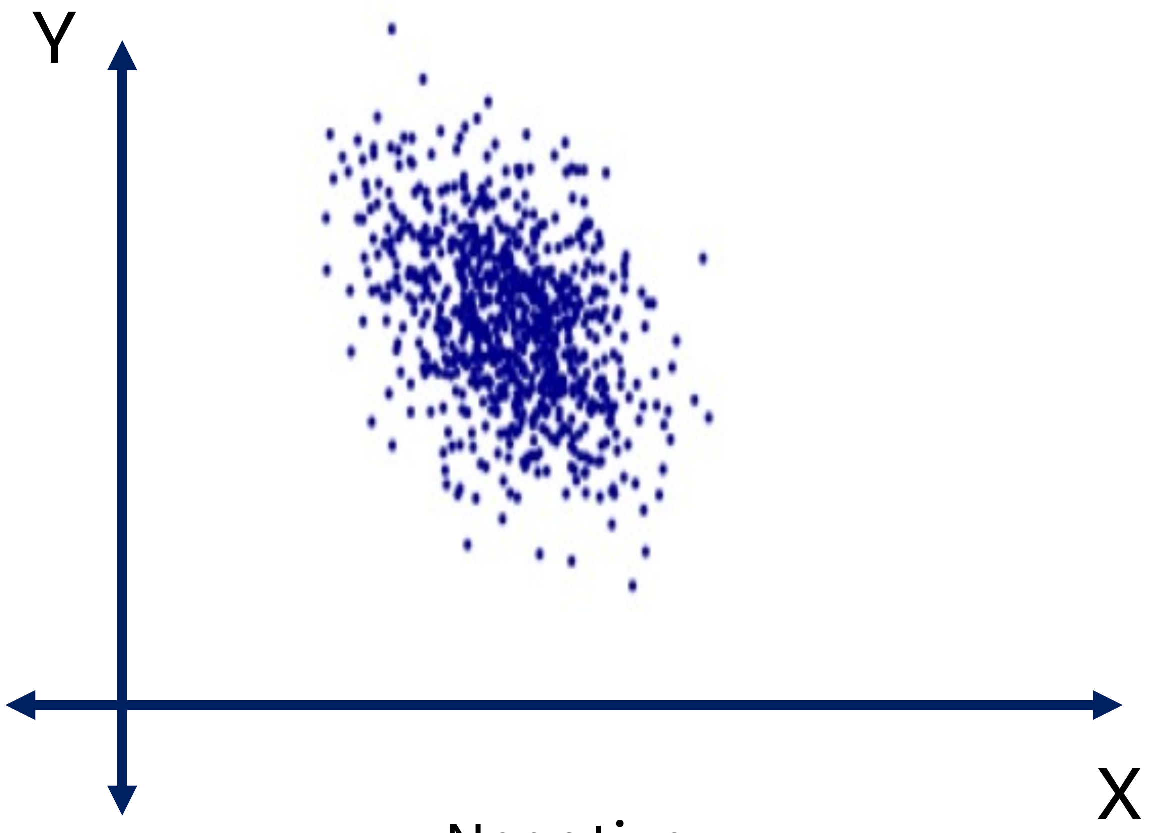
When one increases the other does, and vice versa.



Positive correlation

CORRELATION

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

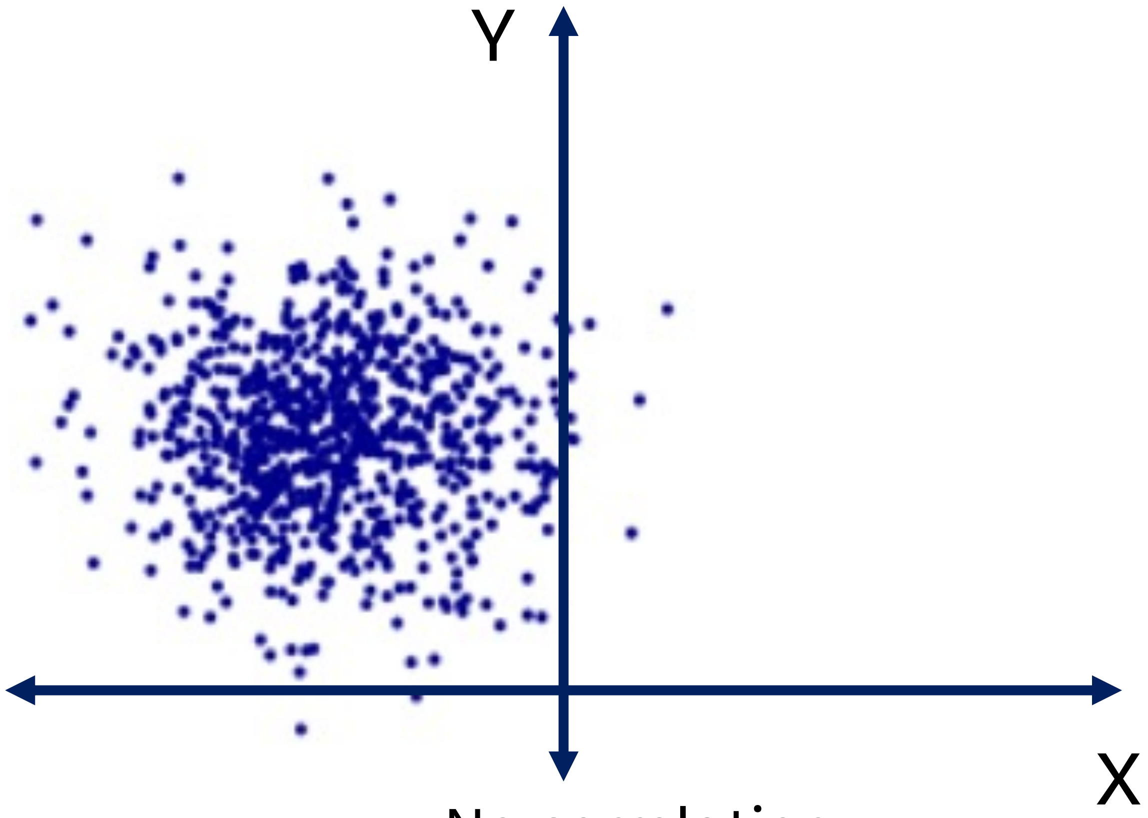


Negative correlation

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.

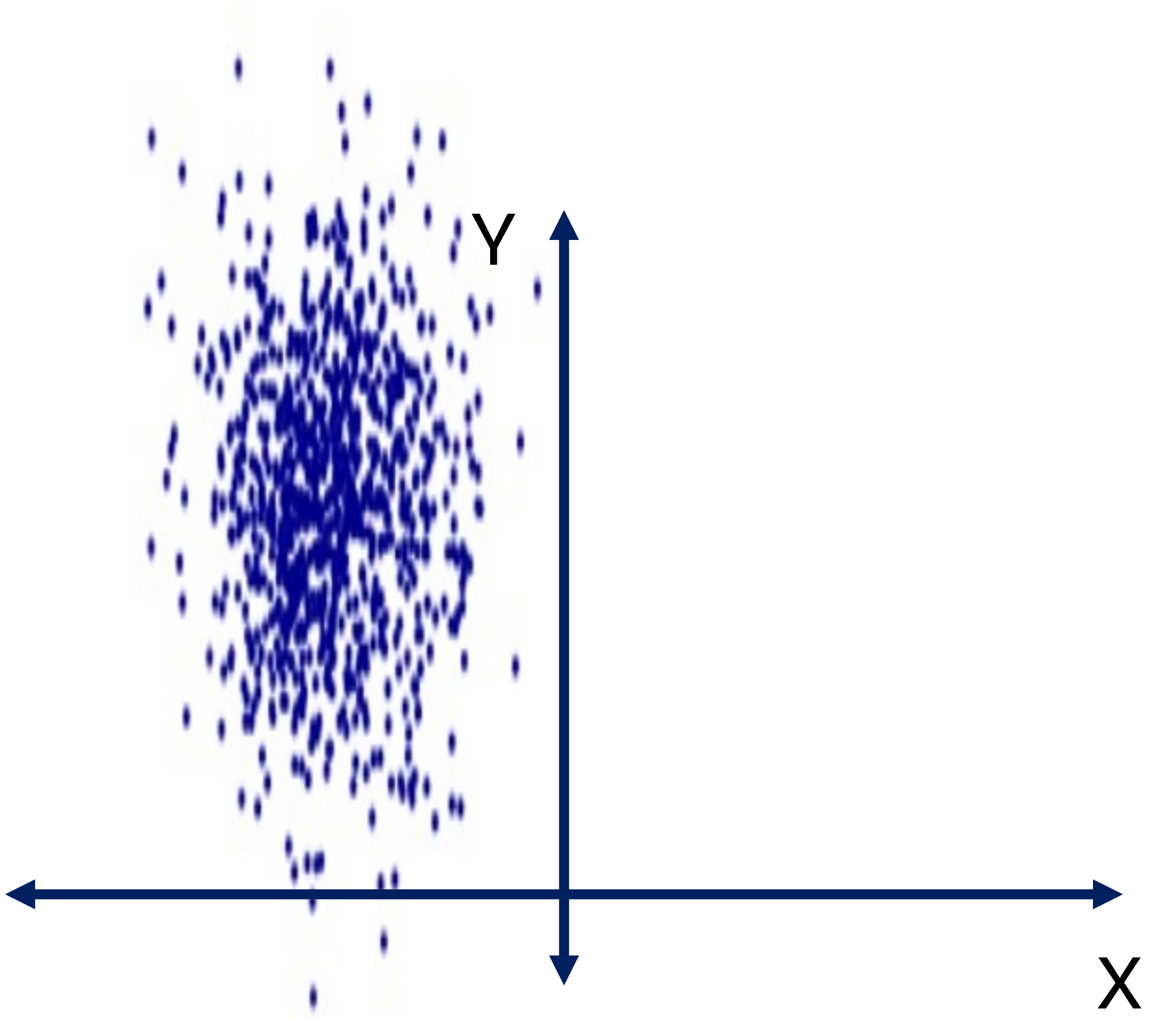


No correlation

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.

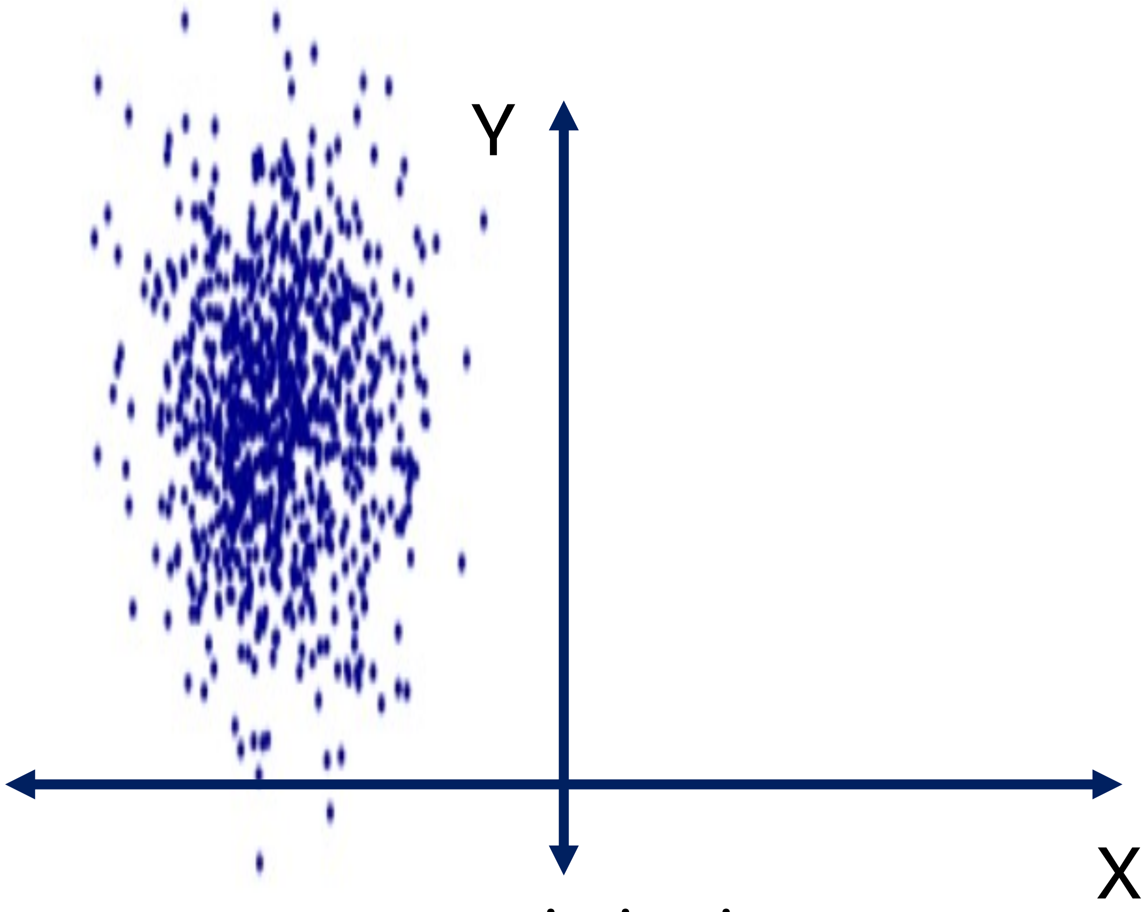


Correlated??????

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.

