

# 13: Statistics on Multiple Random Variables

---

Jerry Cain  
April 29<sup>th</sup>, 2024

[Lecture Discussion on Ed](#)



# Coupon Collecting

# Coupon collecting and server requests

The **coupon collector's problem** in probability theory:

- You buy boxes of cereal.
- There are  $k$  different types of coupons
- For each box you buy, you "collect" a coupon of type  $i$ .

1. How many coupons do you expect after buying  $n$  boxes of cereal?



What is the expected number of servers utilized after  $n$  requests?

Servers  
requests  
 $k$  servers  
request to  
server  $i$



- \* 52% of Amazon profits
- \*\* more profitable than Amazon's North America commerce operations

[source](#)

# Computer cluster utilization

$$E \left[ \sum_{i=1}^n X_i \right] = \sum_{i=1}^n E[X_i]$$

Consider a computer cluster with  $k$  servers. We send  $n$  requests.

- Requests independently go to server  $i$  with probability  $p_i$
- Let  $X = \#$  servers that receive  $\geq 1$  request.

What is  $E[X]$ ?



# Computer cluster utilization

$$E \left[ \sum_{i=1}^n X_i \right] = \sum_{i=1}^n E[X_i]$$

Consider a computer cluster with  $k$  servers. We send  $n$  requests.

- Requests independently go to server  $i$  with probability  $p_i$
- Let  $X = \#$  servers that receive  $\geq 1$  request.

What is  $E[X]$ ?

1. Define additional random variables.

2. Solve.

Let:  $A_i =$  event that server  $i$   
receives  $\geq 1$  request  
 $X_i =$  indicator for  $A_i$

$$\begin{aligned} P(A_i) &= 1 - P(\text{no requests to } i) \\ &= 1 - (1 - p_i)^n \end{aligned}$$

$$\begin{aligned} E[X_i] &= P(A_i) = 1 - (1 - p_i)^n \\ E[X] &= E \left[ \sum_{i=1}^k X_i \right] = \sum_{i=1}^k E[X_i] = \sum_{i=1}^k (1 - (1 - p_i)^n) \\ &= \sum_{i=1}^k 1 - \sum_{i=1}^k (1 - p_i)^n = k - \sum_{i=1}^k (1 - p_i)^n \end{aligned}$$

Note:  $A_i$  are dependent!

# Coupon collecting problems: Hash tables

The **coupon collector's problem** in probability theory:

- You buy boxes of cereal.
- There are  $k$  different types of coupons
- For each box you buy, you "collect" a coupon of type  $i$ .

1. How many coupons do you expect after buying  $n$  boxes of cereal?



What is the expected number of utilized servers after  $n$  requests?

2. How many boxes do you expect to buy until you have one of each coupon?



What is the expected number of strings to hash until each bucket has  $\geq 1$  string?

<u>Servers</u>	<u>Hash Tables</u>
requests	strings
$k$ servers	$k$ buckets
request to server $i$	hashed to bucket $i$

# Hash Tables

$$E \left[ \sum_{i=1}^n X_i \right] = \sum_{i=1}^n E[X_i]$$

Consider a hash table with  $k$  buckets.

- Strings are equally likely to get hashed into any bucket (independently).
- Let  $Y = \#$  strings to hash until each bucket  $\geq 1$  string.

What is  $E[Y]$ ?

1. Define additional random variables.

How should we define  $Y_i$  such that  $Y = \sum_i Y_i$ ?

2. Solve.

# Hash Tables

$$E \left[ \sum_{i=1}^n X_i \right] = \sum_{i=1}^n E[X_i]$$

Consider a hash table with  $k$  buckets.

- Strings are equally likely to get hashed into any bucket (independently).
- Let  $Y = \#$  strings to hash until each bucket  $\geq 1$  string.

What is  $E[Y]$ ?

1. Define additional random variables.

Let:  $Y_i = \#$  of trials needed to get success after  $i$ -th success

- Success: hash string to previously empty bucket
- If  $i$  non-empty buckets:  $P(\text{success}) = \frac{k-i}{k}$

2. Solve.

$$P(Y_i = n) = \left(\frac{i}{k}\right)^{n-1} \left(\frac{k-i}{k}\right)$$

$$\text{Equivalently, } Y_i \sim \text{Geo} \left( p = \frac{k-i}{k} \right) \quad E[Y_i] = \frac{1}{p} = \frac{k}{k-i}$$



# Hash Tables

$$E \left[ \sum_{i=1}^n X_i \right] = \sum_{i=1}^n E[X_i]$$

Consider a hash table with  $k$  buckets.

- Strings are equally likely to get hashed into any bucket (independently).
- Let  $Y = \#$  strings to hash until each bucket  $\geq 1$  string.

What is  $E[Y]$ ?

1. Define additional random variables. Let:  $Y_i = \#$  of trials to needed get success after  $i$ -th success

$$Y_i \sim \text{Geo} \left( p = \frac{k-i}{k} \right), \quad E[Y_i] = \frac{1}{p} = \frac{k}{k-i}$$

2. Solve.  $Y = Y_0 + Y_1 + \dots + Y_{k-1}$

$$E[Y] = E[Y_0] + E[Y_1] + \dots + E[Y_{k-1}]$$

$$= \frac{k}{k} + \frac{k}{k-1} + \frac{k}{k-2} + \dots + \frac{k}{1} = k \left[ \frac{1}{k} + \frac{1}{k-1} + \dots + 1 \right] = O(k \log k)$$



