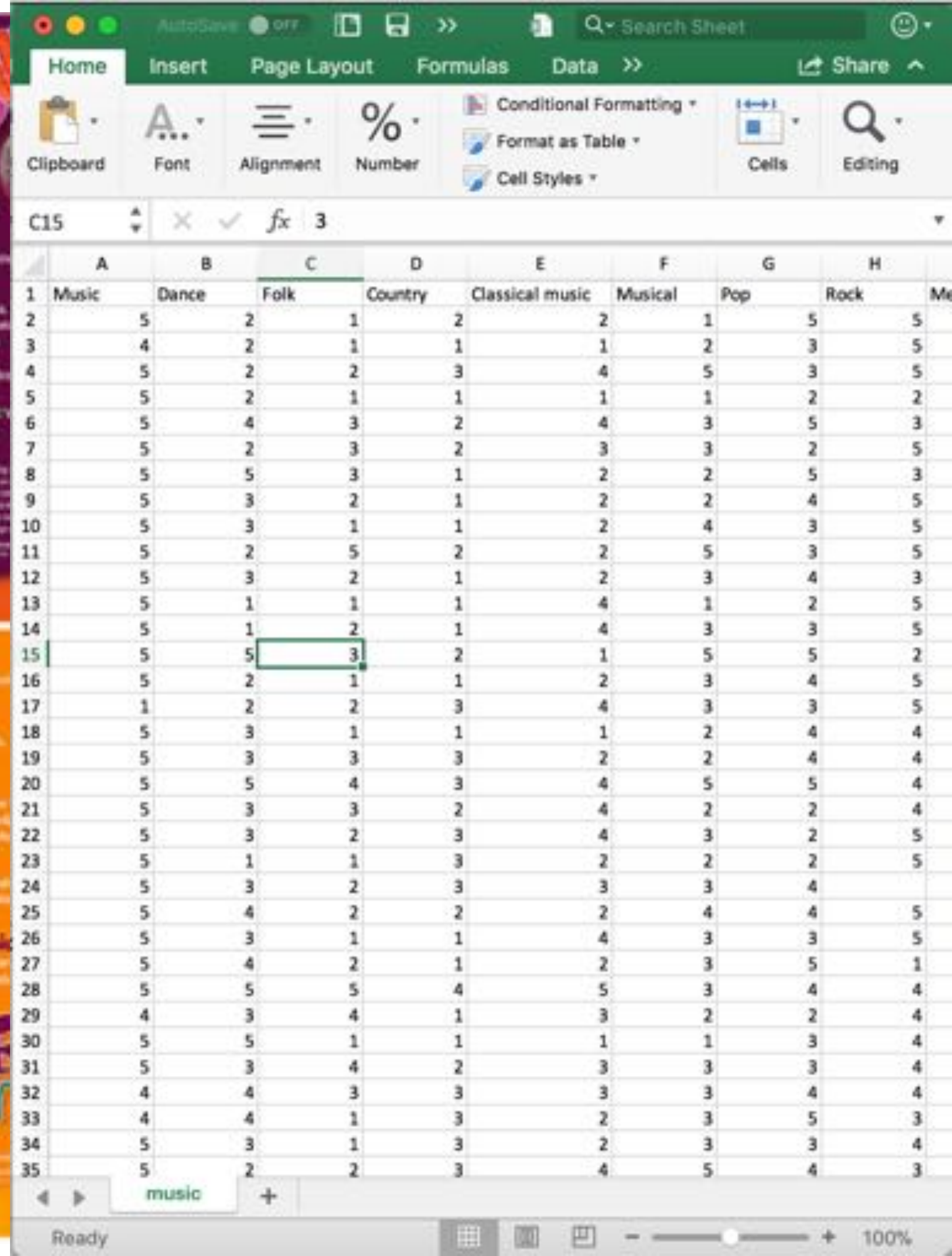
"True friendship comes when ~~the silence between two people is comfortable.~~"

Your random variables are correlated

Covariance and Correlation

Chris Piech

CS109, Stanford University



	A	B	C	D	E	F	G	H
1	Music	Dance	Folk	Country	Classical music	Musical	Pop	Rock
2	5	2	1	2	2	1	5	5
3	4	2	1	1	1	2	3	5
4	5	2	2	3	4	5	3	5
5	5	2	1	1	1	1	2	2
6	5	4	3	2	4	3	5	3
7	5	2	3	2	3	3	2	5
8	5	5	3	1	2	2	5	3
9	5	3	2	1	2	2	4	5
10	5	3	1	1	2	4	3	5
11	5	2	5	2	2	5	3	5
12	5	3	2	1	2	3	4	3
13	5	1	1	1	4	1	2	5
14	5	1	2	1	4	3	3	5
15	5	5	3	2	1	5	5	2
16	5	2	1	1	2	3	4	5
17	1	2	2	3	4	3	3	5
18	5	3	1	1	1	2	4	4
19	5	3	3	3	2	2	4	4
20	5	5	4	3	4	5	5	4
21	5	3	3	2	4	2	2	4
22	5	3	2	3	4	3	2	5
23	5	1	1	3	2	2	2	5
24	5	3	2	3	3	3	4	
25	5	4	2	2	2	4	4	5
26	5	3	1	1	4	3	3	5
27	5	4	2	1	2	3	5	1
28	5	5	5	4	5	3	4	4
29	4	3	4	1	3	2	2	4
30	5	5	1	1	1	1	3	4
31	5	3	4	2	3	3	3	4
32	4	4	3	3	3	3	4	4
33	4	4	1	3	2	3	5	3
34	5	3	1	3	2	3	3	4
35	5	2	2	3	4	5	4	3

Joint Random Variables

- [x] Use a joint table, density function or CDF to solve probability question
- [x] Think about **conditional** probabilities with joint variables (which might be continuous)
- [x] Use and find **expectation** of multiple RVS
- [x] Use and find **independence** of multiple RVS
- [x] What happens when you **add** random variables?
- [ ] How do multiple variables **covary**?

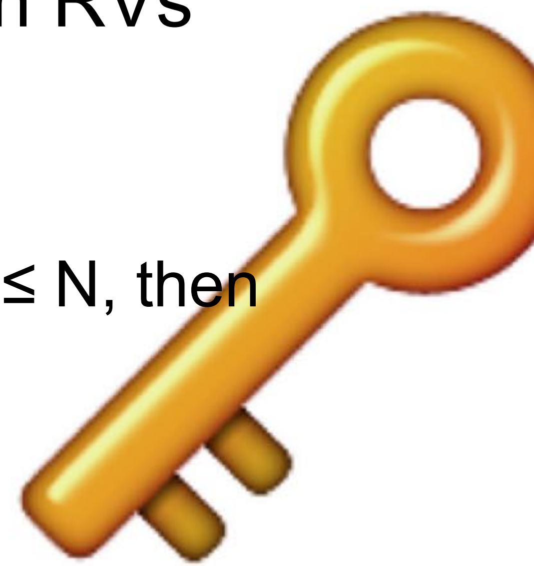
Reference: Sum of Independent RVs

- Let X and Y be independent Binomial RVs
  - $X \sim \text{Bin}(n_1, p)$ and $Y \sim \text{Bin}(n_2, p)$
  - $X + Y \sim \text{Bin}(n_1 + n_2, p)$
  - More generally, let $X_i \sim \text{Bin}(n_i, p)$ for $1 \le i \le N$, then

$$\left(\sum_{i=1}^{N} X_i\right) \sim \text{Bin}\left(\sum_{i=1}^{N} n_i,\ p\right)$$

- Let X and Y be independent Poisson RVs
  - $X \sim \text{Poi}(\lambda_1)$ and $Y \sim \text{Poi}(\lambda_2)$
  - $X + Y \sim \text{Poi}(\lambda_1 + \lambda_2)$
  - More generally, let $X_i \sim \text{Poi}(\lambda_i)$ for $1 \le i \le N$, then

$$\left(\sum_{i=1}^{N} X_i\right) \sim \text{Poi}\left(\sum_{i=1}^{N} \lambda_i\right)$$



Convolution of Probability Distributions

We talked about sum of Binomial and Poisson…who's missing from this party?

Uniform, Normal.

Summation: not just for the 1%

What about the general case?

Were talking about the sum of uniforms

sum.py

```python
import random

def main():
    x = random.random()
    y = random.random()
    z = x + y
    print(z)

if __name__ == '__main__':
    main()
```

The Insight to Convolution Proofs

$$P(X + Y = n)?$$

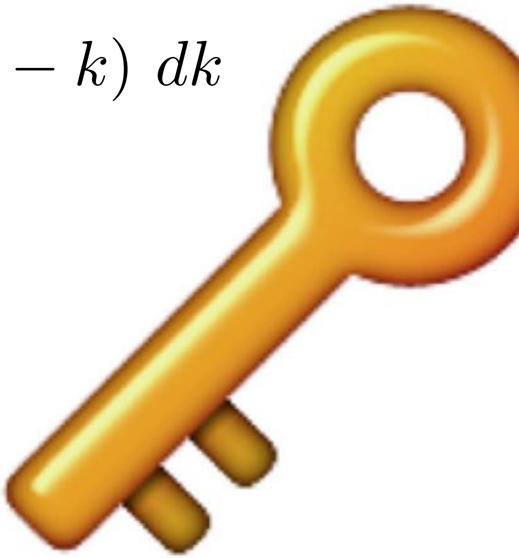
X	Y	k	
0	n	0	$P(X = 0, Y = n)$
1	n - 1	1	$P(X = 1, Y = n - 1)$
2	n - 2	2	$P(X = 2, Y = n - 2)$
	$\cdots$		
n	0	n	$P(X = n, Y = 0)$

$$P(X + Y = n) = \sum_{k=0}^{n} P(X = k, Y = n - k)$$

The Insight to Convolution Proofs

$$P(X+Y=\alpha) = \sum_{k=0}^{\alpha} P(X=k, Y=\alpha-k)$$

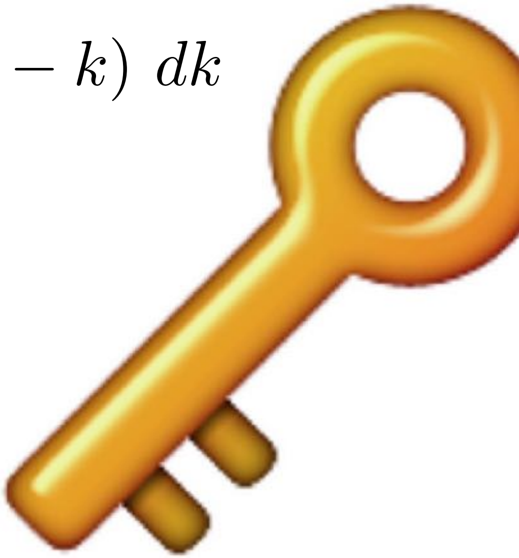
$$f(X+Y=\alpha) = \int_{k=-\infty}^{\infty} f(X=k, Y=\alpha-k)\ dk$$



The Insight to Convolution Proofs

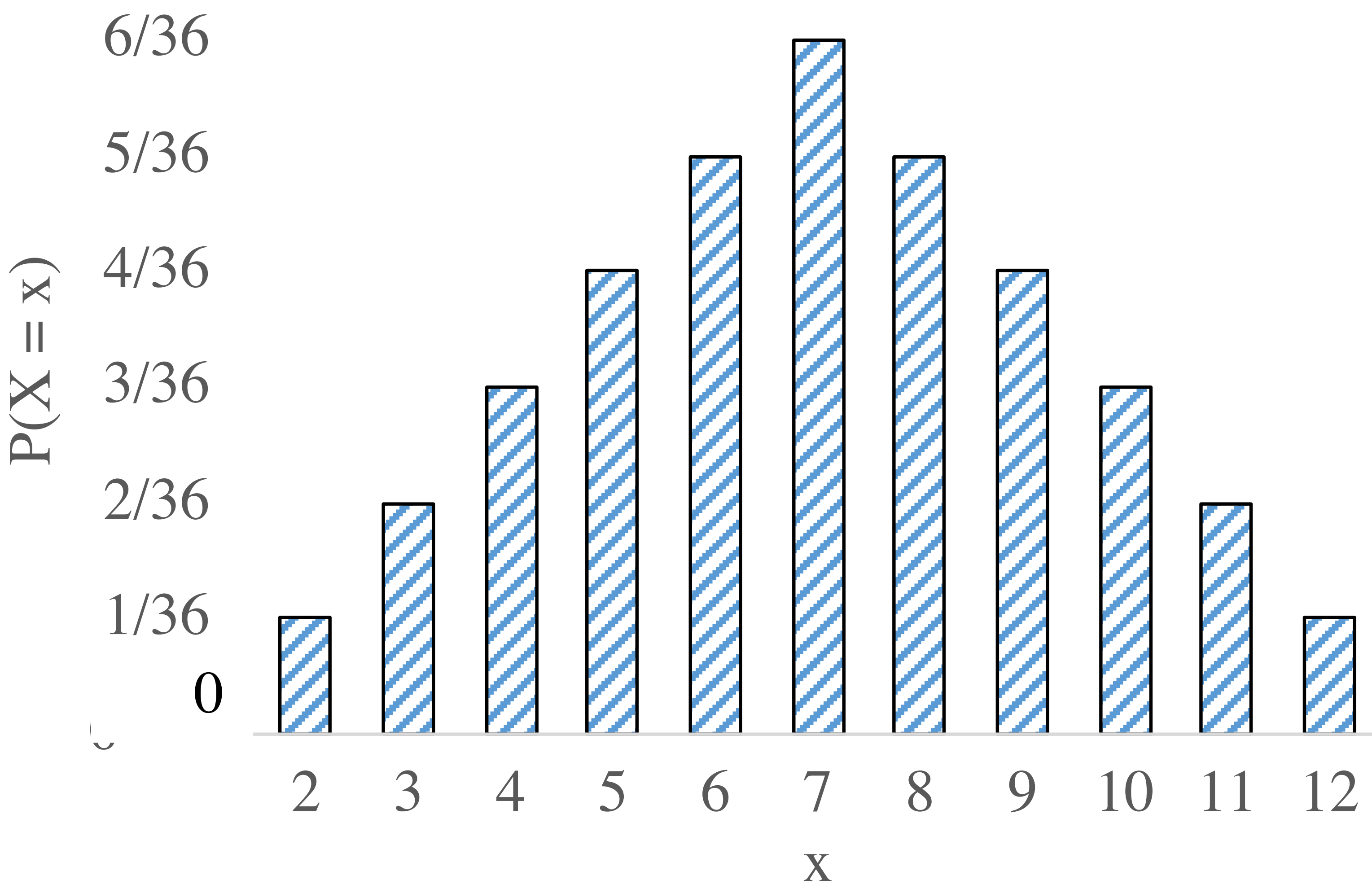
$$P(X+Y=\alpha) = \sum_{k=0}^{\alpha} P(X=k, Y=\alpha-k)$$

$$f_{X+Y}(\alpha) = \int_{k=-\infty}^{\infty} f(X=k, Y=\alpha-k)\ dk$$



Sum of Two Dice

Let X be the value of the sum of two dice
(aka two independent random variables)



Sum of Independent Uniforms

$X \sim \text{Uni}(0,1) \quad Y \sim \text{Uni}(0,1)$

X and Y are independent

$$f_{X+Y}(\alpha)?$$

$$f_{X+Y}(\alpha) = \int_{k=-\infty}^{\infty} f(X = k, Y = \alpha - k)\ dk$$

$$f_{X+Y}(\alpha) = \int_{k=-\infty}^{\infty} f(X = k) f(Y = \alpha - k)\ dk$$

Sum of Independent Uniforms

$X \sim \text{Uni}(0, 1) \quad Y \sim \text{Uni}(0, 1)$

X and Y are independent

$$f_{X+Y}(\alpha)?$$

$$f_{X+Y}(\alpha) = \int_{k=-\infty}^{\infty} f(X = k) f(Y = \alpha - k) \; dk$$

Sum of Independent Uniforms

$X \sim \text{Uni}(0, 1) \quad Y \sim \text{Uni}(0, 1)$

X and Y are independent

$$f_{X+Y}(\alpha)?$$

$$f_{X+Y}(\alpha) = \int_{k=-\infty}^{\infty} f(X = k) f(Y = \alpha - k) \; dk$$

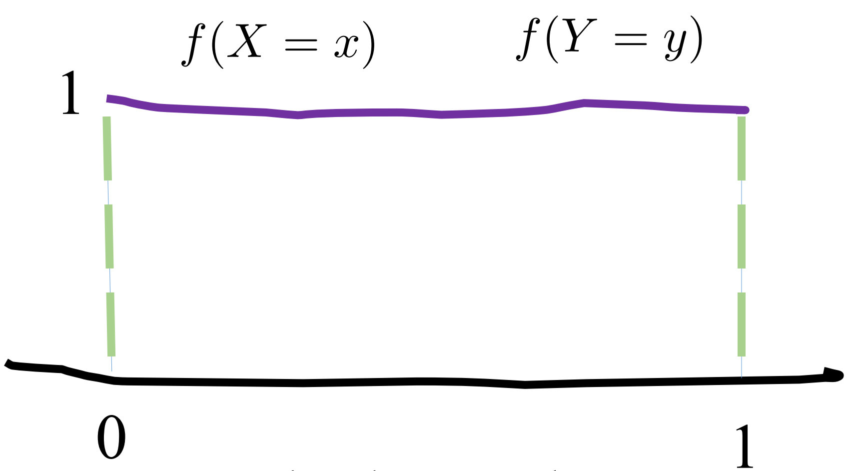
Sum of Independent Uniforms

$X \sim \text{Uni}(0,1)$ $\quad Y \sim \text{Uni}(0,1)$

X and Y are independent

$$f_{X+Y}(\alpha)?$$

$$f_{X+Y}(\alpha) = \int_{k=-\infty}^{\infty} f(X=k) f(Y=\alpha-k)\, dk$$



For both X and Y

Sum of Independent Uniforms

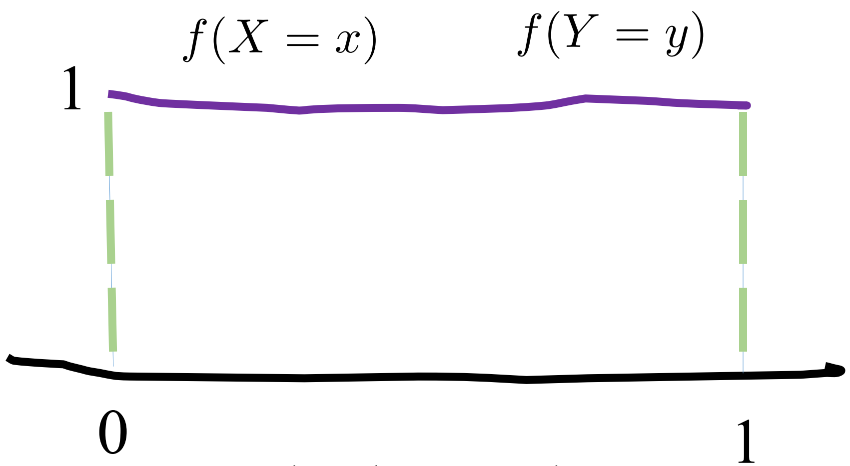
$X \sim \mathrm{Uni}(0,1) \qquad Y \sim \mathrm{Uni}(0,1)$

X and Y are independent

$$f_{X+Y}(\alpha)?$$

$$f_{X+Y}(\alpha) = \int_{k=-\infty}^{\infty} f(X=k)f(Y=\alpha-k)\; dk$$

ISNT THIS JUST ONE!?!?!!!??!