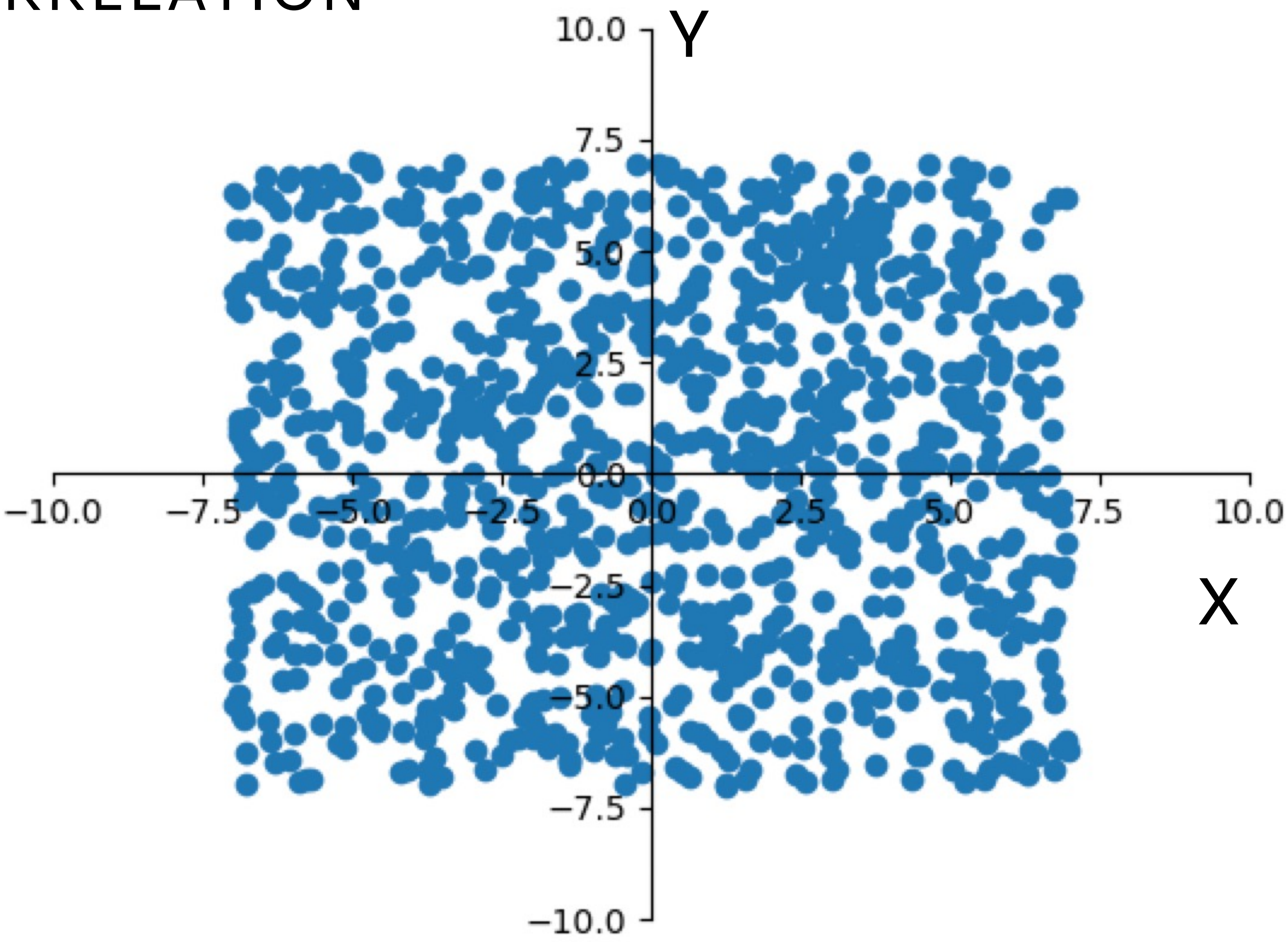


More variation in Y than X but still no correlation

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.

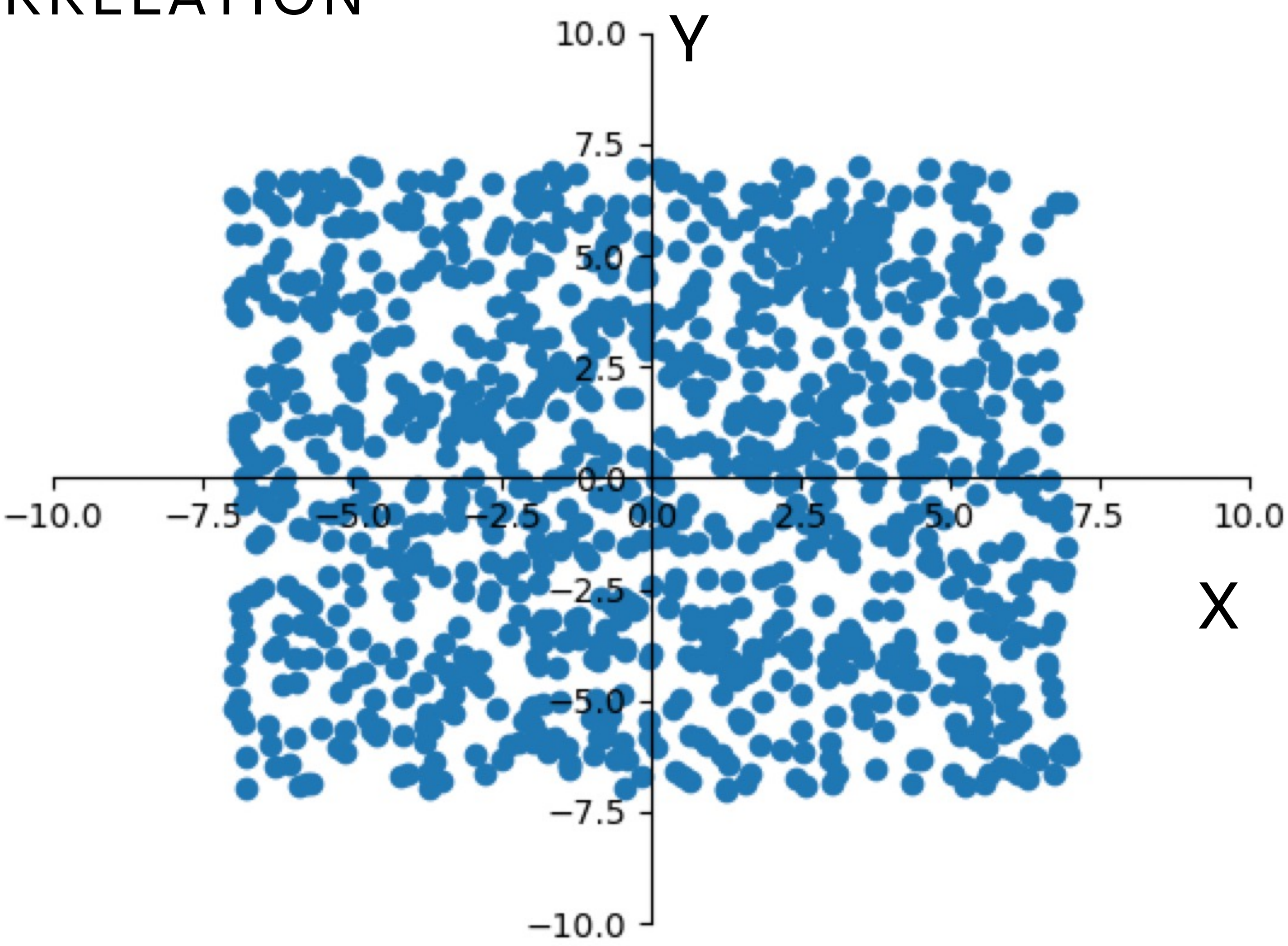


Correlated?????

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.

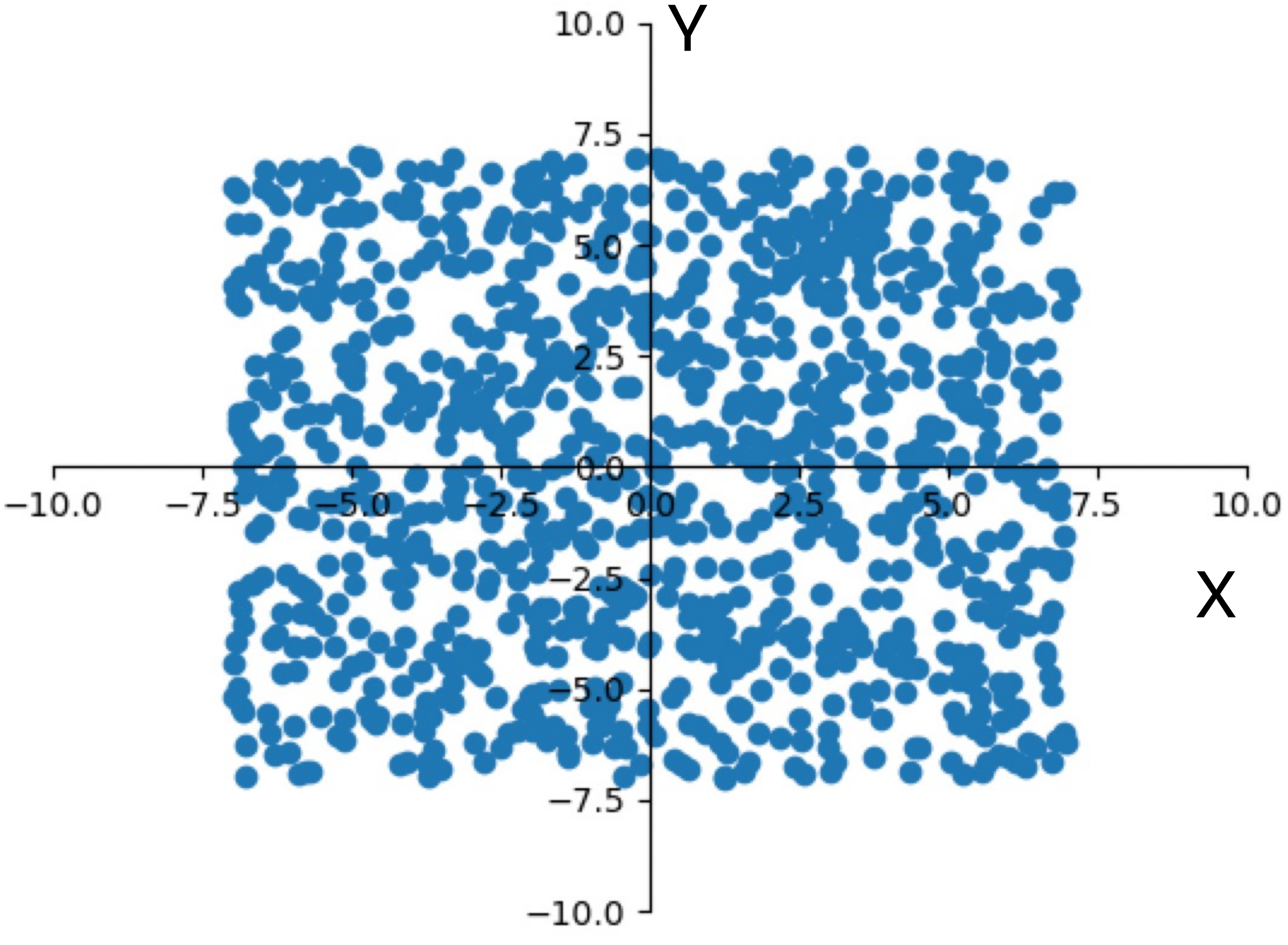


Uniformly distributed in X and Y, but they don’t move together so NO CORRELATION!

CORRELATION

Correlation is qualitative.

We want some way to quantify correlation.



No correlation

COVARIANCE

Let X and Y be two random variables, covariance is

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

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

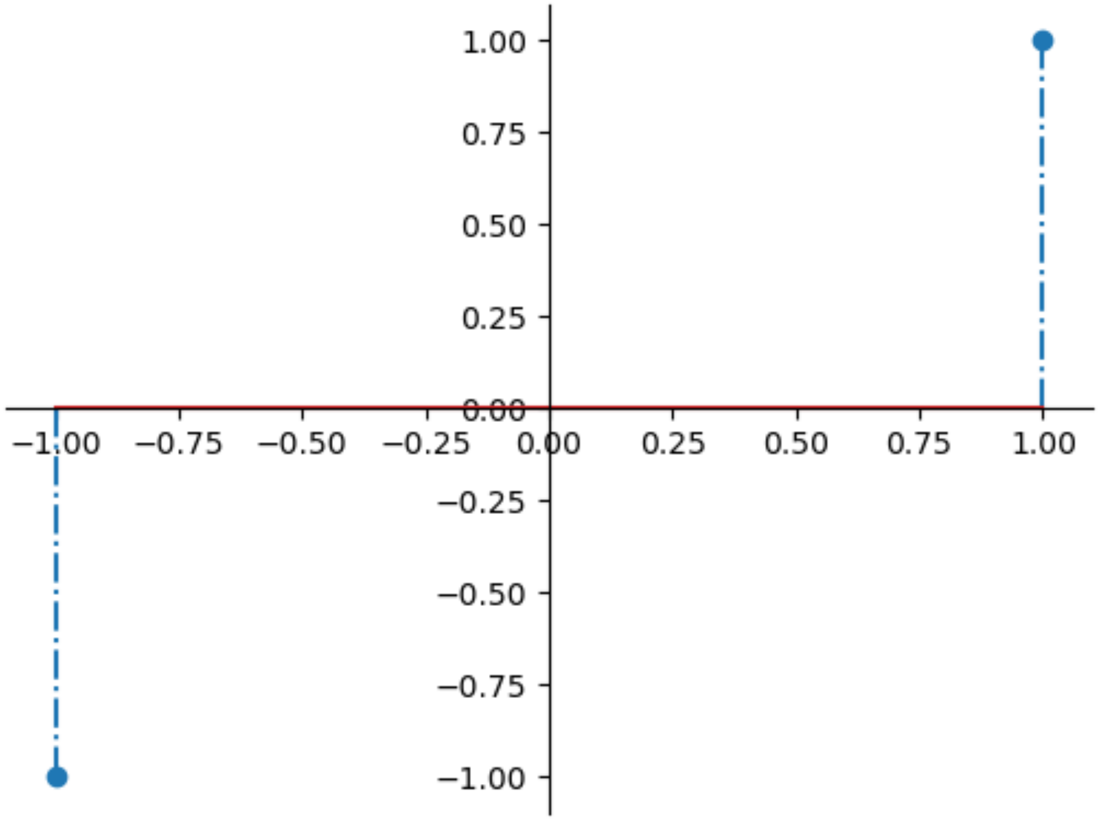
$$cov(X,Y) = E[XY]$$

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

let S denote the samples drawn from X and Y.

$$S = [(x_1, y_1), (x_2, y_2)] = [(-1,-1), (1,1)]$$

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



REMINDER! Specifically chose mean = 0 to simplify equation.

COVARIANCE

Let X and Y be two random variables, covariance is

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

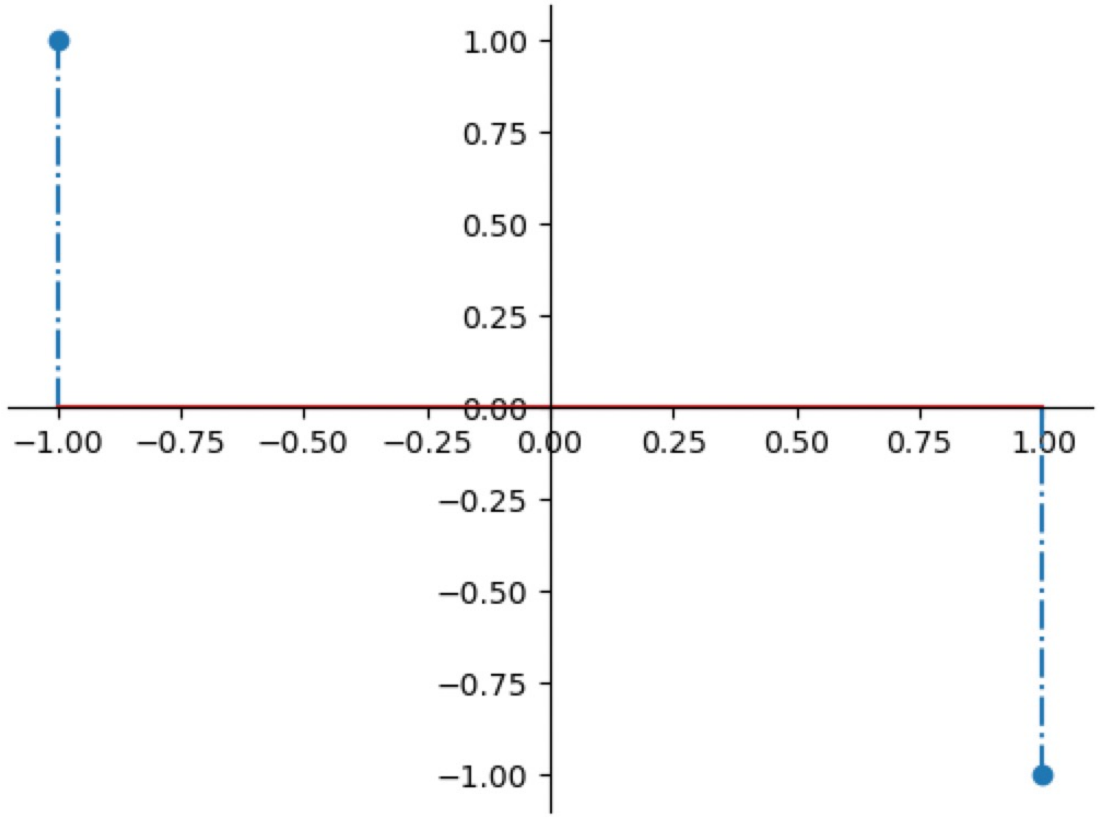
If $\mu_x = 0$ and $\mu_y = 0$ then this simplifies to

$$cov(X,Y) = E[XY]$$

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

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

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



REMINDER! Specifically chose mean = 0 to simplify equation.

