THE UNIVERSITY OF TEXAS AT AUSTIN
Department of Statistics and Data Sciences
College of Natural Sciences

SDS 321: Introduction to Probability and Statistics
Lecture 16: Covariance and Correlation

Purnamrita Sarkar
Department of Statistics and Data Science
The University of Texas at Austin

www.cs.cmu.edu/∼psarkar/teaching

Multiple random variables

- Often, we are interested in multiple random variables.
- These variables may be *dependent* or *independent*
- Recall, two random variables X and Y are independent if
  For all x, y

Discrete case

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

Continuous case

$$f_{X,Y}(x, y) = f_X(x)f_Y(y)$$

Multiple random variables

- Often, we are interested in multiple random variables.
- These variables may be *dependent* or *independent*
- Recall, two random variables X and Y are independent if
  For all x, y

Discrete case | **Continuous case**

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

$$f_{X,Y}(x,y) = f_X(x)f_Y(y)$$

- If $p_Y(y) > 0$ (discrete case)$/f_Y(y) > 0$ (continuous case), this gives us a more interpretable definition

Discrete case | **Continuous case**

$$p_{X|Y}(x|y) = p_X(x)$$

$$f_{X|Y}(x|y) = f_X(x)$$

- i.e. knowing $Y = y$ tells us nothing about X.

Expectations and variances of functions of multiple random variables

- A function $Z = g(X, Y)$ of two (or more) random variables is still a random variable.
- We can extend our definitions of expectation and variance to incorporate such random variables (discrete case omitted for space):

Continuous case

$$E[g(X, Y)] = \iint_{(x,y)} g(x, y) f_{X,Y}(x, y) dx\, dy$$

$$\text{var}(g(X, Y)) = \iint_{(x,y)} (g(x, y) - E[g(X, Y)]) f_{X,Y}(x, y) dx\, dy$$

$$= E[g(X, Y)^2] - E[g(X, Y)]^2$$

- If g is a linear function, e.g. $g(X, Y) = aX + bY + c$, we have

$$E[aX + bY + c] = aE[X] + bE[Y] + c$$

... regardless of whether X and Y are independent

Covariance

- The **covariance** of two random variables X and Y is given by

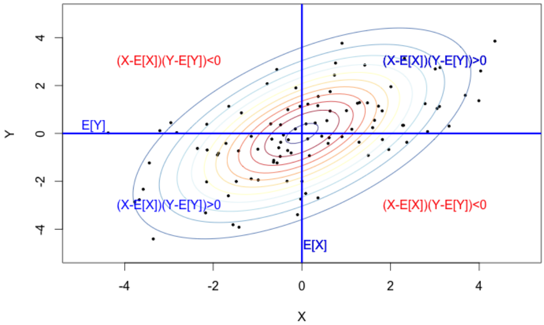
$$\text{cov}(X,Y) = E\left[(X - E[X])(Y - E[Y])\right]$$

- We can simplify this a little

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

- It is a measure of how much X and Y change together.
- A **positive** covariance means that, if $X > E[X]$, we are likely to have $Y > E[Y]$
- A **negative** covariance means that, if $X > E[X]$, we are likely to have $Y < E[Y]$.

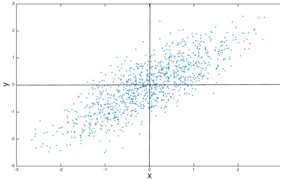
Covariance

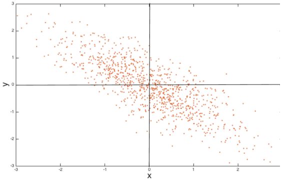


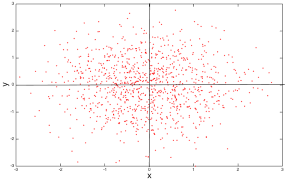
- A **positive** covariance means that we have most mass in the upper right and lower left quadrants.
- A **negative** covariance means that we have most mass in the upper left and lower right quadrants.
- A **zero** covariance means that we have about an equal mass in the upper left and upper right quadrants.

