LECTURE 12: Sums of independent random variables; Covariance and correlation

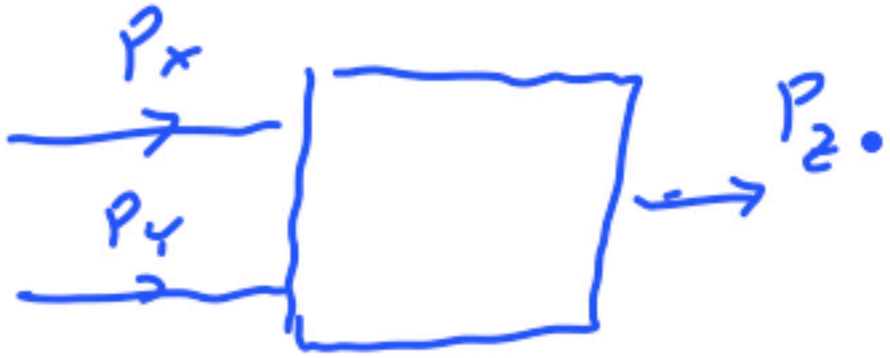
- The PMF/PDF of $X + Y$ (X and Y independent)
  - the discrete case
  - the continuous case
  - the mechanics
  - the sum of independent normals
- Covariance and correlation
  - definitions
  - mathematical properties
  - interpretation

The distribution of $X+Y$: the discrete case

$$\boxed{p_Z(z) = \sum_x p_X(x)\, p_Y(z-x)}$$

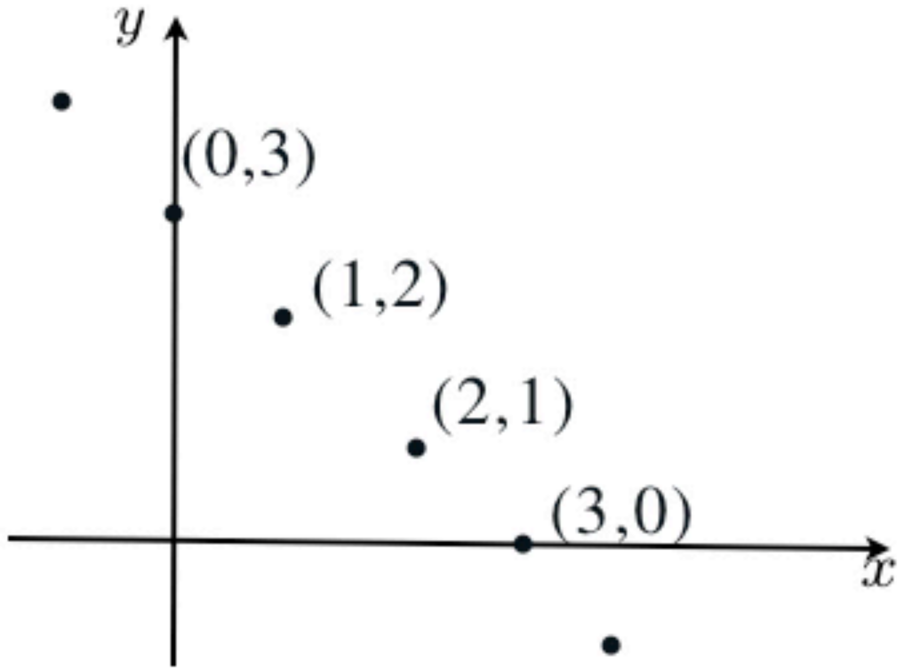
- $Z = X + Y$; $\quad X, Y$ independent, discrete