# Covariance

## Statistics of sums of RVs

---

For any random variables  $X$  and  $Y$ ,

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

$$\text{Var}(X + Y) = ?$$

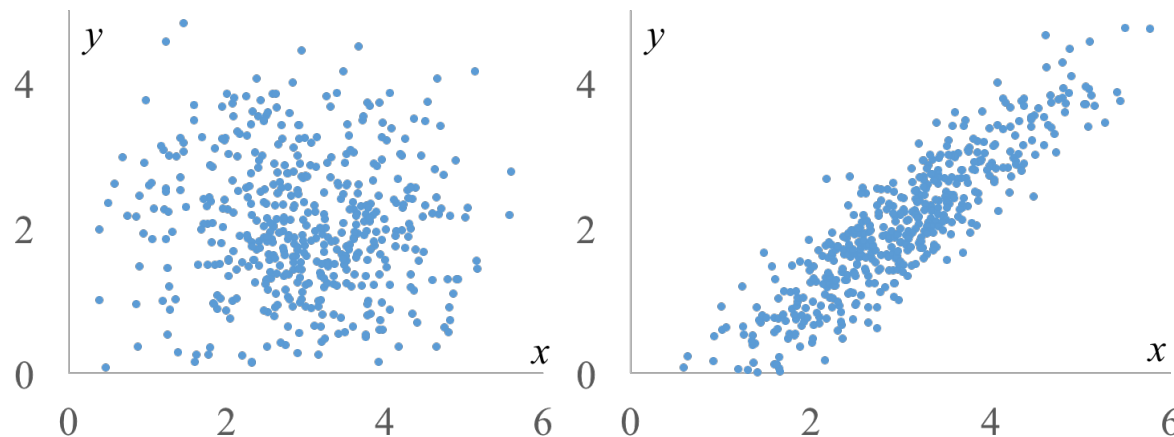
But first, a new statistic!

# Spot the difference

Compare/contrast the following two distributions:

Assume all points are equally likely.

$$P(X = x, Y = y) = \frac{1}{N}$$



Both distributions have the same  $E[X]$ ,  $E[Y]$ ,  $\text{Var}(X)$ , and  $\text{Var}(Y)$

Difference: how the two variables vary with each other.

# Covariance

The **covariance** of two variables  $X$  and  $Y$  is:

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

Proof of second part (rewriting  $E[X]$ ,  $E[Y]$  as  $\mu_X$ ,  $\mu_Y$  to emphasize that they're each constants):

$$\begin{aligned}\text{Cov}(X, Y) &= E[(X - E[X])(Y - E[Y])] = E[(X - \mu_X)(Y - \mu_Y)] \\ &= E[XY - \mu_Y X - \mu_X Y + \mu_X \mu_Y] \\ &= E[XY] - E[\mu_Y X] - E[\mu_X Y] + E[\mu_X \mu_Y] \\ &= E[XY] - \mu_X \mu_Y - \mu_X \mu_Y + \mu_X \mu_Y \\ &= E[XY] - \mu_X \mu_Y = E[XY] - E[X]E[Y]\end{aligned}$$

(linearity of expectation)

( $\mu_X$ ,  $\mu_Y$  are constants)

# Covariance

---

The **covariance** of two variables  $X$  and  $Y$  is:

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

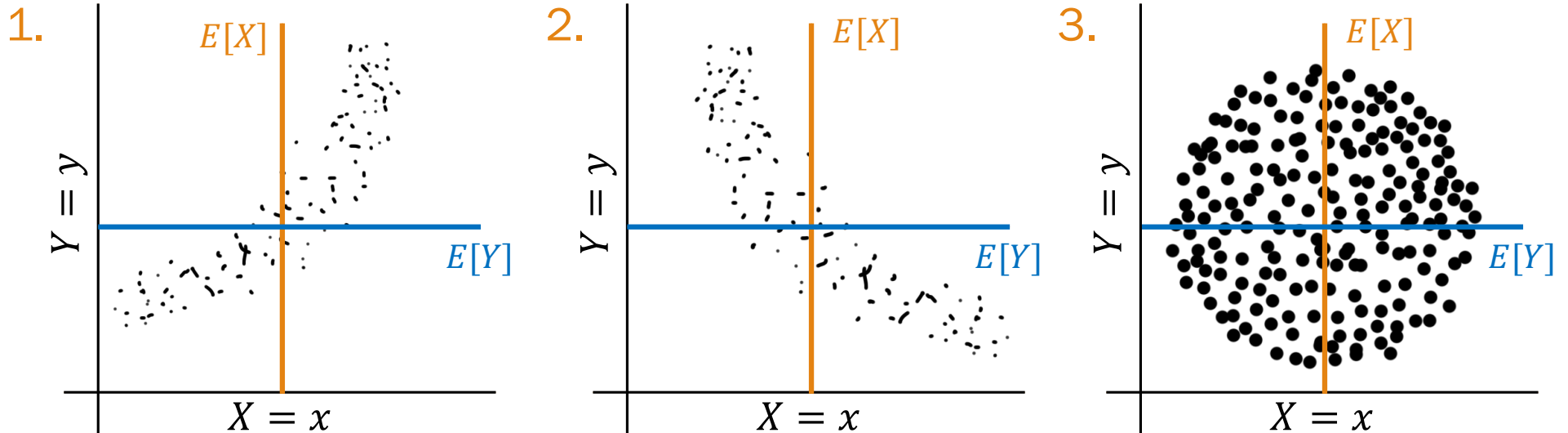
**Covariance** measures how one random variable varies with a second.

- Outside temperature and utility bills have a **negative** covariance.
- Handedness and musical ability have near **zero** covariance.
- Product demand and price have a **positive** covariance.