Covariance

- We are plotting two random variables X and Y below. Which one corresponds to a positive, negative or zero covariance?







Covariance properties

- $Cov(X, a) = 0$ where a is a constant.
- $Cov(aX, bY) = abcov(X, Y)$
- $Cov(X + Y, Z) = Cov(X, Z) + Cov(Y, Z)$
- $Cov(X, X) = Var(X)$

Example: Discrete case

- I flip a fair coin 5 times. Let $X = 1$ if the first coin flip is heads, and 0 otherwise. Let Y be the total number of heads. What is the correlation between X and Y?
- What is $E[X]$? What is $E[Y]$?

Example: Discrete case

- I flip a fair coin 5 times. Let $X = 1$ if the first coin flip is heads, and 0 otherwise. Let Y be the total number of heads. What is the correlation between X and Y?
- What is $E[X]$? $1/2$ What is $E[Y]$?

Example: Discrete case

- I flip a fair coin 5 times. Let $X = 1$ if the first coin flip is heads, and 0 otherwise. Let Y be the total number of heads. What is the correlation between X and Y?
- What is $E[X]$? $1/2$ What is $E[Y]$? $5/2$

Example: Discrete case

- I flip a fair coin 5 times. Let $X = 1$ if the first coin flip is heads, and 0 otherwise. Let Y be the total number of heads. What is the correlation between X and Y?
- What is $E[X]$? $1/2$ What is $E[Y]$? $5/2$

$$E[XY] = \sum_{x=0}^{1}\sum_{y=0}^{5} xyp_{X,Y}(x,y) = \sum_{x=0}^{1}\sum_{y=0}^{5} xyp_X(x)p_{Y|X}(y|x)$$

$$= \frac{1}{2} \underbrace{\sum_{y=0}^{5} yP(Y = y|X = 1)}_{\text{conditional expectation of } Y \text{ given } X = 1}$$

Example: Discrete case

- I flip a fair coin 5 times. Let $X = 1$ if the first coin flip is heads, and 0 otherwise. Let Y be the total number of heads. What is the correlation between X and Y?
- What is $E[X]$? 1/2 What is $E[Y]$? 5/2

$$E[XY] = \sum_{x=0}^{1} \sum_{y=0}^{5} xy p_{X,Y}(x,y) = \sum_{x=0}^{1} \sum_{y=0}^{5} xy p_X(x) p_{Y|X}(y|x)$$

$$= \frac{1}{2} \underbrace{\sum_{y=0}^{5} yP(Y = y|X = 1)}_{\text{conditional expectation of } Y \text{ given } X=1}$$

- If $X = 1$, then $Y - 1$ is a $Binomial(4, 1/2)$ random variable. So, the sum is just the expectation of a $Binomial(4, 1/2)$ random variable plus 1, i.e. $\frac{4}{2} + 1 = 3$.

Example: Discrete case

- I flip a fair coin 5 times. Let $X = 1$ if the first coin flip is heads, and 0 otherwise. Let Y be the total number of heads. What is the correlation between X and Y?
- What is $E[X]$? 1/2 What is $E[Y]$? 5/2

$$E[XY] = \sum_{x=0}^{1}\sum_{y=0}^{5} xyp_{X,Y}(x,y) = \sum_{x=0}^{1}\sum_{y=0}^{5} xyp_X(x)p_{Y|X}(y|x)$$

$$= \frac{1}{2} \underbrace{\sum_{y=0}^{5} yP(Y = y|X = 1)}_{\text{conditional expectation of } Y \text{ given } X = 1}$$

- If $X = 1$, then $Y - 1$ is a *Binomial*$(4, 1/2)$ random variable. So, the sum is just the expectation of a *Binomial*$(4, 1/2)$ random variable plus 1, i.e. $\frac{4}{2} + 1 = 3$.
- So, $E[XY] = \frac{3}{2}$