For both X and Y

Sum of Independent Uniforms

$X \sim \text{Uni}(0,1) \qquad Y \sim \text{Uni}(0,1)$

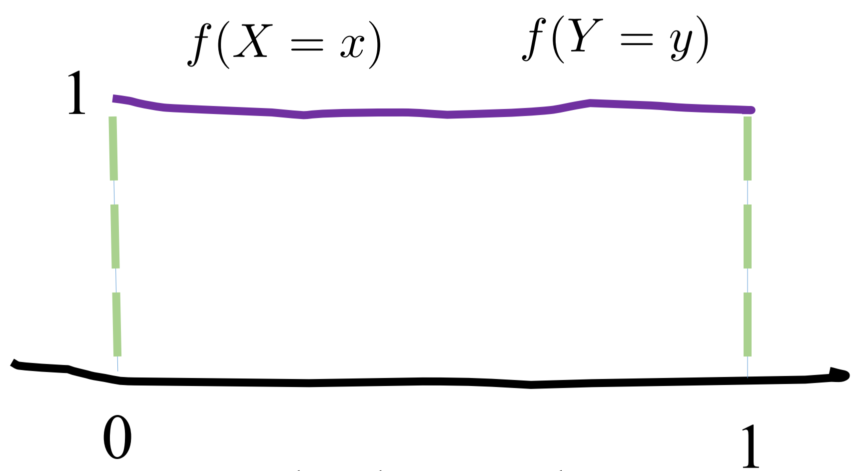
X and Y are independent

$f_{X+Y}(\alpha)?$

$$f_{X+Y}(\alpha) = \int_{k=-\infty}^{\infty} f(X=k)f(Y=\alpha-k)\ dk$$

For these values of k, the densities of f_X and f_Y are 1

$$\boxed{0 < k < 1 \qquad 0 < \alpha - k < 1}$$



For both X and Y

Sum of Independent Uniforms

$X \sim \mathrm{Uni}(0,1) \quad Y \sim \mathrm{Uni}(0,1)$

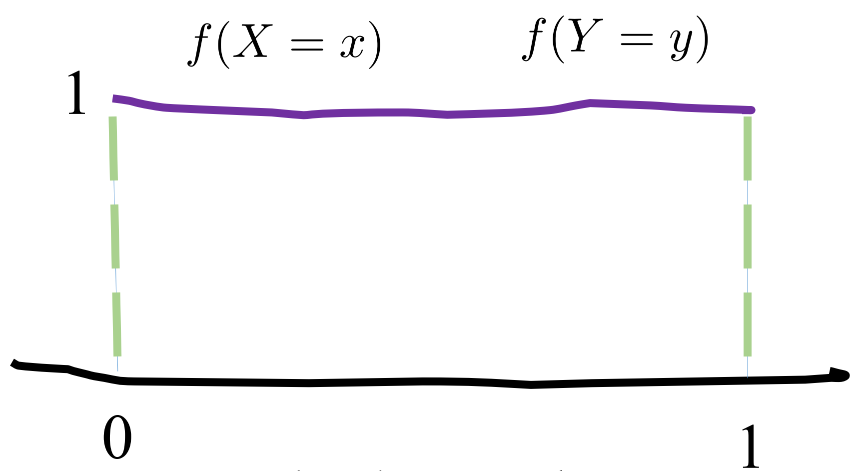
X and Y are independent

$$f_{X+Y}(\alpha)?$$

$$f_{X+Y}(\alpha) = \int_{k=-\infty}^{\infty} f(X=k)f(Y=\alpha-k)\ dk$$

For these values of k, the densities of f_X and f_Y are 1

$$\boxed{0 < k < 1 \qquad -\alpha < -k < 1-\alpha}$$



For both X and Y

Sum of Independent Uniforms

$X \sim \text{Uni}(0,1)$ $\quad$ $Y \sim \text{Uni}(0,1)$

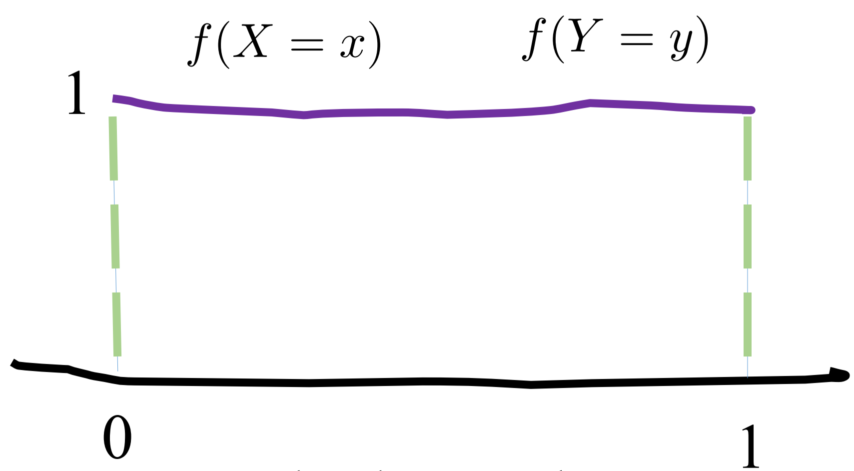
X and Y are independent

$$f_{X+Y}(\alpha)?$$

$$f_{X+Y}(\alpha) = \int_{k=-\infty}^{\infty} f(X=k)f(Y=\alpha-k)\, dk$$

For these values of k, the densities of f_X and f_Y are 1

$$\boxed{0 < k < 1 \qquad \alpha - 1 < k < \alpha}$$



For both X and Y

$\alpha = ½$

$X \sim \text{Uni}(0, 1) \quad Y \sim \text{Uni}(0, 1)$

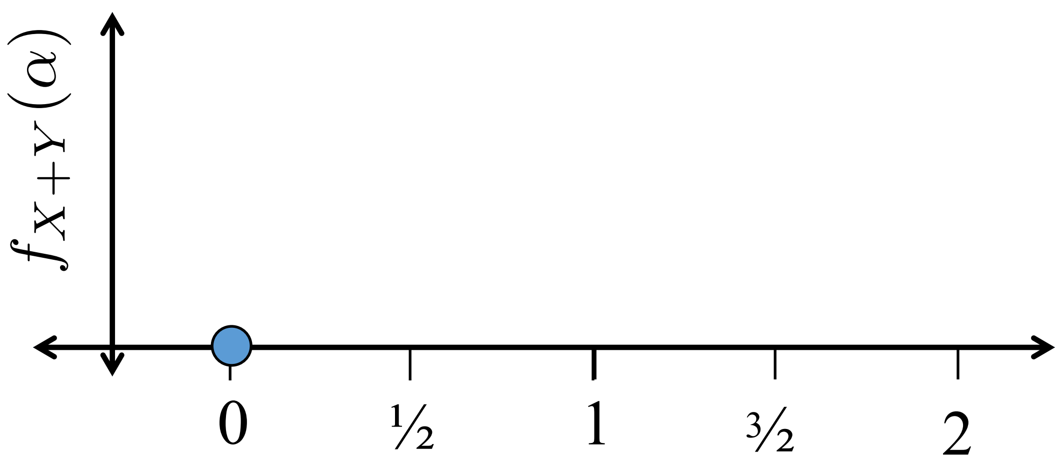
X and Y are independent

$f_{X+Y}(\alpha)?$

$$f_{X+Y}(\alpha) = \int_{k=-\infty}^{\infty} f(X = k) f(Y = \alpha - k) \; dk$$

For these values of k, the densities of f_X and f_Y are 1

$$\boxed{0 < k < 1 \qquad \alpha - 1 < k < \alpha}$$



$\alpha = ½$

$X \sim \text{Uni}(0, 1) \quad Y \sim \text{Uni}(0, 1)$

X and Y are independent

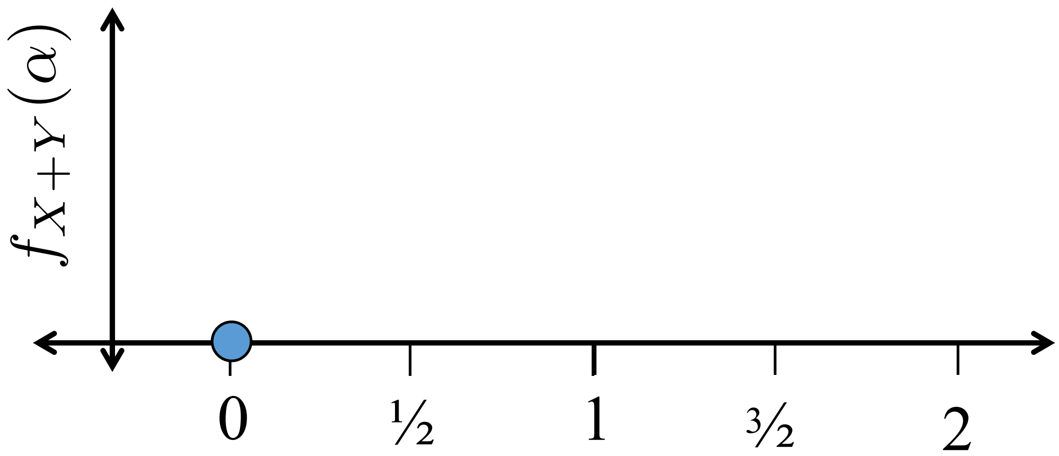
$f_{X+Y}(\alpha)$?

$$f_{X+Y}(1/2) = \int_{k=-\infty}^{\infty} f(X = k) f(Y = 1/2 - k) \, dk$$

α = 1/2

For these values of k, the densities of f_X and f_Y are 1

$$0 < k < 1 \qquad \alpha - 1 < k < \alpha$$



$\alpha = ½$

$X \sim \text{Uni}(0,1) \quad Y \sim \text{Uni}(0,1)$

X and Y are independent

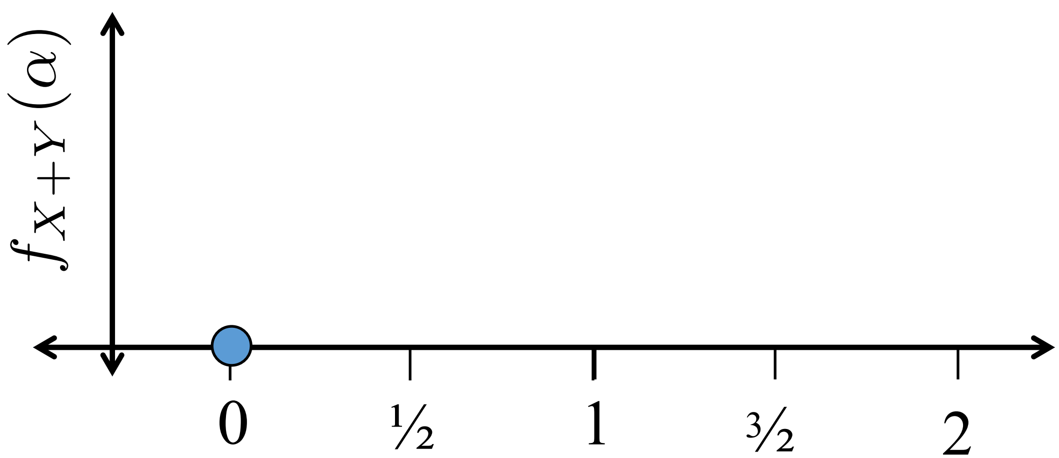
$$f_{X+Y}(\alpha)?$$

$$f_{X+Y}(1/2) = \int_{k=-\infty}^{\infty} f(X=k) f(Y=1/2-k)\, dk$$

$\alpha = 1/2$

For these values of k, the densities of f_X and f_Y are 1

$$\boxed{0 < k < 1 \qquad -1/2 < k < 1/2}$$



$\alpha = ½$

$X \sim \mathrm{Uni}(0,1) \qquad Y \sim \mathrm{Uni}(0,1)$

X and Y are independent

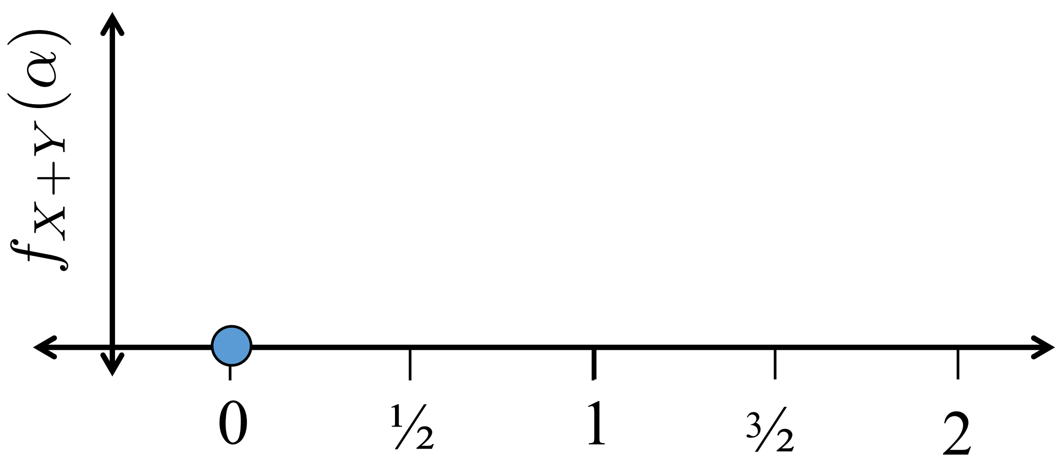
$$f_{X+Y}(\alpha)?$$

$$f_{X+Y}(1/2) = \int_{k=0}^{1/2} f(X=k)f(Y=1/2-k)\ dk$$

α = 1/2

For these values of k, the densities of f_X and f_Y are 1

$0 < k < 1 \qquad -1/2 < k < 1/2$



$\alpha = ½$

$X \sim \text{Uni}(0,1) \quad Y \sim \text{Uni}(0,1)$

X and Y are independent

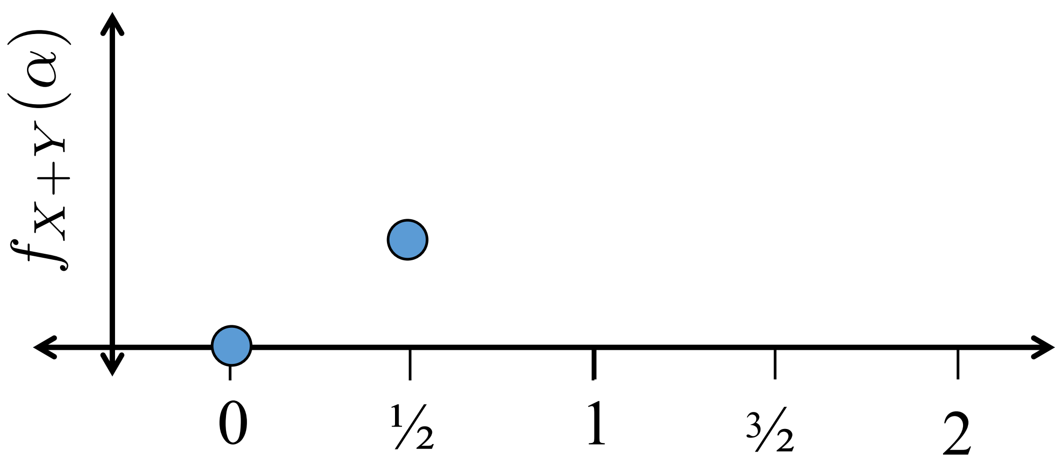
$f_{X+Y}(\alpha)$?