COVARIANCE

Let X and Y be two random variables, covariance is

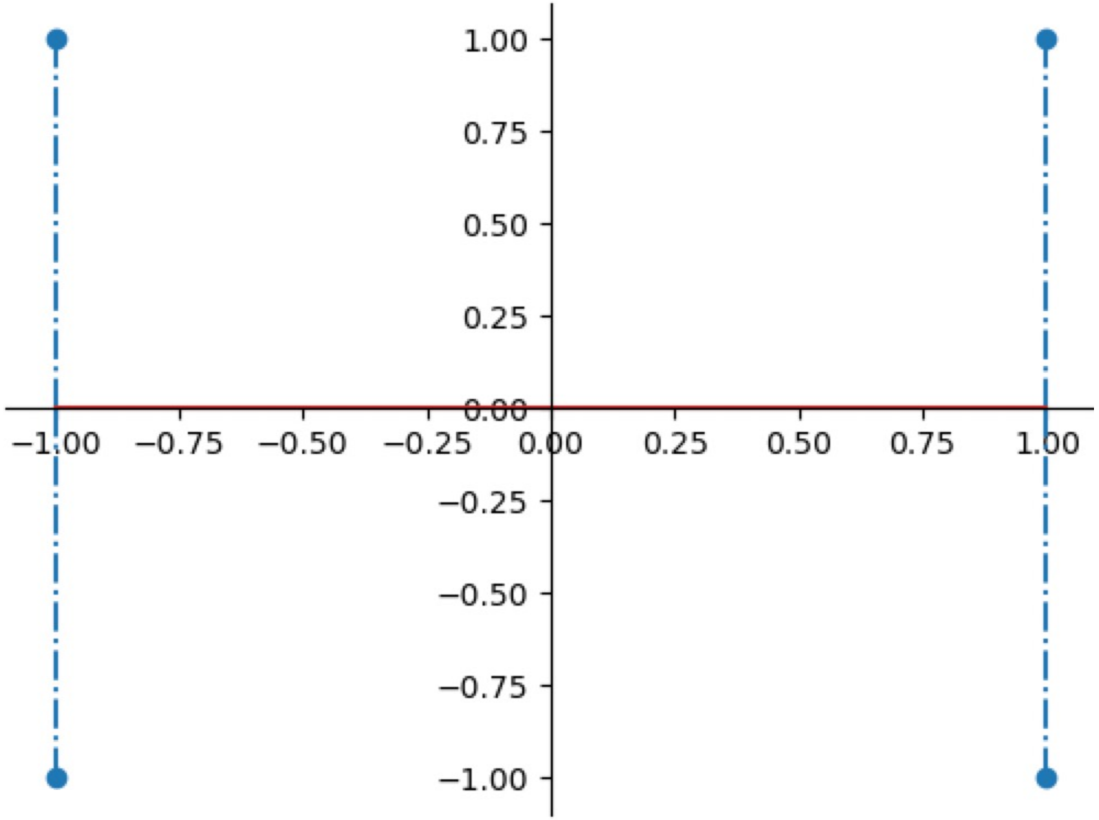
$$cov(X,Y) = E[(X - \mu_x)(Y - \mu_Y)]$$

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

$$cov(X,Y) = E[XY]$$

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

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



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

REMINDER! Specifically chose mean = 0 to simplify equation.

COVARIANCE

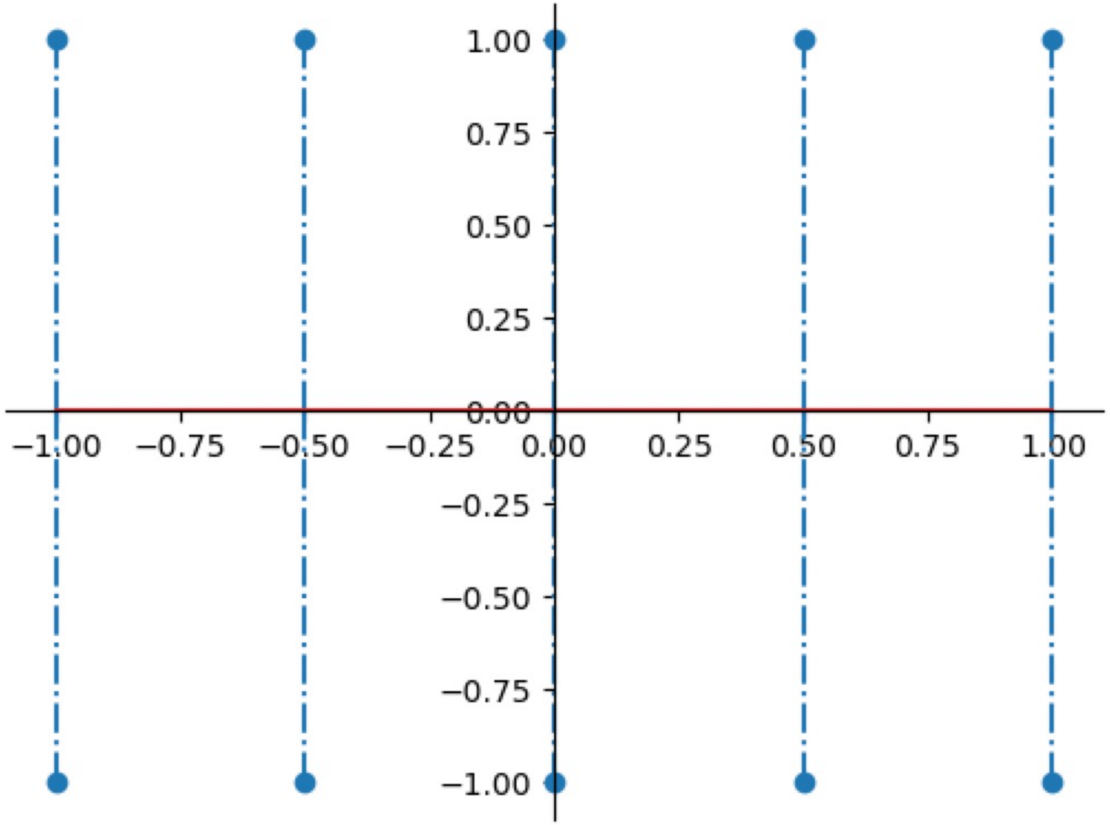
Let X and Y be two random variables, covariance is

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

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

$$cov(X,Y) = E[XY]$$

$$E[XY] = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i$$



$$E[XY] = \frac{1}{10}[(-1 \cdot -1)+(-1 \cdot 1)+(-\frac{1}{2} \cdot \frac{1}{2})+(-\frac{1}{2} \cdot -\frac{1}{2})+\cdots+(1 \cdot 1)+(1 \cdot -1)] = 0$$

COVARIANCE

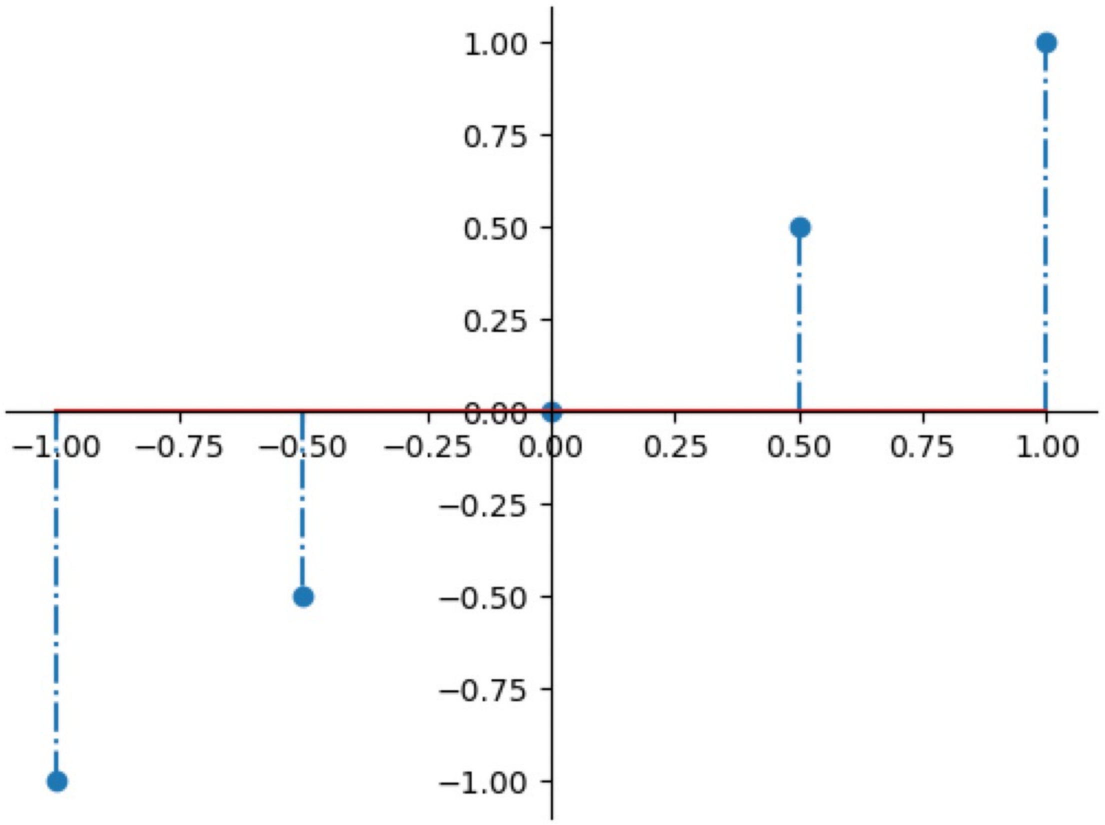
Let X and Y be two random variables, covariance is

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

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

$$cov(X,Y) = E[XY]$$

$$E[XY] = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i$$



$$E[XY] = \frac{1}{5}[(-1 \cdot -1) + (-\frac{1}{2} \cdot -\frac{1}{2}) + (0 \cdot 0) + (\frac{1}{2} \cdot \frac{1}{2}) + (1 \cdot 1)] = ???$$

COVARIANCE

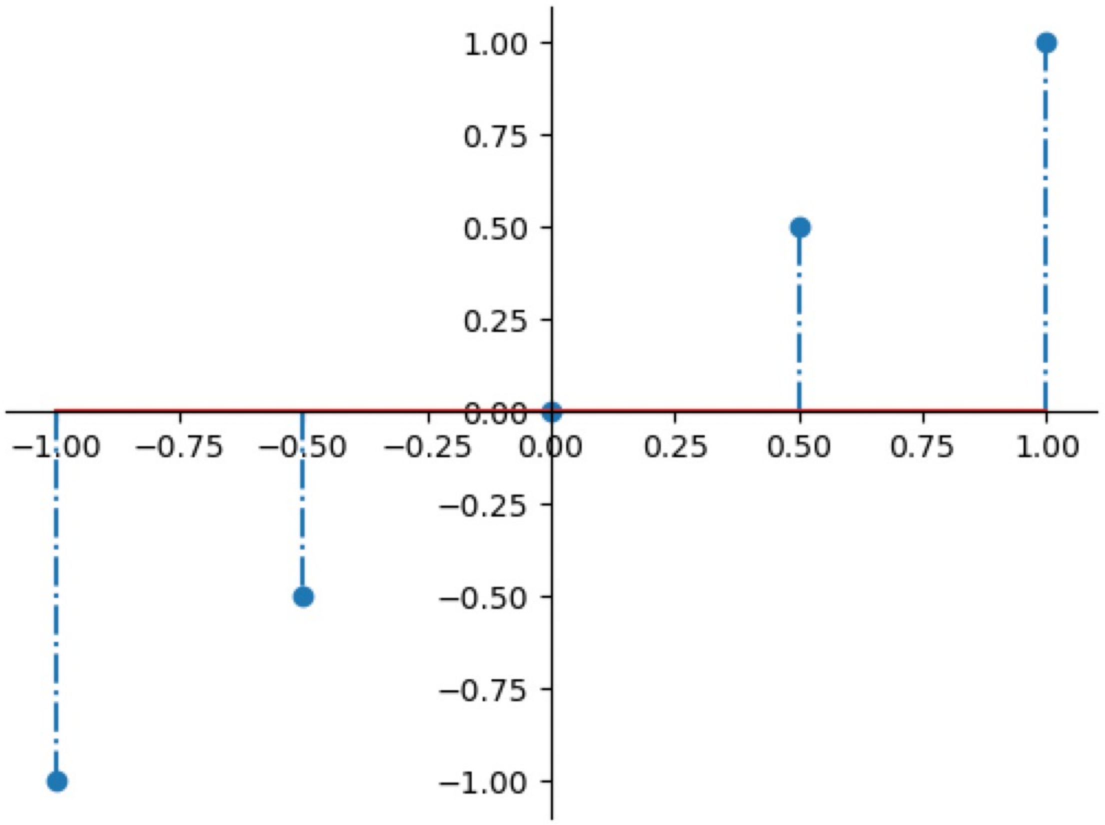
Let X and Y be two random variables, covariance is

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

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

$$cov(X,Y) = E[XY]$$

$$E[XY] = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i$$



$$E[XY] = \frac{1}{5}[(-1 \cdot -1) + (-\frac{1}{2} \cdot -\frac{1}{2}) + (0 \cdot 0) + (\frac{1}{2} \cdot \frac{1}{2}) + (1 \cdot 1)] = \frac{2.5}{5} = \frac{1}{2}$$