Example: Discrete case

- I flip a fair coin 5 times. Let $X = 1$ if the first coin flip is heads, and 0 otherwise. Let Y be the total number of heads. What is the correlation between X and Y?
- What is $E[X]$? $1/2$ What is $E[Y]$? $5/2$

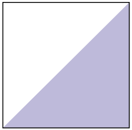
$$E[XY] = \sum_{x=0}^{1}\sum_{y=0}^{5} xyp_{X,Y}(x,y) = \sum_{x=0}^{1}\sum_{y=0}^{5} xyp_X(x)p_{Y|X}(y|x)$$

$$= \frac{1}{2} \underbrace{\sum_{y=0}^{5} yP(Y = y|X = 1)}_{\text{conditional expectation of } Y \text{ given } X = 1}$$

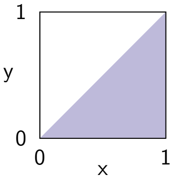
- If $X = 1$, then $Y - 1$ is a *Binomial*$(4, 1/2)$ random variable. So, the sum is just the expectation of a *Binomial*$(4, 1/2)$ random variable plus 1, i.e. $\frac{4}{2} + 1 = 3$.
- So, $E[XY] = \frac{3}{2}$
- So, $\text{cov}(X, Y) = E[XY] - E[X]E[Y] = \frac{3}{2} - \frac{1}{2}\frac{5}{2} = \frac{1}{4}$

Example: Continuous case

- Let $f_{X,Y}(x,y) = \begin{cases} 2 & 0 \leq y \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$
- What is $\text{cov}(X, Y)$?



Example: Continuous case

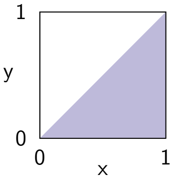


- Let $f_{X,Y}(x,y) = \begin{cases} 2 & 0 \leq y \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$
- What is $\mathrm{cov}(X, Y)$?

  - We first need to calculate the expectation of X and Y.

Example: Continuous case

- Let $f_{X,Y}(x,y) = \begin{cases} 2 & 0 \leq y \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$
- What is $\text{cov}(X,Y)$?



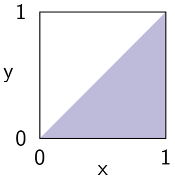
- We first need to calculate the expectation of X and Y.
- The marginal PDF of X and Y are:

$$f_X(x) =$$

$$f_Y(y) =$$

Example: Continuous case

- Let $f_{X,Y}(x,y) = \begin{cases} 2 & 0 \leq y \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$
- What is $\text{cov}(X,Y)$?



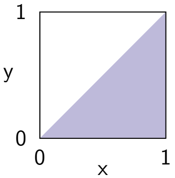
  - We first need to calculate the expectation of X and Y.
  - The marginal PDF of X and Y are:

$$f_X(x) = \int_0^1 f_{X,Y}(x,y)dy = 2\int_0^x dy = 2x \qquad 0 \leq x \leq 1$$

$$f_Y(y) =$$

Example: Continuous case

- Let $f_{X,Y}(x,y) = \begin{cases} 2 & 0 \leq y \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$
- What is $\text{cov}(X,Y)$?



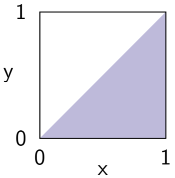
  - We first need to calculate the expectation of X and Y.
  - The marginal PDF of X and Y are:

$$f_X(x) = \int_0^1 f_{X,Y}(x,y)dy = 2\int_0^x dy = 2x \qquad 0 \leq x \leq 1$$

$$f_Y(y) = \int_0^1 f_{X,Y}(x,y)dx = 2\int_y^1 dx = 2(1-y) \qquad 0 \leq y \leq 1$$

Example: Continuous case

- Let $f_{X,Y}(x,y) = \begin{cases} 2 & 0 \leq y \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$
- What is $\text{cov}(X, Y)$?