# Feel the covariance

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

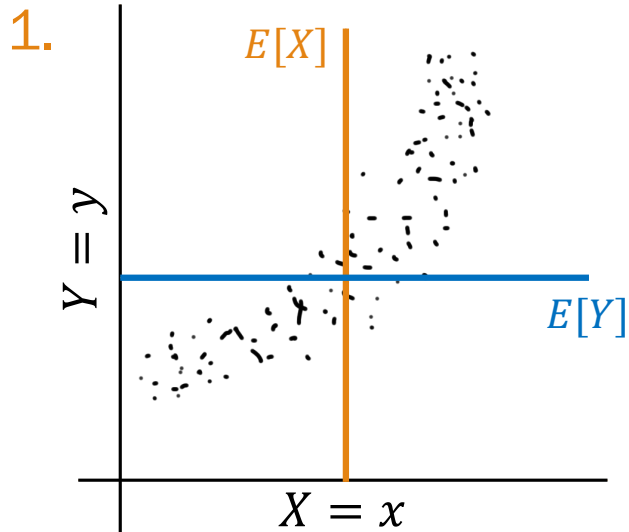
Is the covariance positive, negative, or zero?



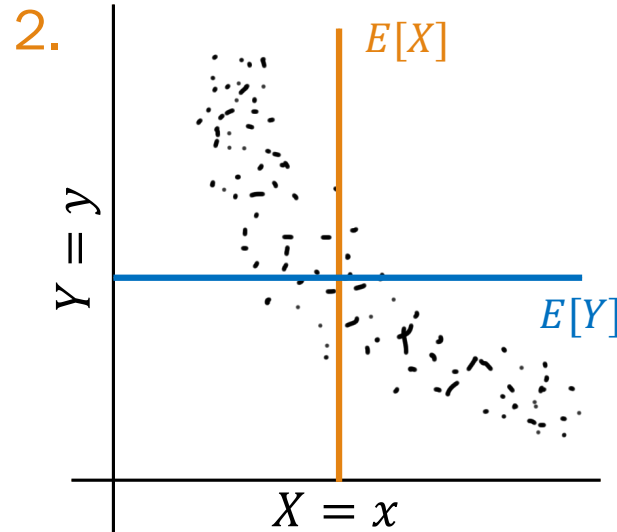
# Feel the covariance

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

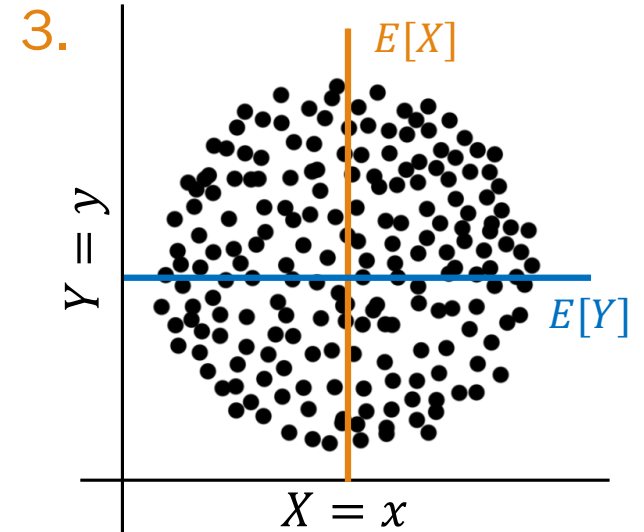
Is the covariance positive, negative, or zero?



positive



negative



zero



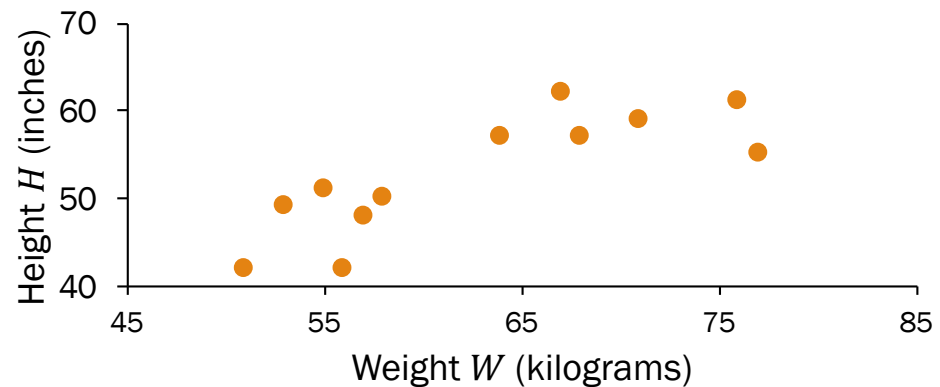
# Covarying humans

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

Weight (kg)	Height (in)	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

What is the covariance of weight  $W$  and height  $H$ ?

$$\begin{aligned} \text{Cov}(W, H) &= E[WH] - E[W]E[H] \\ &= 3355.83 - (62.75)(52.75) \\ &\text{(positive)} = 45.77 \end{aligned}$$



Covariance > 0: one variable ↑, other variable ↑

$$\begin{aligned} E[W] &= 62.75 \\ E[H] &= 52.75 \\ E[WH] &= 3355.83 \end{aligned}$$

# Properties of Covariance

---

The covariance of two variables  $X$  and  $Y$  is:

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

Properties:

1.  $\text{Cov}(X, Y) = \text{Cov}(Y, X)$
2.  $\text{Var}(X) = E[X^2] - (E[X])^2 = E[XX] - E[X]E[X] = \text{Cov}(X, X)$
3. Covariance of sums = sum of all pairwise covariances (proof left to you)  
 $\text{Cov}(X_1 + X_2, Y_1 + Y_2) = \text{Cov}(X_1, Y_1) + \text{Cov}(X_2, Y_1) + \text{Cov}(X_1, Y_2) + \text{Cov}(X_2, Y_2)$
4. Covariance under linear transformation:  $\text{Cov}(aX + b, Y) = a\text{Cov}(X, Y)$

# Zero covariance does not imply independence

---

Let  $X$  take on values  $\{-1,0,1\}$   
with equal probability  $1/3$ .

Define  $Y = \begin{cases} 1 & \text{if } X = 0 \\ 0 & \text{otherwise} \end{cases}$

What is the joint PMF of  $X$  and  $Y$ ?

# Zero covariance does not imply independence

Let  $X$  take on values  $\{-1, 0, 1\}$  with equal probability  $1/3$ .

