Consider n independent tosses of coin with heads probability p.
Let X and Y be numbers of heads / tails and look at correlation coefficient of X and Y. Note that $X+Y=n$ so $E[X]+E[Y]=n$.

Thus, $X-E[X] = -(Y-E[Y])$.

now to do computation,

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

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

Suppose we define matrix of covariances, so (i,j) entry is $\text{cov}(X_i, X_j)$.

Let $\begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}$ have covariance matrix $\begin{bmatrix} 5 & 2 & 0 \\ 2 & 5 & 2 \\ 0 & 2 & 5 \end{bmatrix}$

(a) find $\text{cov}(X_1+X_2, X_1+X_3)$

$$= \text{cov}(X_1,X_1) + \text{cov}(X_1,X_3) + \text{cov}(X_2,X_1) + \text{cov}(X_2,X_3)$$
$$= 5+0+2+2 = 9.$$

(b) find a so $X_2 - aX_1$ is uncorrelated with X_1

$$\text{cov}(X_2-aX_1, X_1) = \text{cov}(X_2,X_1) - a\,\text{cov}(X_1,X_1) = 2-5a$$

which is zero for $a = \frac{2}{5}$.

MMSE estimation

Given data X, we want to infer a random variable Y that cannot be observed directly.

- In communication, from received signal X, infer transmitted signal Y.
- From noisy measurements, infer value Y of true quantity.
  → wastewater treatment.
- From noisy measurements of past positions of vehicle X, infer present position in inertial guidance system Y.
- From observations X on temperature, pressure at several locations, infer future temperature Y at a given site.
- From measurements from array of sensors, infer location of callers in cellular telephony environment.

Want to find an estimator $\hat{Y}(X)$ or $\hat{Y} = g(X)$ that minimizes mean square error (MSE) among all possible functions of data:

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

best estimator, MMSE estimator, is called $g^*(X)$.

Derive the estimator:

$$\cancel{E[(Y-\hat{Y})^2] = E[(Y - E[Y|X] + E[Y|X] - \hat{Y})^2]}$$

$$\cancel{= E[(Y - E[Y|X])]^2 + 1}$$

Intuitively it seems $E[Y|X]$, the conditional is a good thing to use, ie. the mean conditioned on the given data.