- We first need to calculate the expectation of X and Y.
- The marginal PDF of X and Y are:

$$f_X(x) = \int_0^1 f_{X,Y}(x,y)dy = 2\int_0^x dy = 2x \qquad 0 \leq x \leq 1$$

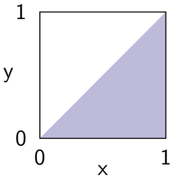
$$f_Y(y) = \int_0^1 f_{X,Y}(x,y)dx = 2\int_y^1 dx = 2(1-y) \qquad 0 \leq y \leq 1$$

- So the expectations are:

$$E[X] =$$

$$E[Y] =$$

Example: Continuous case



- Let $f_{X,Y}(x,y) = \begin{cases} 2 & 0 \le y \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$
- What is $\text{cov}(X, Y)$?

  - We first need to calculate the expectation of X and Y.
  - The marginal PDF of X and Y are:

$$f_X(x) = \int_0^1 f_{X,Y}(x,y)dy = 2\int_0^x dy = 2x \qquad 0 \le x \le 1$$

$$f_Y(y) = \int_0^1 f_{X,Y}(x,y)dx = 2\int_y^1 dx = 2(1-y) \qquad 0 \le y \le 1$$

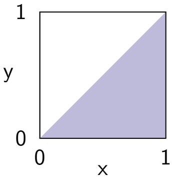
  - So the expectations are:

$$E[X] = \int_0^1 2x^2 dx = 2/3$$

$$E[Y] =$$

Example: Continuous case

- Let $f_{X,Y}(x,y) = \begin{cases} 2 & 0 \leq y \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$
- What is $\text{cov}(X,Y)$?



- We first need to calculate the expectation of X and Y.
- The marginal PDF of X and Y are:

$$f_X(x) = \int_0^1 f_{X,Y}(x,y)dy = 2\int_0^x dy = 2x \qquad 0 \leq x \leq 1$$

$$f_Y(y) = \int_0^1 f_{X,Y}(x,y)dx = 2\int_y^1 dx = 2(1-y) \qquad 0 \leq y \leq 1$$

- So the expectations are:

$$E[X] = \int_0^1 2x^2 dx = 2/3$$

$$E[Y] = \int_0^1 (2y - 2y^2)dy = 1/3$$

Example: Continuous case

- We next need to calculate $E[XY]$.
- This is just the expectation of a function of two random variables

$$E[XY] = \int_0^1 \int_0^1 xyf_{X,Y}(x, y)dxdy$$

Example: Continuous case

- We next need to calculate $E[XY]$.
- This is just the expectation of a function of two random variables

$$
\begin{aligned}
E[XY] &= \int_0^1 \int_0^1 xyf_{X,Y}(x,y)dxdy \\
&= \int_{x=0}^1 \int_{y=0}^x 2xydx\, dy
\end{aligned}
$$

Example: Continuous case

- We next need to calculate $E[XY]$.
- This is just the expectation of a function of two random variables

$$
\begin{aligned}
E[XY] &= \int_0^1 \int_0^1 xyf_{X,Y}(x,y)dxdy \\
&= \int_{x=0}^1 \int_{y=0}^x 2xydx\,dy \\
&= \int_{x=0}^1 x(\int_{y=0}^x 2ydy)\,dx
\end{aligned}
$$

Example: Continuous case

- We next need to calculate $E[XY]$.
- This is just the expectation of a function of two random variables

$$
\begin{aligned}
E[XY] &= \int_0^1 \int_0^1 xyf_{X,Y}(x,y)dxdy \\
&= \int_{x=0}^1 \int_{y=0}^x 2xydx\, dy \\
&= \int_{x=0}^1 x(\int_{y=0}^x 2ydy)\, dx \\
&= \int_{x=0}^1 x \left[y^2\right]_0^x dx = \int_{x=0}^1 x^3 dx = 1/4
\end{aligned}
$$

