Will Monroe
July 26, 2017

with materials by
Mehran Sahami
and Chris Piech

Covariance and correlation

Announcement: Problem Set #4

Due **this Monday** before class (12:30pm).

More algorithm analysis, and detecting an impostor coin flipper!

```
HTHHHTHHHTHTHTTTTHTTHTHTH
THHTHTHTHHTHTHTHTHTHTHTHT
HHHTHTHTHTHTHHHTHT
```

Discrete conditional distributions

The value of a random variable, **conditioned on the value of some other random variable**, has a probability distribution.

$$p_{X|Y}(x,y) = \frac{P(X=x, Y=y)}{P(Y=y)}$$

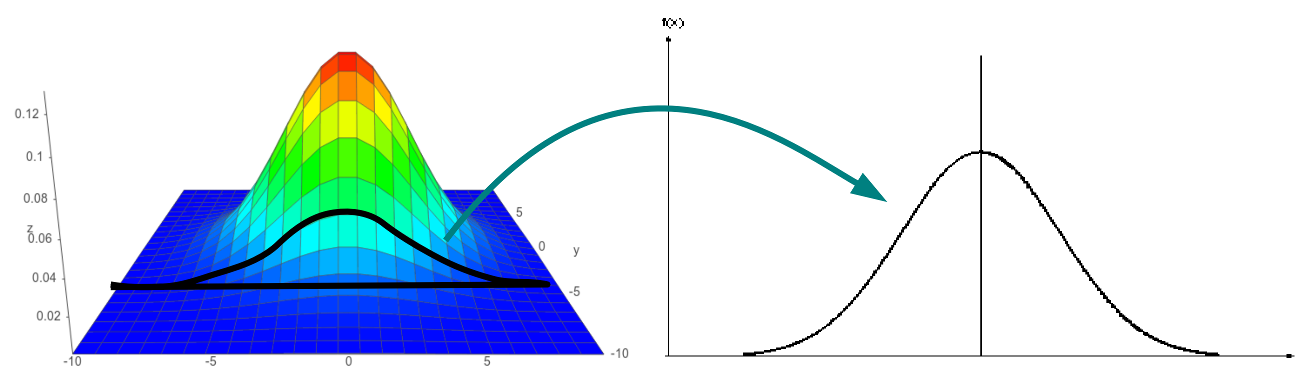
$$= \frac{p_{X,Y}(x,y)}{p_Y(y)}$$

PDF	Single	In a relationship	It's complicated / Other	TOTALS
Freshman	0.00	0.00	0.00	0.00
Sophomore	0.06	0.00	0.00	0.06
Junior	0.19	0.19	0.13	0.50
Senior	0.00	0.00	0.00	0.00
Grad student / Other	0.38	0.06	0.00	0.44
TOTALS	0.63	0.25	0.13	1.00

Continuous conditional distributions

The value of a random variable, **conditioned on the value of some other random variable**, has a probability distribution.

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

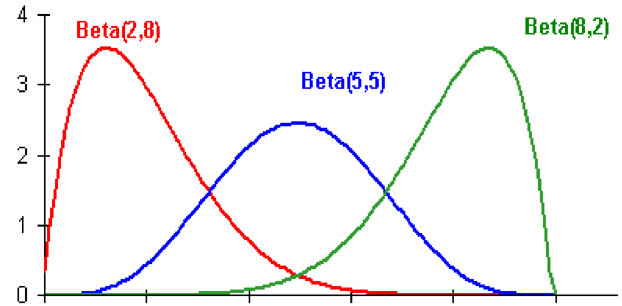


Beta random variable

An **beta** random variable models the **probability** of a trial's success, given previous trials. The PDF/CDF let you compute **probabilities of probabilities**!

$$X \sim \text{Beta}(a, b)$$

$$f_X(x) = \begin{cases} C\, x^{a-1}(1-x)^{b-1} & \text{if } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$



Beta: Fact sheet

number of **successes** + 1

$$X \sim \text{Beta}(a, b)$$

probability of success

number of **failures** + 1

PDF: $$f_X(x) = \begin{cases} C\, x^{a-1}(1-x)^{b-1} & \text{if } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

expectation: $$E[X] = \frac{a}{a+b}$$

variance: $$\text{Var}(X) = \frac{ab}{(a+b)^2(a+b+1)}$$

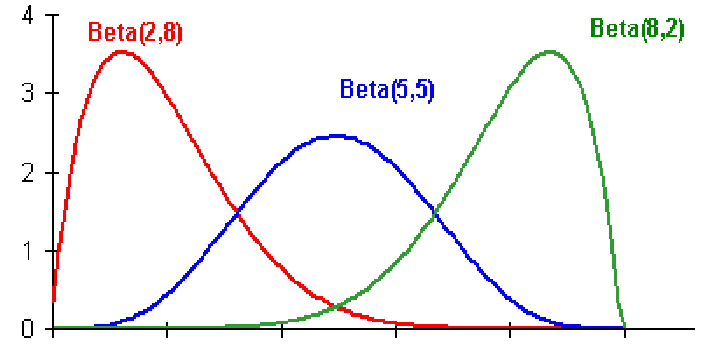
Subjective priors

X | A ~ Beta(a + 1, N − a + 1)
"posterior"

X ~ Beta(1, 1)
"prior"

$$f_{X|A}(x|a) = \frac{P(A=a|X=x)f_X(x)}{P(A=a)}$$

How did we decide on
Beta(1, 1) for the prior?



Beta(1, 1): "we haven't seen any rolls yet."
Beta(4, 1): "we've seen 3 sixes and 0 non-sixes."
Beta(2, 6): "we've seen 1 six and 5 non-sixes."

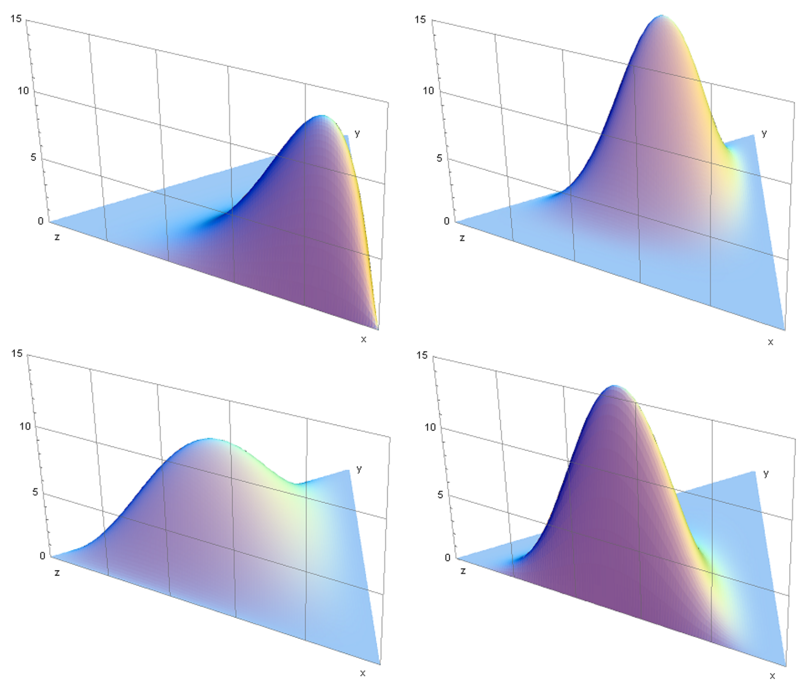
Beta prior = "imaginary" previous trials

Beta calculator

Advanced: Dirichlet distribution

Beta is the distribution ("conjugate prior") for the p in the **Bernoulli** and **binomial**.