$$f_{X+Y}(1/2) = \int_{k=0}^{1/2} 1 \; dk = 0.5$$

α = 1/2

For these values of k, the densities are 1

$$0 < k < 1 \qquad -1/2 < k < 1/2$$



$0 < \alpha < 1$

$X \sim \text{Uni}(0,1) \qquad Y \sim \text{Uni}(0,1)$

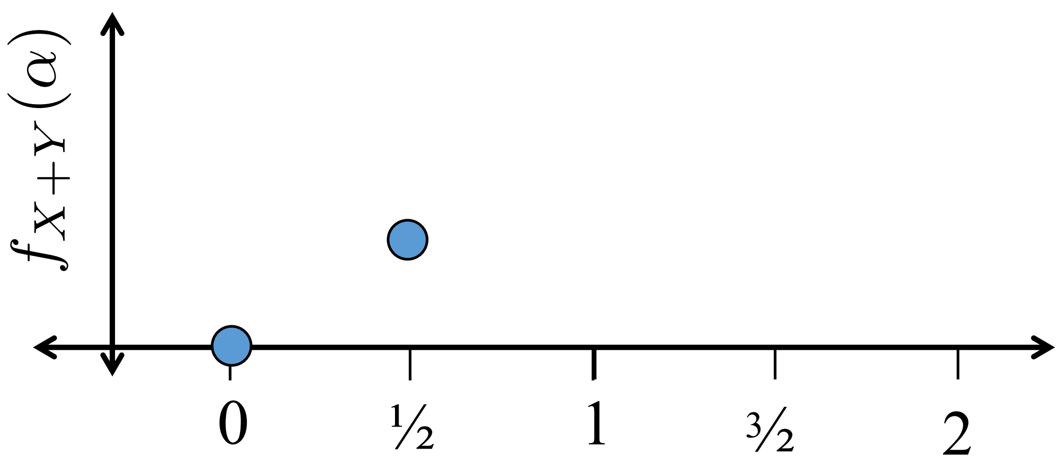
X and Y are independent

$$f_{X+Y}(\alpha)?$$

$$f_{X+Y}(\alpha) = \int_{k=-\infty}^{\infty} f(X=k)f(Y=\alpha-k)\; dk$$

For these values of k, the densities are 1

$$\boxed{0 < k < 1 \qquad \alpha - 1 < k < \alpha}$$



$0 < \alpha < 1$

$X \sim \mathrm{Uni}(0,1) \quad Y \sim \mathrm{Uni}(0,1)$

X and Y are independent

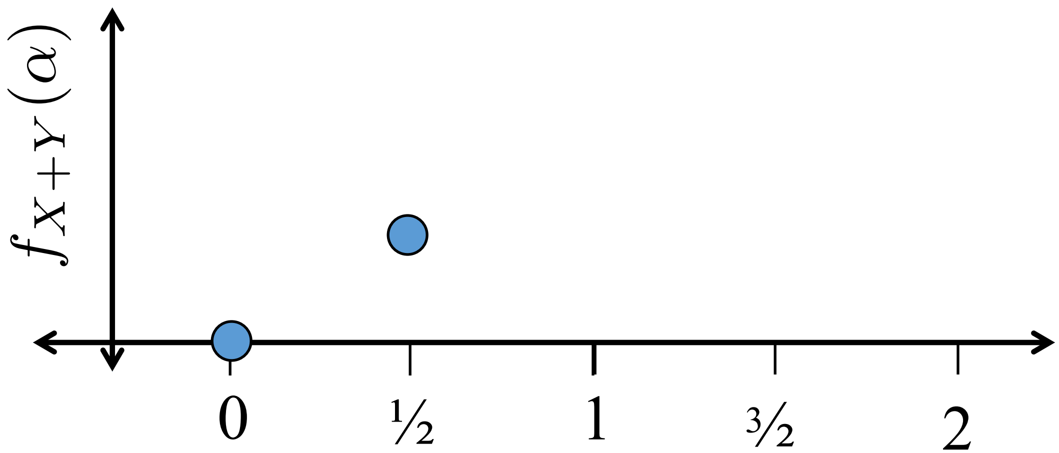
$$f_{X+Y}(\alpha)?$$

$$f_{X+Y}(\alpha) = \int_{k=-\infty}^{\infty} f(X=k)f(Y=\alpha-k)\ dk$$

For these values of k, the densities are 1

$0 < k < 1 \qquad \alpha - 1 < k < \alpha$

$0 < k < \alpha$



$0 < \alpha < 1$

$X \sim \text{Uni}(0,1) \quad Y \sim \text{Uni}(0,1)$

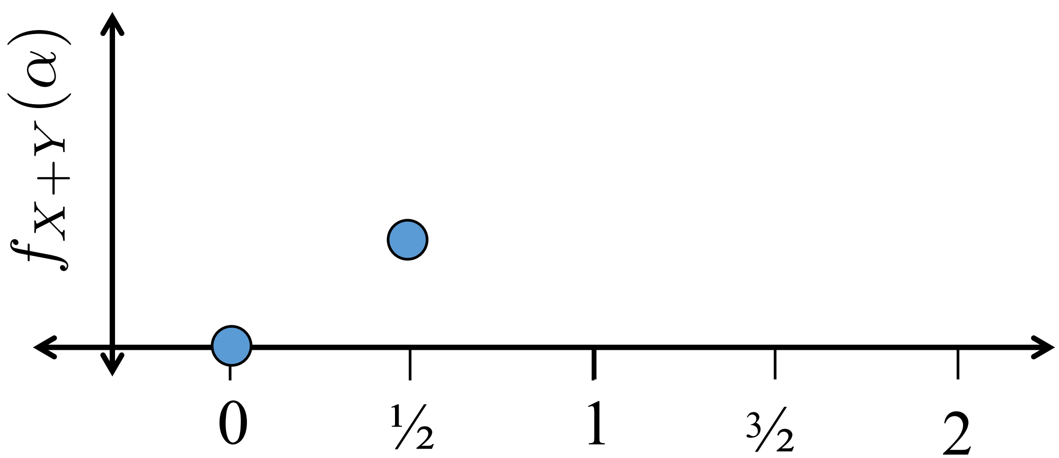
X and Y are independent

$$f_{X+Y}(\alpha)?$$

$$f_{X+Y}(\alpha) = \int_{k=0}^{\alpha} f(X=k) f(Y=\alpha-k)\, dk$$

For these values of k, the densities are 1

$\boxed{0 < k < 1 \qquad \alpha - 1 < k < \alpha}$ $\qquad 0 < k < \alpha$



$0 < \alpha < 1$

$X \sim \text{Uni}(0, 1) \quad Y \sim \text{Uni}(0, 1)$

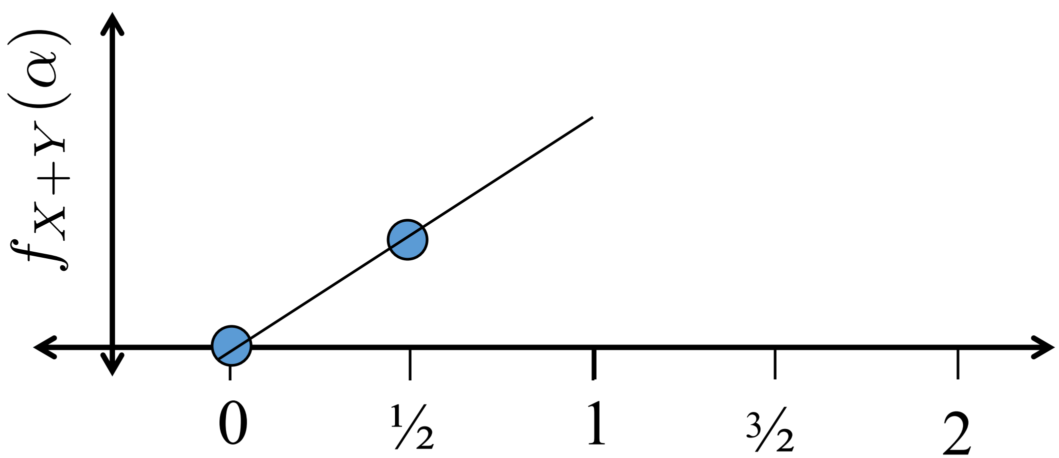
X and Y are independent

$f_{X+Y}(\alpha)$?

$$f_{X+Y}(\alpha) = \int_{k=0}^{\alpha} 1 \; dk = \alpha$$

For these values of k, the densities are 1

$$\boxed{0 < k < 1 \qquad \alpha - 1 < k < \alpha} \qquad 0 < k < \alpha$$



$1 < \alpha < 2$

$X \sim \text{Uni}(0,1) \quad Y \sim \text{Uni}(0,1)$

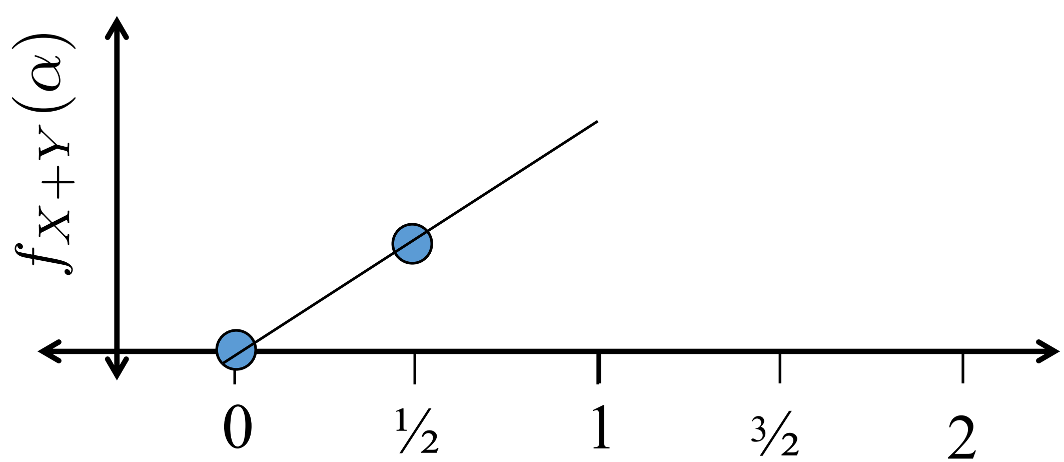
X and Y are independent

$f_{X+Y}(\alpha)?$

$$f_{X+Y}(\alpha) = \int_{k=-\infty}^{\infty} f(X=k)f(Y=\alpha-k)\;dk$$

For these values of k, the densities are 1

$$0 < k < 1 \qquad \alpha - 1 < k < \alpha$$



$1 < \alpha < 2$

$X \sim \mathrm{Uni}(0,1) \qquad Y \sim \mathrm{Uni}(0,1)$

X and Y are independent

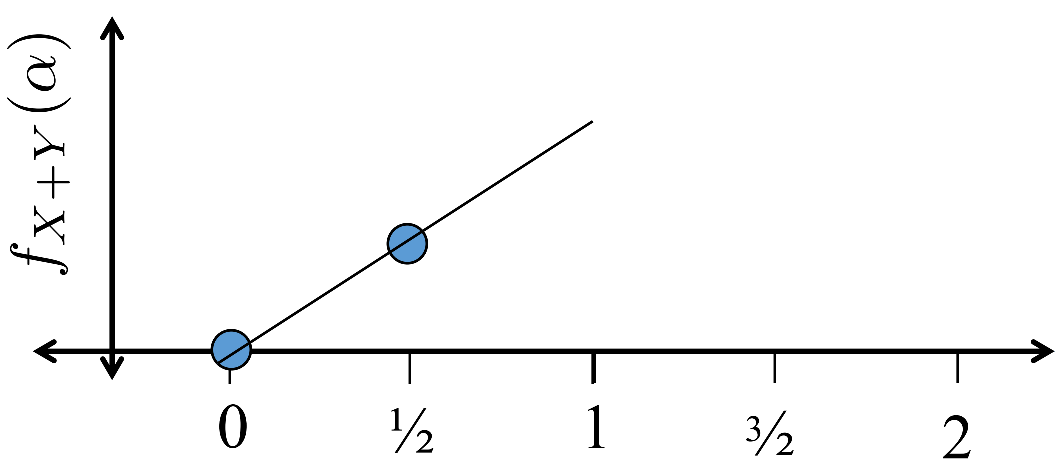
$f_{X+Y}(\alpha)$?