  $g(X,Y)$ known PMFs



$p_Z(3) = \dots + P(X=0, Y=3) + P(X=1, Y=2) + \dots$

$= \dots + p_X(0)\, p_Y(3) + p_X(1)\, p_Y(2) + \dots$

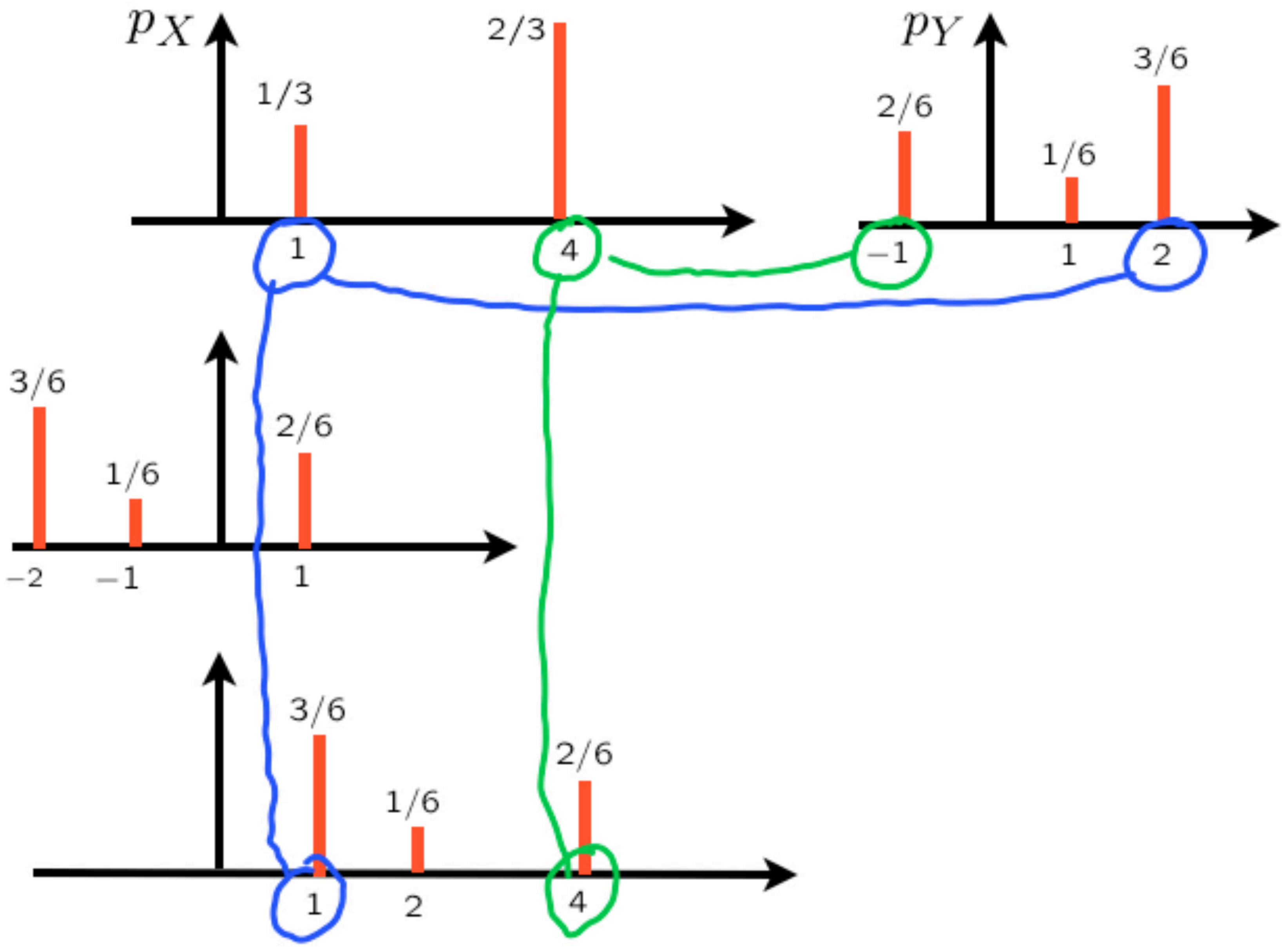


$$p_Z(z) = \sum_x P(X = x, Y = z - x)$$

$$= \sum_x p_X(x)\, p_Y(z-x)$$

Discrete convolution mechanics

$$p_Z(z) = \sum_x p_X(x)\, p_Y(z-x)$$



- To find $p_Z(3)$:
  - Flip (horizontally) the PMF of Y
  - Put it underneath the PMF of X
  - Right-shift the flipped PMF by 3
  - Cross-multiply and add
  - Repeat for other values of z

