13: Statistics of Multiple RVs

Jerry Cain
April 25, 2022

Table of Contents

Expectation of Common RVs

Linearity of Expectation is useful

Expectation is a linear mathematical operation. If $X = \sum_{i=1}^{n} X_i$:

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

- Even if you don't know the **distribution** of X (e.g., because the joint distribution of $(X_1, \ldots, X_n)$ is unknown), you can still compute **expectation** of X.

- Problem-solving key: Define X_i such that $X = \sum_{i=1}^{n} X_i$

Most common use cases:
- $E[X_i]$ easy to calculate
- Sum of dependent RVs

Expectations of common RVs: Binomial

Review

$X \sim \text{Bin}(n, p) \quad E[X] = np$

# of successes in n independent trials with probability of success p

Recall: $\text{Bin}(1, p) = \text{Ber}(p)$

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

Let $X_i = i$th trial is heads
$X_i \sim \text{Ber}(p), E[X_i] = p$

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

Expectations of common RVs: Negative Binomial

$Y \sim \text{NegBin}(r, p) \quad E[Y] = \frac{r}{p}$

# of independent trials with probability of success p until r successes

Recall: $\text{NegBin}(1, p) = \text{Geo}(p)$

$$Y = \sum_{i=1}^{?} Y_i$$

1. How should we define Y_i?
2. How many terms are in our summation?

Expectations of common RVs: Negative Binomial

$Y \sim \text{NegBin}(r, p) \quad E[Y] = \frac{r}{p}$

# of independent trials with probability of success p until r successes

Recall: $\text{NegBin}(1, p) = \text{Geo}(p)$

Y_i = #trial to produce ith success since previous one

$$Y \overset{?}{=} \sum_{i=1}^{r} Y_i$$

Let Y_i = # trials to get ith success (after $(i-1)$th success)

$Y_i \sim \text{Geo}(p), E[Y_i] = \frac{1}{p}$

$$E[Y] = E\left[\sum_{i=1}^{r} Y_i\right] = \sum_{i=1}^{r} E[Y_i] = \sum_{i=1}^{r} \frac{1}{p} = \frac{r}{p}$$

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.

Servers

requests

k servers

request to server 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?

* 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 ≥ 1 request.

What is $E[X]$?

Computer cluster utilization

$X_i = \begin{cases} 1 & \text{if } A_i = H \\ 0 & \text{if } A_i = 0 \end{cases}$

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

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 ≥ 1 request.

What is $E[X]$?

1. Define additional random variables.
2. Solve.

Let: A_i = event that server i receives ≥ 1 request

X_i = indicator for A_i

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

Note: A_i are dependent!

$$E[X_i] = P(A_i) = 1 - (1 - p_i)^n$$

$$E[X] = E\left[\sum_{i=1}^{k} X_i\right] = \sum_{i=1}^{k} E[X_i] = \sum_{i=1}^{k} (1 - (1 - p_i)^n)$$

$$= \sum_{i=1}^{k} 1 - \sum_{i=1}^{k} (1 - p_i)^n = k - \sum_{i=1}^{k} (1 - p_i)^n$$

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.

Servers	Hash Tables
requests	strings
k servers	k buckets
request to server i	hashed to bucket 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 ≥ 1 string?

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 ≥ 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 ≥ 1 string.

What is $E[Y]$?

1. Define additional random variables.

Let: Y_i = # of trials 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.

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

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 ≥ 1 string.

What is $E[Y]$?