COVARIANCE

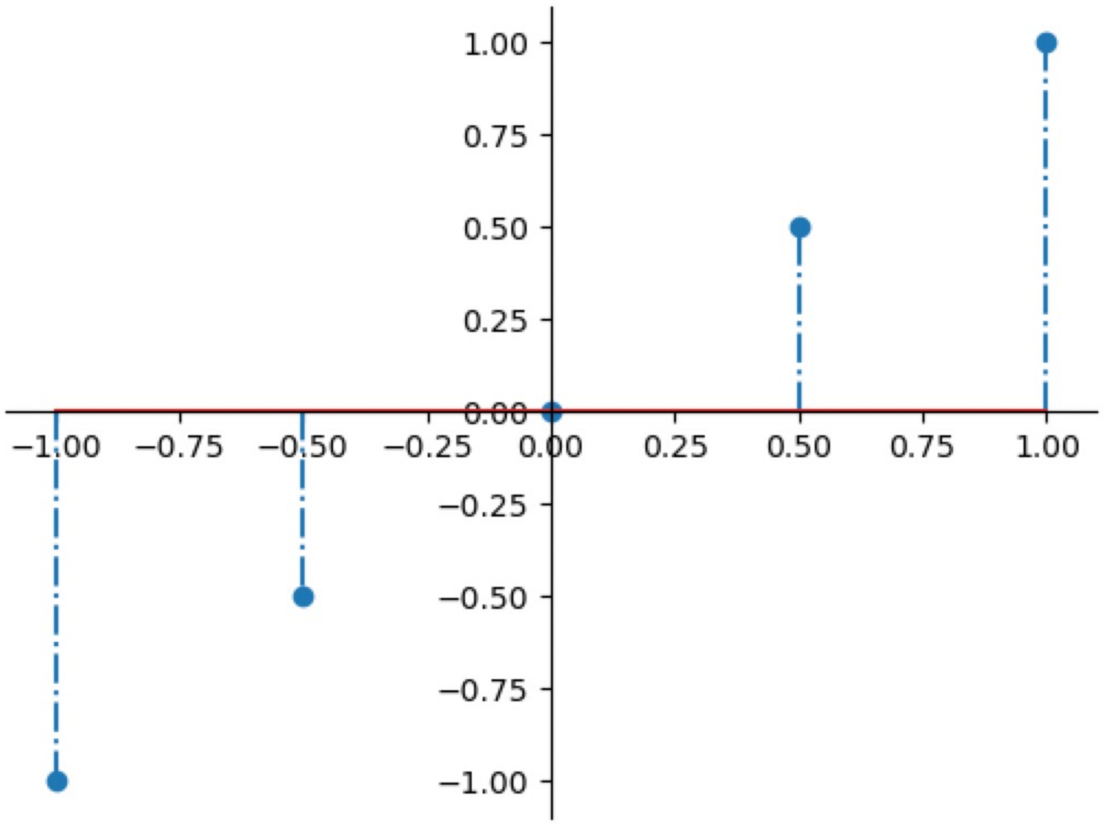
Let X and Y be two random variables, covariance is

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

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

$$cov(X,Y) = E[XY]$$

$$E[XY] = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i$$



$$E[XY] = \frac{1}{5}[(-1\cdot-1)+(-\frac{1}{2}\cdot-\frac{1}{2})+(0\cdot0)+(\frac{1}{2}\cdot\frac{1}{2})+(1\cdot1)] = \frac{2.5}{5} = \frac{1}{2}$$

Why? Why not 1?

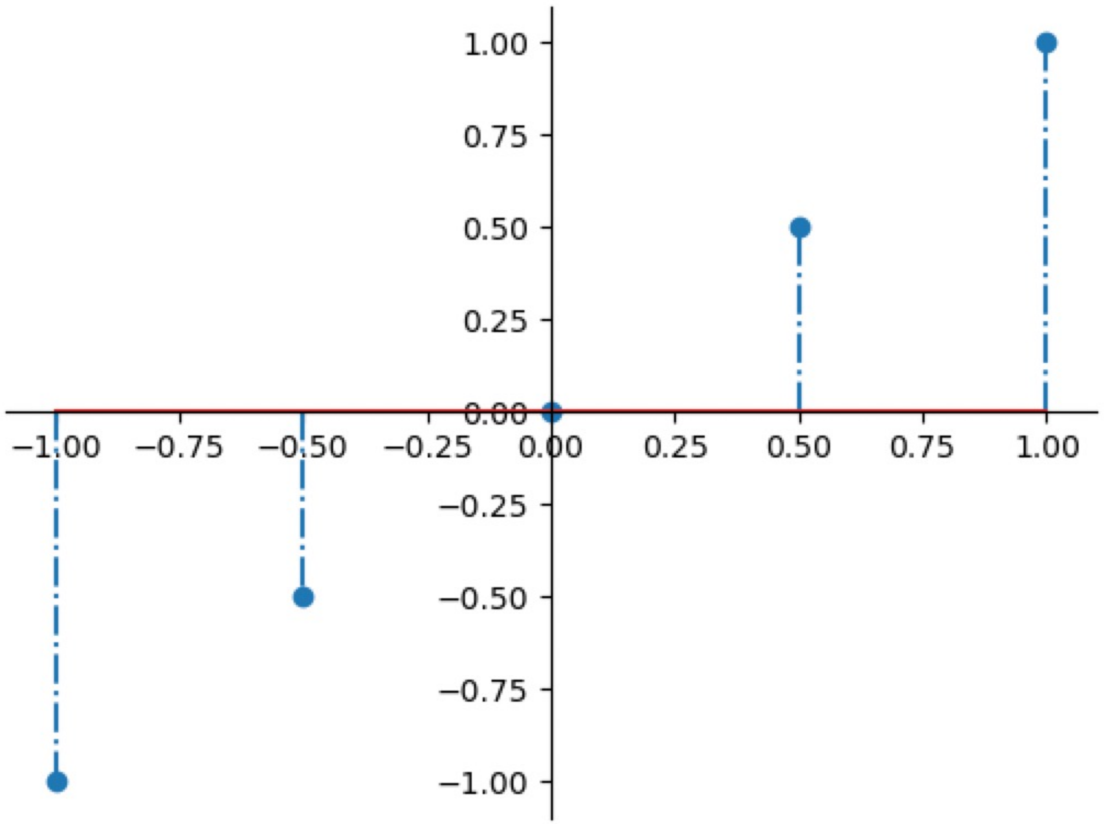
COVARIANCE

$$E[XY] = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i$$

When X and Y are equal, they square.

The impact of a sample grows with the square of the distance from the mean (here mean is 0).

Numbers farther out have greater impact on E[XY].



$$E[XY] = \frac{1}{5}[(-1\cdot-1)+(-\frac{1}{2}\cdot-\frac{1}{2})+(0\cdot0)+(\frac{1}{2}\cdot\frac{1}{2})+(1\cdot1)] = \frac{2.5}{5} = \frac{1}{2}$$

COVARIANCE

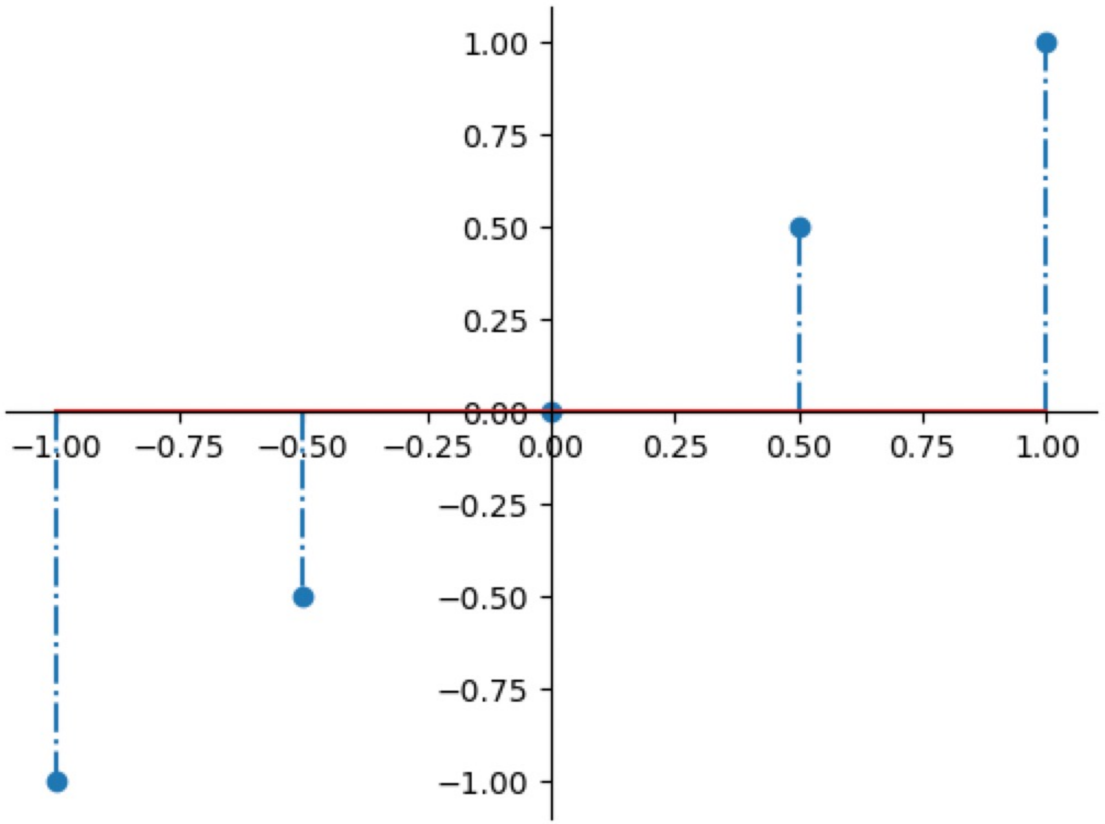
$$E[XY] = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i$$

Variance ALSO grows with the square.

When the mean is zero

$$Var[X] = E[X^2]$$

$$\frac{E[XY]}{Var[X]} = ???$$



$$E[XY] = \frac{\frac{1}{5}[(-1\cdot -1) + (-\frac{1}{2}\cdot -\frac{1}{2}) + (0\cdot 0) + (\frac{1}{2}\cdot\frac{1}{2}) + (1\cdot 1)]}{\frac{1}{5}[(-1\cdot -1) + (-\frac{1}{2}\cdot -\frac{1}{2}) + (0\cdot 0) + (\frac{1}{2}\cdot\frac{1}{2}) + (1\cdot 1)]} = \frac{\frac{1}{2}}{\frac{1}{2}} = 1$$

COVARIANCE

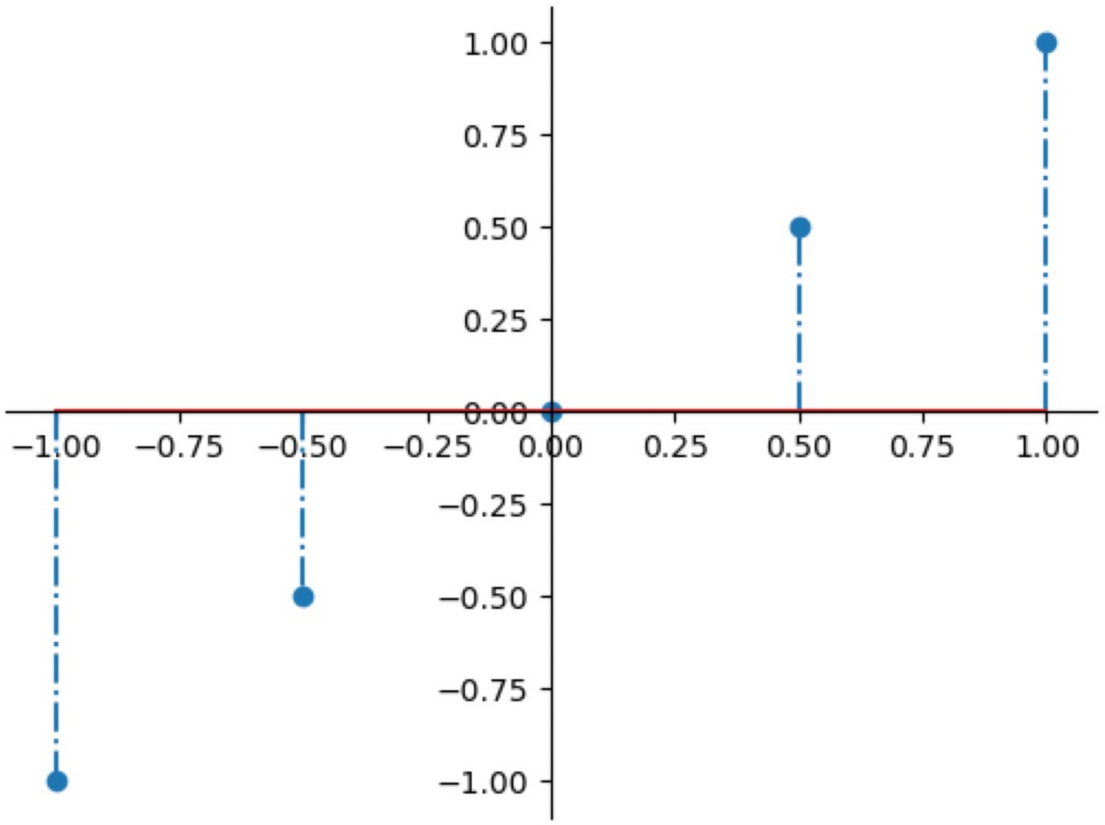
$$E[XY] = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i$$

Variance ALSO grows with the square.

When the mean is zero

$$Var[X] = E[X^2]$$

$$\frac{E[XY]}{Var[X]} = ???$$



But why only Var[X]? Shouldn’t the variation in Y also matter?

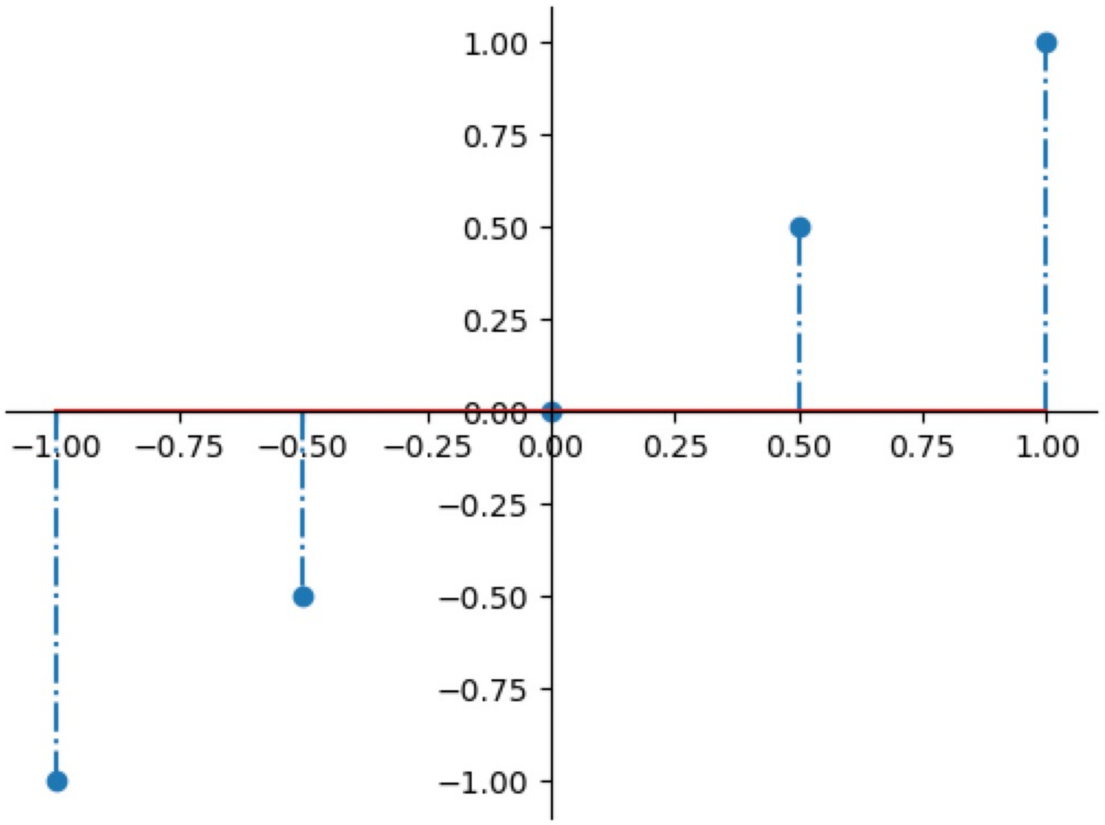
COVARIANCE

$$E[XY] = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i$$

Variance ALSO grows with the square.

