LECTURE 17

Joint Distributions; Covariance and Correlation

CSCI 3022 @ CU Boulder

Maribeth Oscamou

Announcements

No HW due next week

Exam 1 next Friday 10/13 in class

- Scope: Lessons 1-16; HW 1-6; nb 1-6
- You will be provided with the page of Python functions from Quiz 4
- Bring:
  - Buff One Card (required for submitting your exam)
  - Calculator
  - Pencil
  - Optional: Handwritten Crib sheet (can be TWO sided 8.5”x11”)

Exam 1 Review Problems:

- Posted on Canvas Modules.

Review Day in Class next Wednesday:

- Post your questions in this Google doc (available on Piazza) by noon on Wedneday

Today’s Roadmap

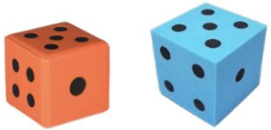
Joint Distributions

Covariance and Correlation

Independence

Joint probability mass functions

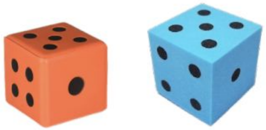
Roll two 6-sided dice, yielding values X and Y.



X random variable	$P(X = 1)$ probability of an event	$P(X = k)$ probability mass function
X, Y random variables	$P(X = 1 \cap Y = 6)$ $P(X = 1, Y = 6)$ new notation: the comma probability of the intersection of two events	$P(X = a, Y = b)$ joint probability mass function

Two dice

Roll two 6-sided dice, yielding values X and Y.



1. What is the joint PMF of X and Y?

$$p_{X,Y}(a,b) = 1/36 \qquad (a,b) \in \{(1,1), \dots, (6,6)\}$$

		X					
		1	2	3	4	5	6
Y	1	1/36	…	…	…	…	1/36
	2	…	…	…	…		
	3	…	…	…	…	…	…
	4	…	…	…	…	…	…
	5	…	…	…	…	…	…
	6	1/36	…	…	…	…	1/36

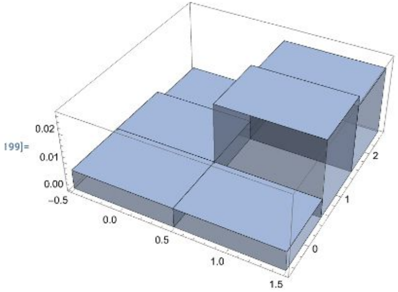
$P(X = 4, Y = 2)$

Probability table
- All possible outcomes for several discrete RVs

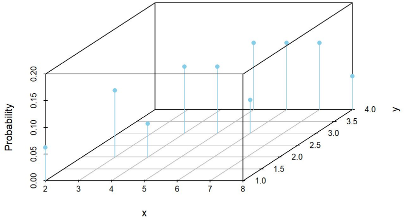
Discrete joint distributions

For two discrete joint random variables X and Y, the joint probability mass function is defined as:

$$p_{X,Y}(a,b) = P(X = a, Y = b)$$



The marginal distributions of the joint PMF are defined as:



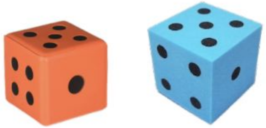
$$p_X(a) = P(X = a) = \sum_y p_{X,Y}(a,y)$$

$$p_Y(b) = P(Y = b) = \sum_x p_{X,Y}(x,b)$$

Use marginal distributions to get a 1-D RV from a joint PMF.

Two dice

Roll two 6-sided dice, yielding values X and Y.



1. What is the joint PMF of X and Y?

$$p_{X,Y}(a,b) = 1/36 \qquad (a,b) \in \{(1,1), \dots, (6,6)\}$$

		X					
		1	2	3	4	5	6
Y	1	1/36	…	…	…	…	1/36
	2	…	…	…	…	…	…
	3	…	…	…	…	…	…
	4	…	…	…	…	…	…
	5	…	…	…	…	…	…
	6	1/36	…	…	…	…	1/36

$P(X = 4, Y = 2)$

If all (X,Y) have equal probability, we can visualize a joint distribution using a scatter plot of the values of (X,Y):

Example:

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?

Example:

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?

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

(X,Y) Scatter Plot:

1. $E[X] =$ $\qquad$ $E[Y] =$

2. $E[XY] =$

Example:

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?

		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. $E[X] =$ $E[Y] =$

$$-1\left(\frac{1}{3}\right) + 0\left(\frac{1}{3}\right) + 1\left(\frac{1}{3}\right) = 0$$

$$0\left(\frac{2}{3}\right) + 1\left(\frac{1}{3}\right) = 1/3$$

2. $E[XY] =$

Example:

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?