Example: Continuous case

- We next need to calculate $E[XY]$.
- This is just the expectation of a function of two random variables

$$
\begin{aligned}
E[XY] &= \int_0^1 \int_0^1 xyf_{X,Y}(x,y)dxdy \\
&= \int_{x=0}^1 \int_{y=0}^x 2xydx\,dy \\
&= \int_{x=0}^1 x(\int_{y=0}^x 2ydy)\,dx \\
&= \int_{x=0}^1 x\left[y^2\right]_0^x dx = \int_{x=0}^1 x^3 dx = 1/4
\end{aligned}
$$

- So, $\mathrm{cov}(X,Y) = E[XY] - E[X]E[Y] = \frac{1}{4} - \frac{1}{3}\cdot\frac{2}{3} = \frac{1}{36}$

Covariance and Independence

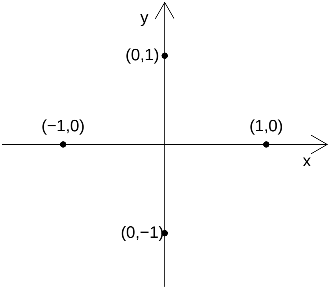
- If two random variables are **independent**, knowing one tells us nothing about the other!
- In this case, $E[XY] = E[X]E[Y]$
- We know that $\text{cov}(X, Y) = E[XY] - E[X]E[Y]$... so if two random variables are independent, their covariance is zero.
- This shouldn't be surprising... we know X can't tell us anything about Y.
- What about the converse? If $\text{cov}(X, Y) = 0$, does that mean that X and Y are independent?

Covariance and Independence

- If two random variables are **independent**, knowing one tells us nothing about the other!
- In this case, $E[XY] = E[X]E[Y]$
- We know that $\text{cov}(X, Y) = E[XY] - E[X]E[Y]$... so if two random variables are independent, their covariance is zero.
- This shouldn't be surprising... we know X can't tell us anything about Y.
- What about the converse? If $\text{cov}(X, Y) = 0$, does that mean that X and Y are independent?
- Another way of asking this is, does $E[XY] = E[X]E[Y]$ imply X and Y are independent?

Covariance and Independence

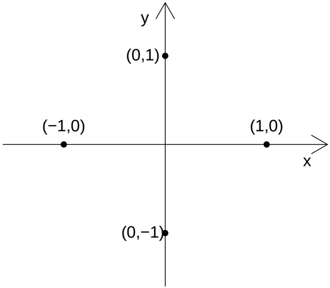
- I start at co-ordinates (0,0). I pick a compass direction (N,S,E,W) uniformly at random, and walk 1 unit in that direction.
- Let (X, Y) be my new coordinates. My sample space is $\{(0,1),(1,0),(0,-1),(-1,0)\}$.



- What are $E[X]$ and $E[Y]$?

Covariance and Independence

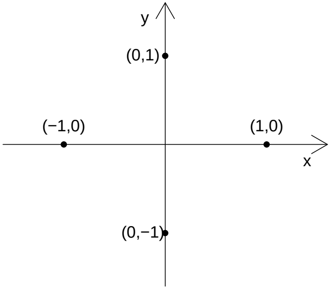
- I start at co-ordinates (0,0). I pick a compass direction (N,S,E,W) uniformly at random, and walk 1 unit in that direction.
- Let (X, Y) be my new coordinates. My sample space is $\{(0,1),(1,0),(0,-1),(-1,0)\}$.



- What are $E[X]$ and $E[Y]$?

Covariance and Independence

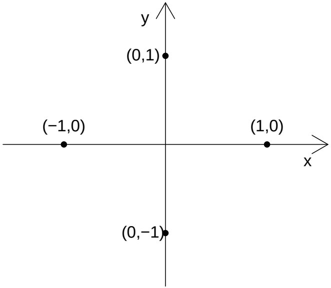
- I start at co-ordinates (0,0). I pick a compass direction (N,S,E,W) uniformly at random, and walk 1 unit in that direction.
- Let (X, Y) be my new coordinates. My sample space is $\{(0,1),(1,0),(0,-1),(-1,0)\}$.