Define  $Y = \begin{cases} 1 & \text{if } X = 0 \\ 0 & \text{otherwise} \end{cases}$

		X			
		-1	0	1	
Y	0	1/3	0	1/3	2/3
	1	0	1/3	0	1/3
		1/3	1/3	1/3	

Marginal PMF of  $Y$ ,  $p_Y(y)$

Marginal PMF of  $X$ ,  $p_X(x)$

1.  $E[X] =$

$E[Y] =$

2.  $E[XY] =$

3.  $\text{Cov}(X, Y) =$

4. Are  $X$  and  $Y$  independent?



# Zero covariance does not imply independence

Let  $X$  take on values  $\{-1, 0, 1\}$  with equal probability  $1/3$ .

Define  $Y = \begin{cases} 1 & \text{if } X = 0 \\ 0 & \text{otherwise} \end{cases}$

		X				
		-1	0	1		
Y	0	1/3	0	1/3	2/3	Marginal PMF of $Y$ , $p_Y(y)$
	1	0	1/3	0	1/3	
		1/3	1/3	1/3	Marginal PMF of $X$ , $p_X(x)$	

$$1. \quad E[X] = -1\left(\frac{1}{3}\right) + 0\left(\frac{1}{3}\right) + 1\left(\frac{1}{3}\right) = 0 \quad E[Y] = 0\left(\frac{2}{3}\right) + 1\left(\frac{1}{3}\right) = 1/3$$

$$2. \quad E[XY] = (-1 \cdot 0)\left(\frac{1}{3}\right) + (0 \cdot 1)\left(\frac{1}{3}\right) + (1 \cdot 0)\left(\frac{1}{3}\right) = 0$$

$$3. \quad \text{Cov}(X, Y) = E[XY] - E[X]E[Y] = 0 - 0(1/3) = 0 \quad \text{! does not imply independence!}$$

$$4. \quad \text{Are } X \text{ and } Y \text{ independent? } \times$$

$$P(Y = 0 | X = 1) = 1 \neq P(Y = 0) = 2/3$$



# Variance of sums of RVs

## Statistics of sums of RVs

---

For any random variables  $X$  and  $Y$ ,

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

$$\text{Var}(X + Y) = \text{Var}(X) + 2 \cdot \text{Cov}(X, Y) + \text{Var}(Y)$$

# Variance of general sum of RVs

For any random variables  $X$  and  $Y$ ,

$$\text{Var}(X + Y) = \text{Var}(X) + 2 \cdot \text{Cov}(X, Y) + \text{Var}(Y)$$

Proof:

$$\text{Var}(X + Y) = \text{Cov}(X + Y, X + Y)$$

$$= \text{Cov}(X, X) + \text{Cov}(X, Y) + \text{Cov}(Y, X) + \text{Cov}(Y, Y)$$

$$= \text{Var}(X) + 2 \cdot \text{Cov}(X, Y) + \text{Var}(Y)$$

$$\text{Var}(X) = \text{Cov}(X, X)$$

covariance of  
all pairs

Symmetry of covariance +  
 $\text{Cov}(X, X) = \text{Var}(X)$

More generally:

$$\text{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \text{Var}(X_i) + 2 \sum_{i=1}^n \sum_{j=i+1}^n \text{Cov}(X_i, X_j) \quad (\text{proof in extra slides})$$



## Statistics of sums of RVs

---

For any random variables  $X$  and  $Y$ ,

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

$$\text{Var}(X + Y) = \text{Var}(X) + 2 \cdot \text{Cov}(X, Y) + \text{Var}(Y)$$

For **independent**  $X$  and  $Y$ ,

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

(Lemma: proof in extra slides)

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$$

# Variance of sum of independent RVs

---

For **independent**  $X$  and  $Y$ ,

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$$

Proof:

- $\text{Cov}(X, Y) = E[XY] - E[X]E[Y]$  def. of covariance  
 $= E[X]E[Y] - E[X]E[Y]$   $X$  and  $Y$  are **independent**  
 $= 0$
- $\text{Var}(X + Y) = \text{Var}(X) + 2 \cdot \text{Cov}(X, Y) + \text{Var}(Y)$   
 $= \text{Var}(X) + \text{Var}(Y)$

# Proving Variance of the Binomial

$$X \sim \text{Bin}(n, p) \quad \text{Var}(X) = np(1 - p)$$

Let  $X = \sum_{i=1}^n X_i$

Let  $X_i = i$ th trial is heads  
 $X_i \sim \text{Ber}(p)$   
 $\text{Var}(X_i) = p(1 - p)$

$X_i$  are **independent**  
(by definition)

$$\begin{aligned} \text{Var}(X) &= \text{Var}\left(\sum_{i=1}^n X_i\right) \\ &= \sum_{i=1}^n \text{Var}(X_i) \\ &= \sum_{i=1}^n p(1 - p) \\ &= np(1 - p) \end{aligned}$$