When the mean is zero

$$Var[X] = E[X^2]$$



$$\frac{E[XY]}{\sqrt{Var[X]}\sqrt{Var[Y]}} = \frac{E[XY]}{\sigma_X \sigma_Y} = \frac{\frac{1}{2}}{\sqrt{\frac{1}{2}}\sqrt{\frac{1}{2}}} = \frac{\frac{1}{2}}{\frac{1}{2}} = 1$$

COVARIANCE

We can adjust the equations to take into account non-zero mean.

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

$$E[XY] = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i \quad \Rightarrow \quad \frac{1}{n}\sum_{i=0}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$$

$$\frac{E[XY]}{\sigma_X \sigma_Y} \quad \Rightarrow \quad \frac{\frac{1}{n}\sum_{i=0}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sigma_X \sigma_Y}$$

COVARIANCE

We can adjust the equations to take into account non-zero mean.

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

$$E[XY] = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i \quad \Rightarrow \quad \frac{1}{n}\sum_{i=0}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$$

$$\left.\frac{E[XY]}{\sigma_X \sigma_Y} \quad \Rightarrow \quad \frac{\frac{1}{n}\sum_{i=0}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sigma_X \sigma_Y}\right\} \text{Pearson Correlation}$$

PEARSON CORRELATION

Correlation Coefficient

a.k.a.,

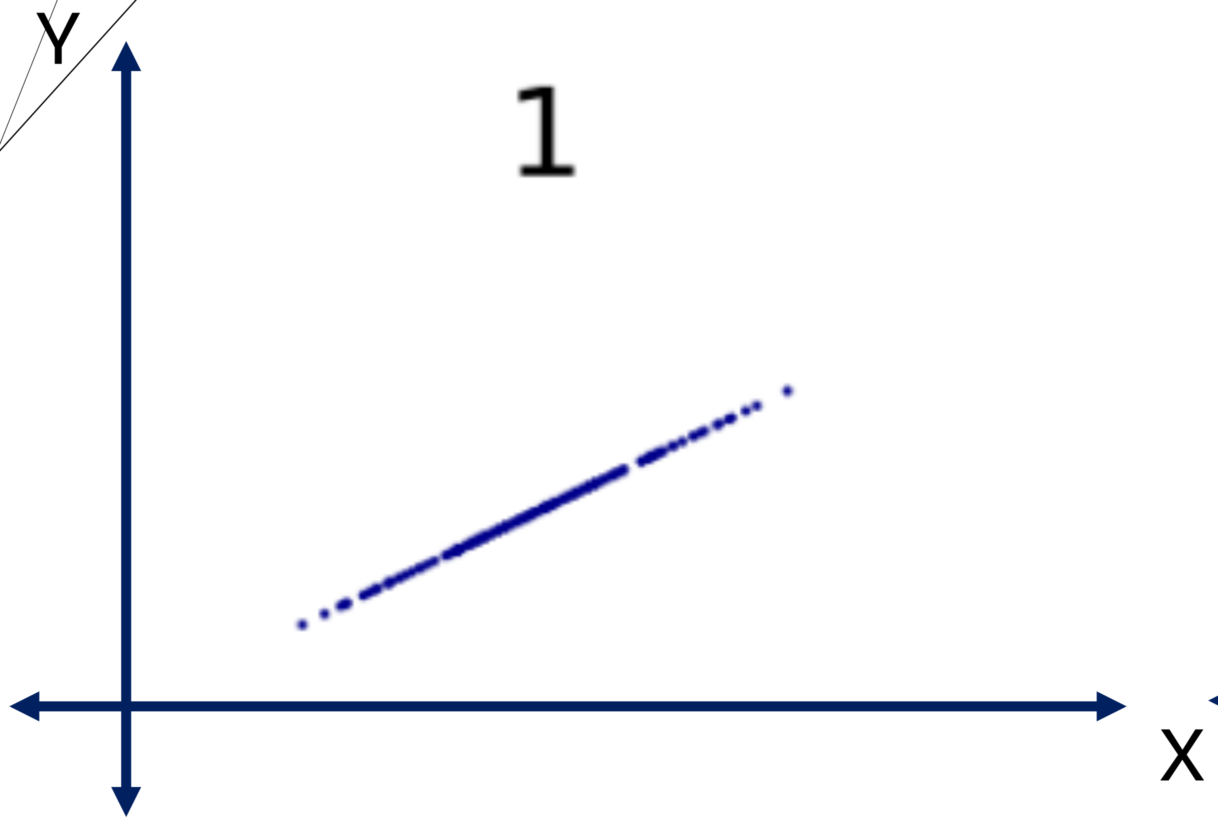
Linear Correlation Coefficient.

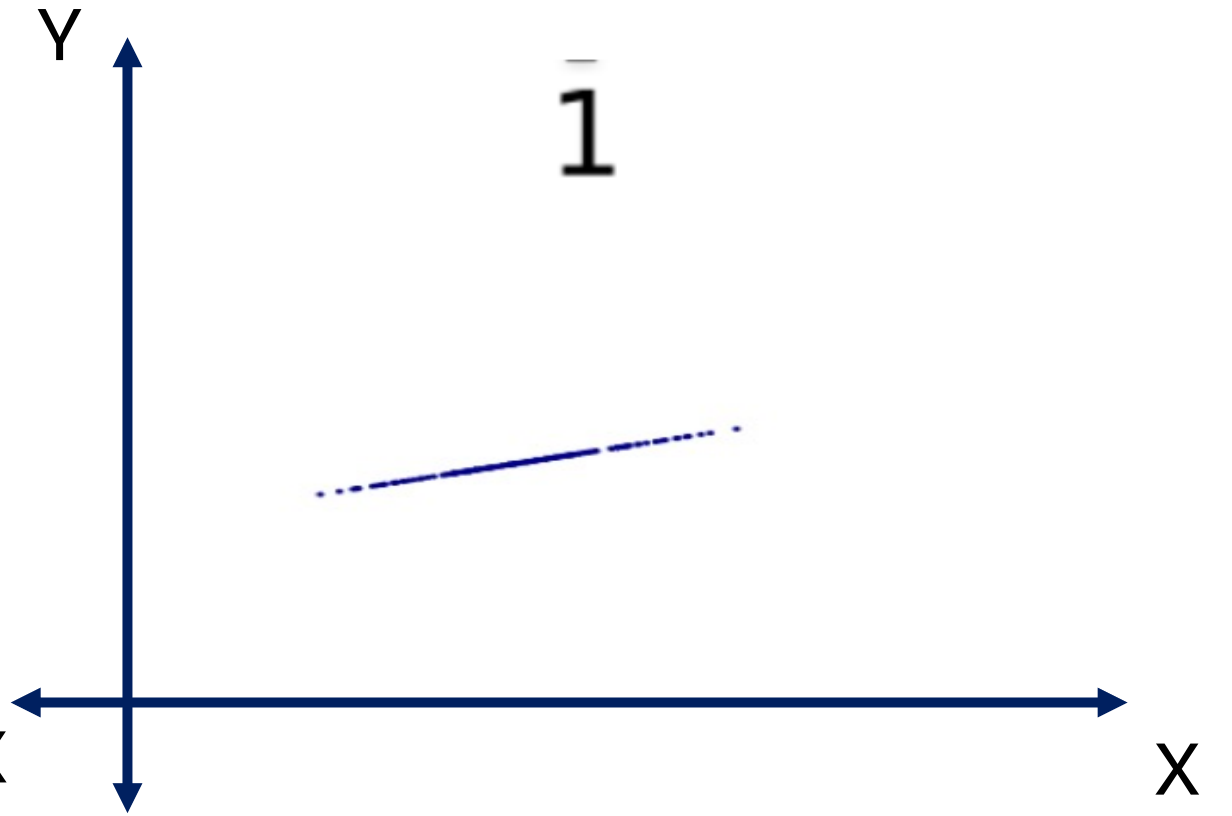
a.k.a.,

the *Pearson Correlation Coefficient*.

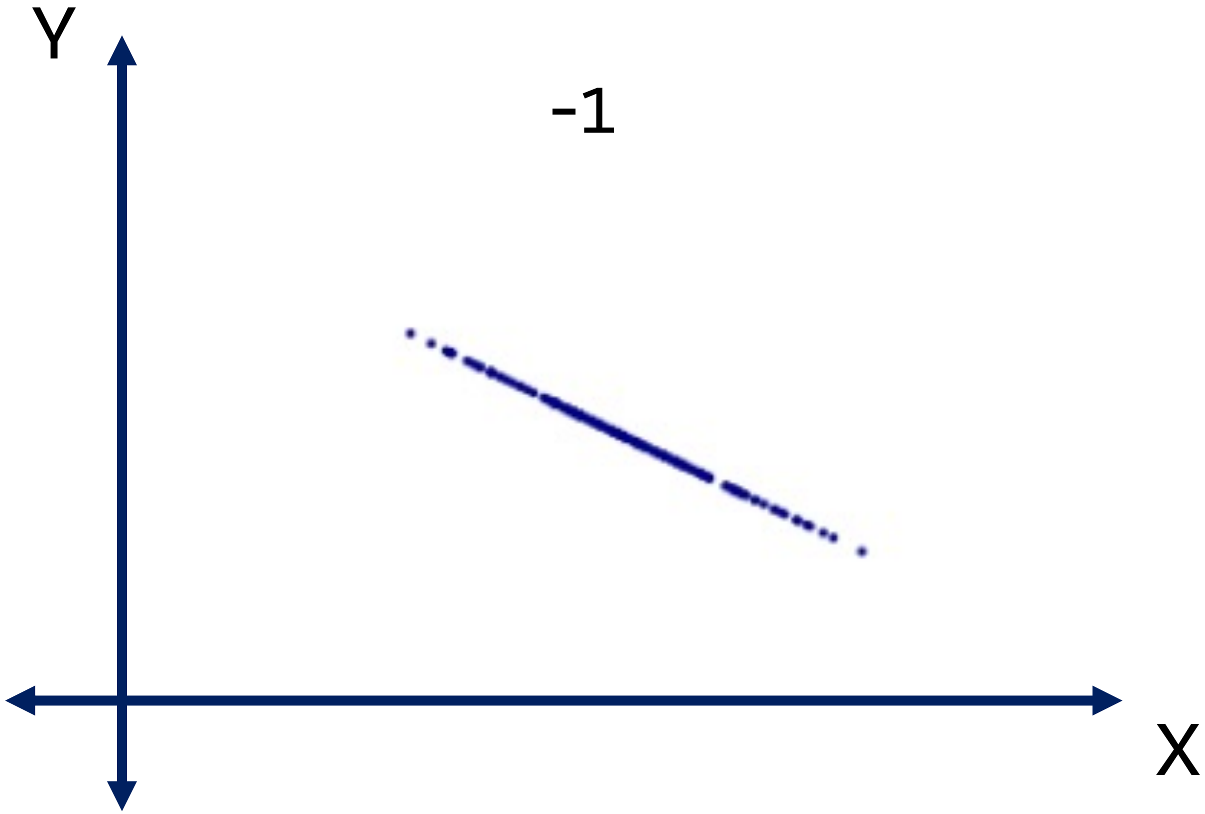
$$r = \frac{\frac{1}{n}\sum_{i=0}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sigma_X \sigma_Y}$$