Dirichlet is the distribution for the p_1, p_2, ... in the **multinomial**.



$$X_1, X_2, \ldots \sim \text{Dir}(a_1, a_2, \ldots)$$

$$f_{X_1, X_2, \ldots}(x_1, x_2, \ldots) = C\, x_1^{a_1 - 1} x_2^{b_2 - 1} \ldots$$

if $0 < \{x_1, x_2, \ldots\} < 1$,

$x_1 + x_2 + \cdots = 1$

(0 otherwise)

Frequentists vs. Bayesians

Frequentist	Bayesian
A probability is the (real or theoretical) result of a number of experiments.	A probability is a belief.
All probabilities are based on objective experiences.	All probabilities are based on subjective priors.

(It's not really a debate anymore—real statisticians / data scientists / machine learning practitioners can and do think both ways!)

image: Eric Kilby

Expectation of a product

If two random variables are **independent**, then the **expectation of their product** equals the **product of their expectations.**

$$X \perp Y \quad \Rightarrow$$

$$E[XY] = E[X]E[Y]$$

$$E[g(X)h(Y)] = E[g(X)]E[h(Y)]$$

Expectation of a product

$$E[g(X)h(Y)] = \int_{-\infty}^{\infty} dy \int_{-\infty}^{\infty} dx\, g(x)h(y)f_{X,Y}(x,y)$$

(independence!)

$$= \int_{-\infty}^{\infty} dy \int_{-\infty}^{\infty} dx\, g(x)h(y)f_X(x)f_Y(y)$$

$$= \int_{-\infty}^{\infty} dy\, h(y)f_Y(y) \int_{-\infty}^{\infty} dx\, g(x)f_X(x)$$

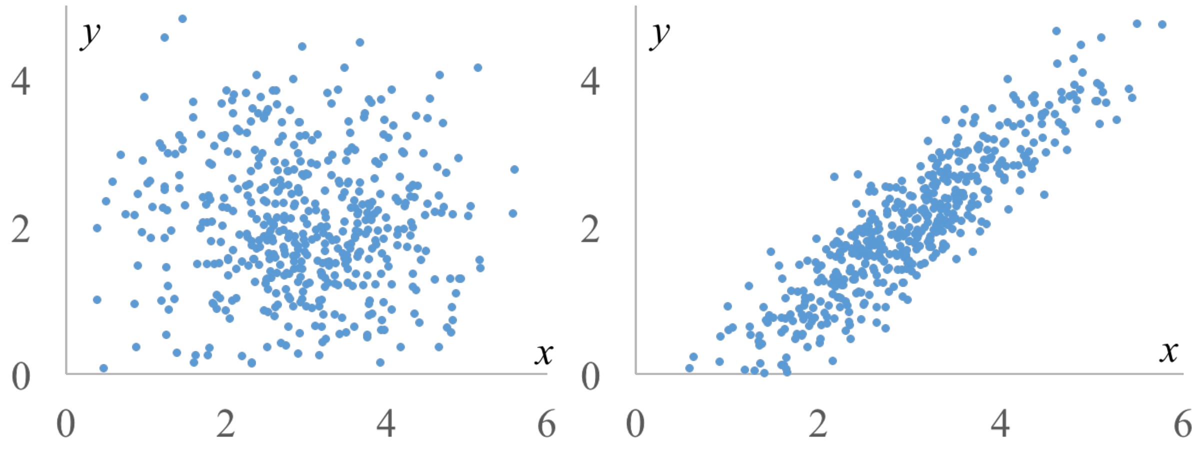
$$= \left(\int_{-\infty}^{\infty} dx\, g(x)f_X(x)\right)\left(\int_{-\infty}^{\infty} dy\, h(y)f_Y(y)\right)$$

$$= E[g(X)]E[h(Y)]$$

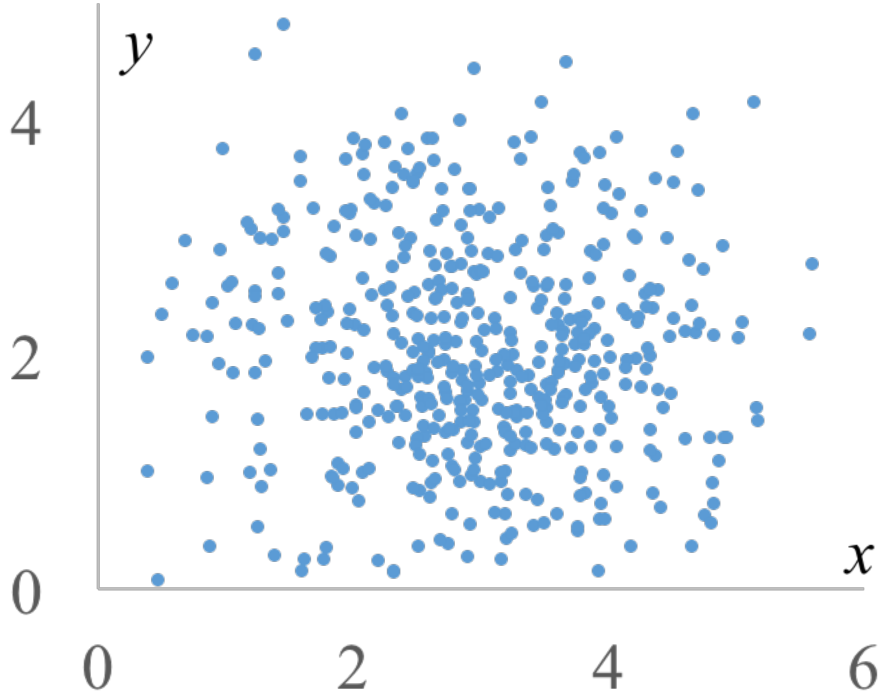
Covariance

The **covariance** of two variables is a measure of how much they **vary together**.