The distribution of $X+Y$: the continuous case

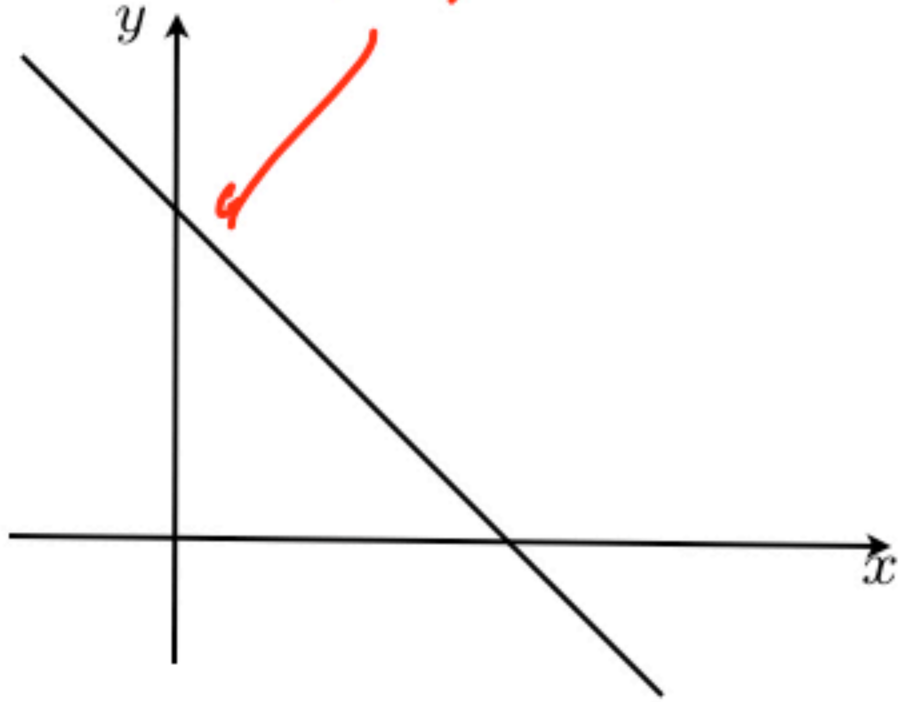
$$\boxed{p_Z(z) = \sum_x p_X(x)\, p_Y(z-x)}$$

- $Z = X + Y$; $\quad X, Y$ independent, continuous

  known PDFs

- $$\boxed{f_Z(z) = \int_{-\infty}^{\infty} f_X(x)\, f_Y(z-x)\, dx}$$

$x+y=\text{constant}$



Conditional on $X = x$: $\quad Z = x + Y \qquad x = 3 \quad Z = Y + 3$

$$f_{Z|X}(z|3) = f_{Y+3|X}(z|3) = f_{Y+3}(z) = f_Y(z-3)$$

$$f_{Z|X}(z|x) = f_Y(z-x) \qquad f_{X+b}(x) = f_X(x-b)$$

Joint PDF of Z and X:

$$f_{X,Z}(x,z) = f_X(x)\, f_Y(z-x)$$

From joint to the marginal: $f_Z(z) = \int_{-\infty}^{\infty} f_{X,Z}(x,z)\, dx$

- Same mechanics as in discrete case (flip, shift, etc.)