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

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

Covariance & Correlation

Joint Distributions:

Covariance and Correlation

Independence

What do the following headlines mean mathematically?

Correlation Found Between Poor School Attendance and Poor Air Quality Days

News Published: October 12, 2020 | Original story from The University of Utah

MONTREAL | News

Correlation between teens' social media time and restrictive eating: Montreal study

Kelly Greig CTV News Montreal Videojournalist
@KellyGreig | Contact

Published Wednesday, January 25, 2023 10:26AM EST
Last Updated Wednesday, January 25, 2023 1:41PM EST

U.S.News 90TH NEWS » News Best Countries Best States Healthiest Communities Opinion Elections Racial Equality in America Photos

Home / News / Newsgram / Study Suggests Correlation Be...

Study Suggests Correlation Between Heart Health and Optimism

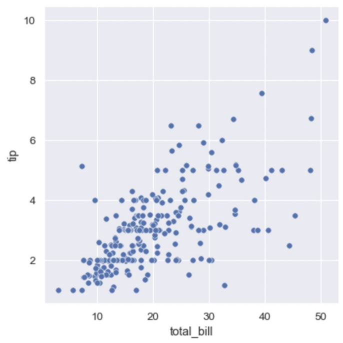
Smile. It'll make your heart happy.

By Andrew Soergel | Jan. 12, 2015, at 2:15 p.m.

Expectation and Covariance

In real life, we often have many RVs interacting at once.

- We've seen some simpler cases
- Computing joint PMFs in general is hard!
- But often you don't need to model joint RVs completely.



Instead, we'll focus next on reporting *statistics* of multiple RVs:

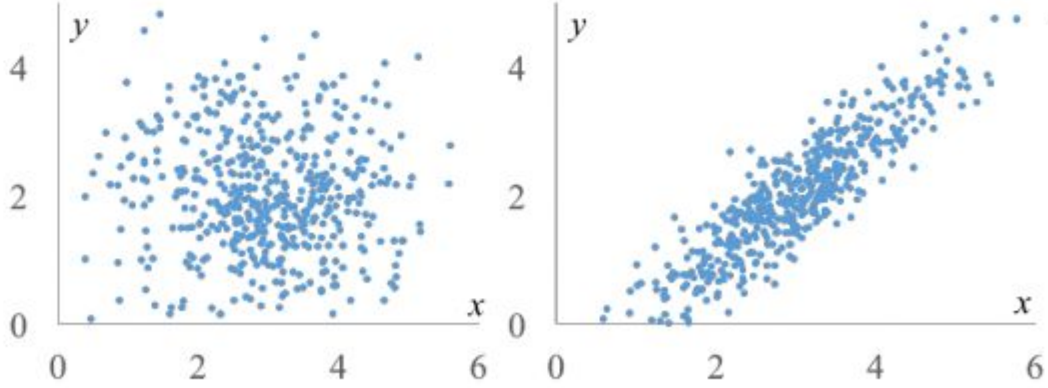
- Expectation of sums (you've seen some of this)
- Covariance: a measure of how two RVs vary with *each other*

Spot the difference

Compare/contrast the following two distributions:

Assume all points are equally likely.

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



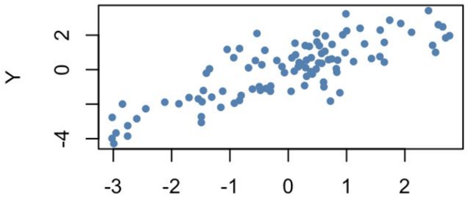
Both distributions have the same $E[X]$, $E[Y]$, Var(X), and Var(Y)

Covariance: The Basics

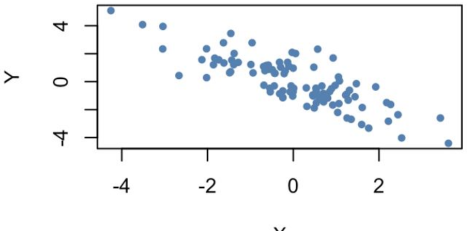
Covariance of X and Y is the expected product of deviations from expectation.

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

- Covariance measures a linear relationship between X and Y.
- A generalization of variance. Note $\text{Cov}(X,X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \text{Var}(X)$.



Positive covariance indicates that one variable increases linearly with the other.