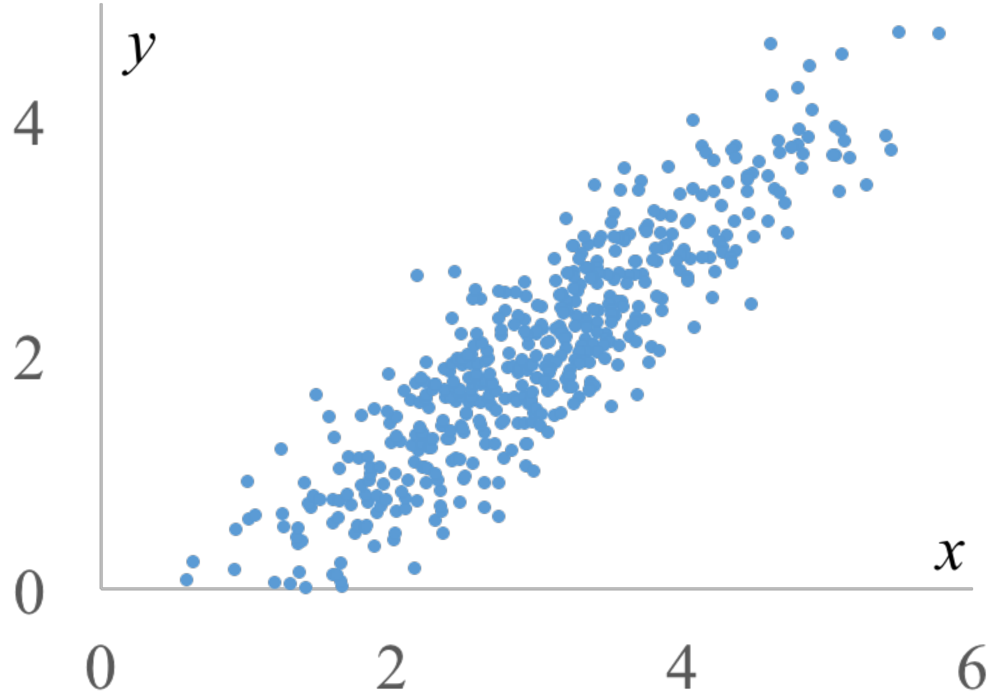
$$\mathrm{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])]$$
$$= E[XY] - E[X]E[Y]$$