The sum of independent normal r.v.'s

$$\boxed{f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)\,dx}$$

- $X \sim N(\mu_x, \sigma_x^2)$, $Y \sim N(\mu_y, \sigma_y^2)$, independent $\qquad Z = X + Y$

$$f_X(x) = \frac{1}{\sqrt{2\pi}\,\sigma_x} e^{-(x-\mu_x)^2/2\sigma_x^2} \qquad f_Y(y) = \frac{1}{\sqrt{2\pi}\,\sigma_y} e^{-(y-\mu_y)^2/2\sigma_y^2}$$

$\underbrace{X+Y}+\underbrace{W}$

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)\,dx$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\,\sigma_x} \exp\left\{-\frac{(x-\mu_x)^2}{2\sigma_x^2}\right\} \frac{1}{\sqrt{2\pi}\,\sigma_y} \exp\left\{-\frac{(z-x-\mu_y)^2}{2\sigma_y^2}\right\} dx$$

(algebra) $$= \frac{1}{\sqrt{2\pi(\sigma_x^2+\sigma_y^2)}} \exp\left\{-\frac{(z-\mu_x-\mu_y)^2}{2(\sigma_x^2+\sigma_y^2)}\right\}$$

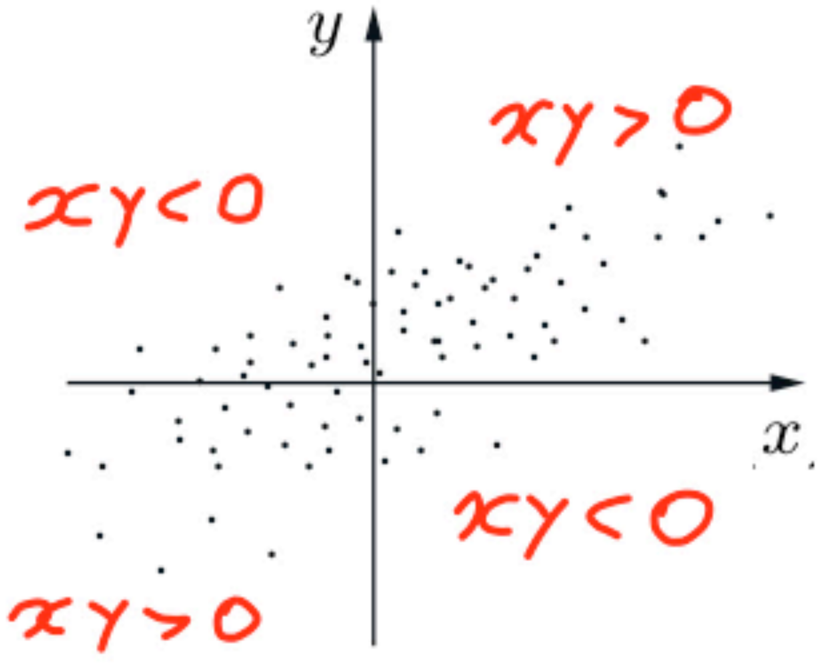
$N(\mu_x+\mu_y, \sigma_x^2+\sigma_y^2)$

The sum of finitely many independent normals is normal

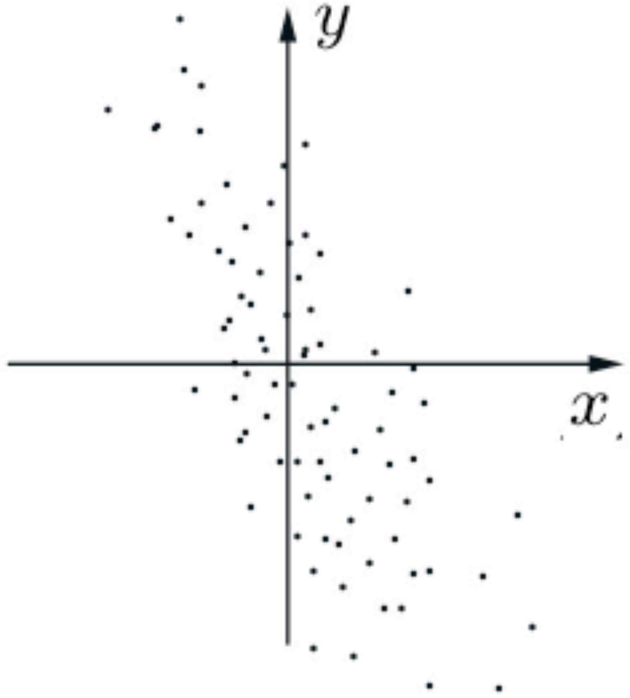