1. Define additional random variables.

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

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

2. Solve.

$$Y = Y_0 + Y_1 + \cdots + Y_{k-1}$$

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

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

discrete math equiv. $\int \frac{1}{x}\,dx$ — Harmonic

$H_k \sim O(\ln k)$

Covariance

Statistics of sums of RVs

For any random variables X and Y,

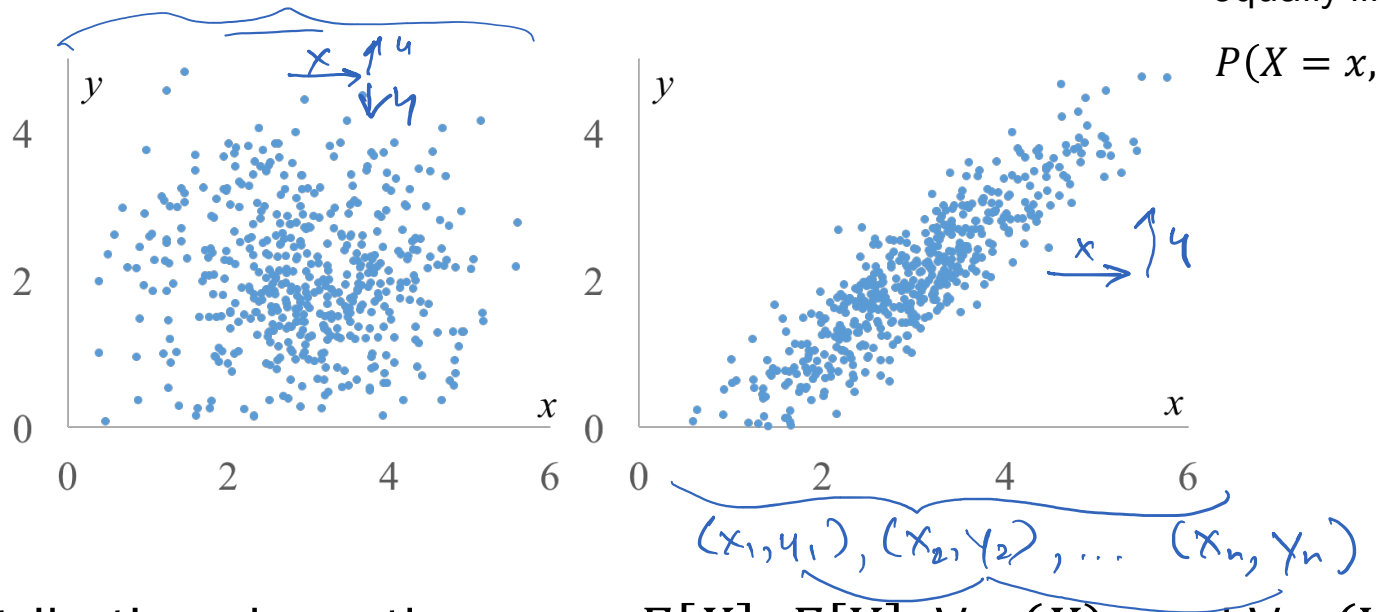
$$E[X + Y] = E[X] + E[Y]$$

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

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:

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

(linearity of expectation)

($E[X]$, $E[Y]$ are scalars)

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.

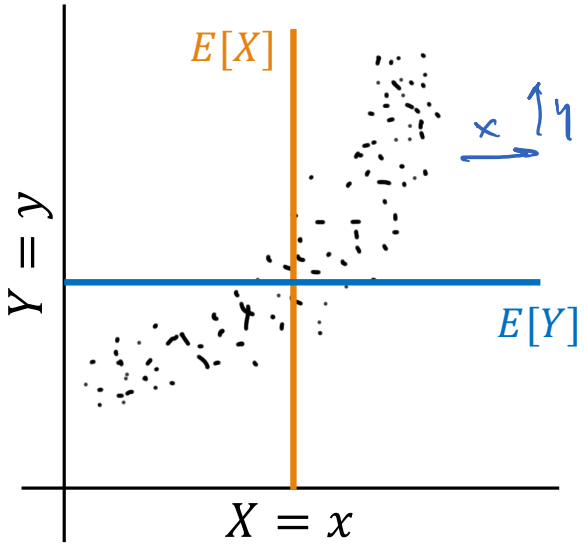
- Outside temperature and utility bills have a negative covariance.
- Handedness and musical ability have near zero covariance.
- Product demand and price have a positive covariance.