A Tale of Two Distributions



$$E[X]=3$$
$$E[Y]=2$$
$$\text{Var}(X)=2.25$$
$$\text{Var}(Y)=1$$

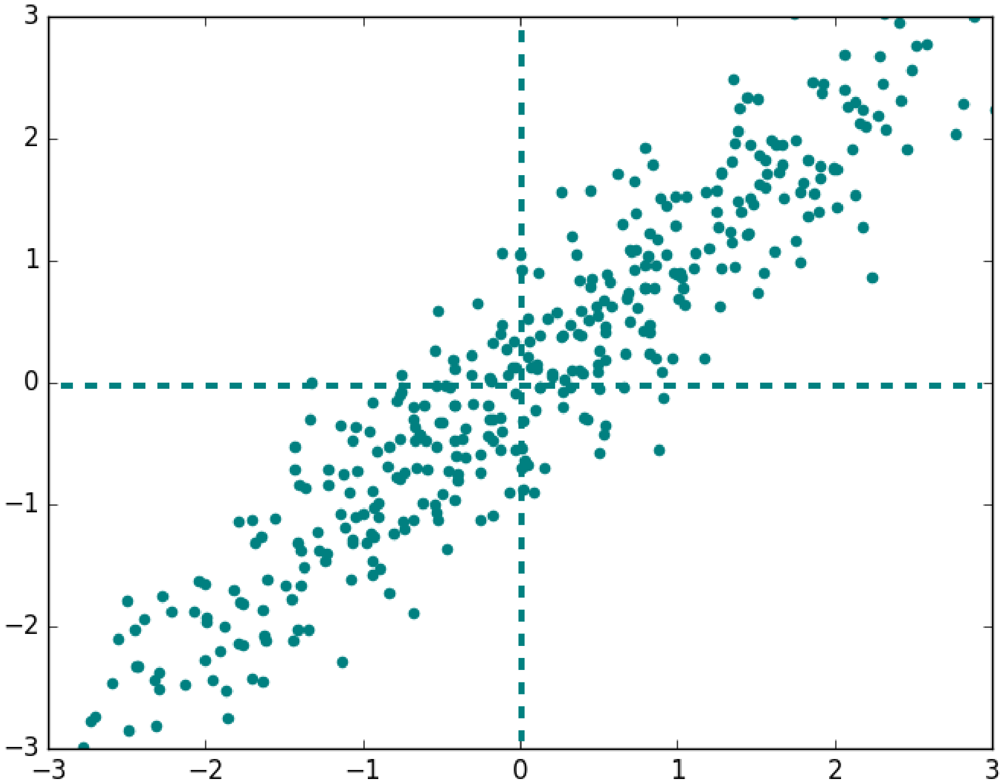


$$E[X]=3$$
$$E[Y]=2$$
$$\text{Var}(X)=2.25$$
$$\text{Var}(Y)=1$$

Two variables playing together

$- \times + = -$ $+ \times + = +$

Y



$- \times - = +$ $+ \times - = -$

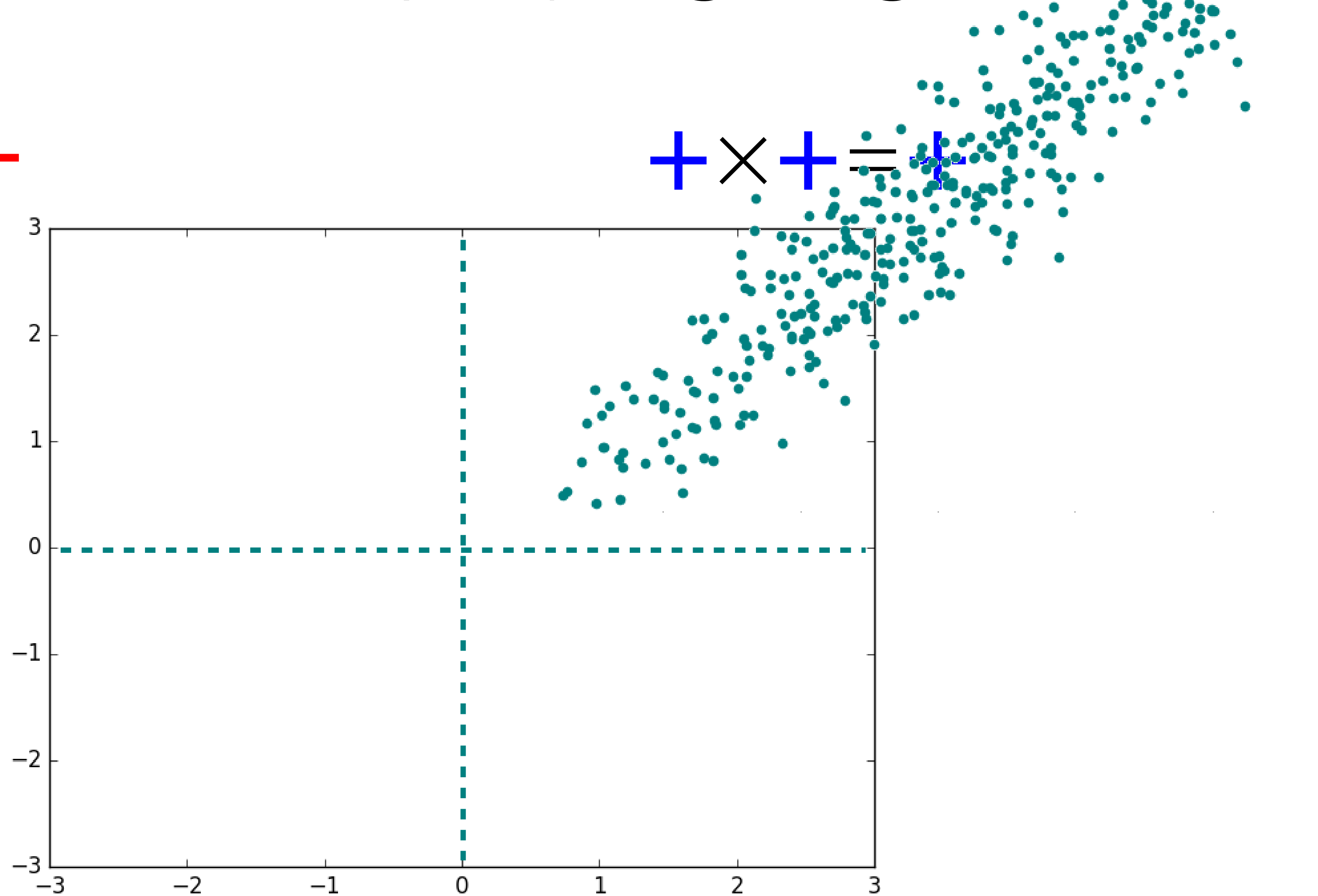
X

$$E[\,X\,Y\,]?$$

Two variables playing together

$- \times + = -$

$+ \times + = +$



Y

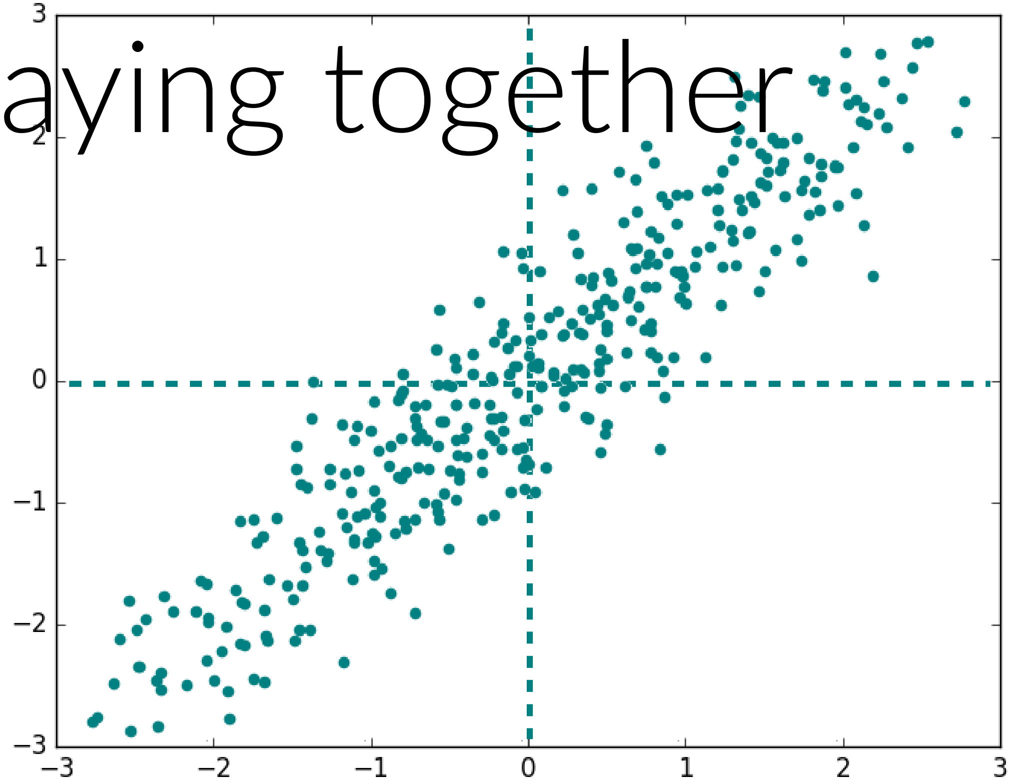
X

$- \times - = +$

$+ \times - = -$

$$E[\,X\,Y\,]?$$

Two variables playing together



Y

$- \times - = +$

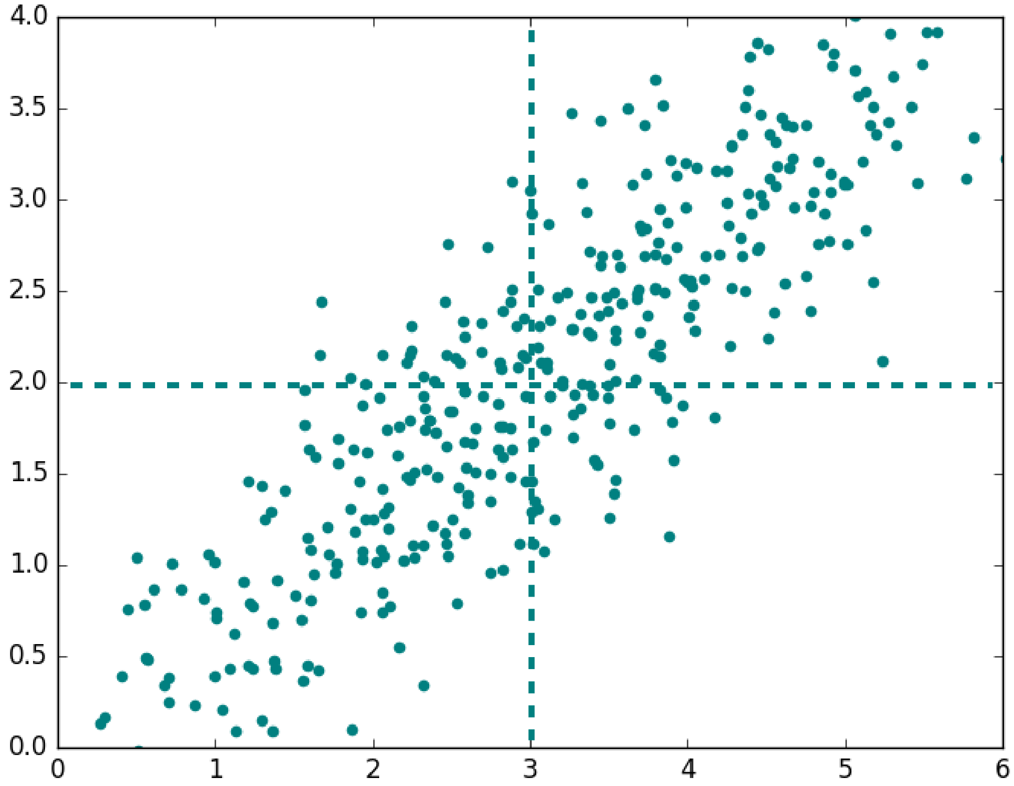
X

$+ \times - =$

$E[XY]$?

Two variables playing together

$- \times + = -$ $\quad$ $+ \times + = +$



Y

$- \times - = +$ $\quad$ $+ \times - = -$

X

$$\mathrm{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])]$$

The easy way to compute covariance

$$\mathrm{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])]$$

$$= E[XY - XE[Y] - E[X]Y + E[X]E[Y]]$$

The easy way to compute covariance

$$\mathrm{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])]$$

$= E[$ [illegible] $- E$ [illegible] $Y + E$ [illegible] $]]$

$= E[$ [illegible] $] - E[$ [illegible] $] - E[E$ [illegible] $Y] + E[E$ [illegible] $]]$

(linearity of expectation!)