Covariance

- Zero-mean, discrete X and Y
  - if independent: $\mathbf{E}[XY] =$

$= E[X]\,E[Y] = 0$



$\mathbf{E}[XY] > 0$



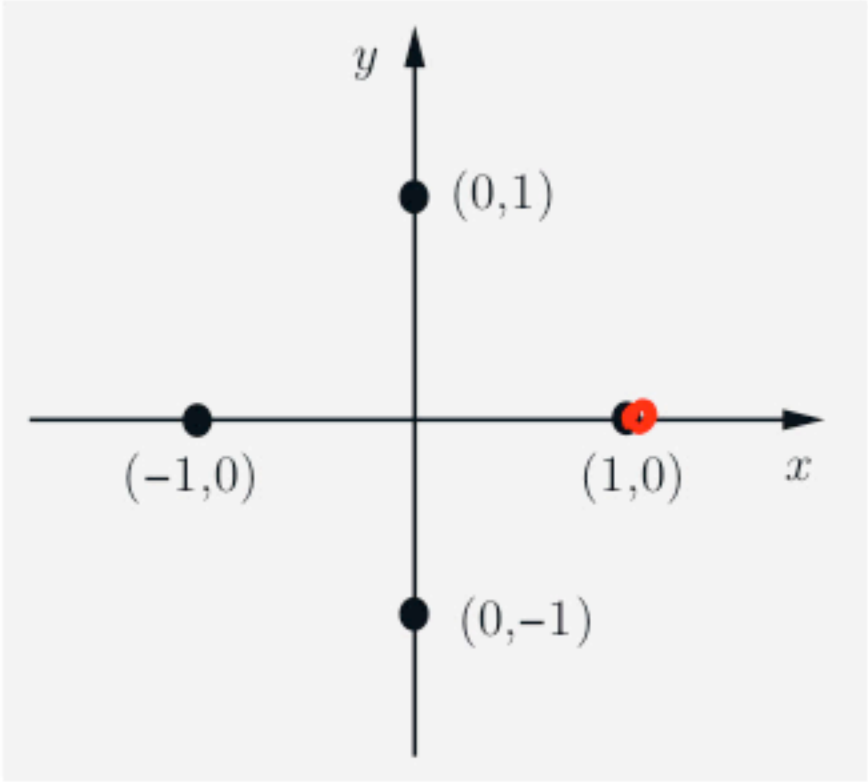
$\mathbf{E}[XY] < 0$

Definition for general case:

$$\boxed{\mathrm{cov}(X,Y) = \mathbf{E}\Big[\underline{(X - \mathbf{E}[X])} \cdot \underline{(Y - \mathbf{E}[Y])}\Big]}$$

and $0 = E\big[(X - E[X])\big]\,E\big[Y - E[Y]\big]$

- independent $\Rightarrow$ $\mathrm{cov}(X,Y) = 0$
  (converse is not true)



$XY = 0$

$\mathrm{Cov} = 0$

$X = 1 \Rightarrow Y = 0$

Covariance properties

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

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

$$\text{cov}(aX + b, Y) =$$

(assume 0 means)

$$= E[(aX+b)Y] = aE[XY] + b\cancel{E[Y]}$$

$$= a \cdot \text{cov}(X,Y)$$

$$\text{cov}(X, Y + Z) = E[X(Y+Z)]$$

$$= E[XY] + E[XZ] = \text{cov}(X,Y) + \text{cov}(X,Z)$$

$$\boxed{\text{cov}(X,Y) = \mathbf{E}\Big[(X - \mathbf{E}[X]) \cdot (Y - \mathbf{E}[Y])\Big]}$$