$X_i$  are **independent**,  
therefore variance of sum  
= sum of variance

Variance of Bernoulli





# Correlation

# Covarying humans

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

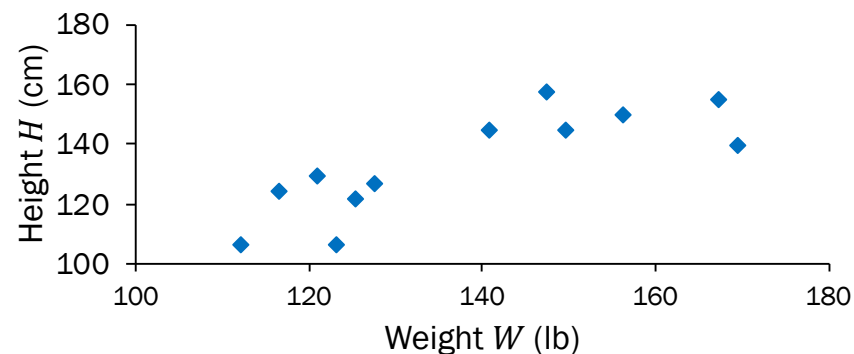
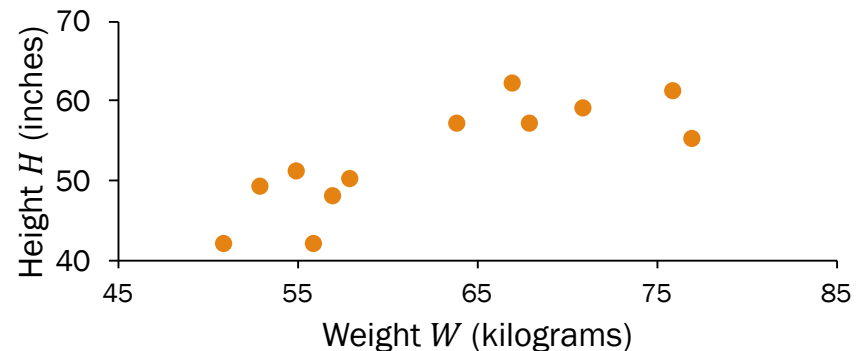
What is the covariance of weight  $W$  and height  $H$ ?

$$\begin{aligned}\text{Cov}(W, H) &= E[WH] - E[W]E[H] \\ &= 3355.83 - (62.75)(52.75) \\ &= 45.77 \text{ (positive)}\end{aligned}$$

What about weight (lb) and height (cm)?

$$\begin{aligned}\text{Cov}(2.20W, 2.54H) &= E[2.20W \cdot 2.54H] - E[2.20W]E[2.54H] \\ &= 18752.38 - (138.05)(133.99) \\ &= 255.06 \text{ (positive)}\end{aligned}$$

⚠ Covariance depends on units!



Sign of covariance (+/-) more meaningful than magnitude

# Correlation

The **correlation** of two variables  $X$  and  $Y$  is:

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

$$\begin{aligned}\sigma_X^2 &= \text{Var}(X), \\ \sigma_Y^2 &= \text{Var}(Y)\end{aligned}$$

- Note:  $-1 \leq \rho(X, Y) \leq 1$
- Correlation measures the **linear relationship** between  $X$  and  $Y$ :

$$\rho(X, Y) = 1 \quad \Rightarrow Y = aX + b, \text{ where } a = \sigma_Y / \sigma_X$$

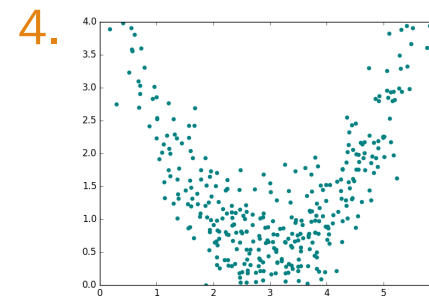
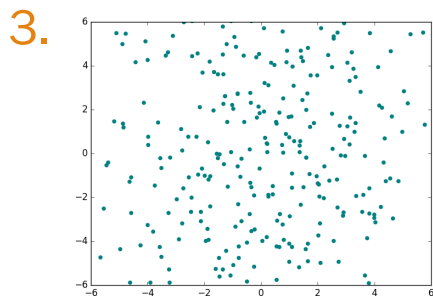
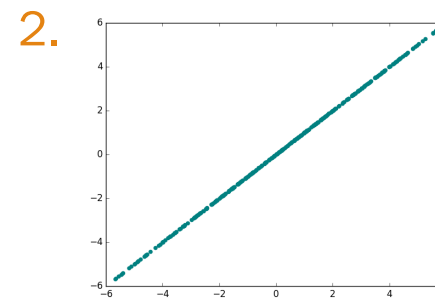
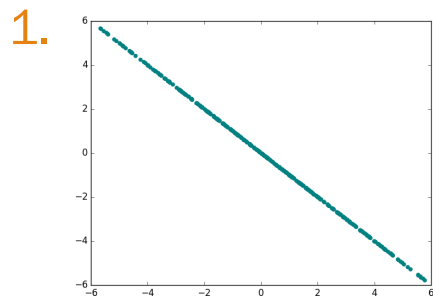
$$\rho(X, Y) = -1 \quad \Rightarrow Y = aX + b, \text{ where } a = -\sigma_Y / \sigma_X$$

$$\rho(X, Y) = 0 \quad \Rightarrow \text{uncorrelated (absence of linear relationship)}$$

# Correlation reps

What is the correlation coefficient  $\rho(X, Y)$ ?

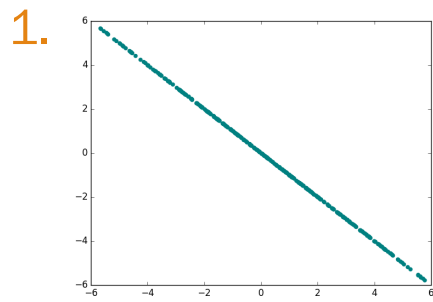
- A.  $\rho(X, Y) = 1$
- B.  $\rho(X, Y) = -1$
- C.  $\rho(X, Y) = 0$
- D. Other



# Correlation reps

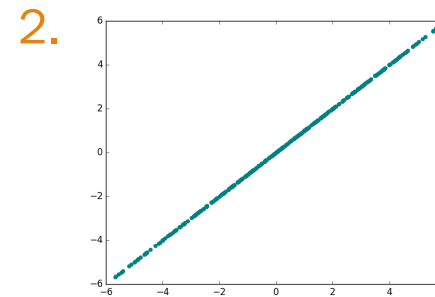
What is the correlation coefficient  $\rho(X, Y)$ ?