PEARSON CORRELATION





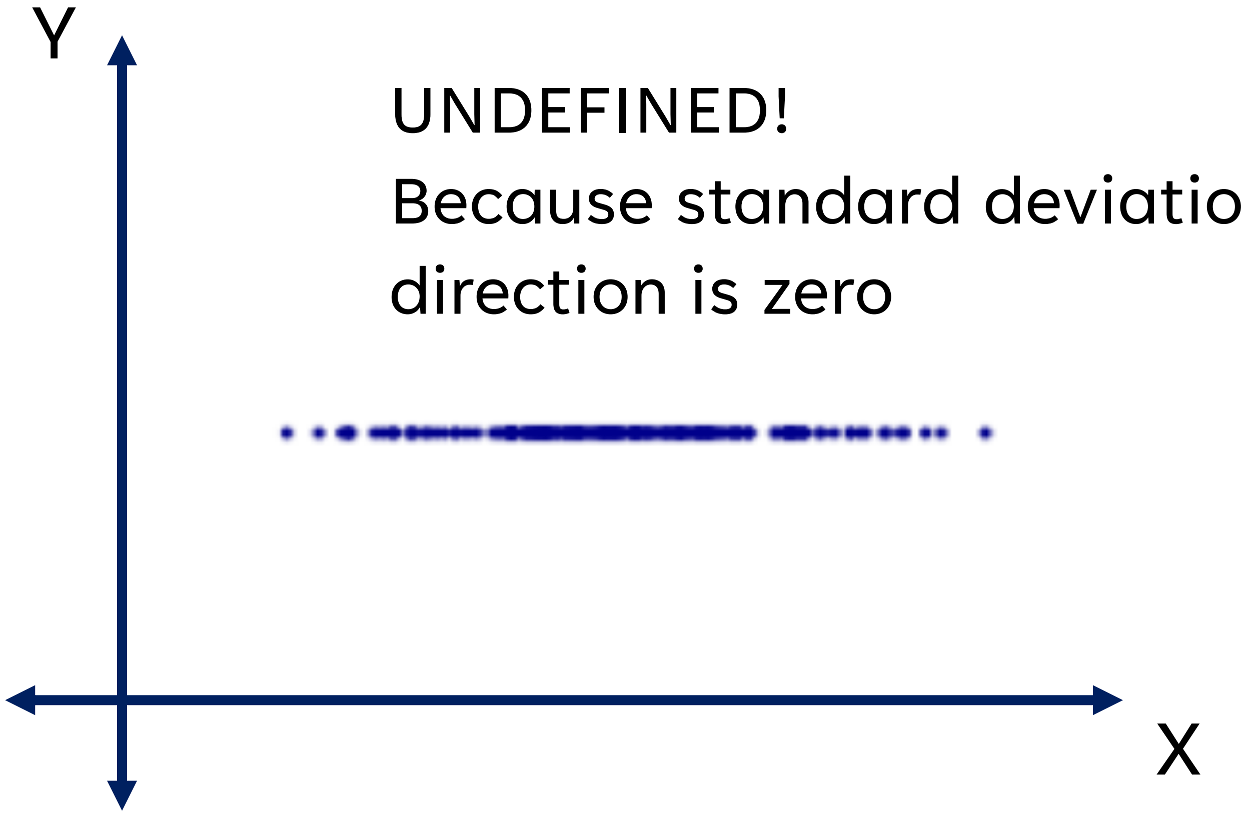
PEARSON CORRELATION



PEARSON CORRELATION

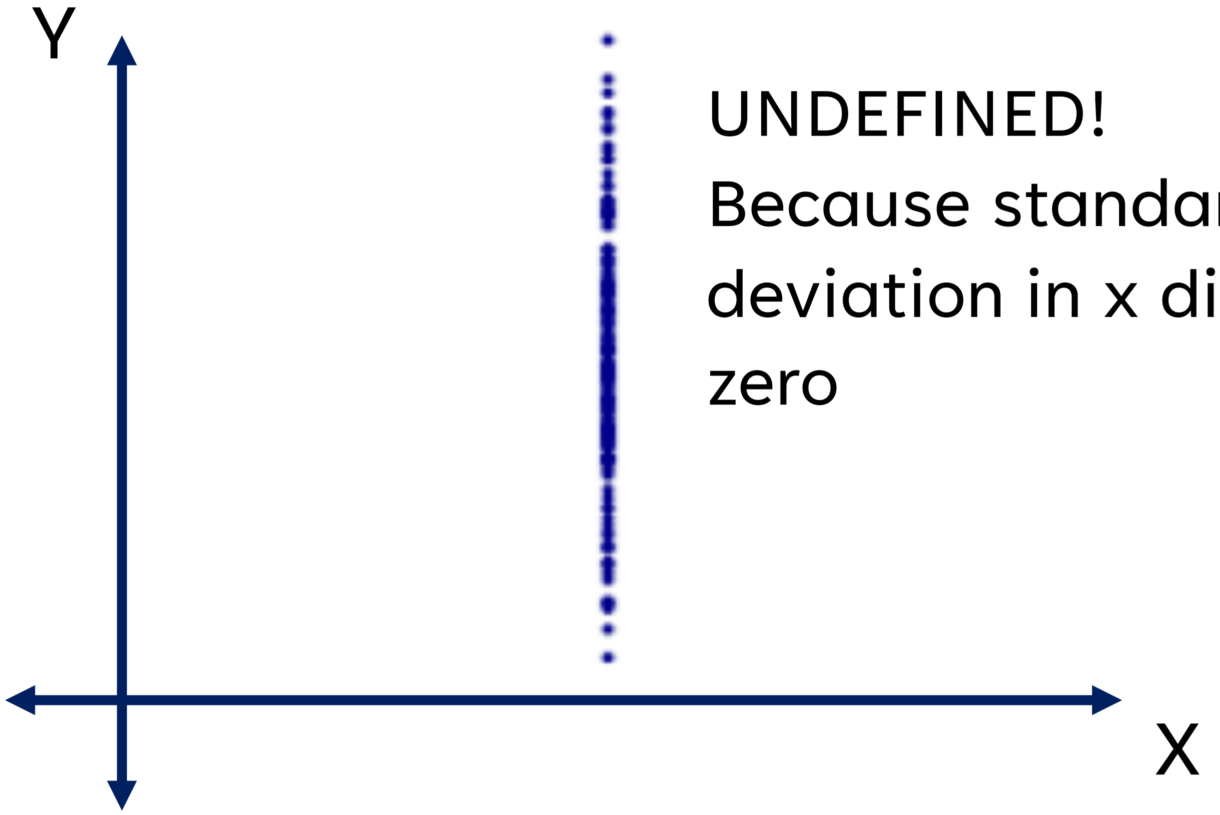
UNDEFINED!

Because standard deviation in y direction is zero



$$r = \frac{\frac{1}{n}\sum_{i=0}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sigma_X \sigma_Y}$$

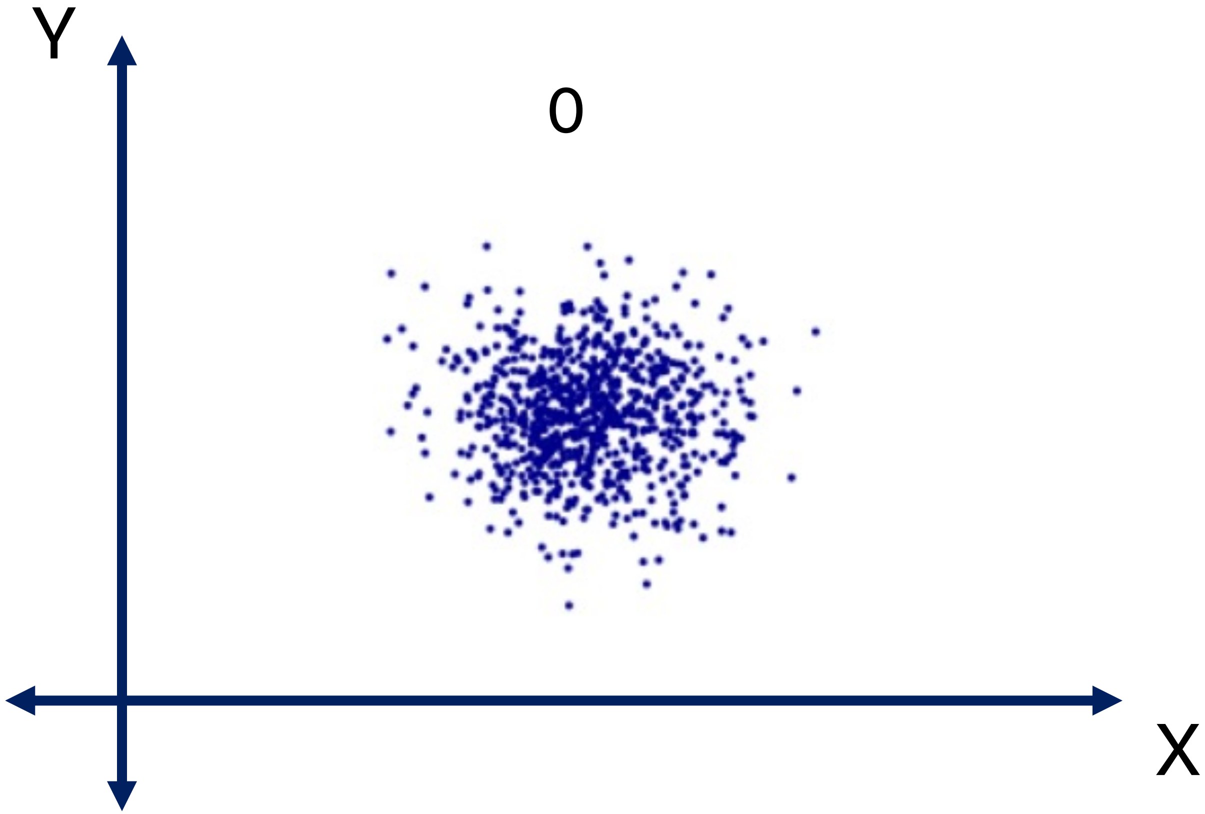
PEARSON CORRELATION



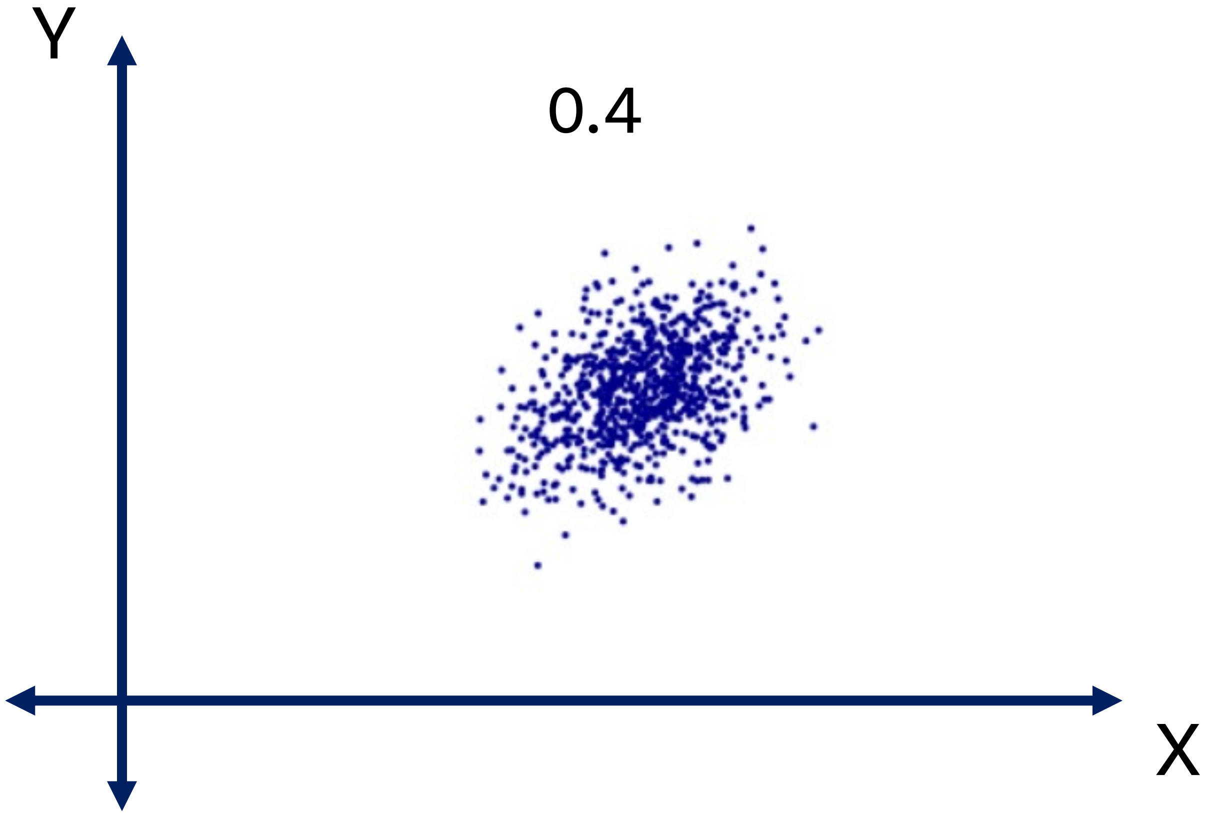
UNDEFINED!
Because standard deviation in x direction is zero

$$r = \frac{\frac{1}{n}\sum_{i=0}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sigma_X \sigma_Y}$$

PEARSON CORRELATION



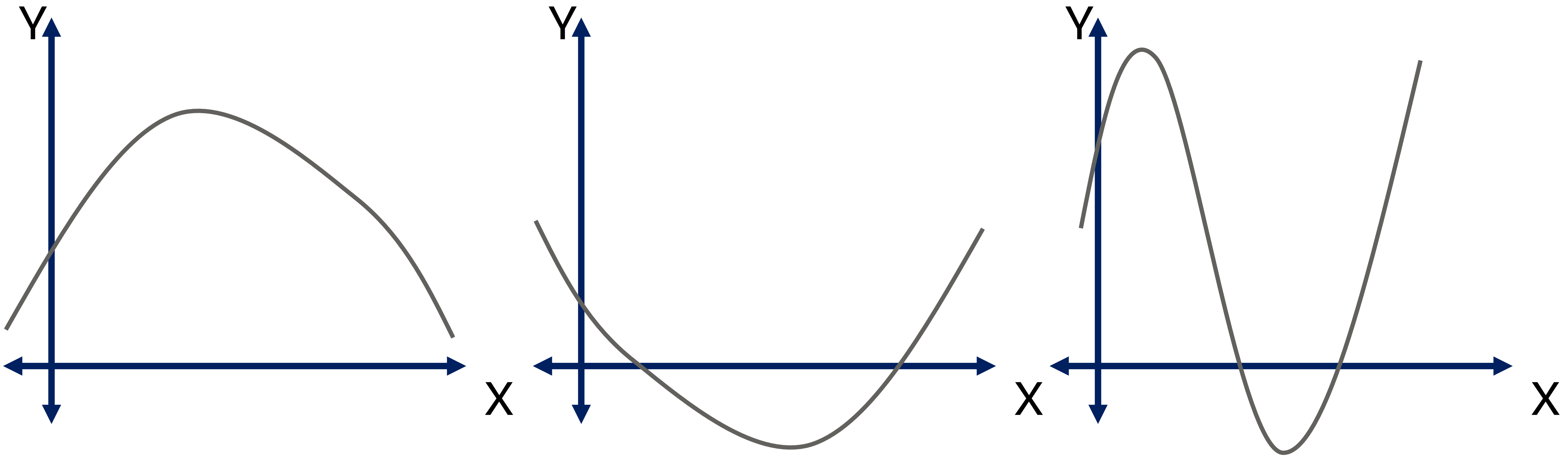
PEARSON CORRELATION



DEPENDENCE

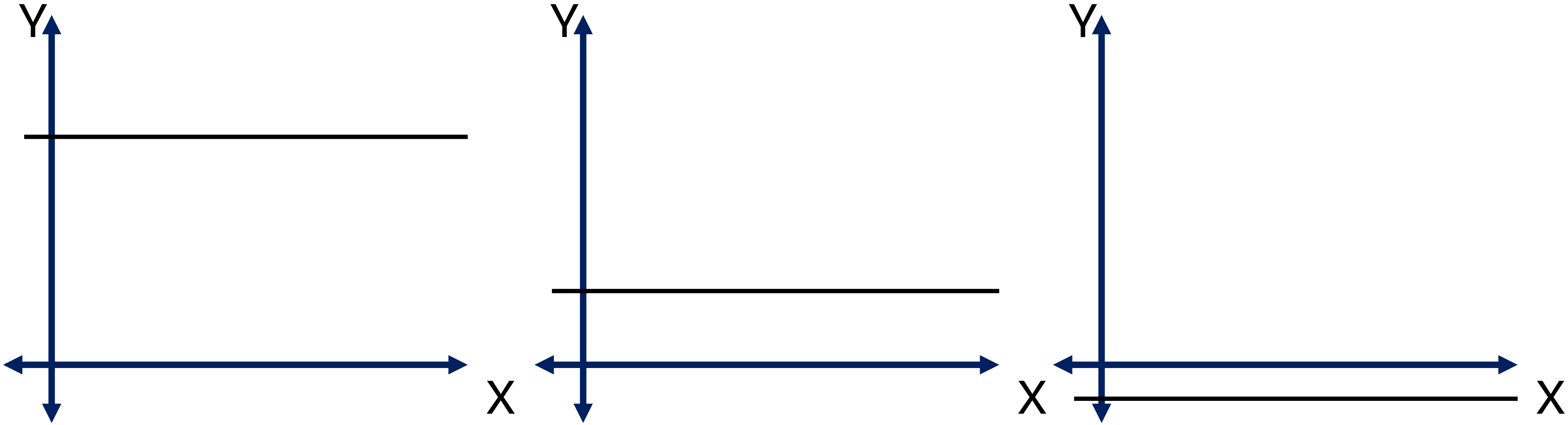
A variable y is dependent on another variable x if y=f(x).