Feel the covariance

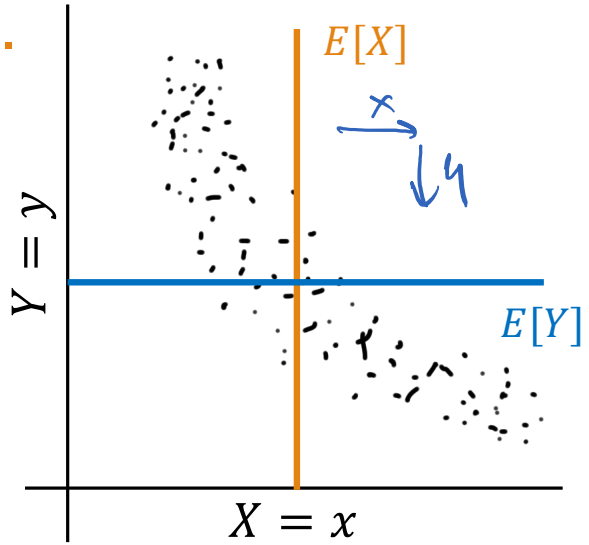
$$\begin{aligned}\text{Cov}(X, Y) &= E[(X - E[X])(Y - E[Y])] \\ &= E[XY] - E[X]E[Y]\end{aligned}$$

Is the covariance positive, negative, or zero?

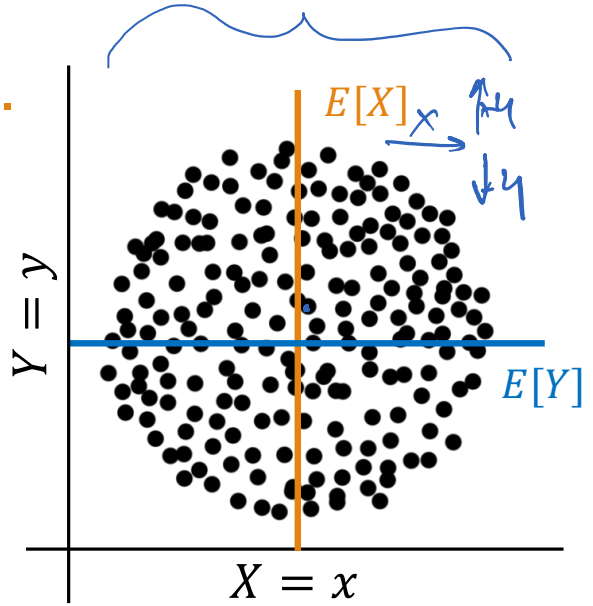
1.

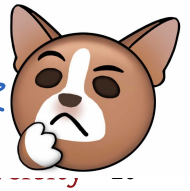


2.



3.