- What are $E[X]$ and $E[Y]$? 0.

Covariance and Independence

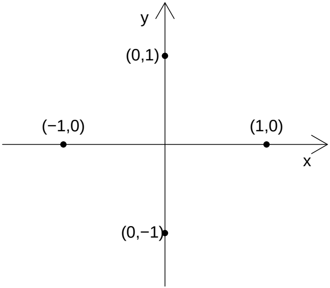
- I start at co-ordinates (0,0). I pick a compass direction (N,S,E,W) uniformly at random, and walk 1 unit in that direction.
- Let (X, Y) be my new coordinates. My sample space is $\{(0,1),(1,0),(0,-1),(-1,0)\}$.



- What are $E[X]$ and $E[Y]$? 0.
- $XY =$

Covariance and Independence

- I start at co-ordinates (0,0). I pick a compass direction (N,S,E,W) uniformly at random, and walk 1 unit in that direction.
- Let (X, Y) be my new coordinates. My sample space is $\{(0,1),(1,0),(0,-1),(-1,0)\}$.



- What are $E[X]$ and $E[Y]$? 0.
- $XY = 0$.
- So, $\text{cov}(X, Y) = 0$.
- But, if I know $X = 1$, then I *must* have $Y = 0$. So, they are not independent!

Independence implies zero correlation... but zero correlation does not imply independence!

Correlation

- We know that the sign of a covariance indicates whether $X - E[X]$ and $Y - E[Y]$ tend to have the same sign.
- The magnitude gives us some indication of the extent to which this is true... but it is hard to interpret.
  - The magnitude depends not just how much X and Y co-vary, but also on how much X and Y deviate from their expected values.
- The **correlation coefficient** $\rho_{X,Y}$(sometimes referred to as the Pearson's correlation coefficient) is a standardized version of the covariance.

$$\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)\text{var}(Y)}}$$

- We always have $-1 \leq \rho_{X,Y} \leq 1$
  - $\rho = 0$ implies zero covariance.
  - $|\rho| = 1$ iff there is a linear relationship between X and Y.

Correlation: Example of $|\rho| = 1$

- We throw a biased coin, with probability of heads p, n times. Let X be the number of heads, and let Y be the number of tails.
- $X = n - Y$
- $E[X] =$

Correlation: Example of $|\rho| = 1$

- We throw a biased coin, with probability of heads p, n times. Let X be the number of heads, and let Y be the number of tails.
- $X = n - Y$
- $E[X] = np$, and $E[Y] = n(1 - p) = n - E[X]$.

Correlation: Example of $|\rho| = 1$

- We throw a biased coin, with probability of heads p, n times. Let X be the number of heads, and let Y be the number of tails.
- $X = n - Y$
- $E[X] = np$, and $E[Y] = n(1 - p) = n - E[X]$.
- $\text{var}(X) = np(1 - p) = \text{var}(Y)$.

Correlation: Example of $|\rho| = 1$

- We throw a biased coin, with probability of heads p, n times. Let X be the number of heads, and let Y be the number of tails.
- $X = n - Y$
- $E[X] = np$, and $E[Y] = n(1 - p) = n - E[X]$.
- $\text{var}(X) = np(1 - p) = \text{var}(Y)$.
- All possible pairs (x, y) must satisfy $x + y = n = E[X] + E[Y]$ So $x - E[X] = -(y - E[Y])$
- Therefore $(x - E[X])(y - E[Y]) = -(x - E[X])^2$.