Meaning y is a function of x.

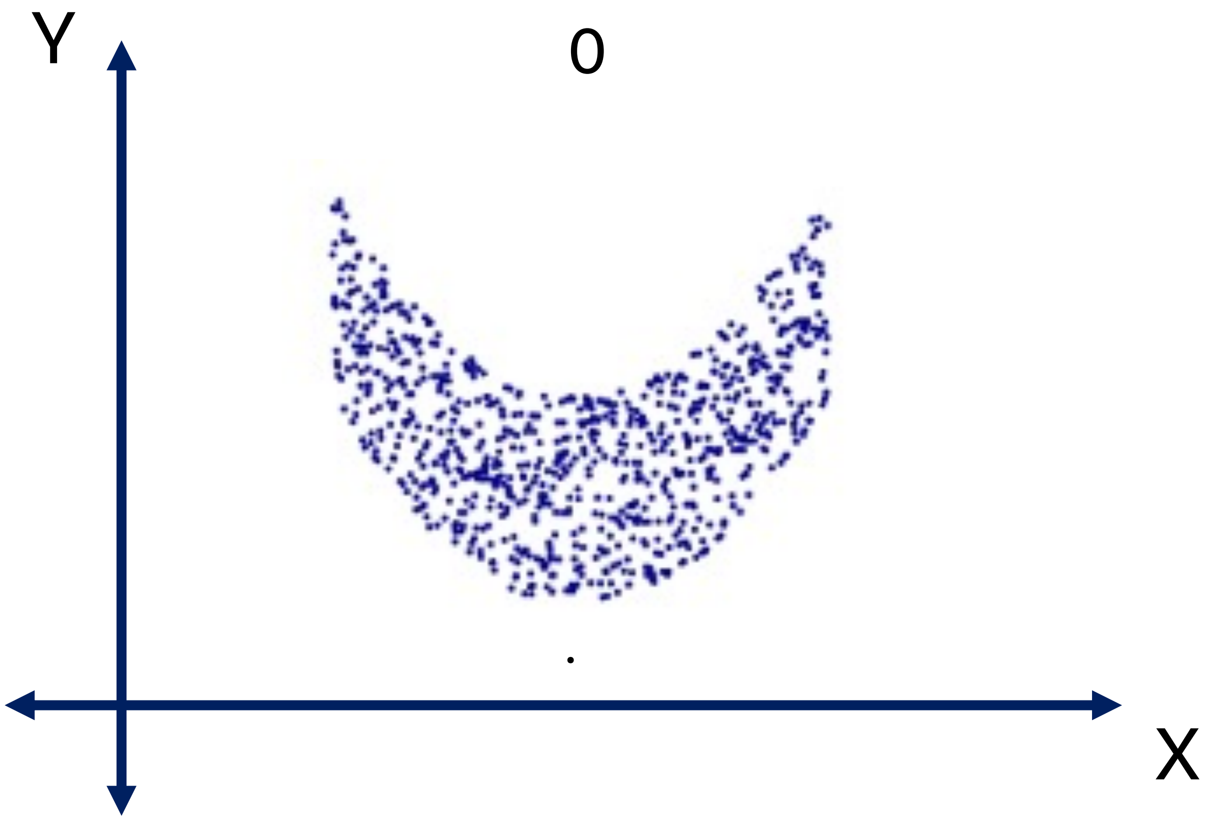


INDEPENDENCE

A variable y is *independent* of x if y remains constant as x changes.



CORRELATION



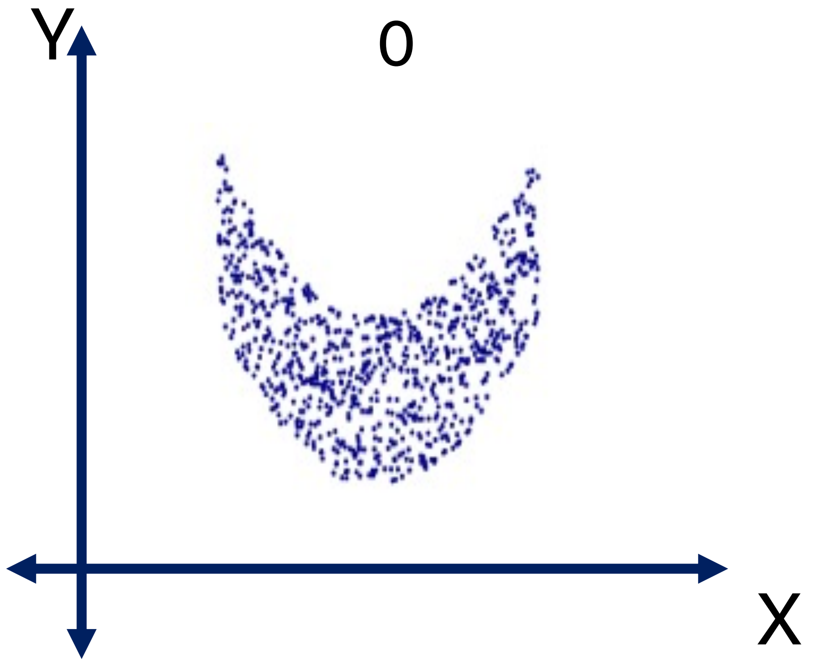
PEARSON CORRELATION COEFFICIENT ONLY CAPTURES LINEAR RELATIONSHIPS

Y is dependent on X, but has zero Pearson Correlation

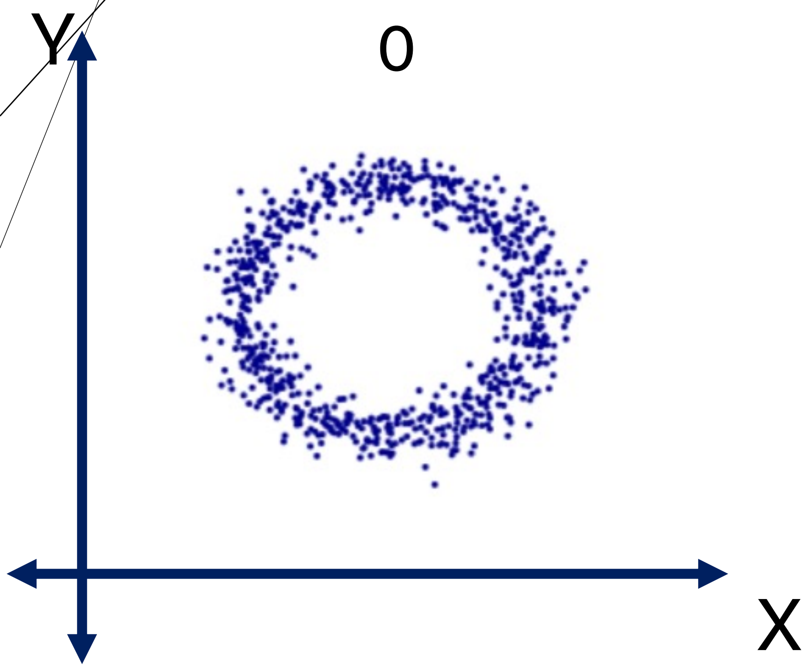
CORRELATION VS. DEPENDENCE VS. INDEPENDENCE

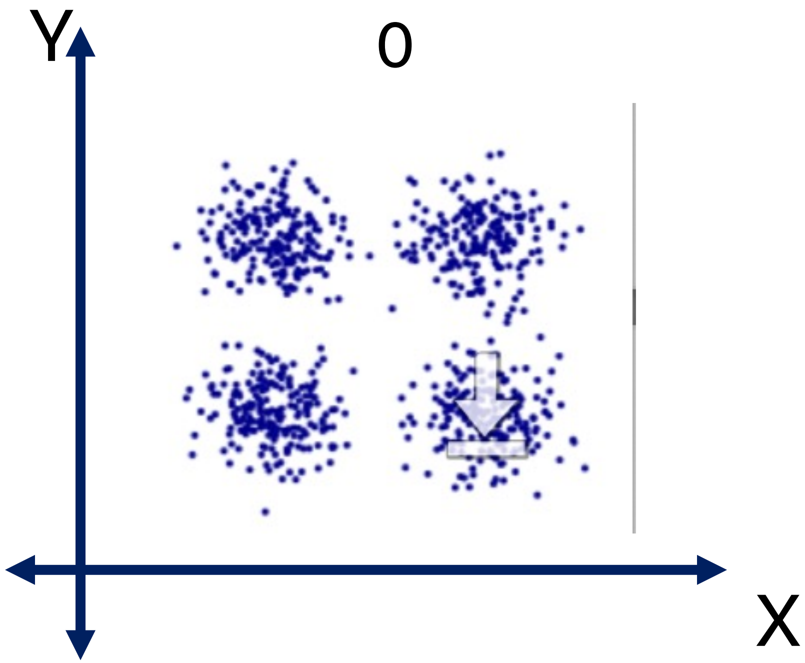
If two random variables are linearly correlated then they are dependent.

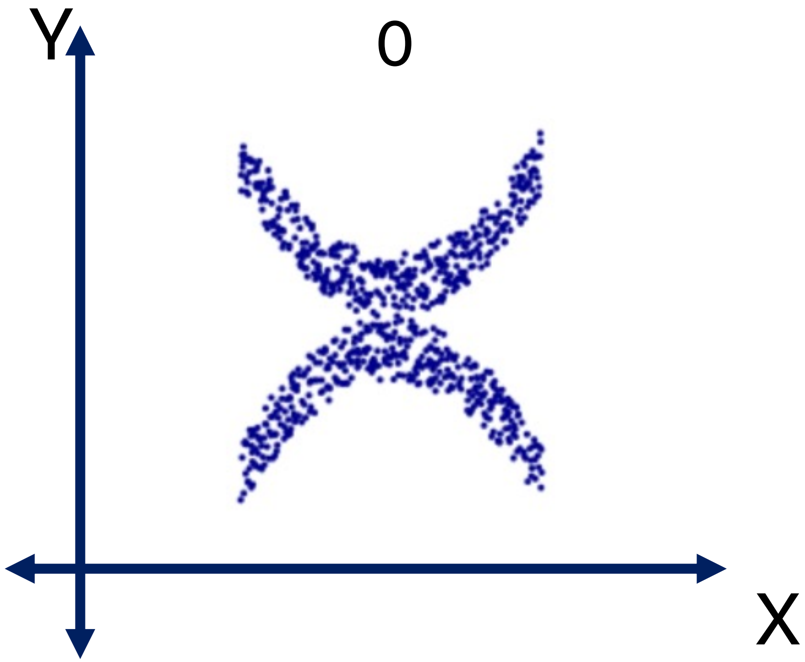
If two random variables are related in a non-linear way, they may have zero correlation and yet still be dependent!