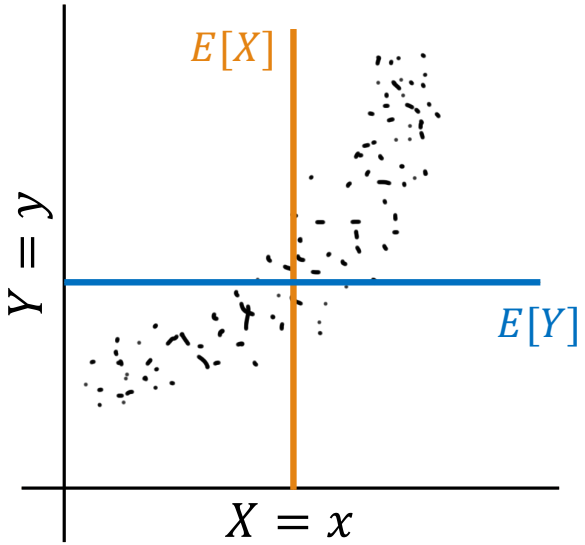


Feel the covariance

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

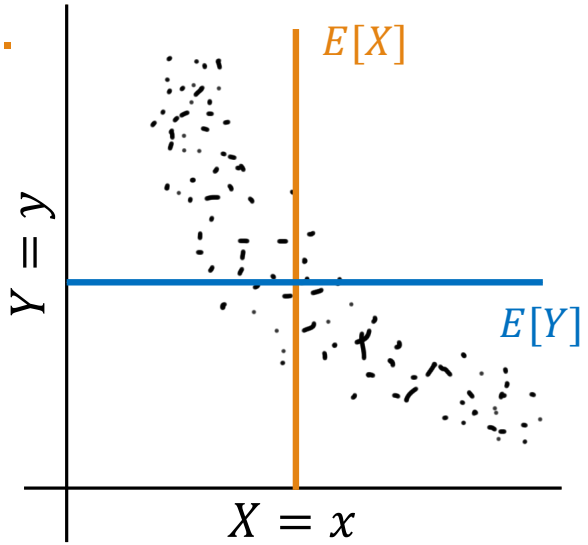
Is the covariance positive, negative, or zero?

1.



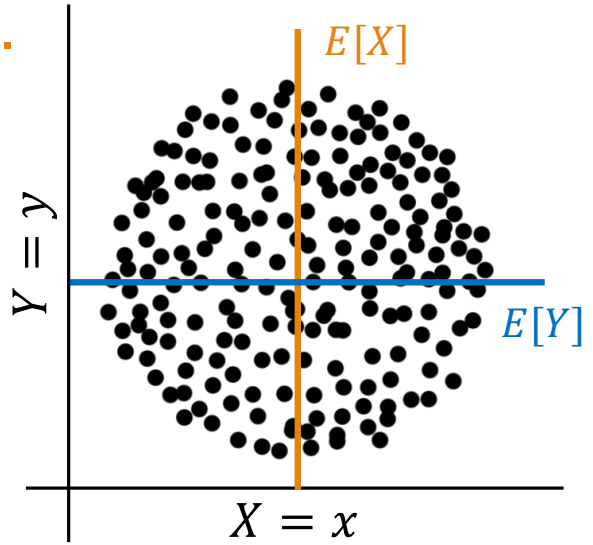
positive

2.



negative

3.



zero

Covarying humans

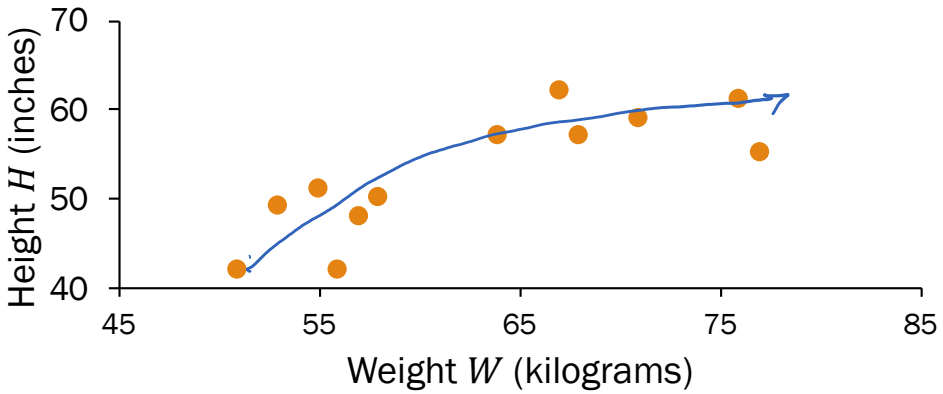
$$\text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])]$$
$$= E[XY] - E[X]E[Y]$$

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?

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



Covariance > 0: one variable ↑, other variable ↑

Properties of Covariance

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

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

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

Properties:

1. $\text{Cov}(X,Y) = \text{Cov}(Y,X)$
2. $\text{Var}(X) = E[X^2] - (E[X])^2 = \text{Cov}(X,X)$
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 is non-linear: $\text{Cov}(aX + b, Y) = a\text{Cov}(X,Y)$

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 $1/3$.