The easy way to compute covariance

$$\text{Cov}(X,Y) = E[(X - E[X])(Y - E[Y])]$$

$$= E[XY - XE[Y] - E[X]Y + E[X]E[Y]]$$

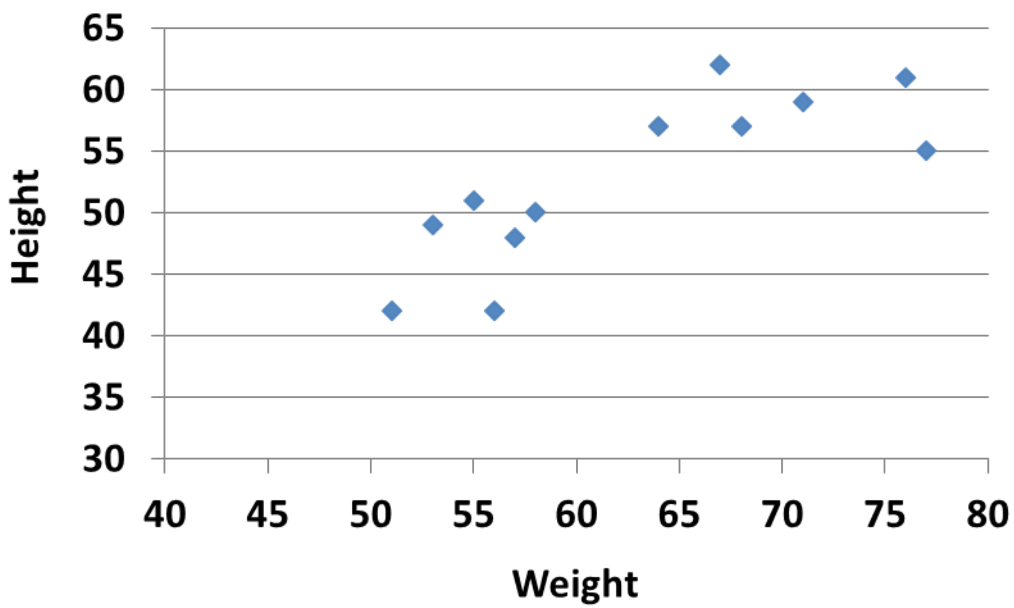
$$= E[XY] - E[XE[Y]] - E[E[X]Y] + E[E[X]E[Y]]$$

(linearity of expectation!)

$$= E[XY] - E[Y]E[X] - E[X]E[Y] + E[X]E[Y]$$

$$= E[XY] - E[X]E[Y]$$

Example: Weight and height data



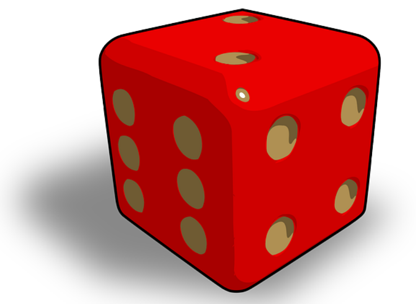
Weight	Height	W · H
64	57	3648
71	59	4189
53	49	2597
67	62	4154
55	51	2805
58	50	2900
77	55	4235
57	48	2736
56	42	2352
51	42	2142
76	61	4636
68	57	3876

$$E[W] = 62.75$$
$$E[H] = 52.75$$
$$E[W \cdot H] = 3355.83$$

$$\mathrm{Cov}(W, H) = 3355.83 - (62.75)(52.75) = 45.77$$

Positive covariance:
Knowing high W makes high H more likely!

Example: Die rolling



Roll a (fair!) 6-sided die.

X = indicator variable for {1, 2, 3, 4}
Y = indicator variable for {3, 4, 5, 6}

$$E[X] = P(\{1,2,3,4\}) = 2/3$$

$$E[Y] = P(\{3,4,5,6\}) = 2/3$$

$$E[XY] = \sum_x \sum_y x\,y\,p_{X,Y}(x,y)$$

$$= 0\cdot 0(0) + 0\cdot 1(1/3) + 1\cdot 0(1/3) + 1\cdot 1(1/3) = 1/3$$

$$\mathrm{Cov}(X,Y) = E[XY] - E[X]E[Y]$$

$$= 1/3 - 4/9 = -1/9$$

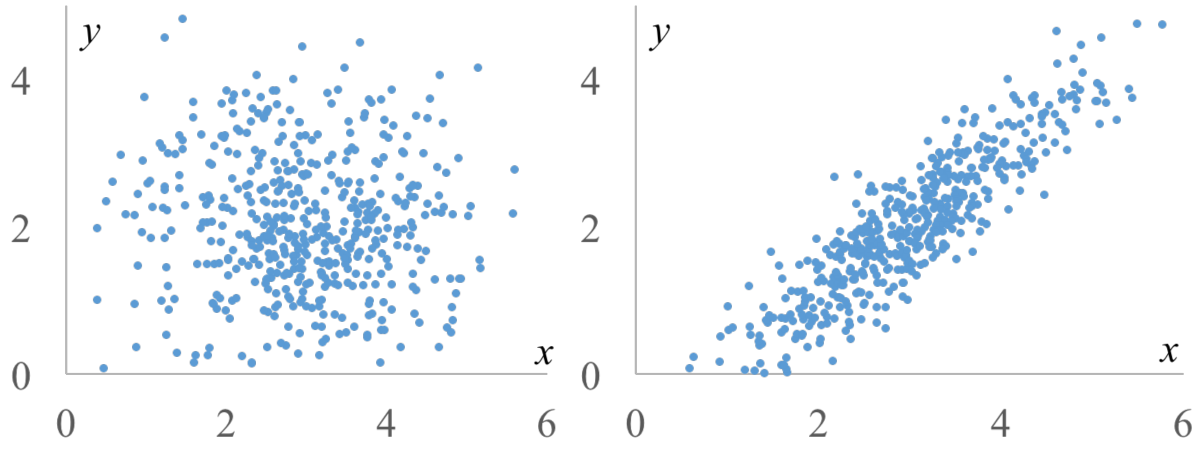
Negative covariance:
Knowing Y = 1 makes X = 1 less likely!

Break time!

Covariance

The **covariance** of two variables is a measure of how much they **vary together**.