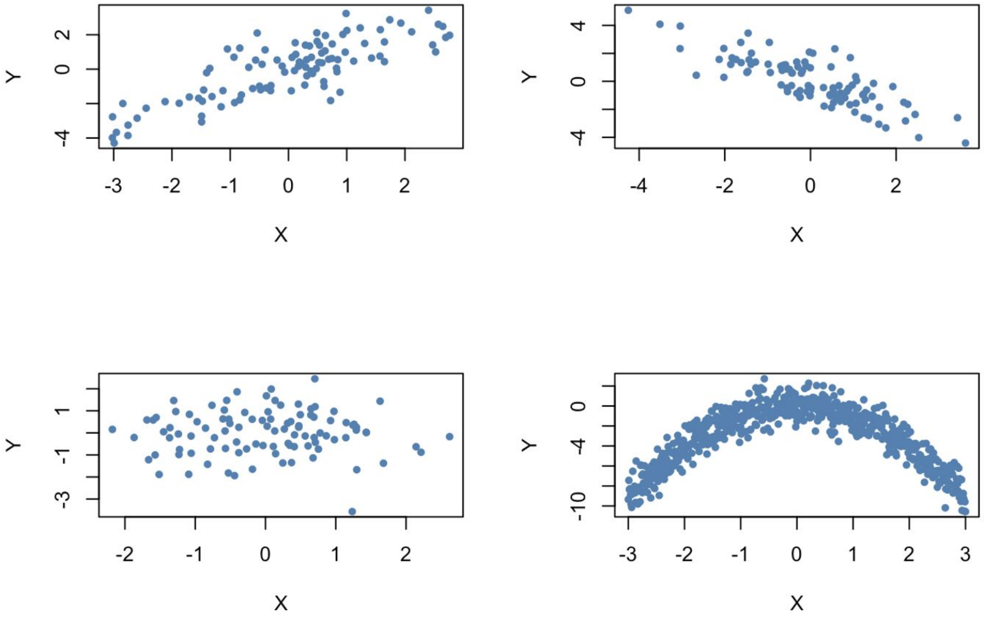
Negative covariance indicates that one variable decreases linearly with the other.

Covariance

Feel the covariance

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

Is the covariance positive, negative, or zero?

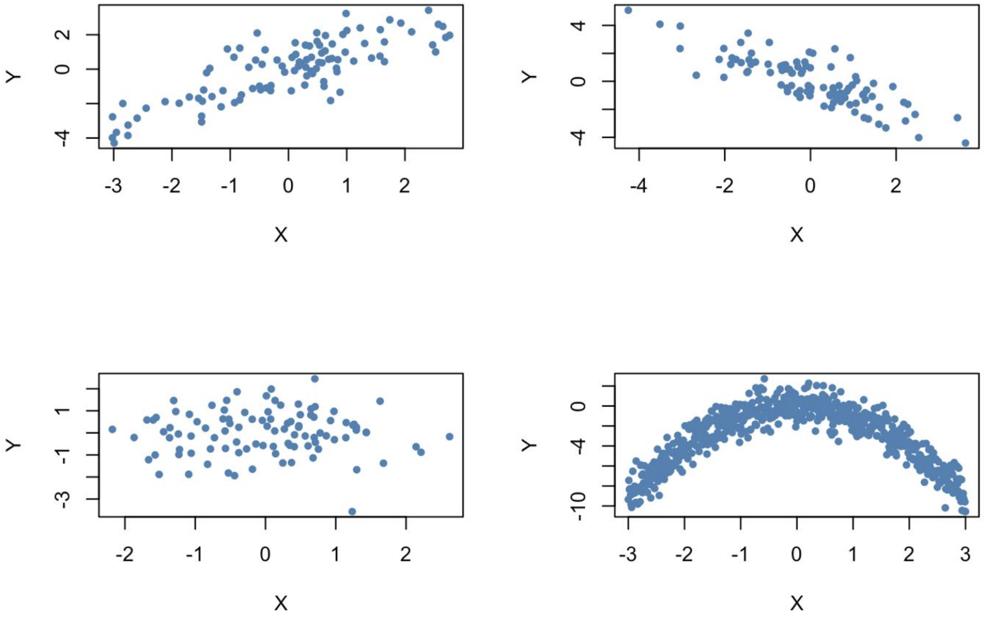


Covariance

Feel the covariance

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

Is the covariance positive, negative, or zero?



$\text{Cov}(X, Y) = 0$ does NOT imply independence of X and Y!

Announcements

Exam 1 Friday 10/13 in class
- Scope: Lessons 1-16; HW 1-6; nb 1-6
- You will be provided with the page of Python functions from Quiz 4
- Bring:
  - Buff One Card (required for submitting your exam)
  - Calculator
  - Pencil
  - Optional: Handwritten Crib sheet (can be TWO sided 8.5”x11”)

Exam 1 Review Problems:
- Posted on Canvas Modules.