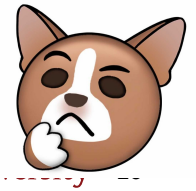
Define $Y = \begin{cases} 1 & \text{if } X = 0 \\ 0 & \text{otherwise} \end{cases}$

Y \ X	-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. $E[X] =$ 0 $E[Y] =$
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 $Y = \begin{cases} 1 & \text{if } X = 0 \\ 0 & \text{otherwise} \end{cases}$

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

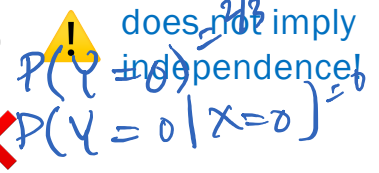
Marginal PMF of Y, $p_Y(y)$

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



4. Are X and Y independent? ✗

$P(Y = 0|X = 1) = 1$

$\neq \quad P(Y = 0) = 2/3$

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:

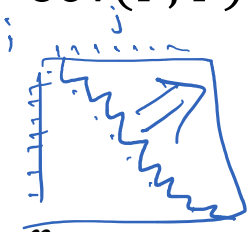
$$\text{Var}(X + Y) = \text{Cov}(X + Y, X + Y)$$

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

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

covariance of all pairs

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



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

More generally:

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

(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,

LOTUS 2D

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

(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:

1. $\text{Cov}(X, Y) = E[XY] - E[X]E[Y]$ — def. of covariance
   $= E[X]E[Y] - E[X]E[Y]$ — X and Y are independent
   $= 0$

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

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

Proving Variance of the Binomial

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

To simplify the algebra a bit, let $q = 1 - p$, so $p + q = 1$.

So:

$$\begin{aligned}
\mathrm{E}(X^2) &= \sum_{k \geq 0}^{n} k^2 \binom{n}{k} p^k q^{n-k} && \text{Definition of Binomial Distribution: } p + q = 1 \\
&= \sum_{k=0}^{n} kn \binom{n-1}{k-1} p^k q^{n-k} && \text{Factors of Binomial Coefficient: } k\binom{n}{k} = n\binom{n-1}{k-1} \\
&= np \sum_{k=1}^{n} k \binom{n-1}{k-1} p^{k-1} q^{(n-1)-(k-1)} && \text{Change of limit: term is zero when } k - 1 = 0 \\
&= np \sum_{j=0}^{m} (j+1) \binom{m}{j} p^j q^{m-j} && \text{putting } j = k - 1, m = n - 1 \\
&= np \left(\sum_{j=0}^{m} j \binom{m}{j} p^j q^{m-j} + \sum_{j=0}^{m} \binom{m}{j} p^j q^{m-j} \right) && \text{splitting sum up into two} \\
&= np \left(\sum_{j=0}^{m} m \binom{m-1}{j-1} p^j q^{m-j} + \sum_{j=0}^{m} \binom{m}{j} p^j q^{m-j} \right) && \text{Factors of Binomial Coefficient: } j\binom{m}{j} = m\binom{m-1}{j-1} \\
&= np \left((n-1)p \sum_{j=1}^{m} \binom{m-1}{j-1} p^{j-1} q^{(m-1)-(j-1)} + \sum_{j=0}^{m} \binom{m}{j} p^j q^{m-j} \right) && \text{Change of limit: term is zero when } j - 1 = 0 \\
&= np\left((n-1)p(p+q)^{m-1} + (p+q)^m \right) && \text{Binomial Theorem} \\
&= np((n-1)p + 1) && \text{as } p + q = 1 \\
&= n^2 p^2 + np(1-p) && \text{by algebra}
\end{aligned}$$

Then:

$$\begin{aligned}
\mathrm{var}(X) &= \mathrm{E}(X^2) - (\mathrm{E}(X))^2 \\
&= np(1-p) + n^2 p^2 - (np)^2 && \text{Expectation of Binomial Distribution: } \mathrm{E}(X) = np \\
&= np(1-p)
\end{aligned}$$

as required.

proofwiki.org

Let's instead prove this using independence and variance!