ALL DEPENDENT BUT NOT LINEARLY CORRELATED

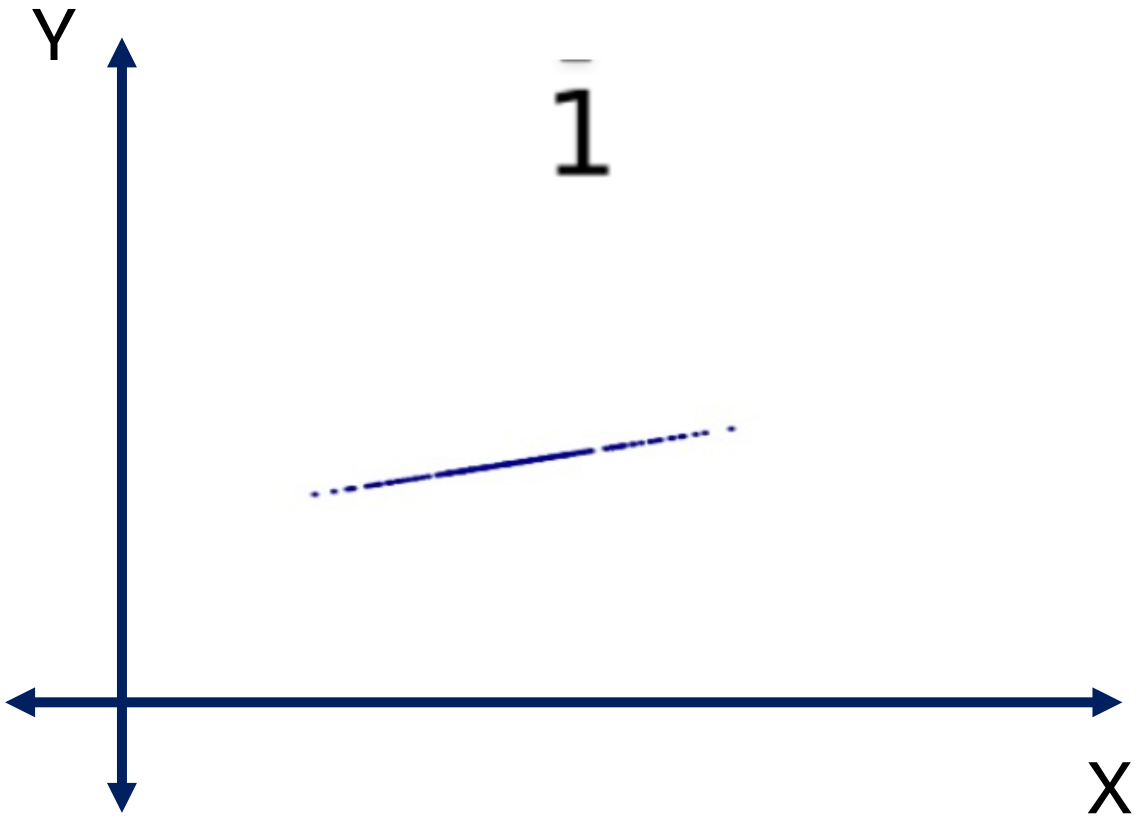






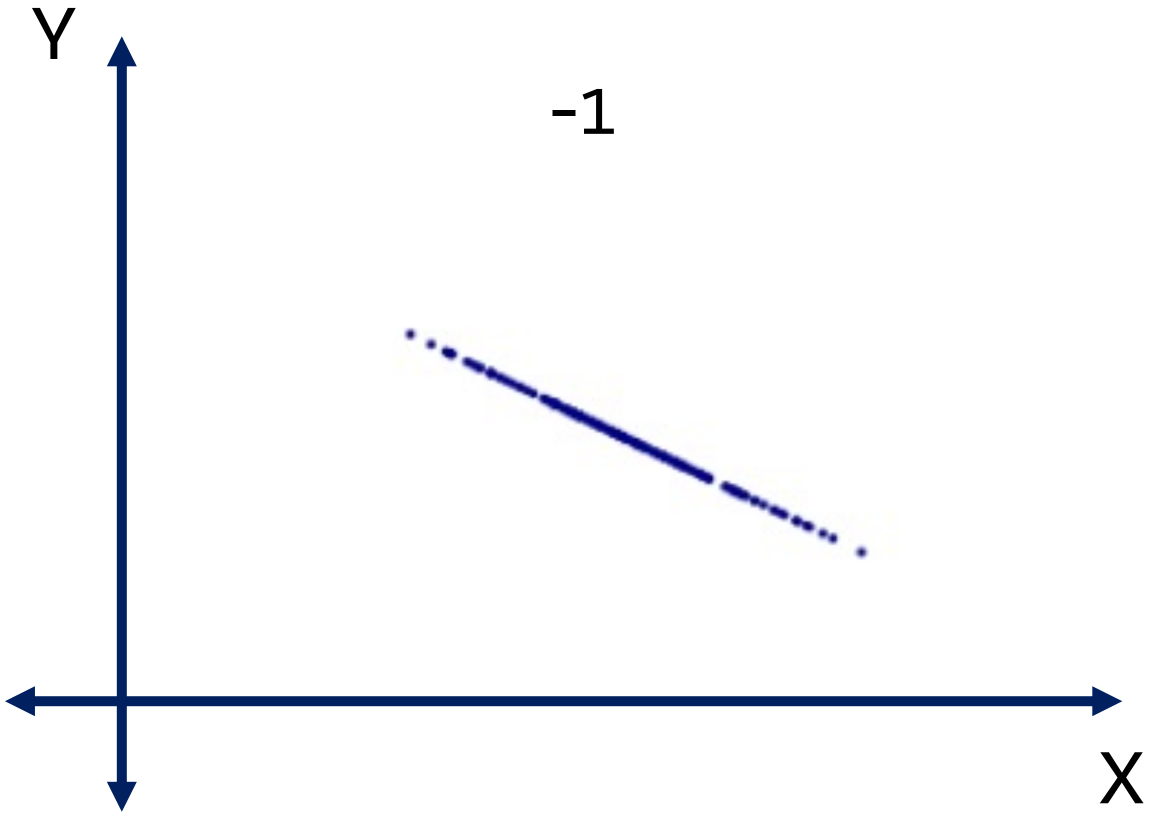
CORRELATION MATRIX

$$\begin{array}{cc} & \begin{array}{cc} X & Y \end{array} \\ \begin{array}{c} X \\ Y \end{array} & \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \end{array}$$



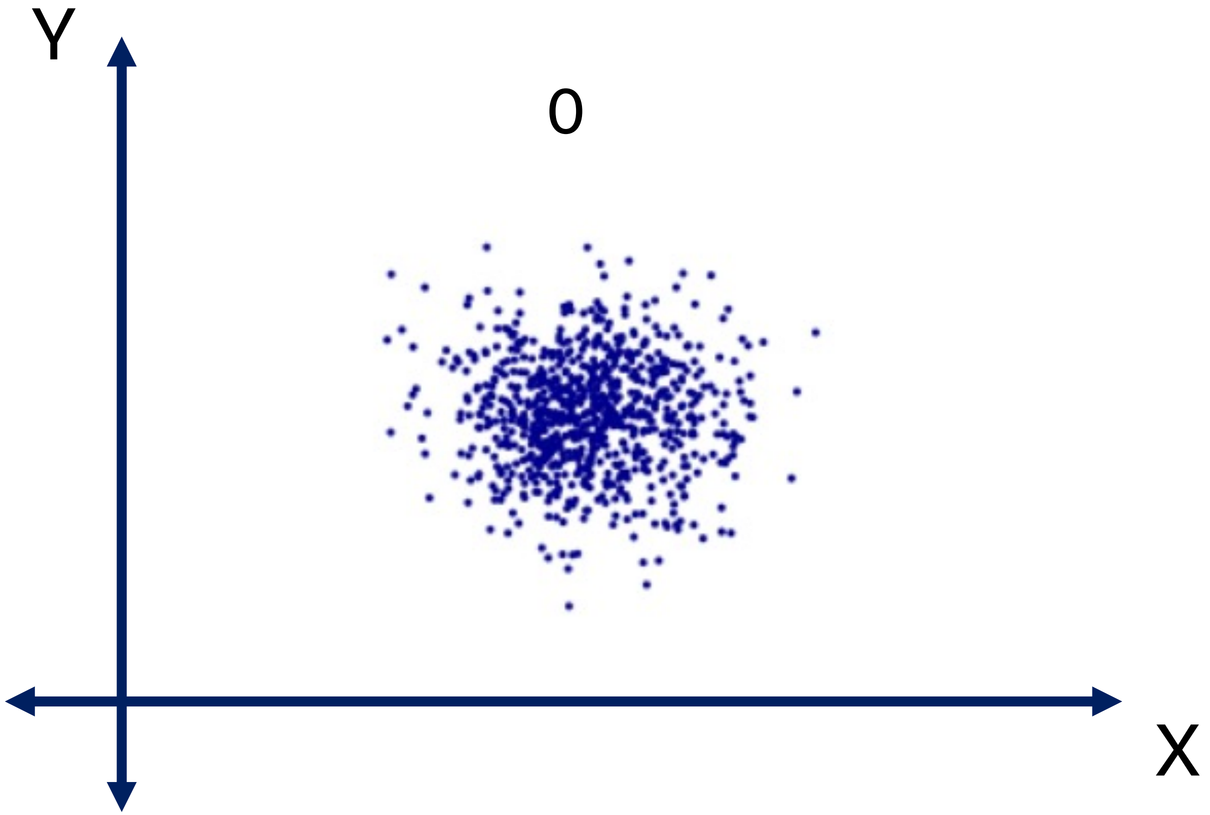
CORRELATION MATRIX

$$\begin{array}{ccc} & X & Y \\ \begin{array}{c} X \\ Y \end{array} & \multicolumn{2}{c}{\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}} \end{array}$$



CORRELATION MATRIX

$$\begin{array}{ccc} & X & Y \\ X & \multicolumn{2}{c}{\multirow{2}{*}{$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$}} \\ Y & & \end{array}$$



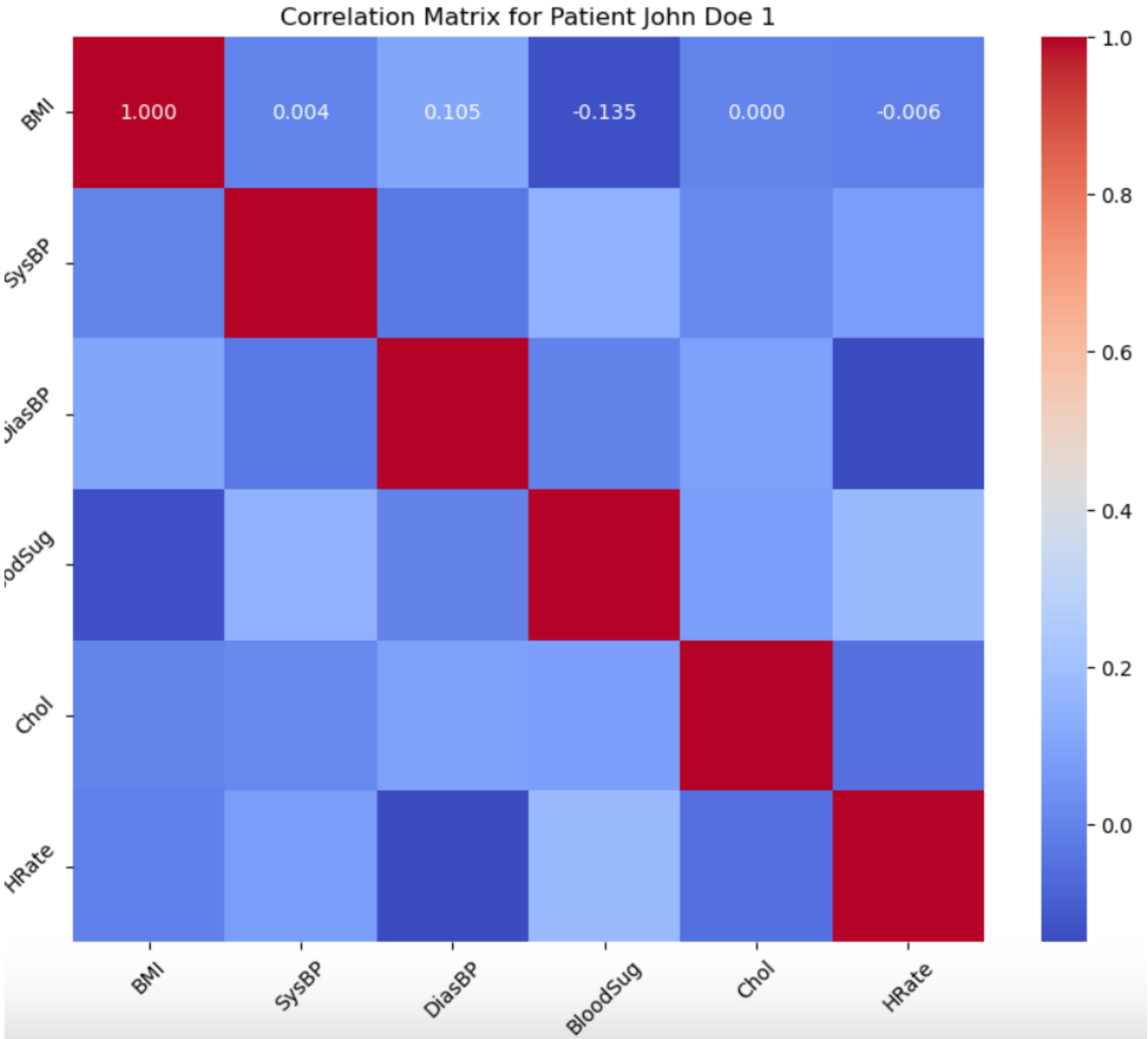
CORRELATION MATRIX

Correlation matrices are often used when there are many random variables and we want to see which ones might correlate with each other.

	BMI	SysBP	DiasBP	BloodSug	Chol	HRate
BMI	1.000	0.004	0.105	-0.135	0.000	-0.006
SysBP	0.004	1.000	-0.028	0.143	0.016	0.079
DiasBP	0.105	-0.028	1.000	-0.004	0.086	-0.149
BloodSug	-0.135	0.143	-0.004	1.000	0.082	0.183
Chol	0.000	0.016	0.086	0.082	1.000	-0.054
HRate	-0.006	0.079	-0.149	0.183	-0.054	1.000

CORRELATION MATRIX

Same correlation matrix represented as a heat map.



CORRELATION MATRIX

T = AT&T

VZ = Verizon

LVLT = Level 3

- Telecomm / Network Infrastructure

Table 1-7. Correlation between telecommunication stock returns

	T	CTL	FTR	VZ	LVLT
T	1.000	0.475	0.328	0.678	0.279
CTL	0.475	1.000	0.420	0.417	0.287
FTR	0.328	0.420	1.000	0.287	0.260
VZ	0.678	0.417	0.287	1.000	0.242
LVLT	0.279	0.287	0.260	0.242	1.000

CORRELATION MATRIX

Plot using both color and shape to denote the strength of the correlation.

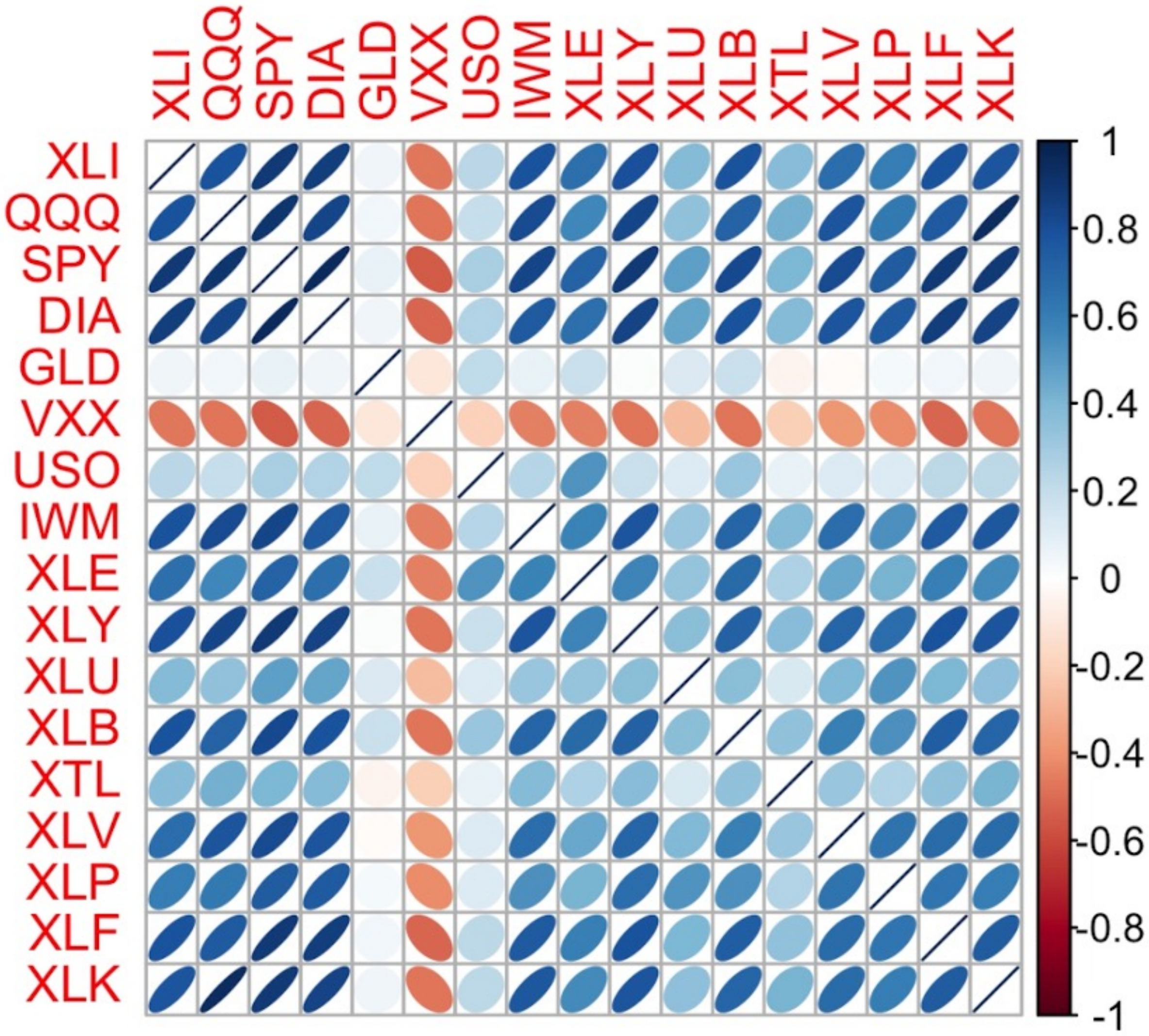


Figure 1-6. Correlation between ETF returns

THANK YOU

David Harrison

Harrison@cs.olemiss.edu