Review Day in Class Wednesday:
- Post your questions in this Google doc (available on Piazza) by noon on Wednesday

Day 2: Remainder of Lec17

Covariance and Correlation

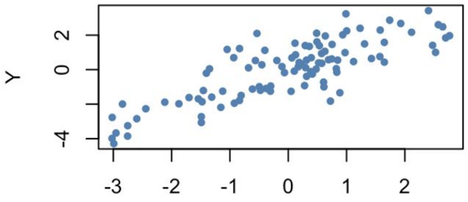
Independence

Covariance: The Basics

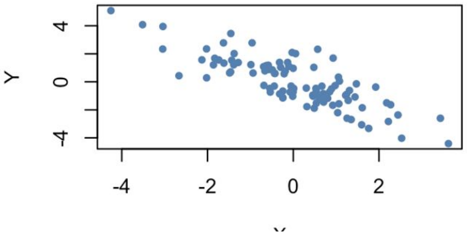
Covariance of X and Y is the expected product of deviations from expectation.

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

- Covariance measures a linear relationship between X and Y.
- A generalization of variance. Note $\text{Cov}(X, X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \text{Var}(X)$.



Positive covariance indicates that one variable increases linearly with the other.



Negative covariance indicates that one variable decreases linearly with the other.

Example:

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

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

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

4. Are X and Y independent?
   $P(Y = 0|X = 1) = 1$
   $\neq \quad P(Y = 0) = 2/3$

Covarying humans

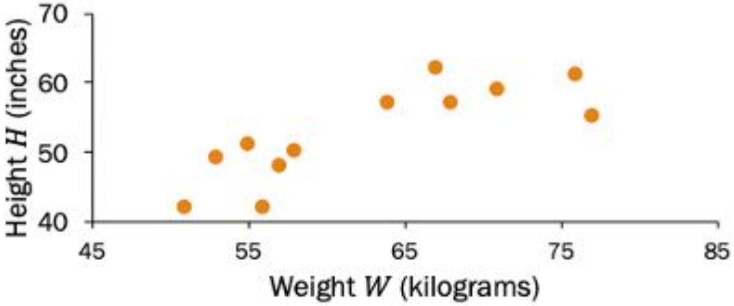
$$\mathrm{Cov}(X,Y) = E[(X - E[X])(Y - E[Y])]$$
$$= E[XY] - E[X]E[Y]$$

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

$E[W] = 62.75$ $E[H] = 52.75$ $E[WH] = 3355.83$

What is the covariance of weight W and height H?

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



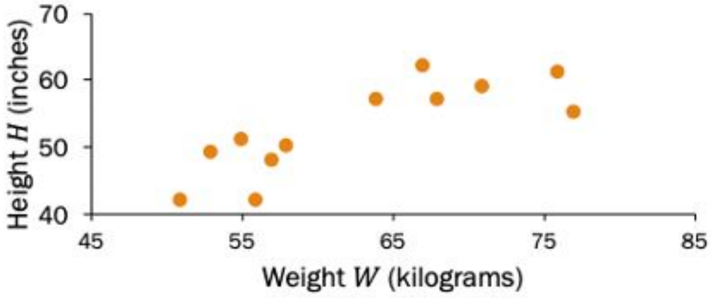
Covariance > 0: one variable ↑, other variable ↑

Covarying humans

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

What is the covariance of weight W and height H?

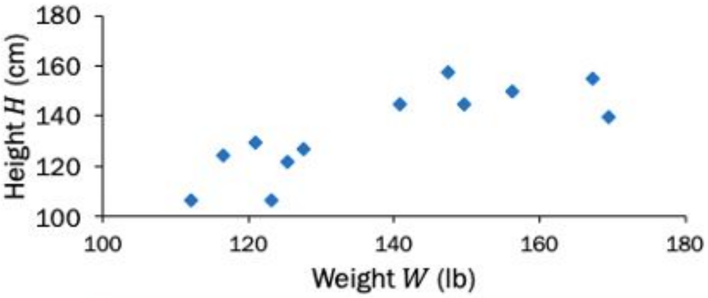
$$\begin{aligned}\text{Cov}(W,H) &= E[WH] - E[W]E[H] \\ &= 3355.83 - (62.75)(52.75) \\ &= 45.77 \quad \text{(positive)}\end{aligned}$$



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

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

⚠ Covariance depends on units!



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

Correlation

- We can interpret covariance by defining **correlation** (yes, *that* correlation!):

$$\rho(X,Y) = \mathbb{E}\left[\underbrace{\left(\frac{X - \mathbb{E}[X]}{\mathrm{SD}(X)}\right)}_{\textbf{standard units} \text{ of } X}\left(\frac{Y - \mathbb{E}[Y]}{\mathrm{SD}(Y)}\right)\right] = \frac{\mathrm{Cov}(X,Y)}{\mathrm{SD}(X)\mathrm{SD}(Y)}$$

Correlation (and therefore covariance) measures a linear relationship between X and Y.