$$f_{X+Y}(\alpha) = \int_{k=-\infty}^{\infty} f(X=k) f(Y=\alpha-k)\ dk$$

For these values of k, the densities are 1

$0 < k < 1 \qquad \alpha - 1 < k < \alpha$

$\alpha - 1 < k < 1$



$1 < \alpha < 2$

$X \sim \text{Uni}(0,1) \qquad Y \sim \text{Uni}(0,1)$

X and Y are independent

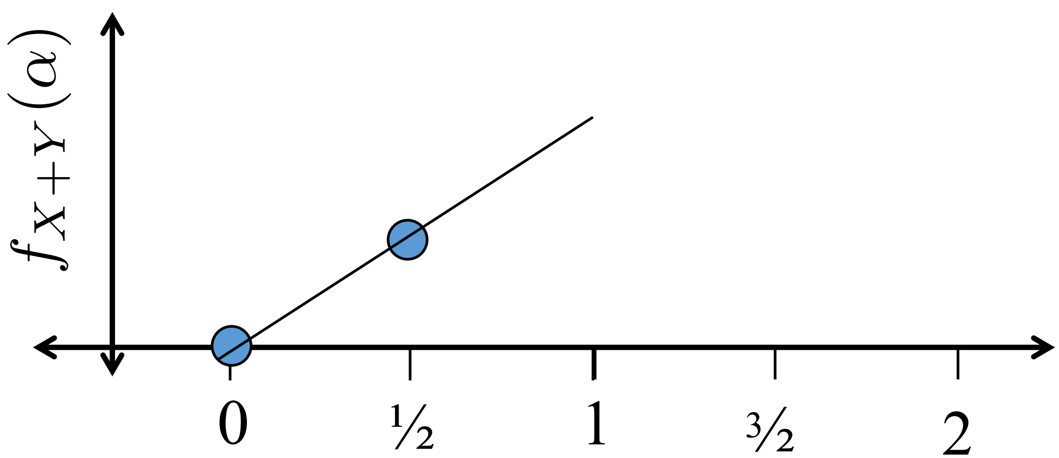
$$f_{X+Y}(\alpha)?$$

$$f_{X+Y}(\alpha) = \int_{k=\alpha-1}^{1} f(X=k) f(Y=\alpha-k)\ dk$$

For these values of k, the densities are 1

$$\boxed{0 < k < 1 \qquad \alpha - 1 < k < \alpha}$$

$$\alpha - 1 < k < 1$$



$1 < \alpha < 2$

$X \sim \mathrm{Uni}(0,1) \quad Y \sim \mathrm{Uni}(0,1)$

X and Y are independent

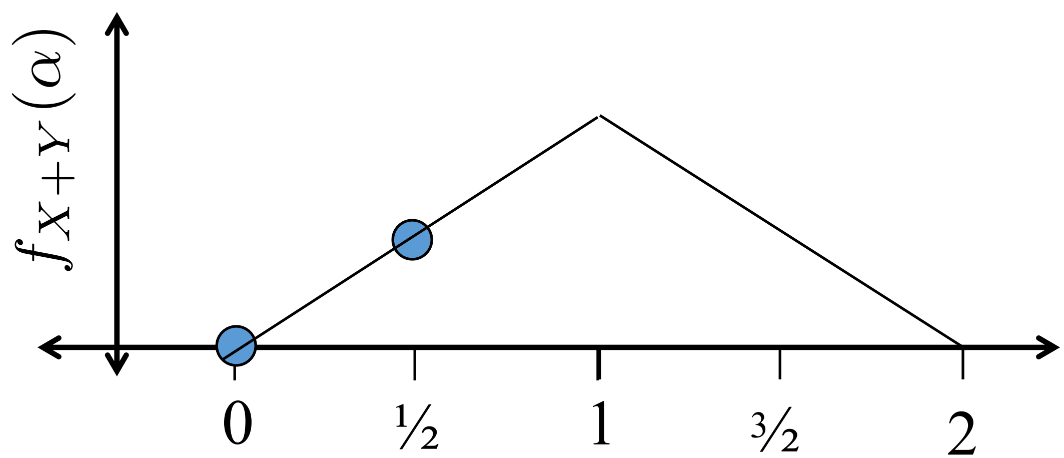
$$f_{X+Y}(\alpha)?$$

$$f_{X+Y}(\alpha) = \int_{k=\alpha-1}^{1} 1\ dk = 2 - \alpha$$

For these values of k, the densities are 1

$$\boxed{0 < k < 1 \qquad \alpha - 1 < k < \alpha}$$

$$\alpha - 1 < k < 1$$



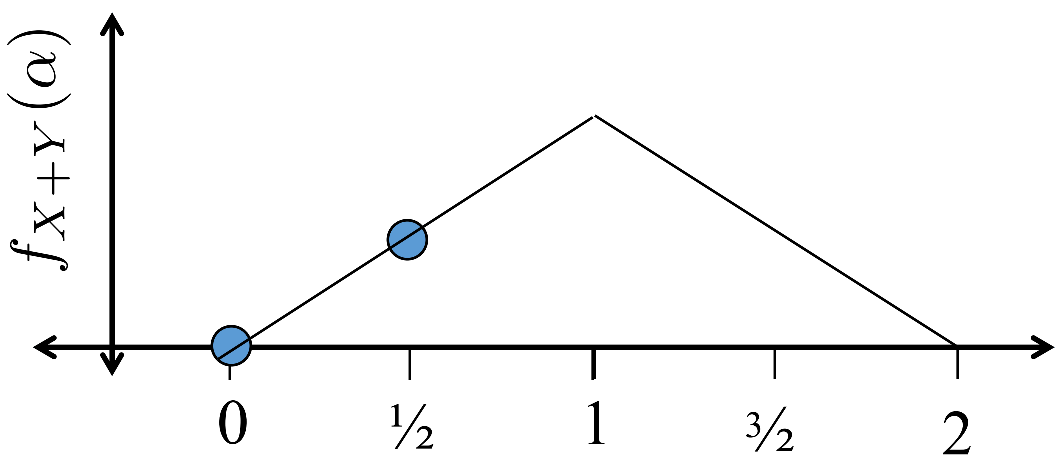
$1 < \alpha < 2$

$X \sim \text{Uni}(0,1) \quad Y \sim \text{Uni}(0,1)$

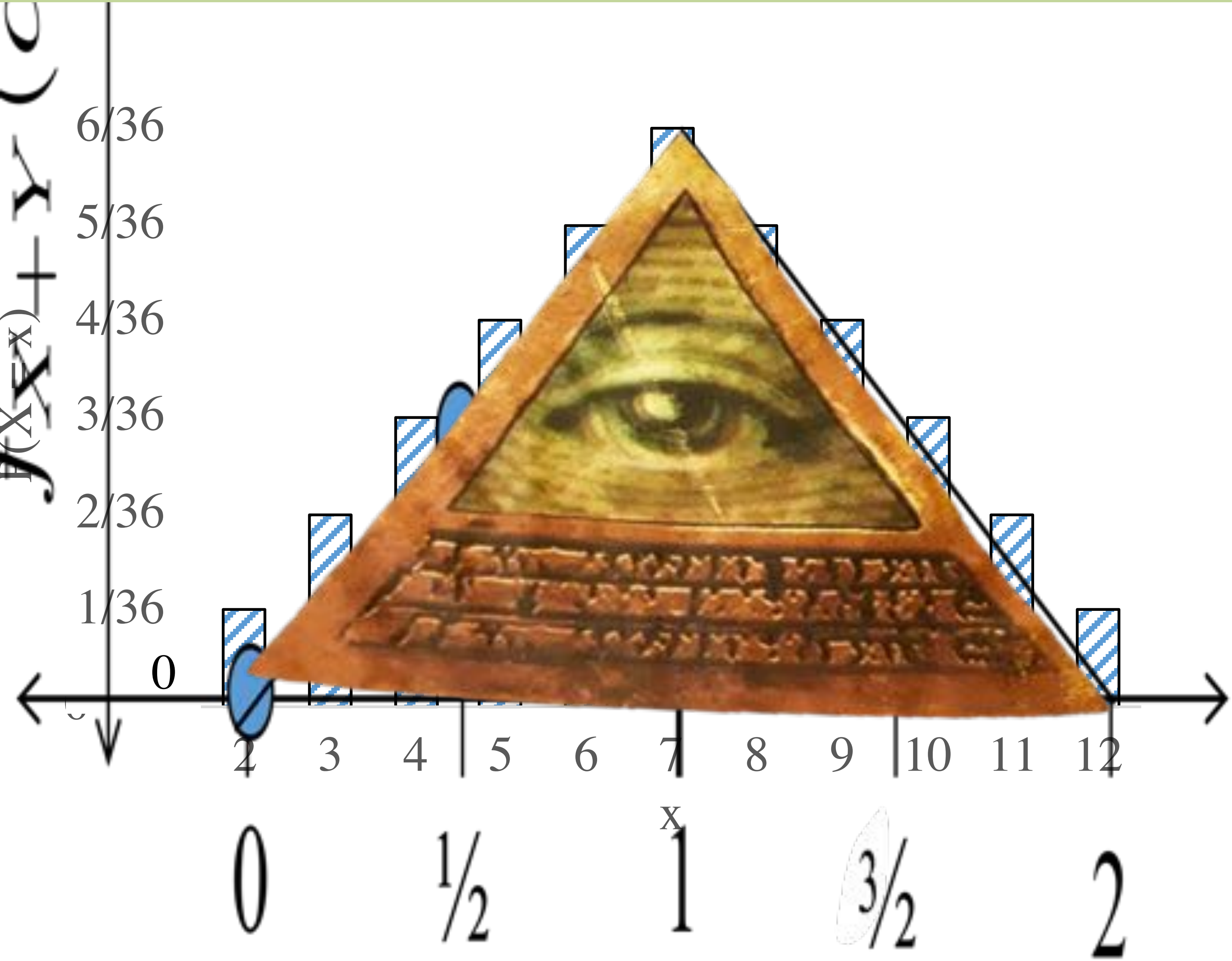
X and Y are independent

$f_{X+Y}(\alpha)?$

$$f_{X+Y}(a) = \begin{cases} a & 0 \le a \le 1 \\ 2-a & 1 < a \le 2 \\ 0 & \text{otherwise} \end{cases}$$



Sum of Uniforms and Sum of Dice



That was hard…

Ready for something easy and truly useful?

Sum of Independent Normals

- Let X and Y be independent random variables
  - $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$
  - $X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$

- Generally, have n independent random variables $X_i \sim N(\mu_i, \sigma_i^2)$ for $i = 1, 2, ..., n$:

$$\left(\sum_{i=1}^{n} X_i\right) \sim N\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$$

Virus Infections

- Say you are working with the WHO to plan a response to a the initial conditions of a virus:
  - Two exposed groups
  - P1: 50 people, each independently infected with $p = 0.1$
  - P2: 100 people, each independently infected with $p = 0.4$
  - Question: Probability of more than 40 infections?

Sanity check: Should we use the Binomial Sum-of-RVs shortcut?

A. YES!
B. NO!
C. Other/none/more

Virus Infections

- Say you are working with the WHO to plan a response to a the initial conditions of a virus:
  - Two exposed groups
  - P1: 50 people, each independently infected with $p = 0.1$
  - P2: 100 people, each independently infected with $p = 0.4$
  - Question: Probability of more than 40 infections?

Virus Infections

- Say you are working with the WHO to plan a response to a the initial conditions of a virus:
  - Two exposed groups
  - P1: 50 people, each independently infected with $p = 0.1$
  - P2: 100 people, each independently infected with $p = 0.4$
  - A = # infected in P1 A ~ Bin(50, 0.1) ≈ X ~ N(5, 4.5)
  - B = # infected in P2 B ~ Bin(100, 0.4) ≈ Y ~ N(40, 24)
  - What is P(≥ 40 people infected)?
  - P(A + B ≥ 40) ≈ P(X + Y ≥ 39.5)
  - X + Y = W ~ N(5 + 40 = 45, 4.5 + 24 = 28.5)

$$P(W \geq 39.5) = P\left(\frac{W - 45}{\sqrt{28.5}} > \frac{39.5 - 45}{\sqrt{28.5}}\right) = 1 - \Phi(-1.03) \approx 0.8485$$

Linear Transform

$$X \sim N(\mu, \sigma^2)$$

$$Y = X + X = 2 \cdot X$$

$$Y \sim N(2\mu, 4\sigma^2)$$

$$Y = X + X = 2 \cdot X$$

$$X + X \sim N(\mu + \mu, \sigma^2 + \sigma^2)$$

$$Y \sim N(2\mu, 2\sigma^2)$$

X is not independent of X

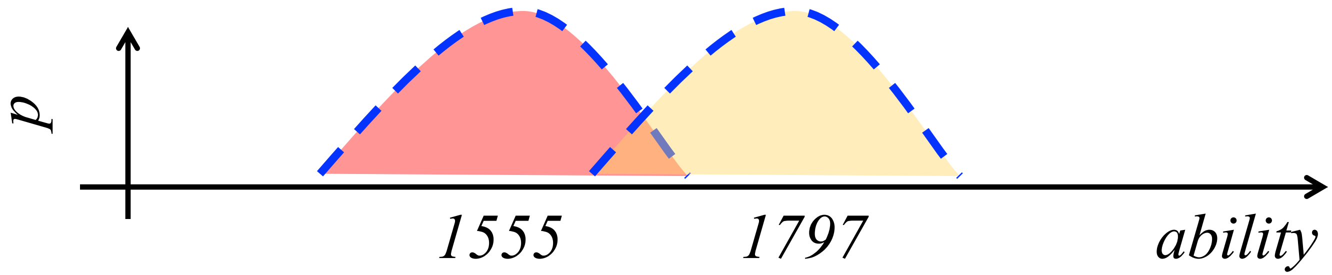
Motivating Idea: Zero Sum Games

How it works:

- Each team has an “ELO” score *S*, calculated based on their past performance.
- Each game, the team has ability $A \sim N(S, 200^2)$
- The team with the higher sampled ability wins.

Arpad Elo

$$A_B \sim \mathcal{N}(1555, 200^2) \qquad A_W \sim \mathcal{N}(1797, 200^2)$$



$$P(\text{Warriors win}) = P(A_W > A_B)$$

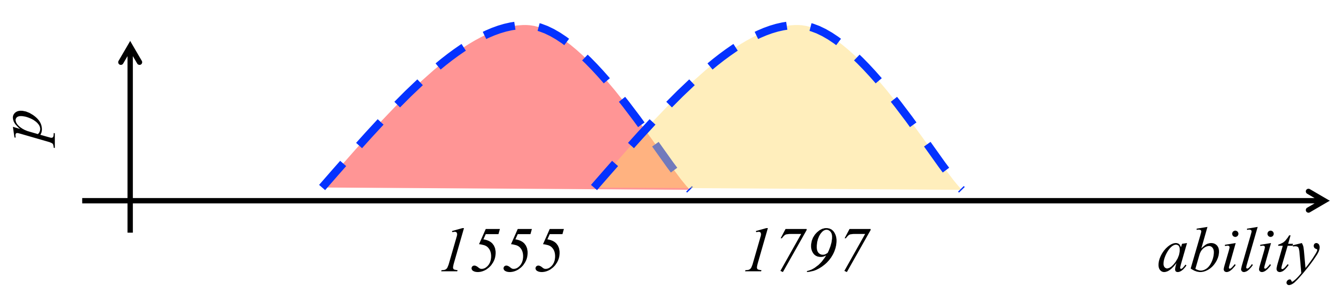
Motivating Idea: Zero Sum Games

$$A_B \sim \mathcal{N}(1555, 200^2) \qquad\qquad A_W \sim \mathcal{N}(1797, 200^2)$$

$$\begin{aligned} P(\text{Warriors win}) = \; & P(A_W > A_B) \\ & = P(A_W - A_B > 0) \end{aligned}$$

$$\begin{aligned} D &= A_W - A_B \\ D &\sim N(\mu = 1795 - 1555, \sigma_2 = 2 \cdot 200^2) \\ &\sim N(\mu = 240, \sigma_2 = 283) \end{aligned}$$

$$P(D > 0) = F_D(0) = 1 - \Phi\left(\frac{0 - 240}{283}\right) \approx 0.65$$



Dance of Covariance

Recall our Ebola Bats

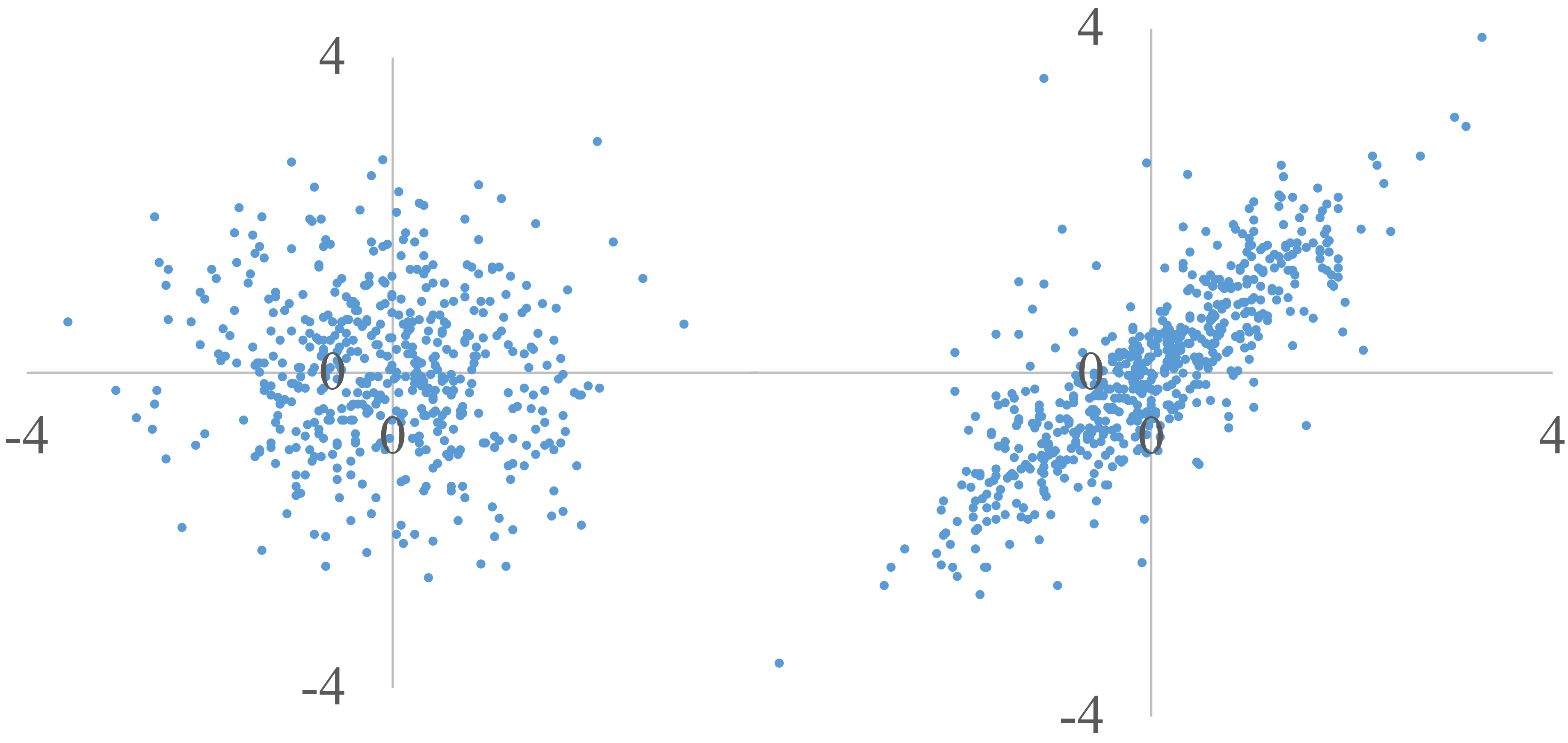
Bat Data

Gene1	Gene2	Gene3	Gene4	Gene5	Trait
TRUE	FALSE	TRUE	TRUE	FALSE	FALSE
FALSE	FALSE	TRUE	TRUE	TRUE	TRUE
TRUE	FALSE	TRUE	FALSE	FALSE	FALSE
TRUE	FALSE	TRUE	TRUE	TRUE	FALSE
FALSE	TRUE	TRUE	TRUE	TRUE	TRUE
FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
TRUE	FALSE	FALSE	TRUE	FALSE	FALSE
TRUE	FALSE	FALSE	TRUE	FALSE	FALSE
TRUE	FALSE	TRUE	FALSE	FALSE	FALSE
FALSE	TRUE	FALSE	TRUE	FALSE	FALSE
TRUE	TRUE	FALSE	TRUE	FALSE	FALSE
TRUE	FALSE	FALSE	TRUE	FALSE	FALSE
TRUE	FALSE	TRUE	TRUE	TRUE	FALSE
FALSE	FALSE	TRUE	TRUE	FALSE	FALSE
TRUE	FALSE	FALSE	TRUE	FALSE	FALSE
TRUE	FALSE	FALSE	TRUE	FALSE	FALSE
		...			
TRUE	FALSE	FALSE	TRUE	FALSE	FALSE