- A.  $\rho(X, Y) = 1$
- B.  $\rho(X, Y) = -1$
- C.  $\rho(X, Y) = 0$
- D. Other



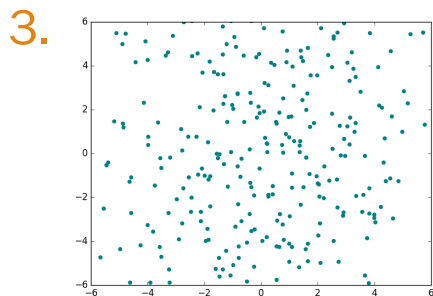
B.  $\rho(X, Y) = -1$

$$Y = -aX + b$$
$$a > 0$$



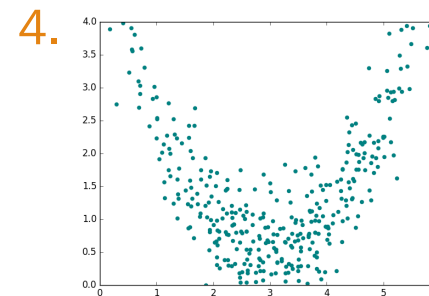
A.  $\rho(X, Y) = 1$

$$Y = aX + b$$
$$a > 0$$



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

“uncorrelated”



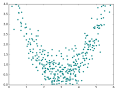
C.  $\rho(X, Y) = 0$

$$Y = X^2$$

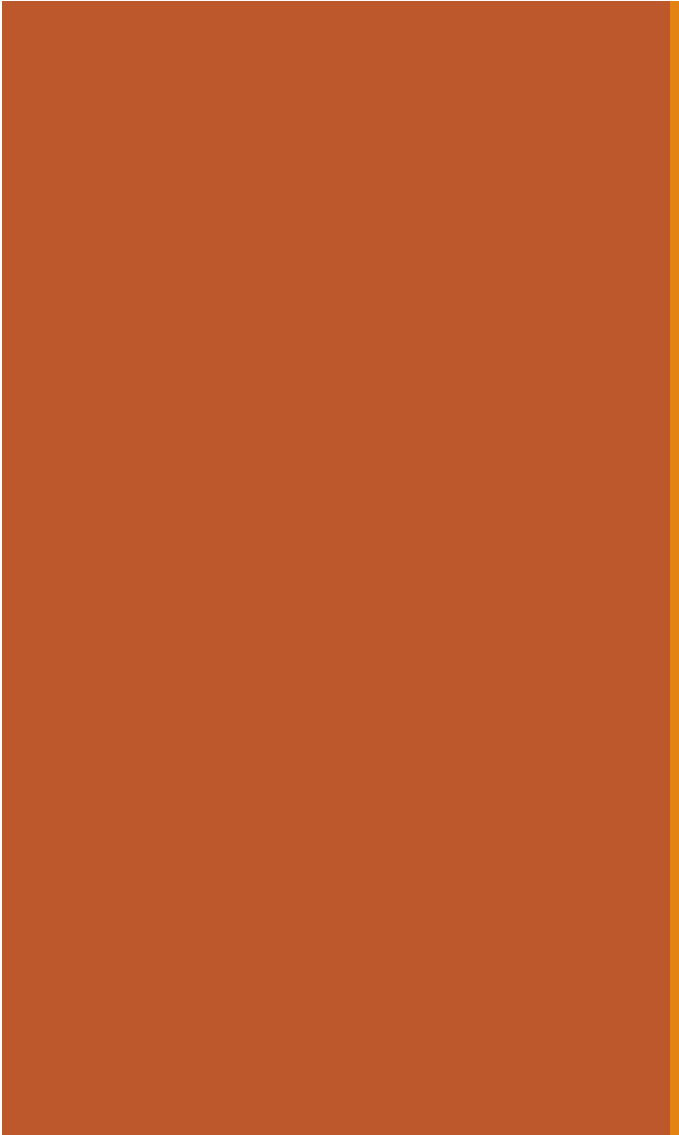
$X$  and  $Y$  can be nonlinearly related even if  $\rho(X, Y) = 0$ .



# Throwback to CS103: Conditional statements

Statement $P \rightarrow Q$ :	Independence $\rightarrow$ No correlation	✓
Contrapositive $\neg Q \rightarrow \neg P$ :	Correlation $\rightarrow$ Dependence	✓ (logically equivalent)
Inverse $\neg P \rightarrow \neg Q$ :	Dependence $\rightarrow$ Correlation	✗ (not always) $Y = X^2$ $\rho(X, Y) = 0$ 
Converse $Q \rightarrow P$ :	No correlation $\rightarrow$ Independence	✗ (not always)

“Correlation does not imply causation”



# Spurious Correlation

# Spurious Correlations

$\rho(X, Y)$  is used a lot to statistically quantify the relationship b/t X and Y.