Covariance and Correlation

Covariance is the expected product of deviations from expectation.

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

- A generalization of variance. Note $\mathrm{Cov}(X, X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathrm{Var}(X)$.
- Interpret by defining **correlation** (yes, *that* correlation!):

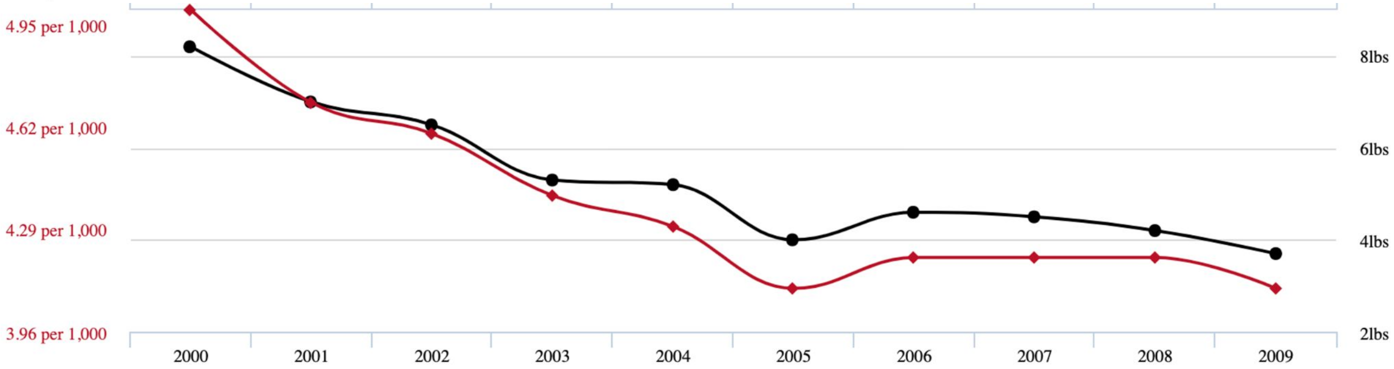
$$\rho(X, Y) = \mathbb{E}\left[\underbrace{\left(\frac{X - \mathbb{E}[X]}{\mathrm{SD}(X)}\right)}_{\text{standard units of } X \text{ (link)}}\left(\frac{Y - \mathbb{E}[Y]}{\mathrm{SD}(Y)}\right)\right] = \frac{\mathrm{Cov}(X, Y)}{\mathrm{SD}(X)\mathrm{SD}(Y)}$$

Correlation (and therefore covariance) measures a **linear relationship** between X and Y.

- If X and Y are correlated, then knowing X tells you something about Y.
- “X and Y are uncorrelated” is the same as “Correlation and covariance equal to 0”
- **Independent X, Y are uncorrelated**, because knowing X tells you nothing about Y.
- The converse is not necessarily true: **X, Y could be uncorrelated but not independent**.

Correlation

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



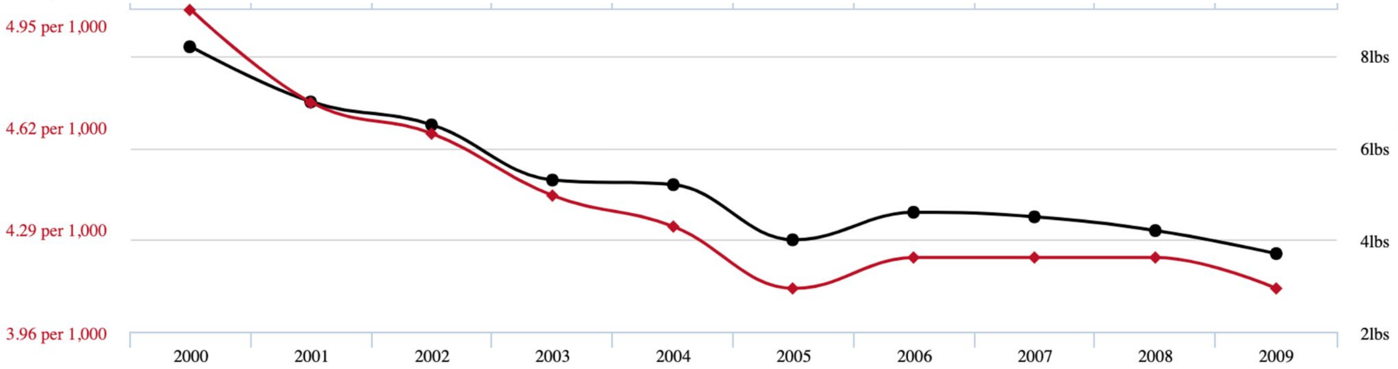
a). Plot (X,Y) Scatter Plot:

b). Are X and Y independent or dependent?

c). Do X and Y have a positive, negative or zero correlation?