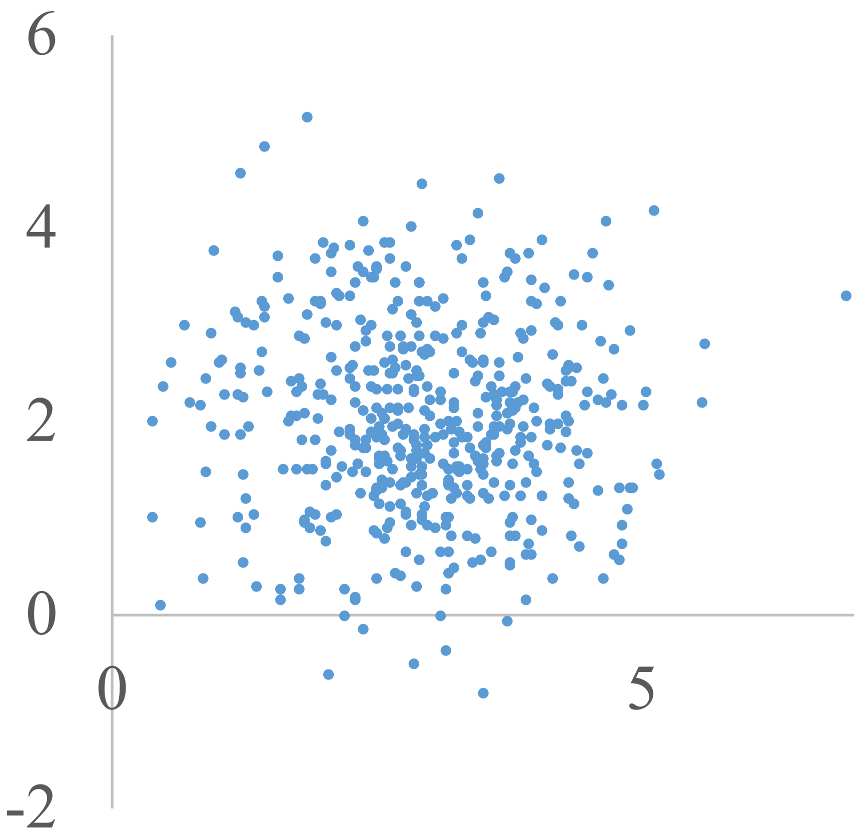
Expression Amount

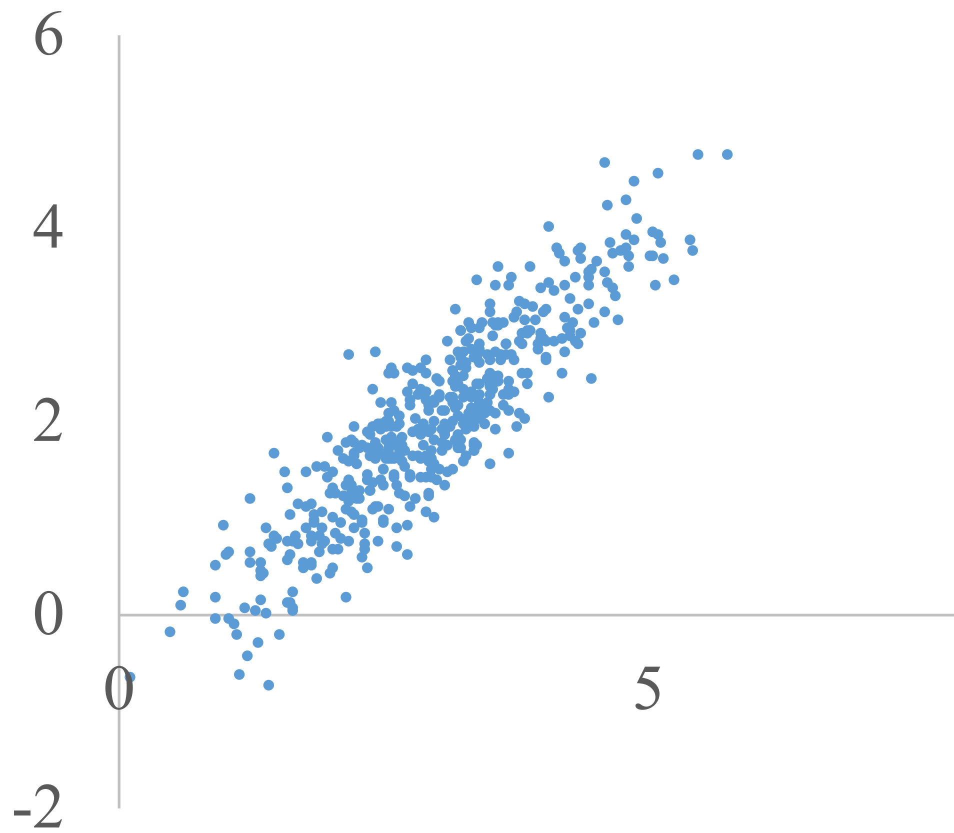
Gene5	Trait
0.76	0.83
0.94	0.85
0.82	0.03
0.94	0.32
0.50	0.10
0.40	0.53
0.90	0.67
0.29	0.71
0.72	0.25
0.15	0.24
0.79	0.98
0.68	0.77
0.71	0.37
0.36	0.18
0.62	0.08
0.59	0.38
0.82	0.76

Spot The Difference

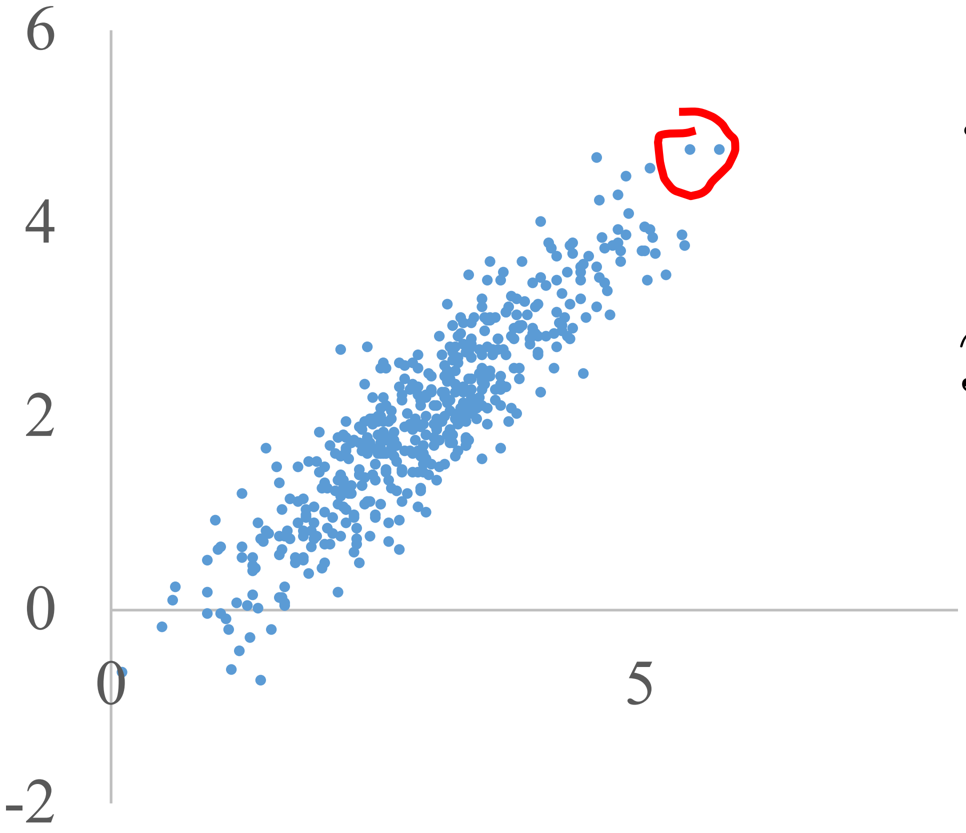


Spot The Difference





Vary Together

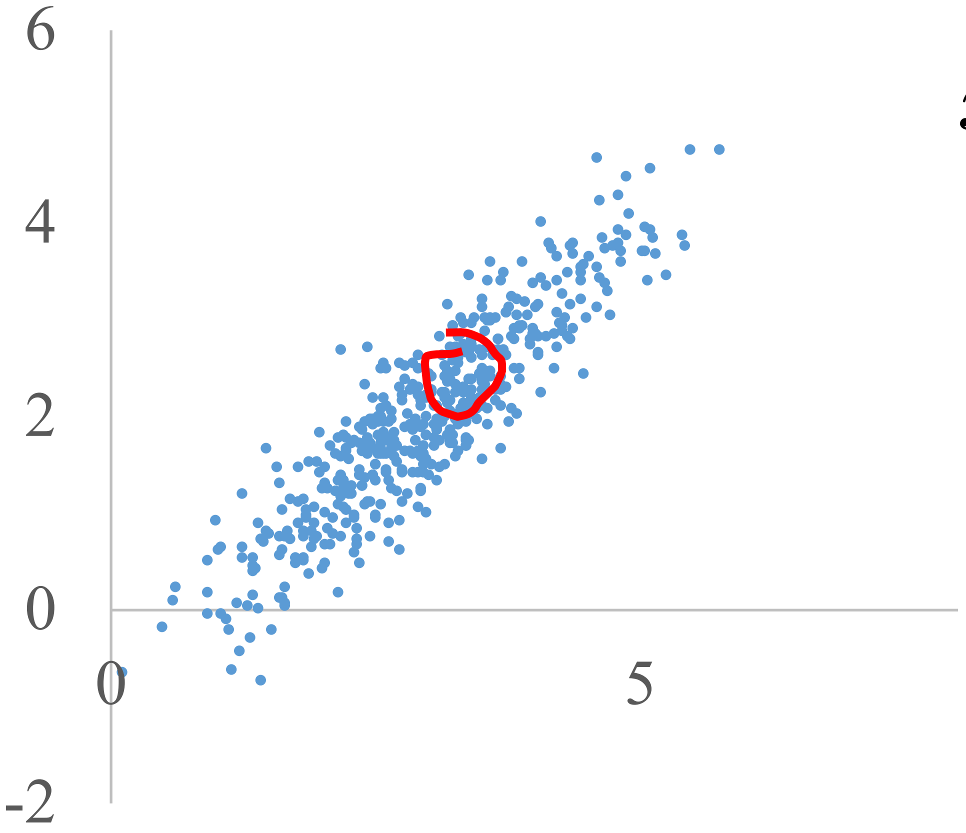


$$x - E[x] = 3$$

$$y - E[y] = 2.6$$

$$(x - E[x])(y - E[y]) = 7.8$$

Vary Together

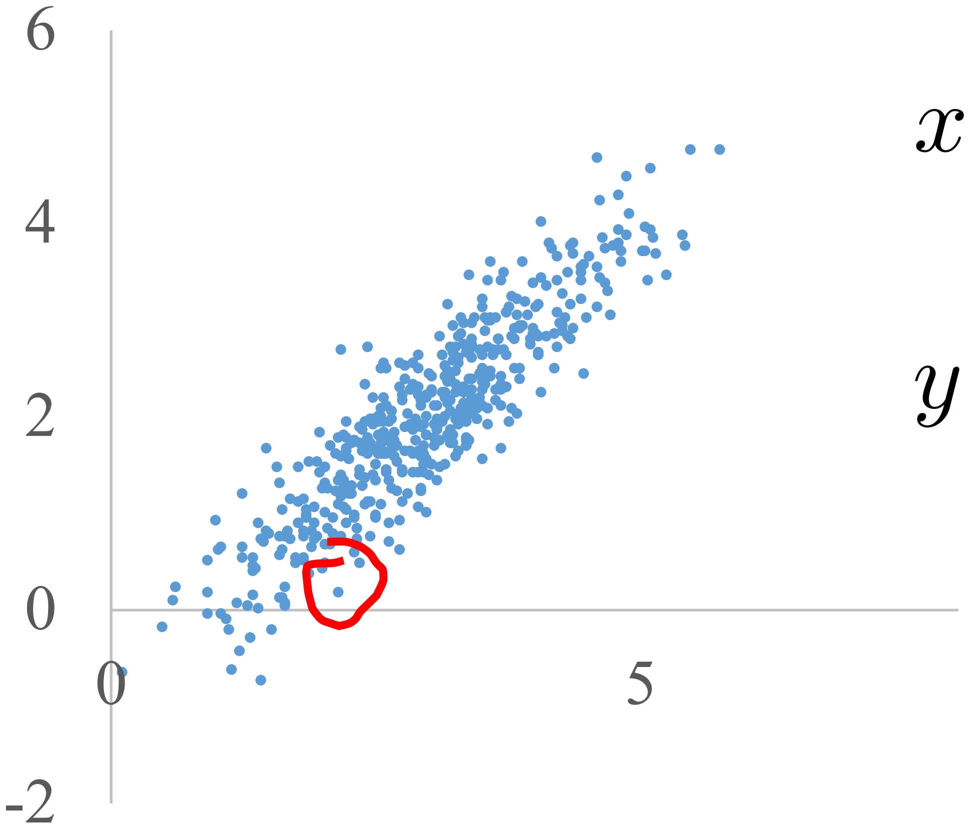


$$x - E[x] \approx 0$$

$$y - E[y] \approx 0$$

$$(x - E[x])(y - E[y]) = 0$$

Vary Together

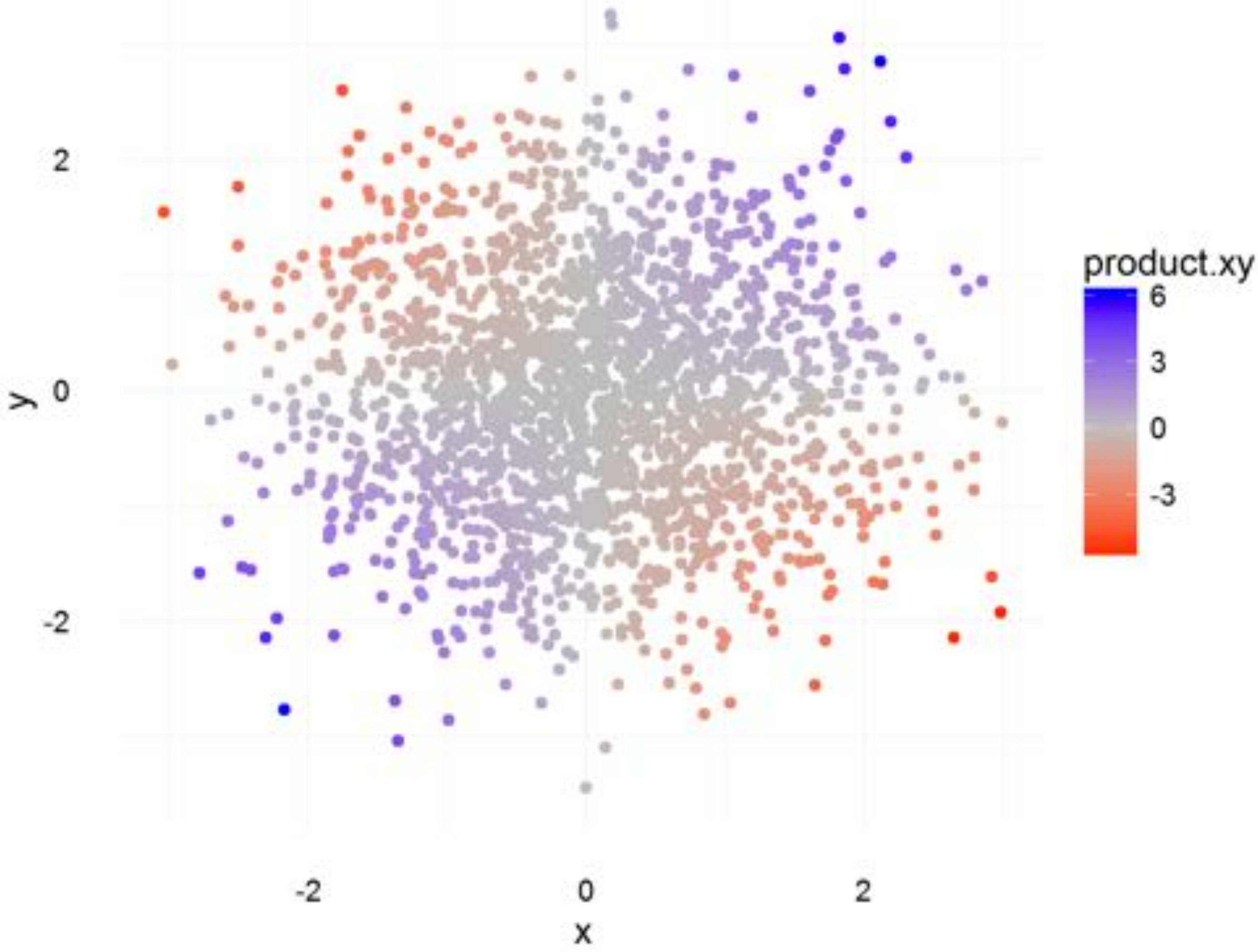


$$x - E[x] = -1.1$$

$$y - E[y] = -2.8$$

$$(x - E[x])(y - E[y]) \approx 3.1$$

Understanding Covariance



The Dance of the Covariance

- Say X and Y are arbitrary random variables
- Covariance of X and Y:

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

x	y	$(x - E[X])(y - E[Y])p(x,y)$
Above mean	Above mean	Positive
Bellow mean	Bellow mean	Positive
Bellow mean	Above mean	Negative
Above mean	Bellow mean	Negative

The Dance of the Covariance

- Say X and Y are arbitrary random variables
- Covariance of X and Y:

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

- Equivalently:

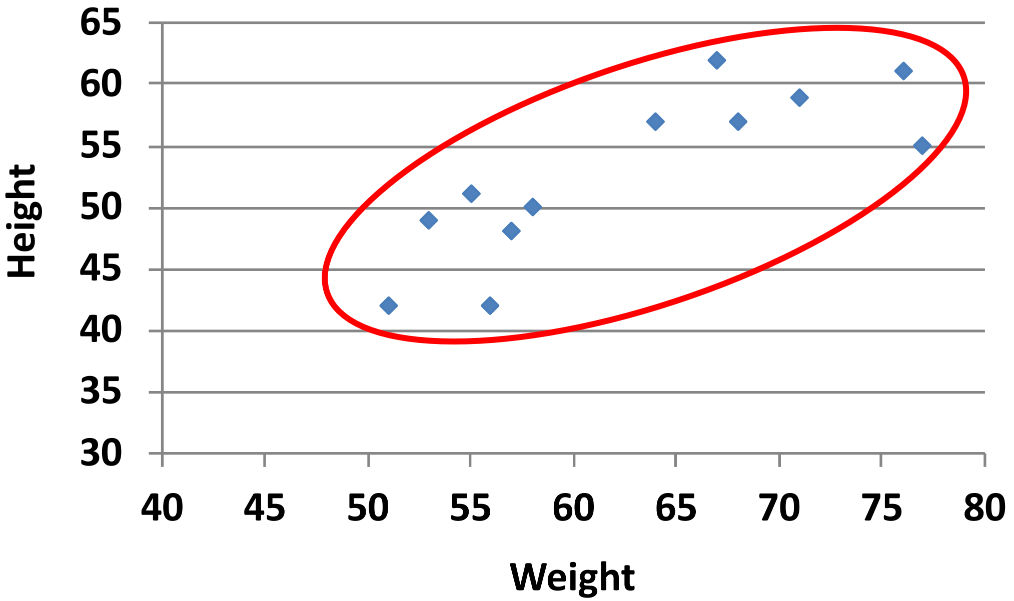
$$\begin{aligned}\mathrm{Cov}(X, Y) &= E[XY - E[X]Y - XE[Y] + E[Y]E[X]] \\ &= E[XY] - E[X]E[Y] - E[X]E[Y] + E[X]E[Y] \\ &= E[XY] - E[X]E[Y]\end{aligned}$$

  - X and Y independent, E[XY] = E[X]E[Y] → Cov(X,Y) = 0
  - But Cov(X,Y) = 0 does **not** imply X and Y independent!

Covariance and Data

- Consider the following data:

Weight	Height	Weight * Height
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