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



Properties of covariance

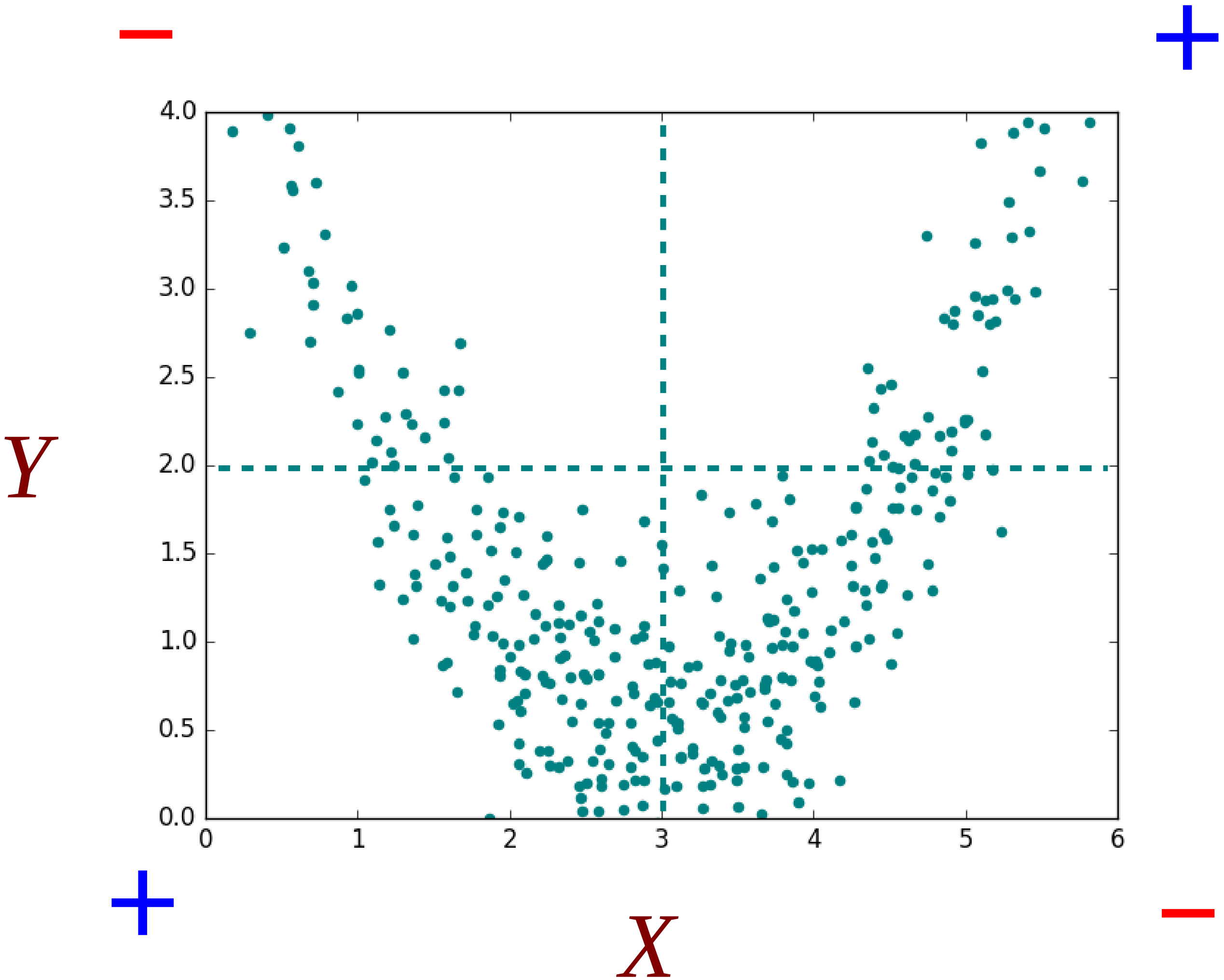
$$\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$$ (symmetric)

$$\mathrm{Cov}(X, X) = E[XX] - E[X]E[X] = \mathrm{Var}(X)$$

$$\mathrm{Cov}(aX + b, Y) = a\,\mathrm{Cov}(X, Y)$$

$$\mathrm{Cov}\left(\sum_i X_i, \sum_j Y_j\right) = \sum_i \sum_j \mathrm{Cov}(X_i, Y_j)$$