Correlation

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

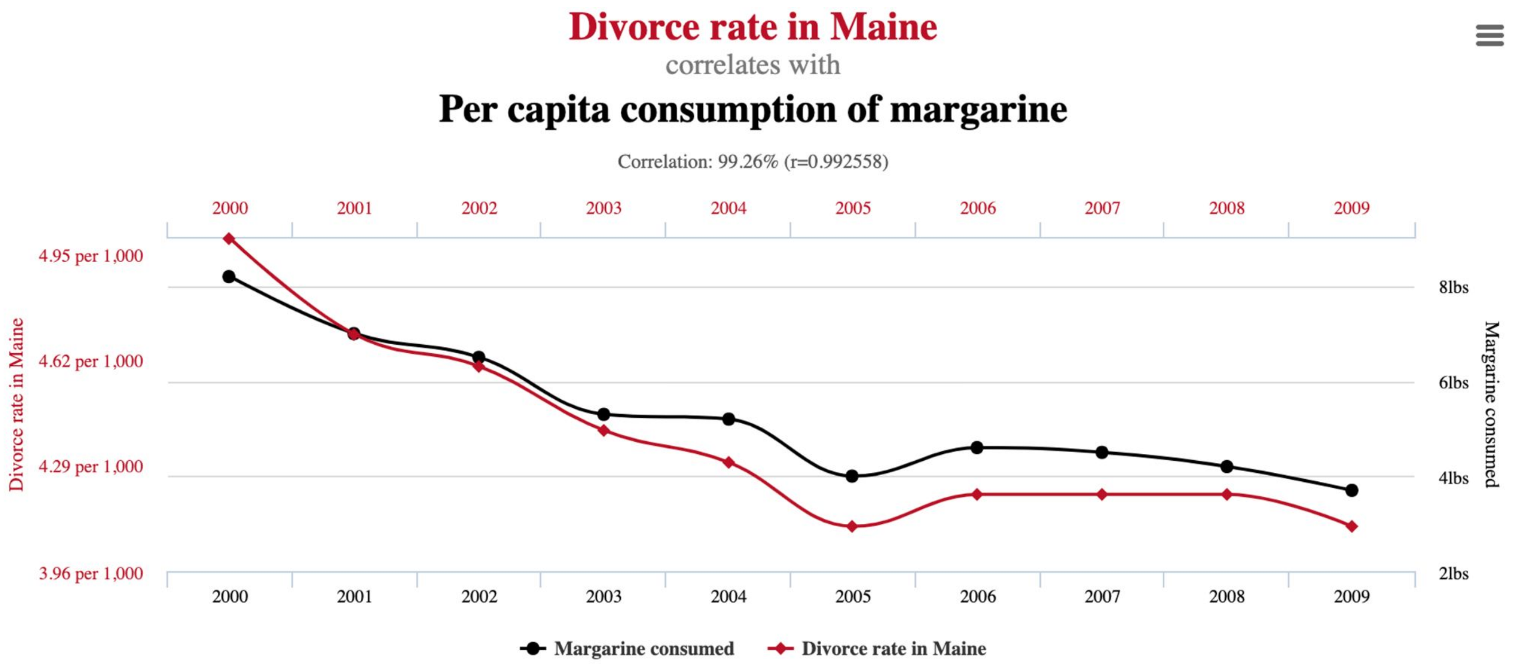


a). Plot (X,Y) Scatter Plot:

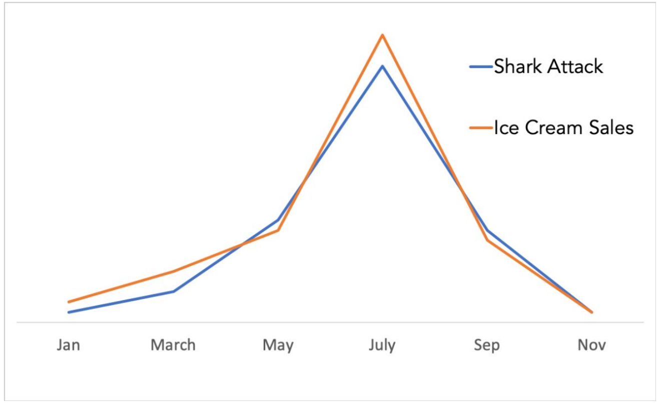
b). Are X and Y independent or dependent?