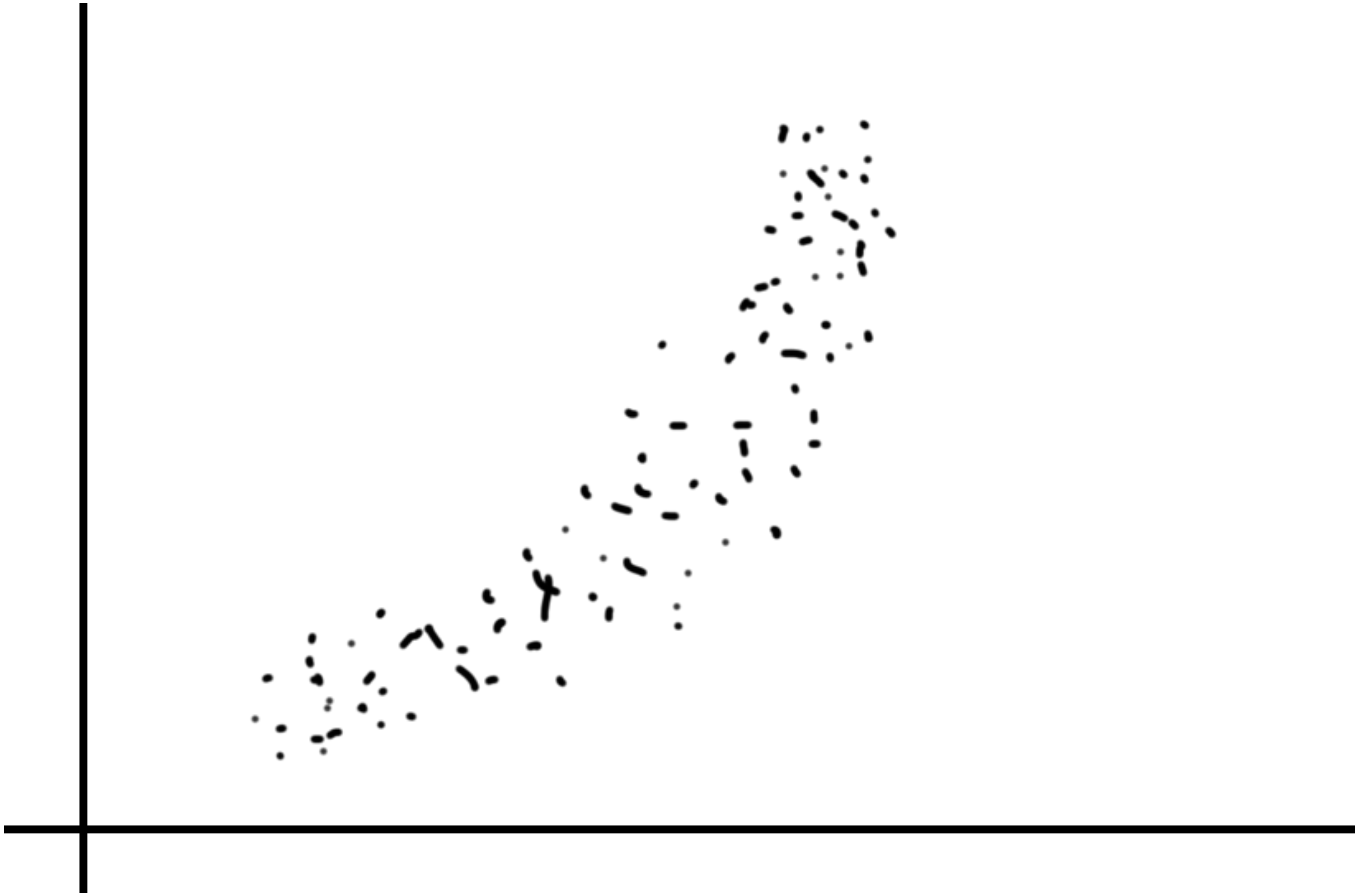


E[W] = 62.75 **E[H] = 52.75** **E[W*H] = 3355.83**

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

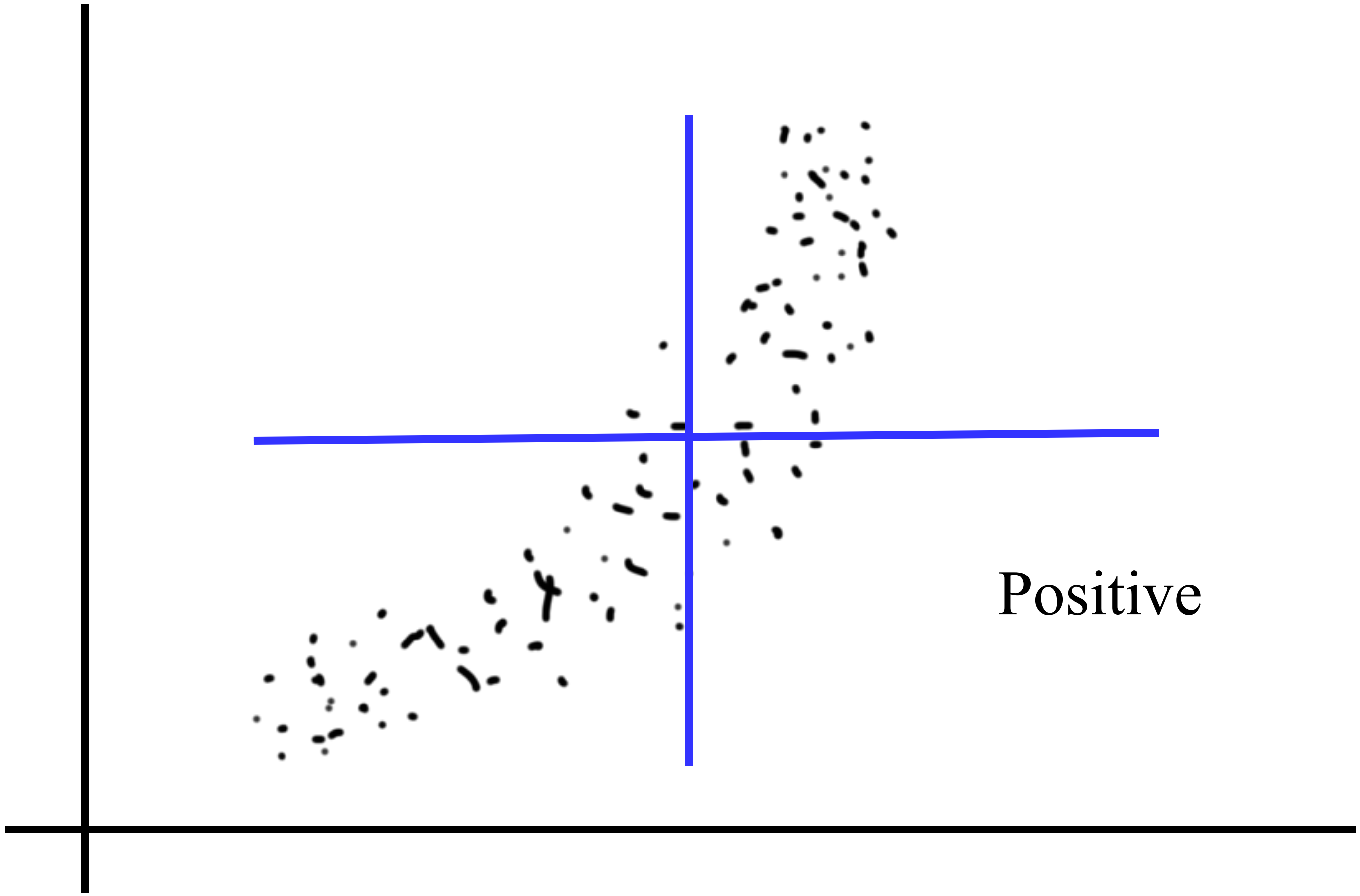
Covariance

Poll: (a) positive, (b) negative, (c) zero



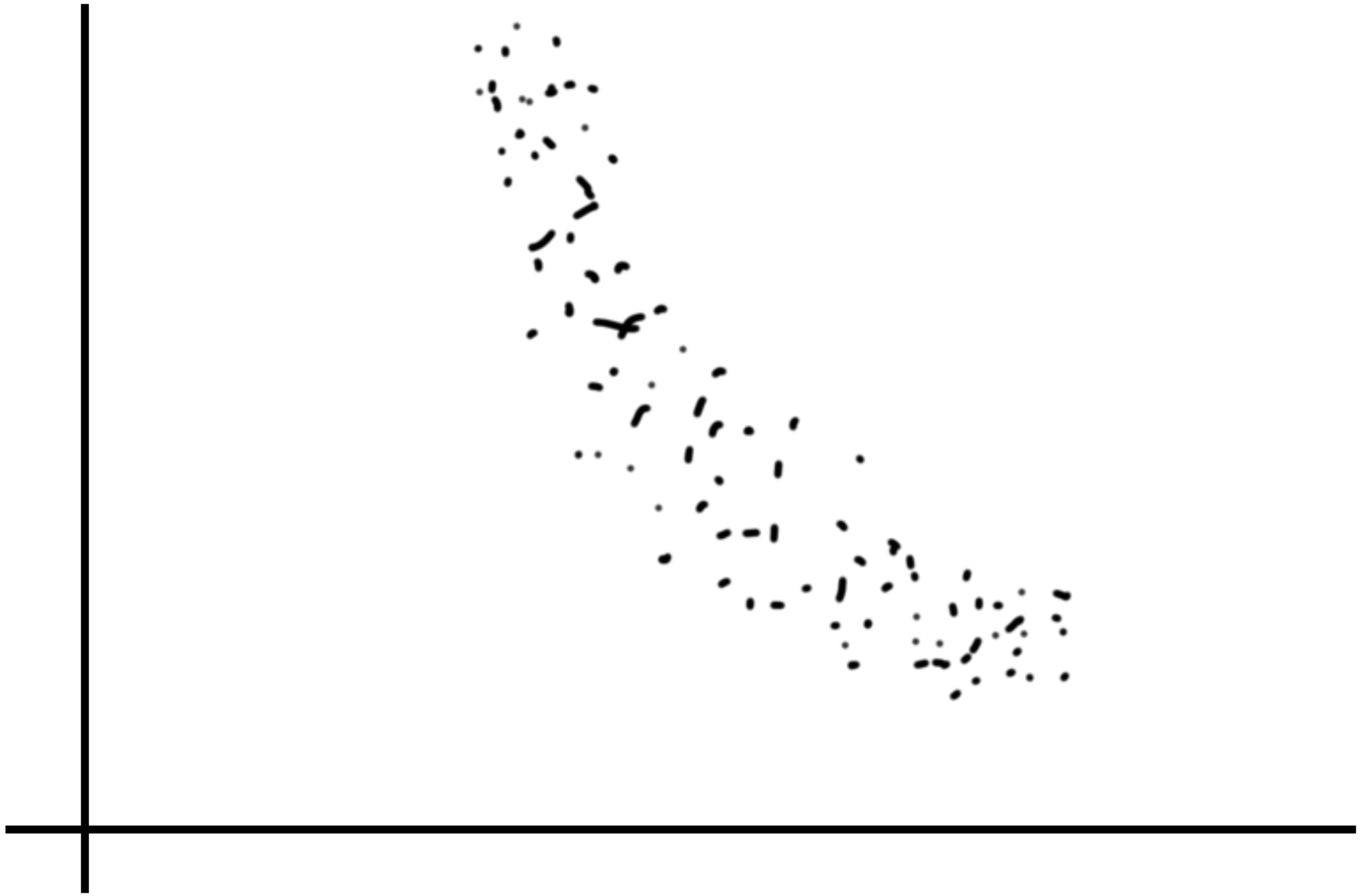
Covariance

Is the Covariance: (a) positive, (b) negative, (c) zero



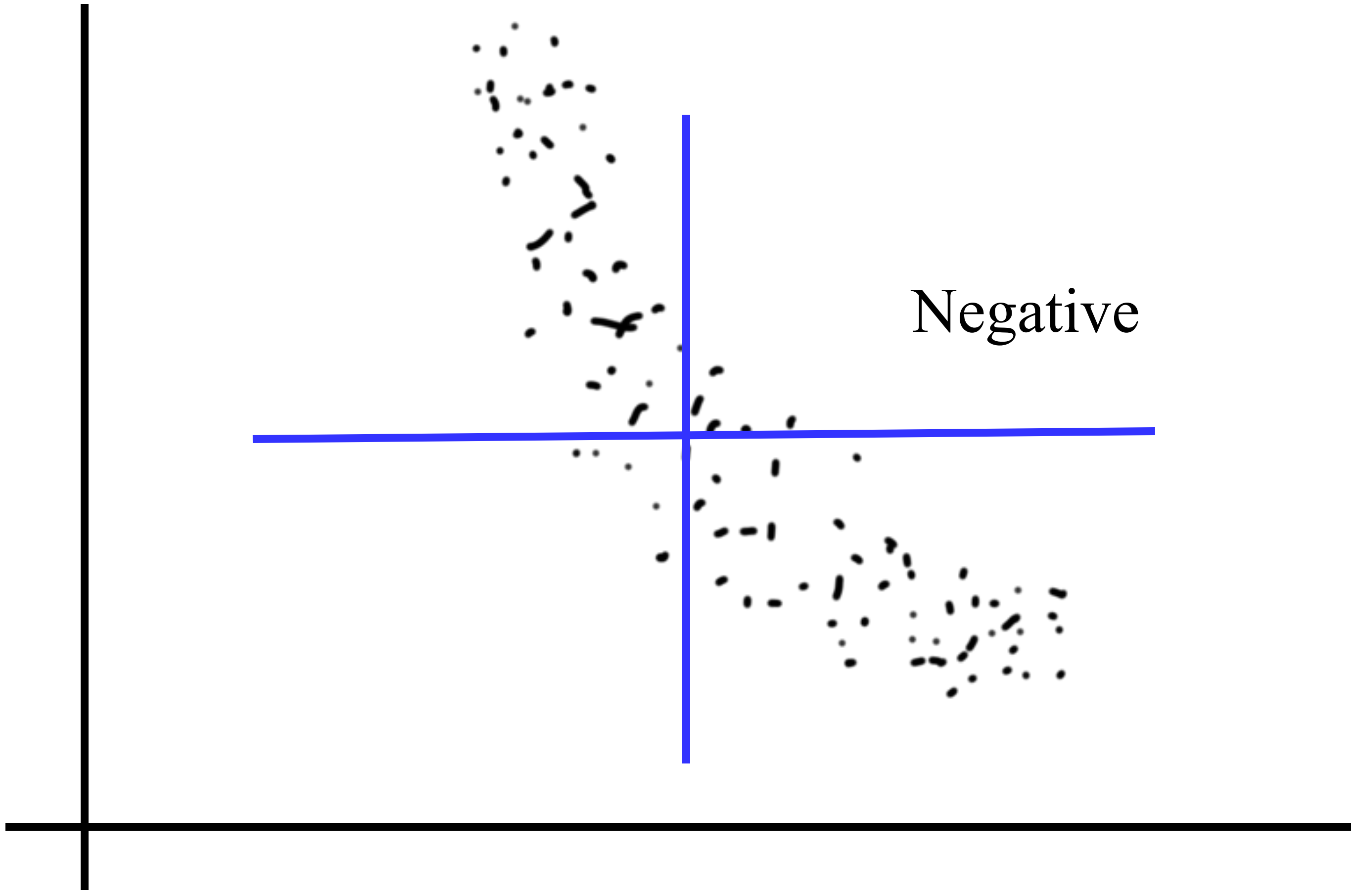
Covariance

Is the Covariance: (a) positive, (b) negative, (c) zero



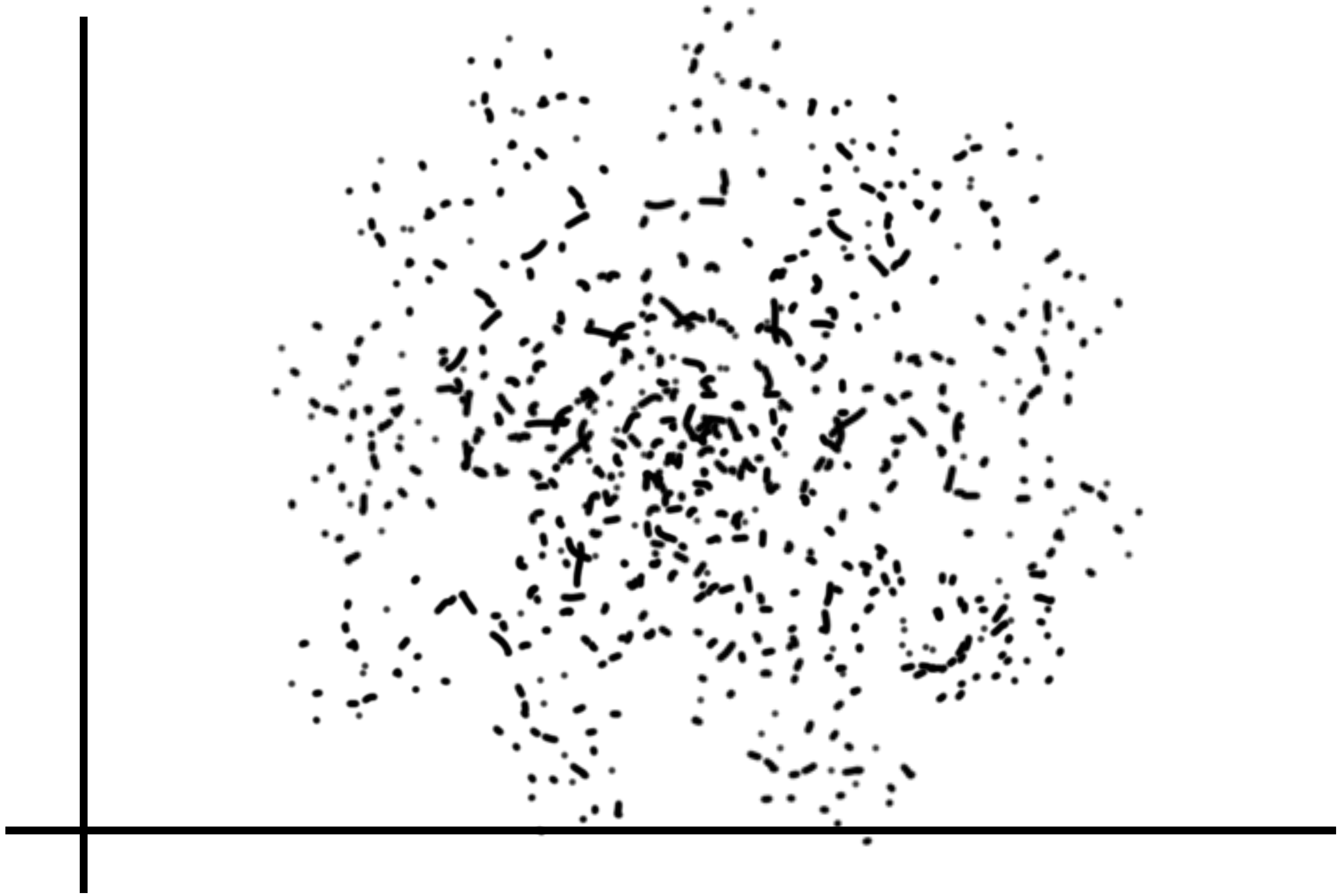
Covariance

Is the Covariance: (a) positive, (b) negative, (c) zero



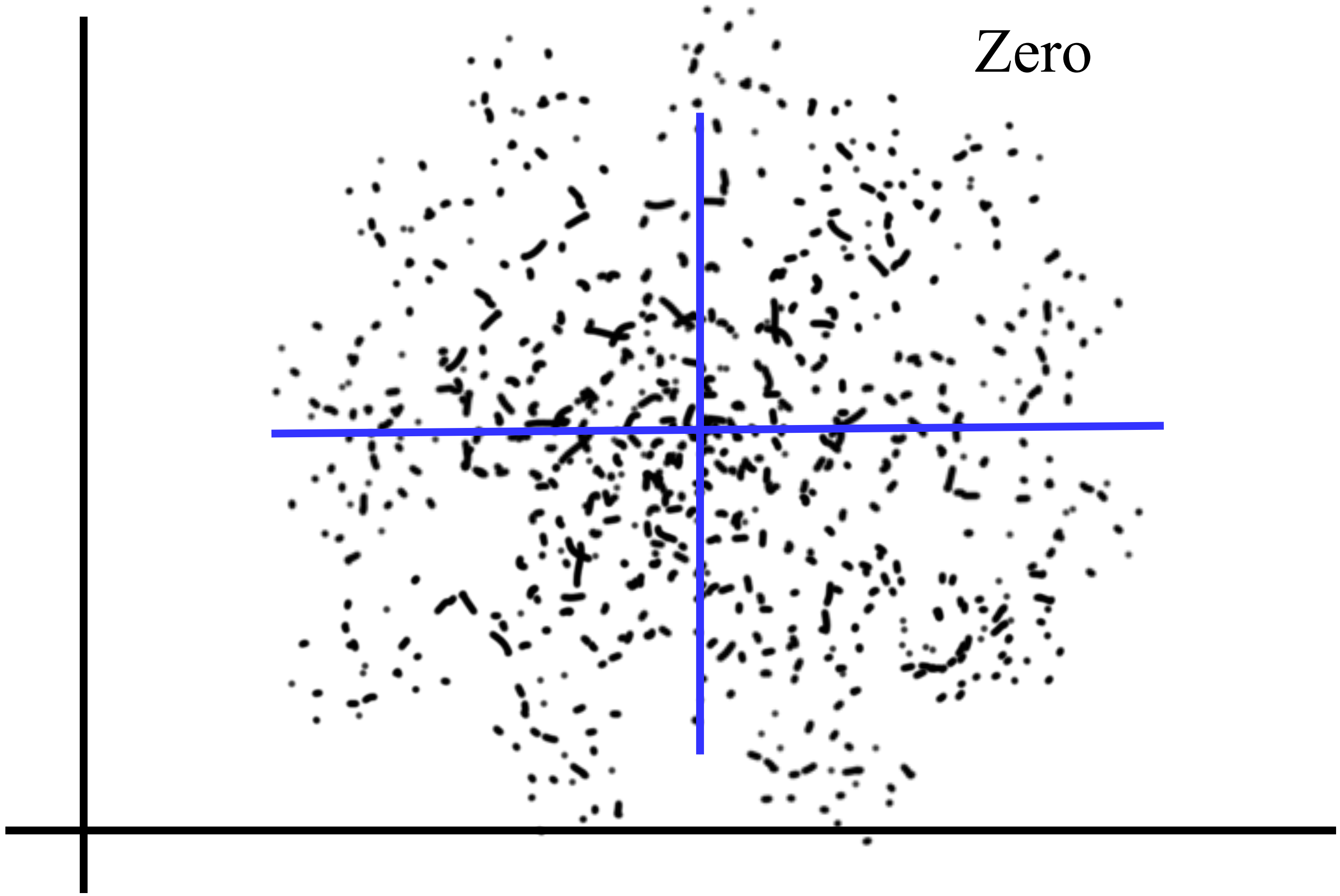
Covariance

Is the Covariance: (a) positive, (b) negative, (c) zero



Covariance

Is the Covariance: (a) positive, (b) negative, (c) zero



Independence and Covariance

- X and Y are random variables with PMF:

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

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

  - E[X] = -1(1/3) + 0(1/3) + 1(1/3) = 0
  - E[Y] = 0(2/3) + 1(1/3) = 1/3
  - Since XY = 0, E[XY] = 0
  - Cov(X, Y) = E[XY] – E[X]E[Y] = 0 – 0 = 0
- But, X and Y are clearly dependent!

Properties of Covariance

- Say X and Y are arbitrary random variables
  - $\text{Cov}(X, Y) = \text{Cov}(Y, X)$
  - $\text{Cov}(X, X) = E[X^2] - E[X]E[X] = \text{Var}(X)$
  - $\text{Cov}(aX + b, Y) = a\text{Cov}(X, Y)$
- Covariance of sums of random variables
  - $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$ are random variables
  - $\text{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} \text{Cov}(X_i, Y_j)$

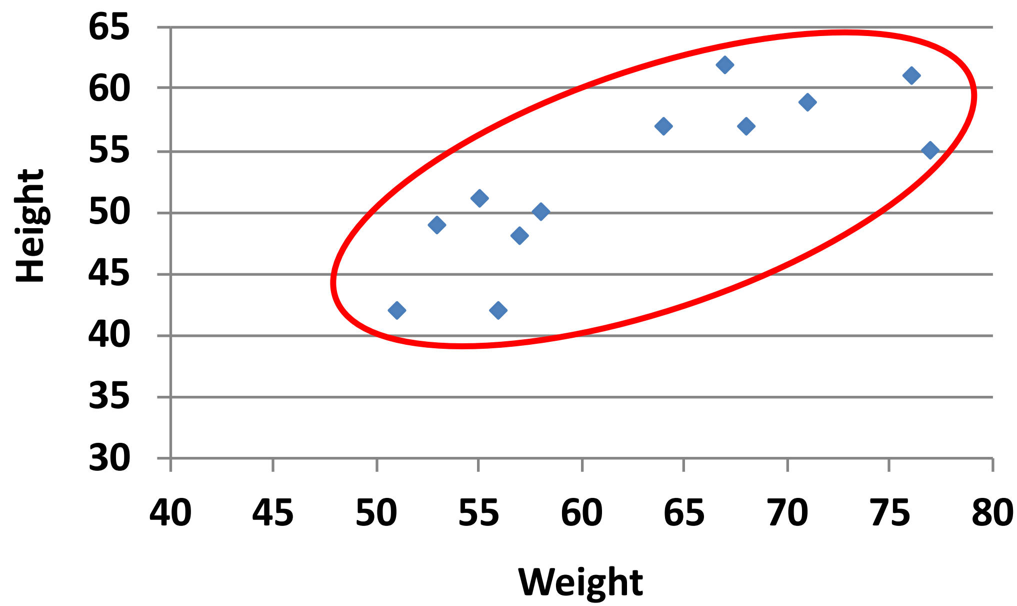
Correlation

What is Wrong With This?

- Consider the following data:

Weight	Height	Weight * Height
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*H] = 3355.83**



Cov(W, H) = E[W*H] – E[W]E[H]
= 3355.83 – (62.75)(52.75)
= 45.77

$$-\mathrm{Std}(X)\mathrm{Std}(Y) \leq \mathrm{Cov}(X,Y) \leq \mathrm{Std}(X)\mathrm{Std}(Y)$$

Viva La Correlatión

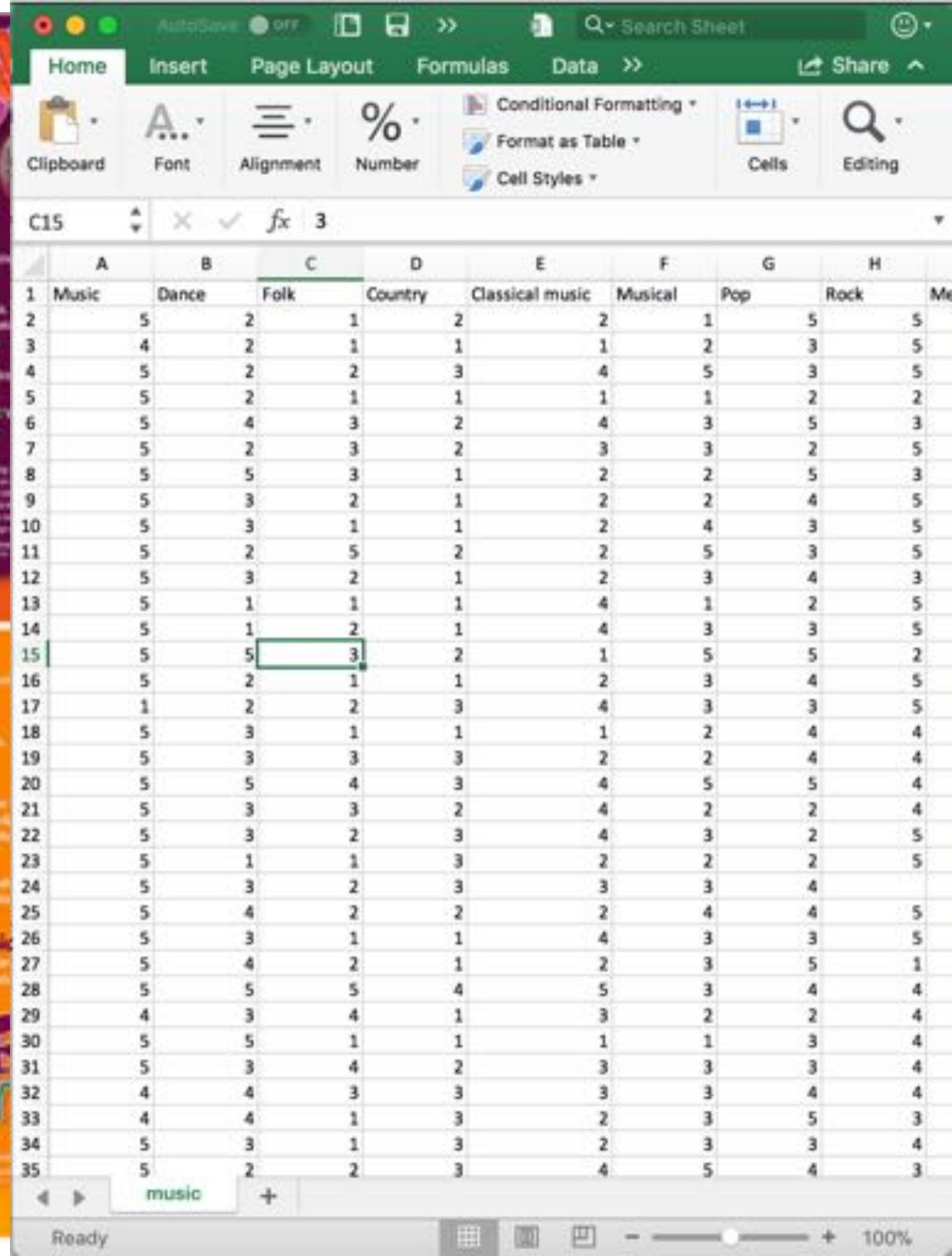
- Say X and Y are arbitrary random variables
  - Correlation of X and Y, denoted $\rho(X, Y)$:
$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var(X)Var(Y)}}}$$
  - Note: $-1 \le \rho(X, Y) \le 1$
  - Correlation measures <u>linearity</u> between X and Y
  - $\rho(X, Y) = 1 \quad \Rightarrow \quad Y = aX + b \quad$ where $a = \sigma_y/\sigma_x$
  - $\rho(X, Y) = -1 \quad \Rightarrow \quad Y = aX + b \quad$ where $a = -\sigma_y/\sigma_x$
  - $\rho(X, Y) = 0 \quad \Rightarrow$ absence of <u>linear</u> relationship
    - But, X and Y can still be related in some other way!
  - If $\rho(X, Y) = 0$, we say X and Y are "uncorrelated"
    - Note: Independence implies uncorrelated, but <u>**not**</u> vice versa!