Covariance = linear dependence

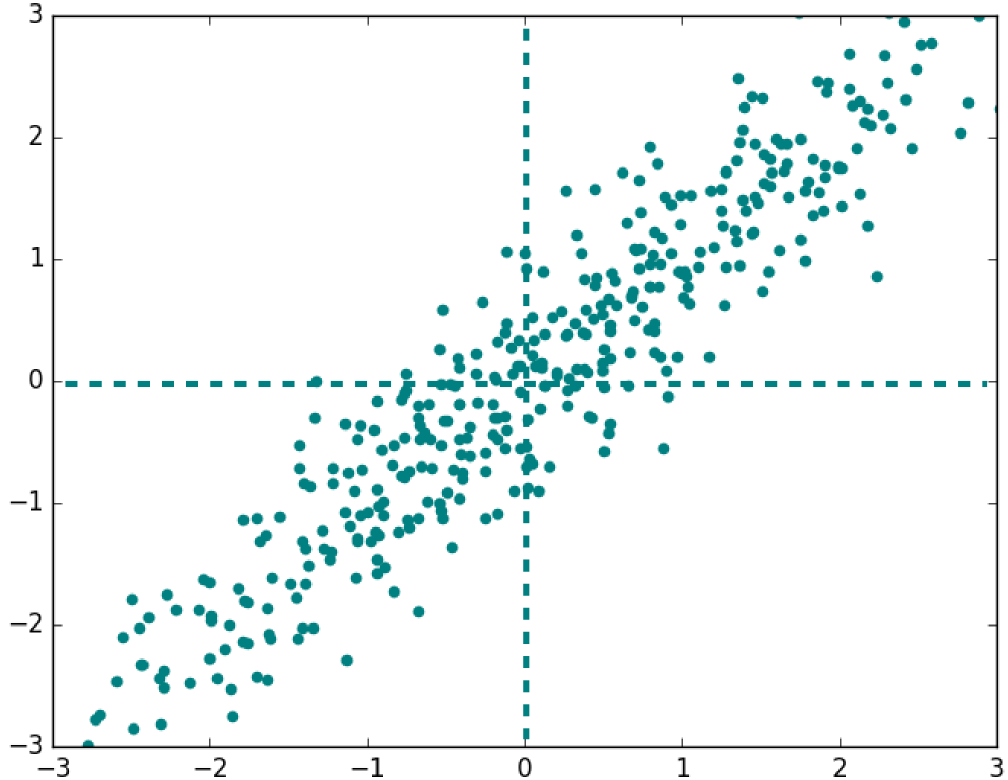


$$\mathrm{Cov}(X, Y) = 0$$

Inflating your covariance

$- \times + = -$ $\qquad$ $+ \times + = +$

Y



X

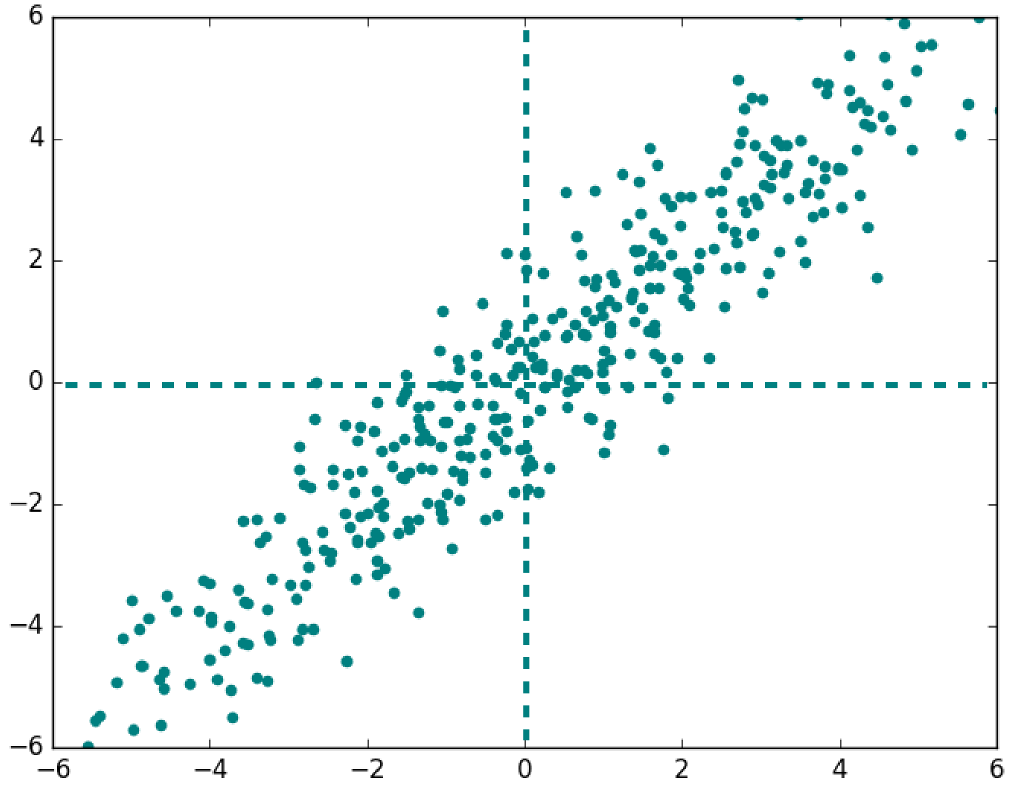
$- \times - = +$ $\qquad$ $+ \times - = -$

$$\mathrm{Cov}(X,Y) = E[(X - E[X])(Y - E[Y])]$$

Inflating your covariance

$- \times + = -$ $\qquad$ $+ \times + = +$

Y



$- \times - = +$ $\qquad$ $+ \times - = -$

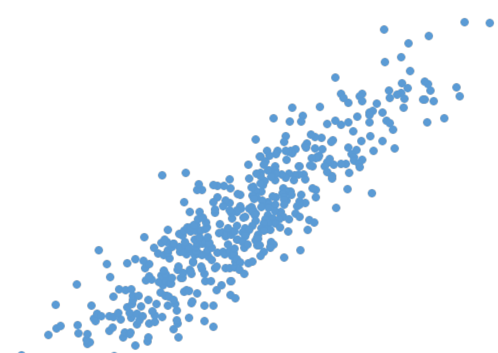
X

$$\mathrm{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])]$$

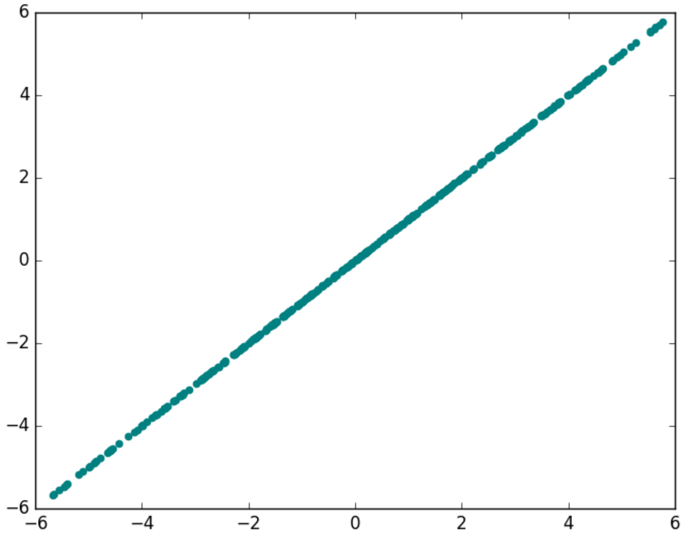
Correlation

The **correlation** of two variables is a measure of the **linear dependence** between them, scaled to always take on values between -1 and 1.

$$\rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}$$



Perfect correlation



Suppose X and Y form a perfect line:

$$Y = aX + b$$

Then

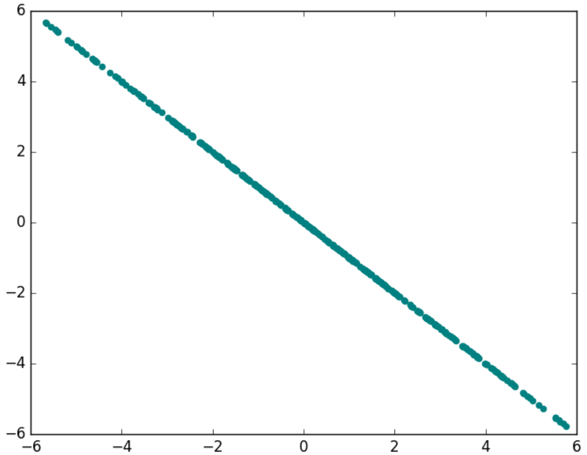
$$\begin{aligned}\mathrm{Cov}(X, Y) &= \mathrm{Cov}(X, aX + b) \\ &= a\,\mathrm{Cov}(X, X) \\ &= a\,\mathrm{Var}(X) \\ &= \pm\sqrt{a^2\,\mathrm{Var}(X)\cdot\mathrm{Var}(X)} \\ &= \pm\sqrt{\mathrm{Var}(Y)\cdot\mathrm{Var}(X)}\end{aligned}$$

Cutting covariance down to size

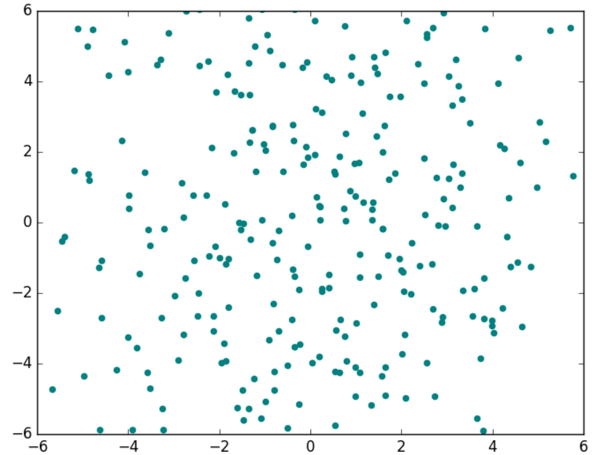
$$\rho(X,Y)=\frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}$$