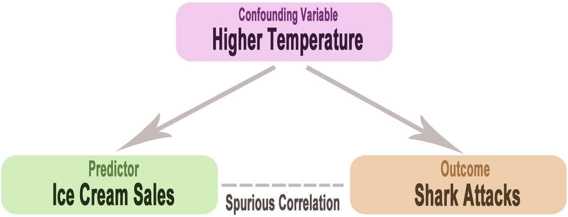
c). Do X and Y have a positive, negative or zero correlation?

“Correlation does not imply causation”



- A spurious correlation (or spuriousness) refers to a connection between two variables that appears to be causal but is not. With spurious correlation, any observed dependencies between variables are either:
  - related to some unseen confounder:

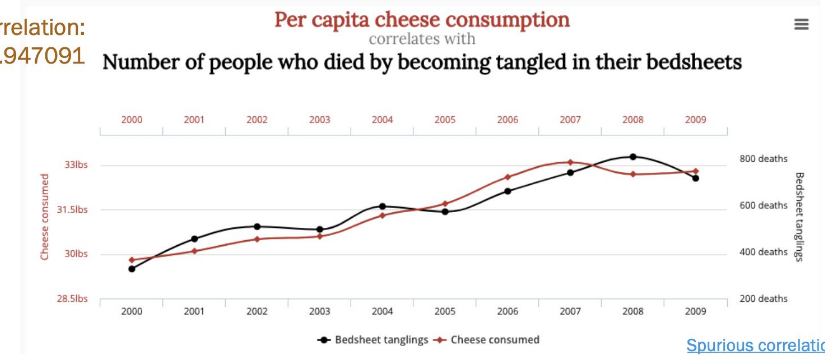




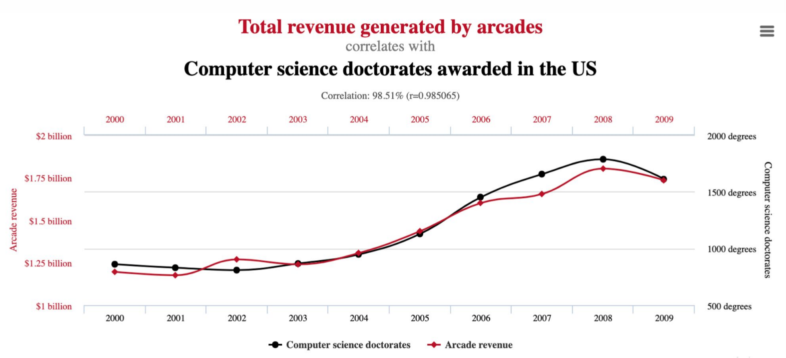
A "Confounding Variable" (also known as a "Confounder") is a variable that simultaneously causes the predictor/independent variable/explanatory variable/treatment (X) and the dependent variable/outcome (Y).

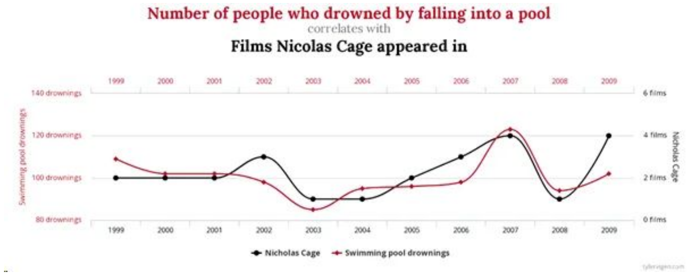
- A spurious correlation refers to a connection between two variables that appears to be causal but is not. With spurious correlation, any observed dependencies between variables are either:
  - both related to some unseen confounder.
  - merely due to chance (especially when you compute millions of correlations among thousands of variables)

Correlation:
0.947091

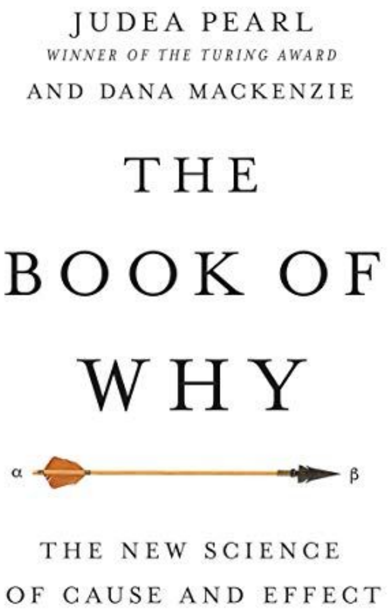


Spurious correlations





“Correlation does not imply causation”