Viva La Correlatión

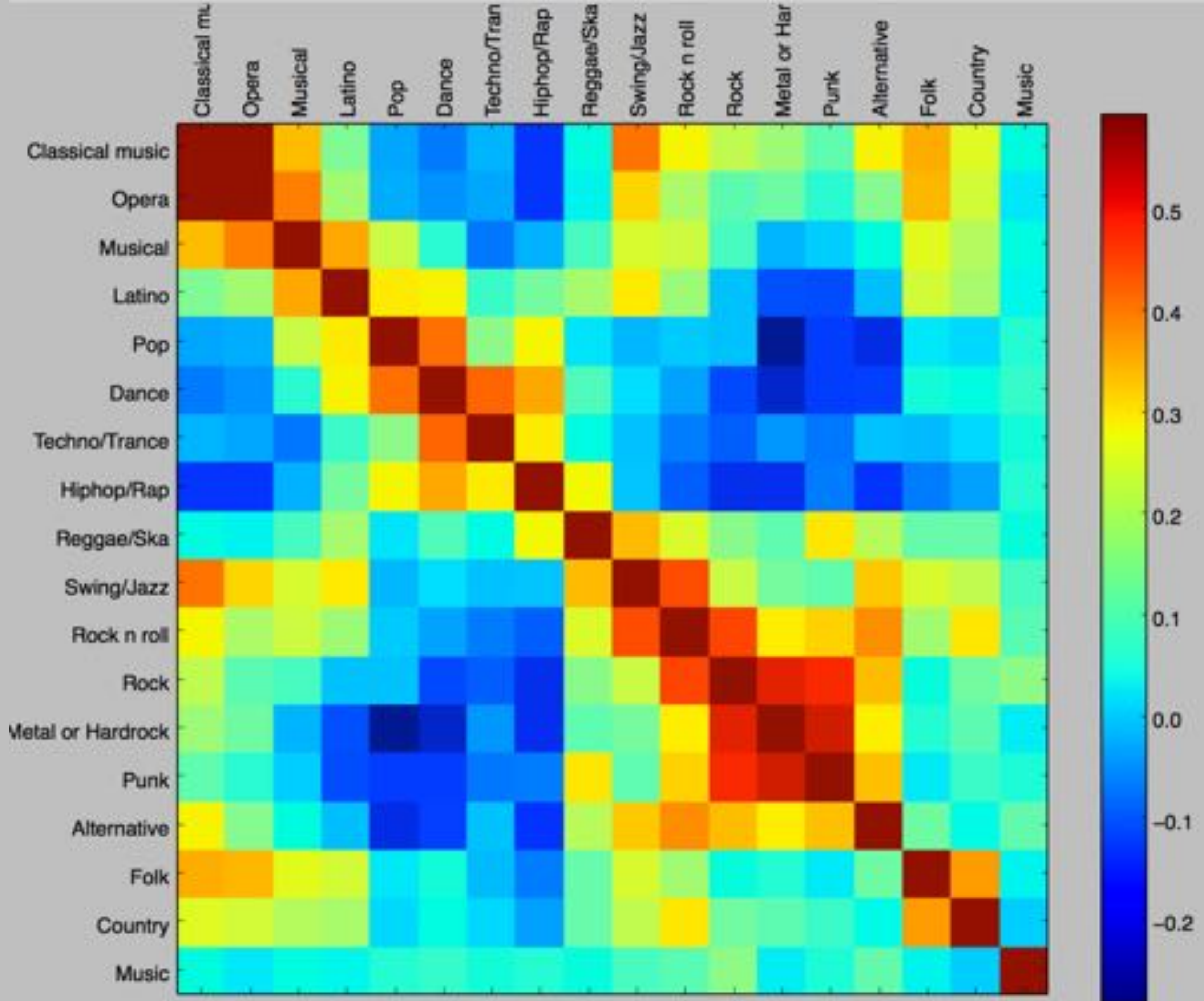
- Say X and Y are arbitrary random variables
  - Correlation of X and Y, denoted $\rho(X, Y)$:

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

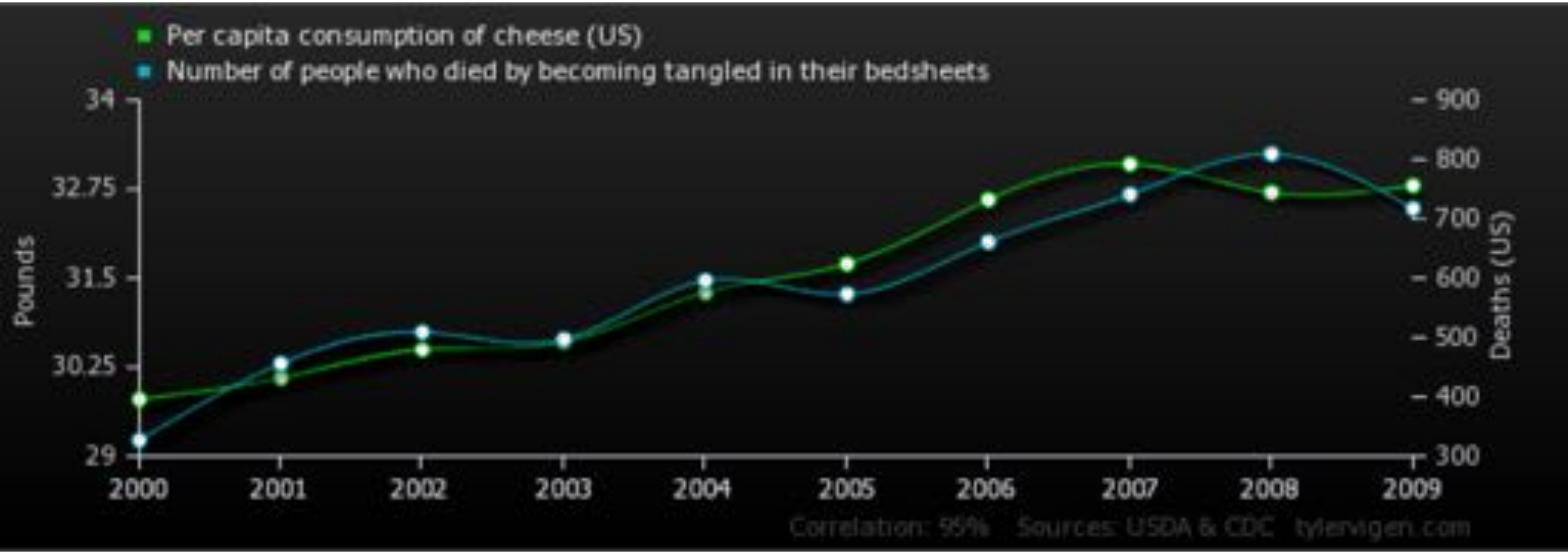
- Say Y = cX. Correlation should be 1.



	A	B	C	D	E	F	G	H
1	Music	Dance	Folk	Country	Classical music	Musical	Pop	Rock
2	5	2	1	2	2	1	5	5
3	4	2	1	1	1	2	3	5
4	5	2	2	3	4	5	3	5
5	5	2	1	1	1	1	2	2
6	5	4	3	2	4	3	5	3
7	5	2	3	2	3	3	2	5
8	5	5	3	1	2	2	5	3
9	5	3	2	1	2	2	4	5
10	5	3	1	1	2	4	3	5
11	5	2	5	2	2	5	3	5
12	5	3	2	1	2	3	4	3
13	5	1	1	1	4	1	2	5
14	5	1	2	1	4	3	3	5
15	5	5	3	2	1	5	5	2
16	5	2	1	1	2	3	4	5
17	1	2	2	3	4	3	3	5
18	5	3	1	1	1	2	4	4
19	5	3	3	3	2	2	4	4
20	5	5	4	3	4	5	5	4
21	5	3	3	2	4	2	2	4
22	5	3	2	3	4	3	2	5
23	5	1	1	3	2	2	2	5
24	5	3	2	3	3	3	4	
25	5	4	2	2	2	4	4	5
26	5	3	1	1	4	3	3	5
27	5	4	2	1	2	3	5	1
28	5	5	5	4	5	3	4	4
29	4	3	4	1	3	2	2	4
30	5	5	1	1	1	1	3	4
31	5	3	4	2	3	3	3	4
32	4	4	3	3	3	3	4	4
33	4	4	1	3	2	3	5	3
34	5	3	1	3	2	3	3	4
35	5	2	2	3	4	5	4	3


Classical mu
Opera
Musical
Latino
Pop
Dance
Techno/Tran
Hiphop/Rap
Reggae/Ska
Swing/Jazz
Rock n roll
Rock
Metal or Har
Punk
Alternative
Folk
Country
Music
Classical music
Opera
Musical
Latino
Pop
Dance
Techno/Trance
Hiphop/Rap
Reggae/Ska
Swing/Jazz
Rock n roll
Rock
Metal or Hardrock
Punk
Alternative
Folk
Country
Music
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2

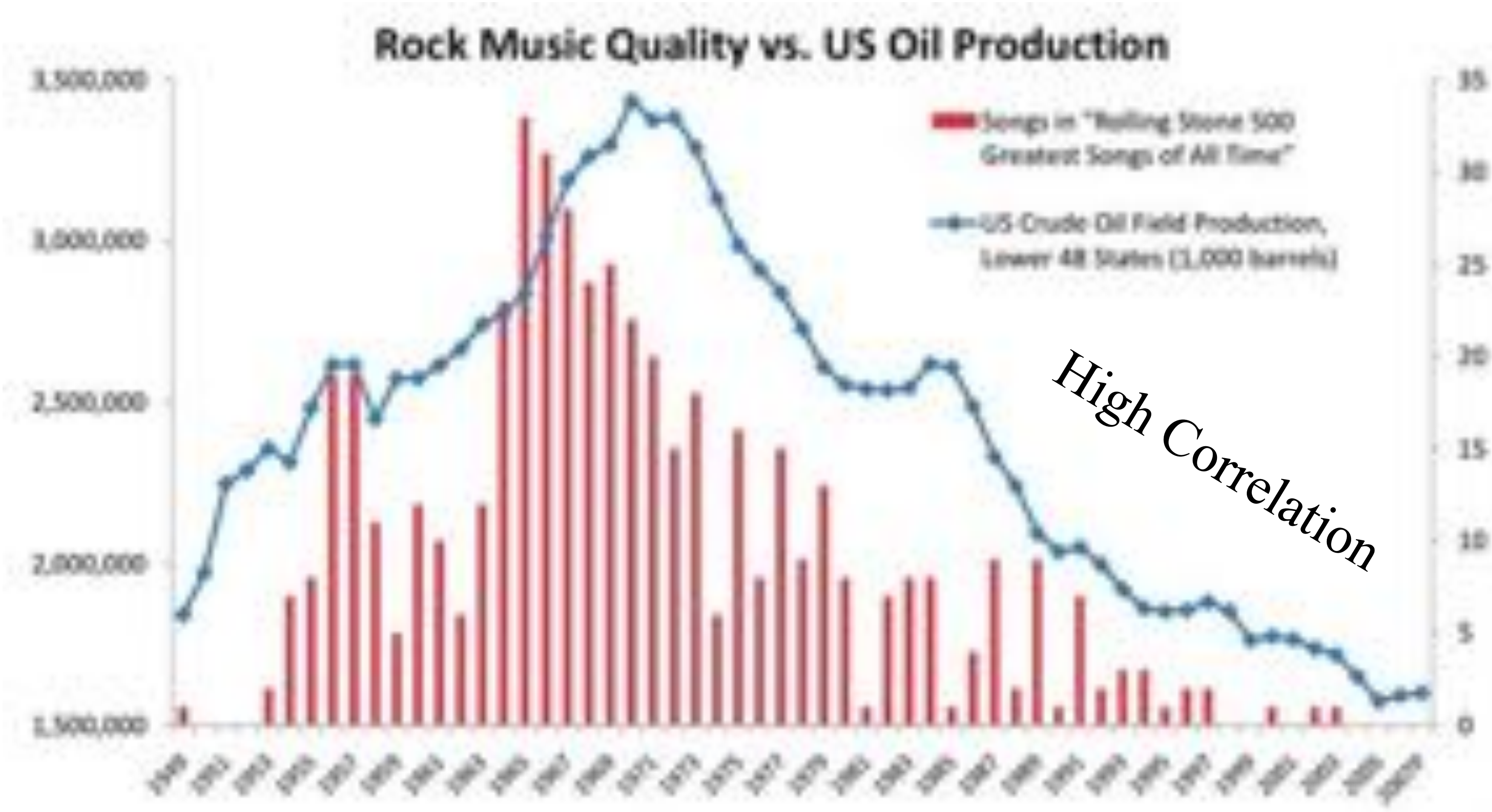
Tell your friends!



	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Per capita consumption of cheese (US) *Pounds (USDA)*	29.8	30.1	30.5	30.6	31.3	31.7	32.6	33.1	32.7	32.8
Number of people who died by becoming tangled in their bedsheets *Deaths (US) (CDC)*	327	456	509	497	596	573	661	741	809	717

Correlation: 0.947091

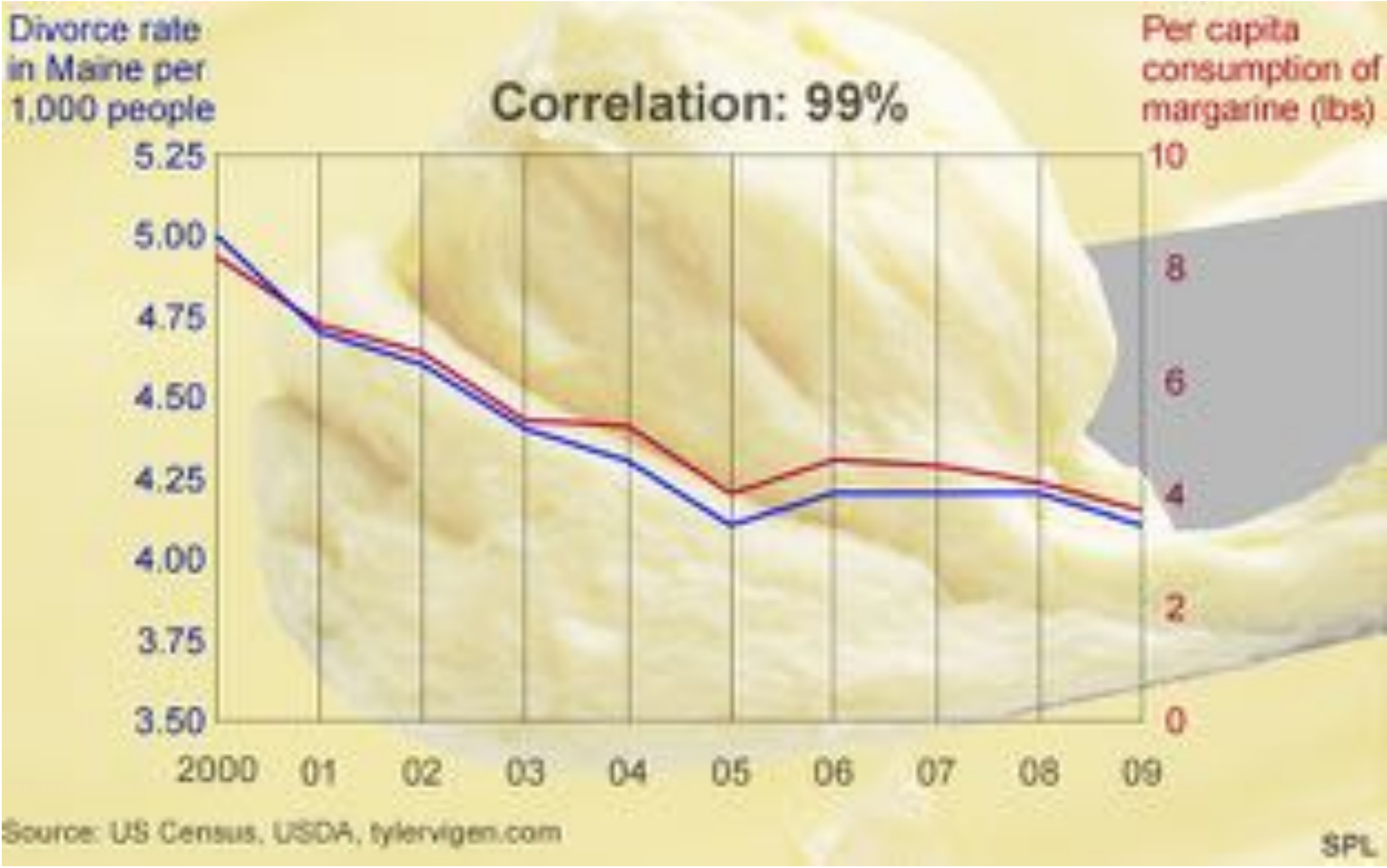
Rock Music Vs Oil?



Hubbert Peak Theory

http://www.aei.org/publication/blog/

Divorce Vs Butter?



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