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)

$$\text{Var}(X) = \text{Var}\left(\sum_{i=1}^{n} X_i\right)$$

$$= \sum_{i=1}^{n} \text{Var}(X_i)$$

$$= \sum_{i=1}^{n} p(1-p)$$

$$= np(1-p)$$

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

Variance of Bernoulli

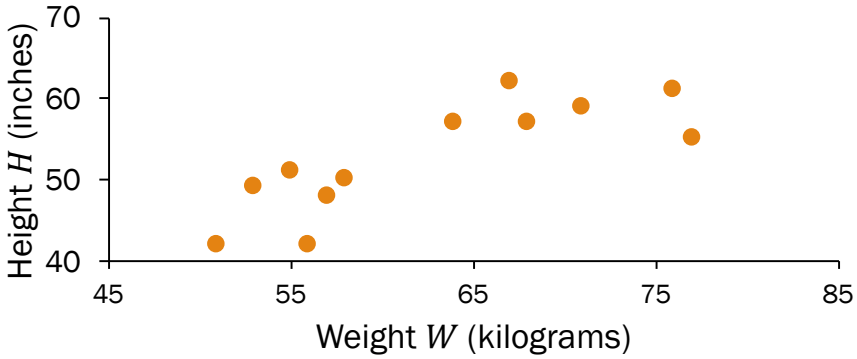
Correlation

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?

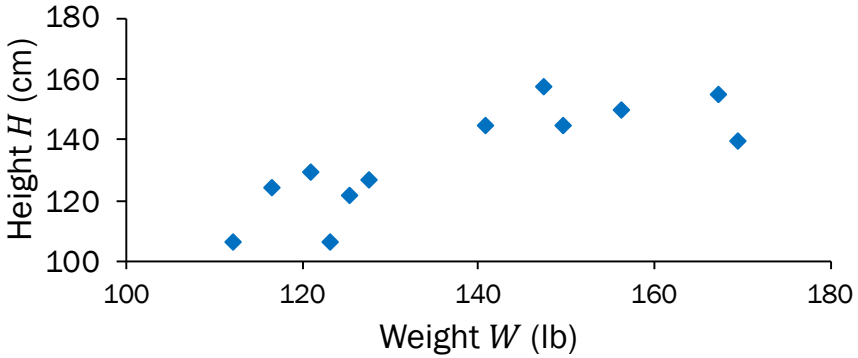
$$\begin{aligned}\text{Cov}(W, H) &= E[WH] - E[W]E[H] \\ &= 3355.83 - (62.75)(52.75) \\ &= 45.77 \quad \text{(positive)}\end{aligned}$$



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

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

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

unitless

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

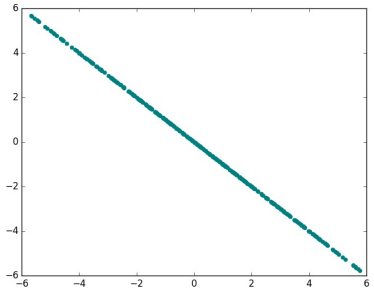
- Note: $-1 \leq \rho(X,Y) \leq 1$
- Correlation measures the **linear relationship** between X and Y:

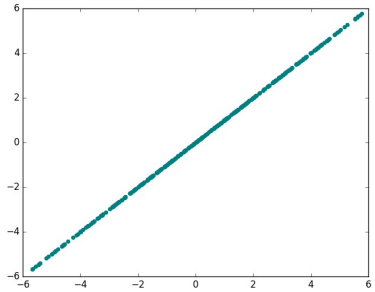
$$\rho(X,Y) = 1 \quad \Longrightarrow Y = aX + b, \text{where } a = \sigma_Y/\sigma_X$$
$$\rho(X,Y) = -1 \quad \Longrightarrow Y = aX + b, \text{where } a = -\sigma_Y/\sigma_X$$
$$\rho(X,Y) = 0 \quad \Longrightarrow \text{“uncorrelated” (absence of linear relationship)}$$