Randomized Controlled Trials vs Observational Study

	Experimental Studies	Observational Studies
Examples	Randomized Clinical Trial	Cohort Case-control Cross-sectional Case Series
Group Assignment is Based On	Researcher assigns groups	“Natural Conditions” (personal preference, genetics, social determinants, environment…)
Use	“Gold Standard” for studying therapeutic interventions (treatments) or prophylactic interventions (prevention)	Associations between health outcomes and exposures. This can include studies on diagnosis, prognosis, etiology or harm

Common terminology examining relationships between variables:

- **association**
  - any relation
  - Link
- correlation
- causation

[Summary] Expectation and Variance for Linear Functions of Random Variables

Let X be a random variable with distribution $P(X = x)$.

$$\mathbb{E}[X] = \sum_x xP(X = x)$$

$$\mathrm{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \quad \text{(definition)}$$

$$= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \quad \text{(easier computation)}$$

Let a and b be scalar values.

$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$$

$$\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$$

Let Y be another random variable.

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

$$\mathrm{Var}(X_1 + X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \underbrace{2\mathrm{Cov}(X, Y)}_{\text{Zero if X, Y independent.}}$$

Addition rule for variance

If X and Y are **uncorrelated** then

$$\mathbb{V}ar(X+Y) = \mathbb{V}ar(X) + \mathbb{V}ar(Y)$$

Therefore, under the same conditions,

$$\mathbb{SD}(X+Y) = \sqrt{\mathbb{V}ar(X) + \mathbb{V}ar(Y)} = \sqrt{(\mathbb{SD}(X))^2 + (\mathbb{SD}(Y))^2}$$

- Think of this as “Pythagorean theorem” for random variables.
- Uncorrelated random variables are like orthogonal vectors.

Independent RV

Independent discrete RVs

Recall the definition of independent events E and F:

$$P(EF) = P(E)P(F)$$

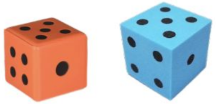
Two discrete random variables X and Y are independent if:

for all x, y:

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

- Intuitively: knowing value of X tells us nothing about the distribution of Y (and vice versa)
- If two variables are not independent, they are called dependent.

Dice (after all this time, still our friends)



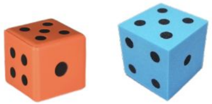
Let: D_1 and D_2 be the outcomes of two rolls
$S = D_1 + D_2$, the sum of two rolls

- Each roll of a 6-sided die is an independent trial.
- Random variables D_1 and D_2 are independent.

1. Are events $(D_1 = 1)$ and $(S = 7)$ independent?
2. Are events $(D_1 = 1)$ and $(S = 5)$ independent?
3. Are random variables D_1 and S independent?

Dice (after all this time, still our friends)

Let: D_1 and D_2 be the outcomes of two rolls
$S = D_1 + D_2$, the sum of two rolls



- Each roll of a 6-sided die is an independent trial.
- Random variables D_1 and D_2 are independent.

1. Are events $(D_1 = 1)$ and $(S = 7)$ independent? ✅
2. Are events $(D_1 = 1)$ and $(S = 5)$ independent? ❌
3. Are random variables D_1 and S independent?

All events $(X = x, Y = y)$ must be independent for X, Y to be independent RVs.

Example:

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

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

3. $\text{Cov}(X, Y) = E[XY] - E[X]E[Y]$
   $= 0 - 0(1/3) = 0$

4. Are X and Y independent?

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

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

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

4. Are X and Y independent?
$P(Y = 0 | X = 1) = 1$
$\neq \quad P(Y = 0) = 2/3$

Equal vs. Identically Distributed vs. IID

Suppose that we have two random variables X and Y.

X and Y are **equal** if:

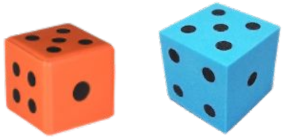
- $X(s) = Y(s)$ for every sample s.
- We write $X = Y$.

X and Y are **identically distributed** if:

- The distribution of X is the same as the distribution of Y
- We say "X and Y are equal in distribution."
- If $X = Y$, then X and Y are identically distributed;
  but the converse is not true (ex: **Y = 7-X**, X is a die)

X and Y are **independent and identically distributed (IID)** if:

- X and Y are identically distributed, *and*
- "X and Y are independent." (Knowing the outcome of X does not influence your belief of the outcome of Y, and vice versa)



IID RVs

LECTURE 17

Joint Distributions; Covariance and Correlation

CSCI 3022 @ CU Boulder

Maribeth Oscamou