Correlation:  
0.947091



# Spurious Correlations

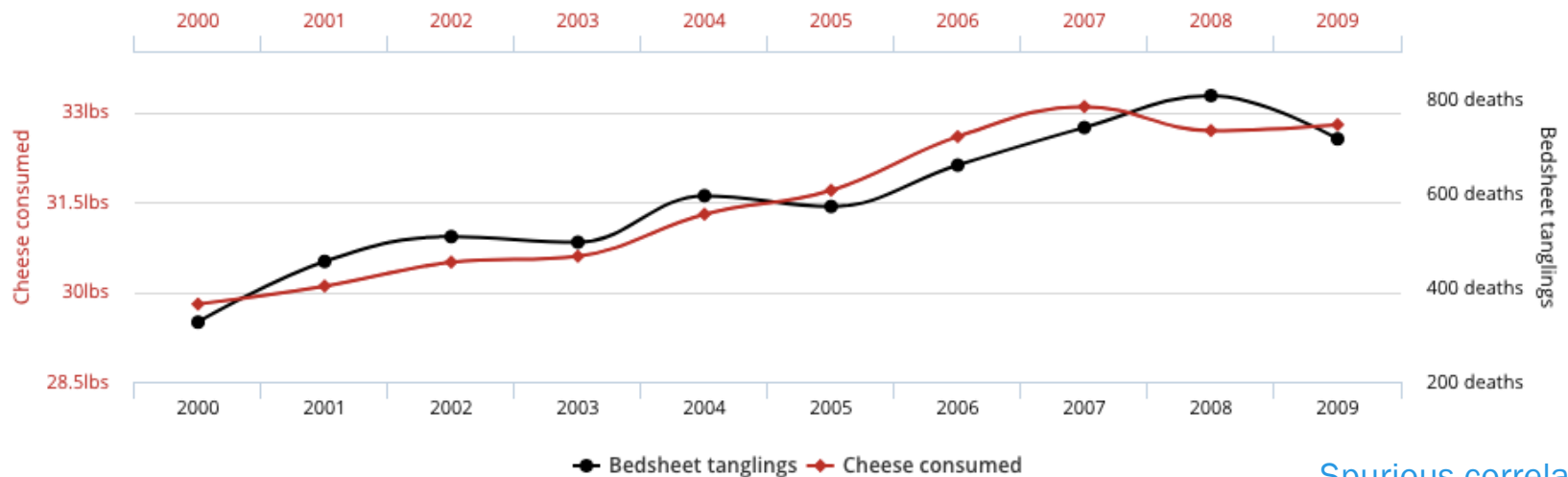
$\rho(X, Y)$  is used a lot to statistically quantify the relationship b/t X and Y.

Correlation:  
0.947091

**Per capita cheese consumption**

correlates with

**Number of people who died by becoming tangled in their bedsheets**

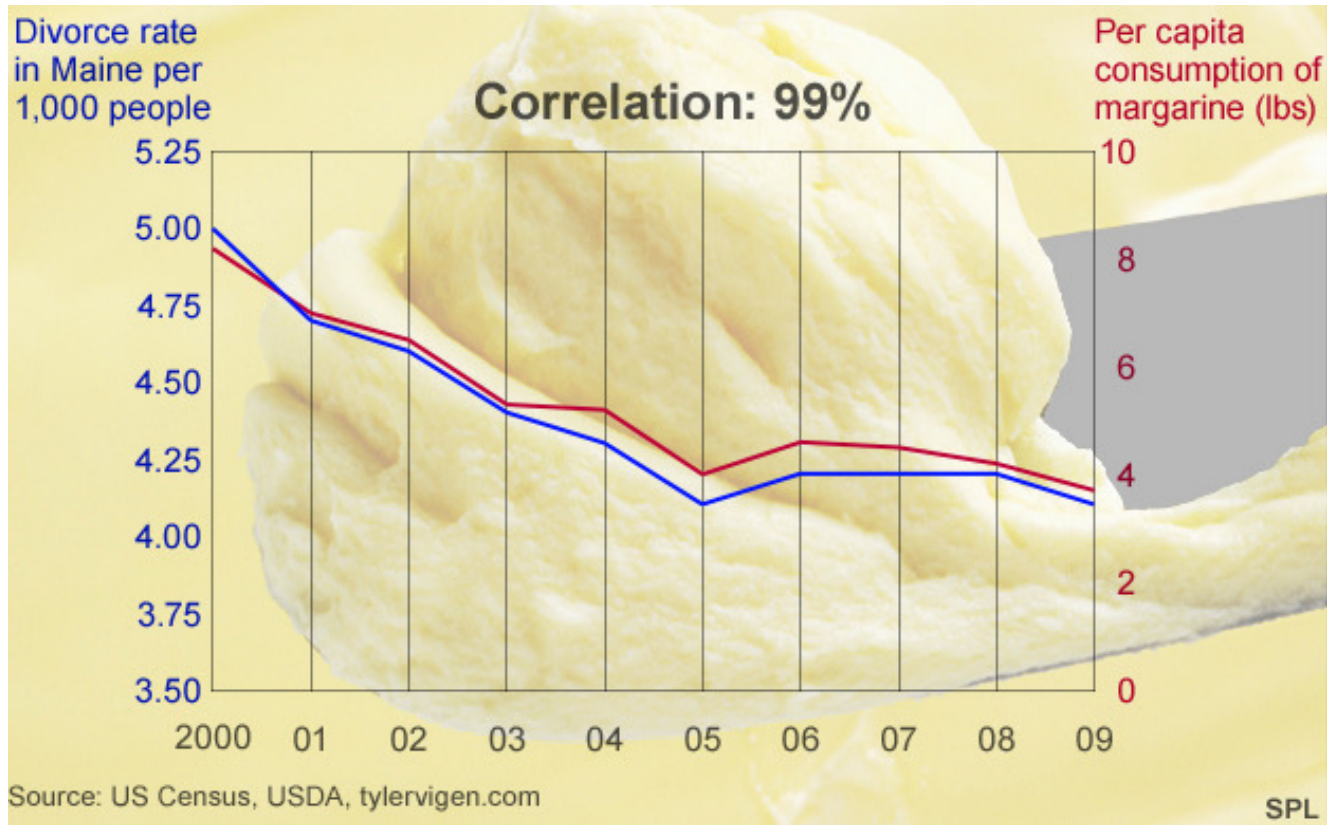


Lisa Yan, Chris Piech, Mehran Sahami, and Jerry Cain, CS109, Spring 2024

[Spurious correlations](#)