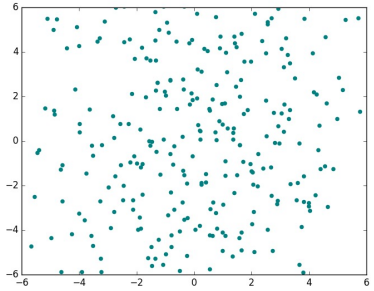
Correlation reps

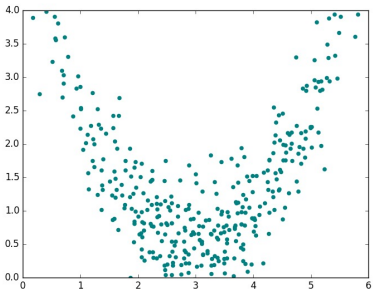
A. $\rho(X,Y) = 1$
B. $\rho(X,Y) = -1$
C. $\rho(X,Y) = 0$
D. Other

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

1. 

2. 

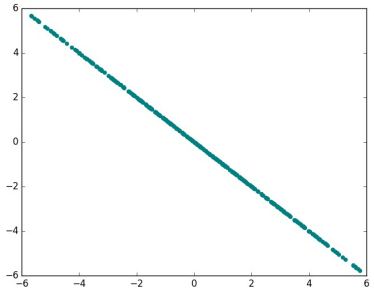
3. 

4. 

Correlation reps

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

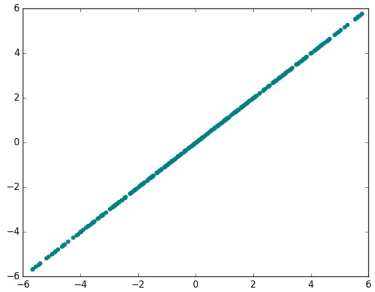
What is the correlation coefficient $\rho(X,Y)$?

1. 

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

$Y = -aX + b$

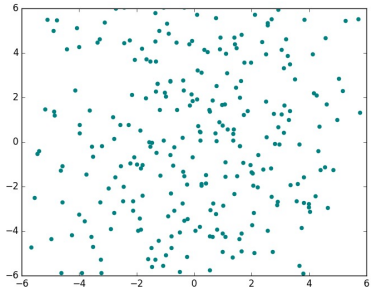
$a > 0$

2. 

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

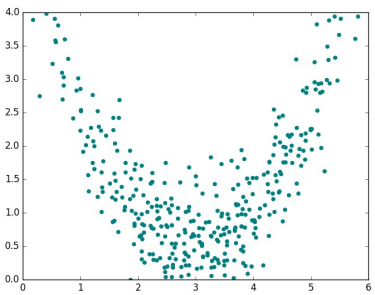
$Y = aX + b$

$a > 0$

3. 

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

"uncorrelated"

4. 

uncorrelated

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

$Y = X^2$

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

Extras

Expectation of product of independent RVs

If X and Y are independent, then

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

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

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

Variance of Sums of Variables

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

Proof:

$$\text{Var}\left(\sum_{i=1}^{n} X_i\right) \overset{\text{Var}(X) = \text{Cov}(X,X)}{=} \text{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i\right) \overset{\text{covariance of all pairs}}{=} \sum_{i=1}^{n}\sum_{j=1}^{n} \text{Cov}(X_i, X_j)$$

$$= \sum_{i=1}^{n} \text{Var}(X_i) + \sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n} \text{Cov}\left(X_i, X_j\right) \quad \text{Symmetry of covariance } \text{Cov}(X,X) = \text{Var}(X)$$

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