Correlation: Example of $|\rho| = 1$

- We throw a biased coin, with probability of heads p, n times. Let X be the number of heads, and let Y be the number of tails.
- $X = n - Y$
- $E[X] = np$, and $E[Y] = n(1-p) = n - E[X]$.
- $\text{var}(X) = np(1-p) = \text{var}(Y)$.
- All possible pairs (x, y) must satisfy $x + y = n = E[X] + E[Y]$ So $x - E[X] = -(y - E[Y])$
- Therefore $(x - E[X])(y - E[Y]) = -(x - E[X])^2$.
- We know that

$$\text{cov}(X, Y) = E[(X - E[X])(Y - E[Y])] = -E[(X - E[X])^2] = -\text{var}(X)$$

Correlation: Example of $|\rho| = 1$

- We throw a biased coin, with probability of heads p, n times. Let X be the number of heads, and let Y be the number of tails.
- $X = n - Y$
- $E[X] = np$, and $E[Y] = n(1-p) = n - E[X]$.
- $\text{var}(X) = np(1-p) = \text{var}(Y)$.
- All possible pairs (x, y) must satisfy $x + y = n = E[X] + E[Y]$ So $x - E[X] = -(y - E[Y])$
- Therefore $(x - E[X])(y - E[Y]) = -(x - E[X])^2$.
- We know that

$$\text{cov}(X, Y) = E[(X - E[X])(Y - E[Y])] = -E[(X - E[X])^2] = -\text{var}(X)$$

- The correlation coefficient is therefore

$$\rho_{X,Y} = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)\text{var}(Y)}} = \frac{-\text{var}(X)}{\sqrt{\text{var}(X)\text{var}(X)}} = -1$$

Correlation: Example of $|\rho| = 1$

- We throw a biased coin, with probability of heads p, n times. Let X be the number of heads, and let Y be the number of tails.
- $X = n - Y$
- $E[X] = np$, and $E[Y] = n(1-p) = n - E[X]$.
- $\text{var}(X) = np(1-p) = \text{var}(Y)$.
- All possible pairs (x, y) must satisfy $x + y = n = E[X] + E[Y]$ So $x - E[X] = -(y - E[Y])$
- Therefore $(x - E[X])(y - E[Y]) = -(x - E[X])^2$.
- We know that

$$\text{cov}(X, Y) = E[(X - E[X])(Y - E[Y])] = -E[(X - E[X])^2] = -\text{var}(X)$$

- The correlation coefficient is therefore

$$\rho_{X,Y} = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)\text{var}(Y)}} = \frac{-\text{var}(X)}{\sqrt{\text{var}(X)\text{var}(X)}} = -1$$

- Remember $X = n - Y$, so they have a linear relationship.

Variance of a sum of random variables

- Earlier in the course, we looked at the variance of the sum of independent random variables.
- Let's now consider the variance of sums of arbitrary random variables:

$$
\begin{aligned}
\text{var}(X+Y) &= E[(X+Y)^2] - (E[X+Y])^2 \\
&= E[X^2 + Y^2 + 2XY] - (E[X]^2 + E[Y]^2 + 2E[X]E[Y]) \\
&= \underbrace{E[X^2] - E[X]^2}_{\text{var}(X)} + \underbrace{E[Y^2] - E[Y]^2}_{\text{var}(Y)} + 2\underbrace{E[XY] - E[X]E[Y]}_{\text{cov}(X,Y)} \\
&= \text{var}(X) + \text{var}(Y) + 2\text{cov}(X,Y)
\end{aligned}
$$

- When X, Y are independent, the variance of the sum is the sum of variances.
- Can be extended to multiple random variables.

$$
\begin{aligned}
\text{var}(X+Y+Z) = {} & \text{var}(X) + \text{var}(Y) + \text{var}(Z) \\
& + 2\text{cov}(X,Y) + 2\text{cov}(Y,Z) + 2\text{cov}(X,Z)
\end{aligned}
$$

Summary

- **Expectation** tells us where we expect our random variable to be, on average.
- **Variance** is a measure of how far away from the expectation we expect it to be.
- If we have two random variables, **covariance** is a measure of the strength and direction of the relationship between them.
- It is often easier to interpret the **correlation coefficient**, a standardized form of the covariance with values between -1 and 1.
- If X and Y are independent, their covariance is zero.
- However, the converse is not always true!