divide by the covariance's maximum value

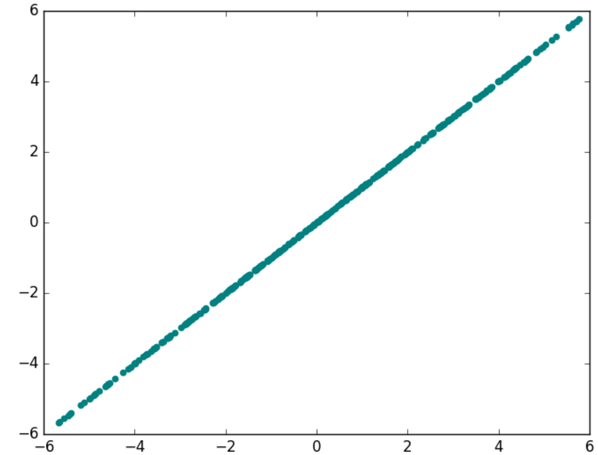
Important correlations



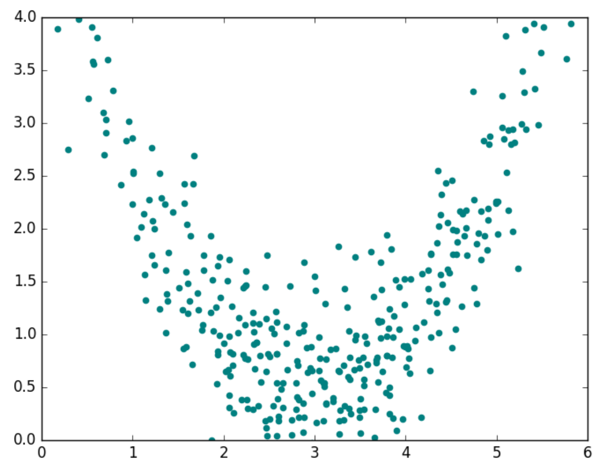
$$\rho(X, Y) = -1$$



$$\rho(X, Y) = 0$$

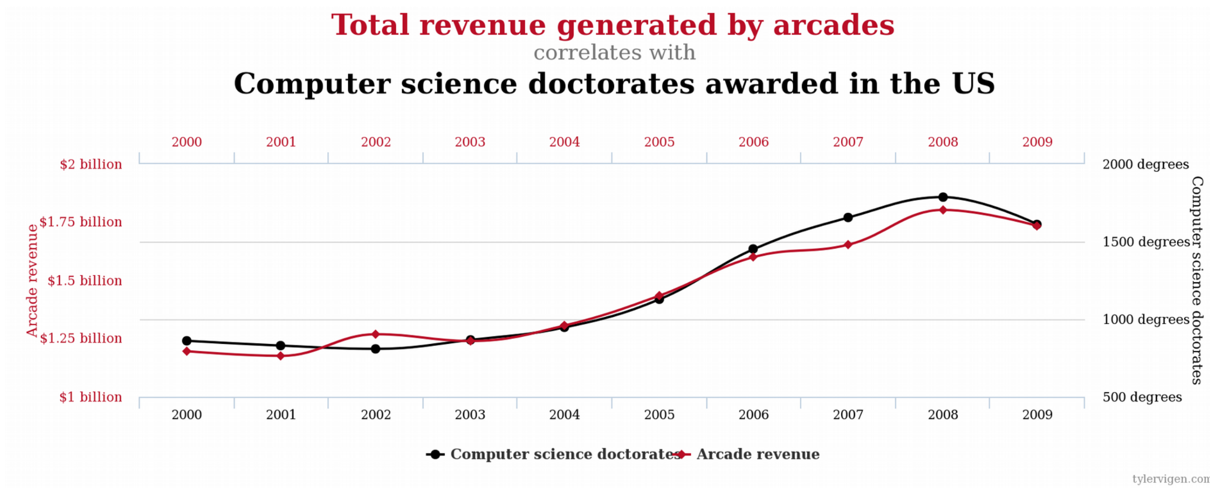


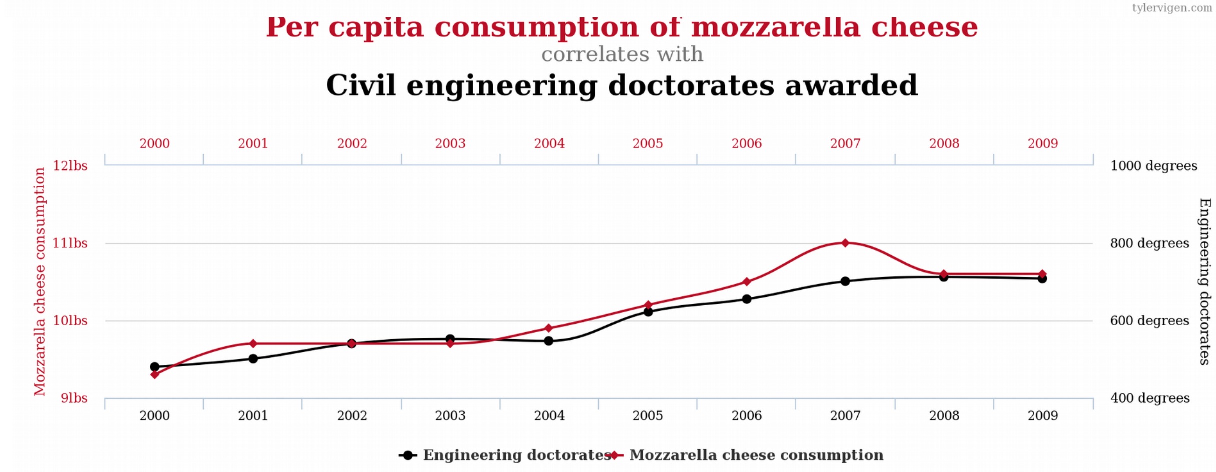
$$\rho(X, Y) = 1$$

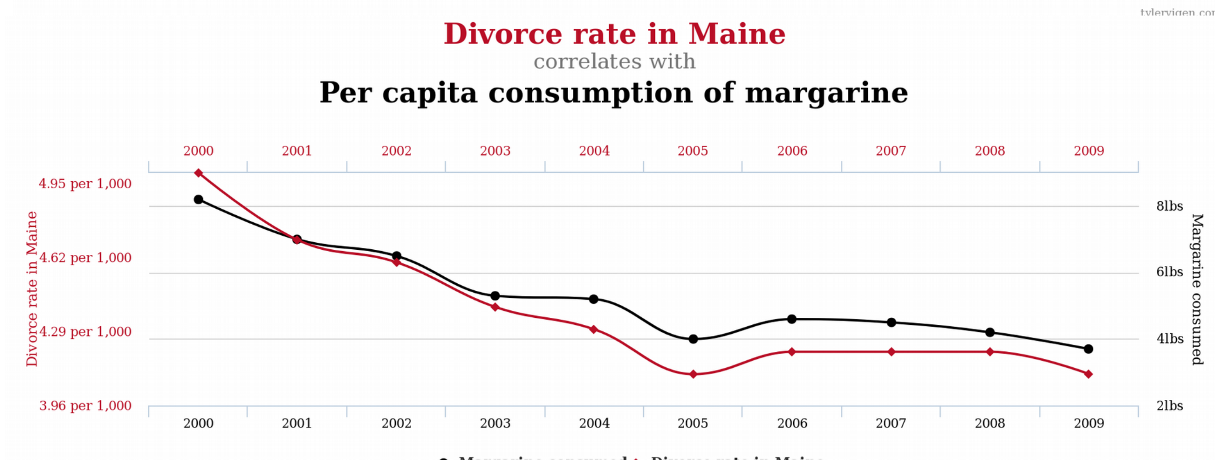


$$\rho(X, Y) = 0$$

Unimportant correlations







"Correlation does not imply causation"

Spurious Correlations
by Tyler Vigen