Stanford University 36

# Divorce vs. Margarine



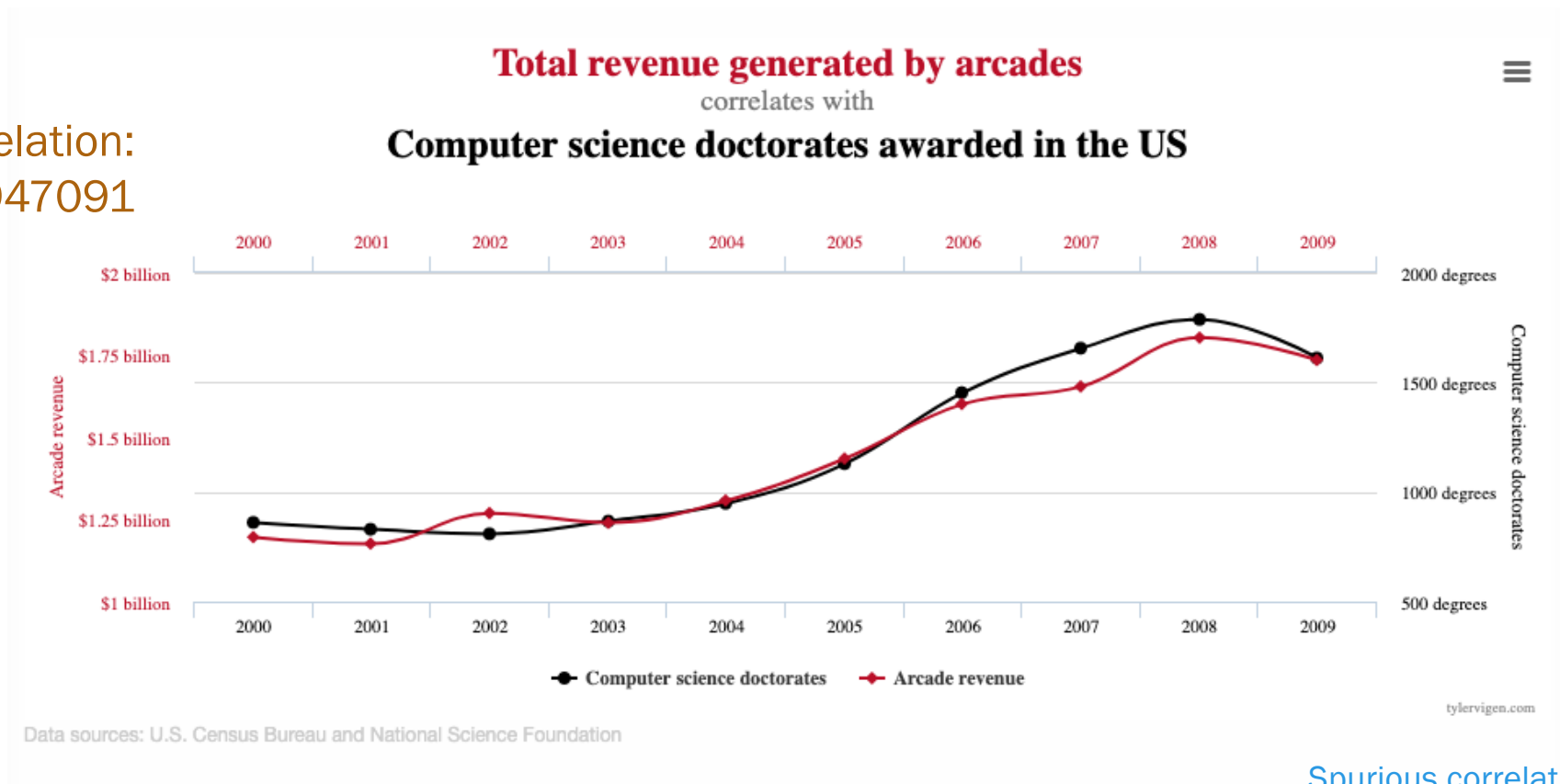
<http://www.bbc.com/news/magazine-27537142>

Lisa Yan, Chris Piech, Mehran Sahami, and Jerry Cain, CS109, Spring 2024

Stanford University

# Arcade revenue vs. CS PhDs

Correlation:  
0.947091





# Extras

# Expectation of product of independent RVs

If  $X$  and  $Y$  are  
**independent**, then

$$\begin{aligned} E[XY] &= E[X]E[Y] \\ E[g(X)h(Y)] &= E[g(X)]E[h(Y)] \end{aligned}$$

Proof:  $E[g(X)h(Y)] = \sum_y \sum_x g(x)h(y)p_{X,Y}(x,y)$  (for continuous proof, replace summations with integrals)

$$= \sum_y \sum_x g(x)h(y)p_X(x)p_Y(y)$$

$X$  and  $Y$  are independent

$$= \sum_y \left( h(y)p_Y(y) \sum_x g(x)p_X(x) \right)$$

Terms dependent on  $y$  are constant in integral of  $x$

$$= \left( \sum_x g(x)p_X(x) \right) \left( \sum_y h(y)p_Y(y) \right)$$

Summations separate

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

Lisa Yan, John's Heintz, Michael Sotham, and Jerry Cain, CS109, Spring 2024



# Variance of Sums of Variables

$$\text{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \text{Var}(X_i) + 2 \sum_{i=1}^n \sum_{j=i+1}^n \text{Cov}(X_i, X_j)$$

Proof:

$$\begin{aligned} \text{Var}\left(\sum_{i=1}^n X_i\right) & \stackrel{\text{Var}(X) = \text{Cov}(X, X)}{=} \text{Cov}\left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i\right) \stackrel{\text{covariance of all pairs}}{=} \sum_{i=1}^n \sum_{j=1}^n \text{Cov}(X_i, X_j) \\ & = \sum_{i=1}^n \text{Var}(X_i) + \sum_{i=1}^n \sum_{j=1, j \neq i}^n \text{Cov}(X_i, X_j) \\ & = \sum_{i=1}^n \text{Var}(X_i) + 2 \sum_{i=1}^n \sum_{j=i+1}^n \text{Cov}(X_i, X_j) \end{aligned}$$

Symmetry of covariance  
 $\text{Cov}(X, X) = \text{Var}(X)$

Adjust summation bounds