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

$$- E[E[X]Y] + E[E[X]E[Y]]$$

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

$$- \cancel{E[X]E[Y]} + \cancel{E[X]E[Y]}$$

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

The variance of a sum of random variables

$$\mathrm{var}(X_1 + X_2) = E\left[\left(X_1 + X_2 - E[X_1 + X_2]\right)^2\right]$$

$$= E\left[\left(\left(X_1 - E[X_1]\right) + \left(X_2 - E[X_2]\right)\right)^2\right]$$

$$= E\left[\left(X_1 - E[X_1]\right)^2 + \left(X_2 - E[X_2]\right)^2 + 2\left(X_1 - E[X_1]\right)\left(X_2 - E[X_2]\right)\right]$$

$$= \mathrm{var}(X_1) + \mathrm{var}(X_2) + 2\,\mathrm{cov}(X_1, X_2).$$

The variance of a sum of random variables

$$\mathrm{var}(X_1 + X_2) = \mathrm{var}(X_1) + \mathrm{var}(X_2) + 2\,\mathrm{cov}(X_1, X_2)$$

$$\mathrm{var}(X_1 + \cdots + X_n) = E\left[(X_1 + \cdots + X_n)^2\right]$$

(assume 0 means)

$$= E\left[\sum_{i=1}^{n} X_i^2 + \underbrace{\sum_{\substack{i=1,\ldots,n\\ j=1,\ldots,n\\ i\neq j}} X_i X_j}_{n^2-n \text{ terms}}\right]$$

$$= \sum_i \mathrm{Var}(X_i) + \sum_{i\neq j} \mathrm{Cov}(X_i, X_j)$$

$$\boxed{\mathrm{var}(X_1 + \cdots + X_n) = \sum_{i=1}^{n} \mathrm{var}(X_i) + \sum_{\{(i,j):\, i\neq j\}} \mathrm{cov}(X_i, X_j)}$$

The Correlation coefficient

- Dimensionless version of covariance:

$$\rho(X,Y) = \mathbf{E}\left[\frac{(X-\mathbf{E}[X])}{\sigma_X}\cdot\frac{(Y-\mathbf{E}[Y])}{\sigma_Y}\right] = \frac{\text{cov}(X,Y)}{\sigma_X\sigma_Y}$$

$$\boxed{-1 \le \rho \le 1}$$

- Measure of the degree of "association" between X and Y

- Independent $\Rightarrow$ $\rho = 0$, "uncorrelated" (converse is not true)

- $\rho(X,X) = \frac{\text{var}(X)}{\sigma_X^2} = 1$

- $|\rho| = 1 \;\Leftrightarrow\; (X-\mathbf{E}[X]) = c(Y-\mathbf{E}[Y])$ (linearly related)

- $\text{cov}(aX+b,Y) = a\cdot\text{cov}(X,Y) \quad\Rightarrow\quad \rho(aX+b,Y) = \frac{a\,\text{cov}(X,Y)}{|a|\,\sigma_X\,\sigma_Y} = \text{sign}(a)\cdot\rho(X,Y)$

Proof of key properties of the correlation coefficient

$$\rho(X,Y) = \mathbf{E}\left[\frac{(X-\mathbf{E}[X])}{\sigma_X}\cdot\frac{(Y-\mathbf{E}[Y])}{\sigma_Y}\right]$$

$$\boxed{-1 \le \rho \le 1}$$

- Assume, for simplicity, zero means and unit variances, so that $\rho(X,Y) = \mathbf{E}[XY]$

$0 \le \mathbf{E}\left[\underline{(X-\rho Y)^2}\right] = E[X^2] - 2\rho E[XY] + \rho^2 E[Y^2]$

$= 1 - 2\rho^2 + \rho^2 = \underline{\underline{1-\rho^2}}$ $\quad 1-\rho^2 \ge 0 \Rightarrow \rho^2 \le 1$

If $|\rho| = 1$, then $X = \rho Y \Rightarrow X = Y$ or $X = -Y$

Interpreting the correlation coefficient

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

- Association does not imply causation or influence

  X: math aptitude

  Y: musical ability

- Correlation often reflects underlying, common, hidden factor

  – Assume, Z, V, W are independent

  $$\rho(x,y) = \frac{1}{\sqrt{2}\cdot\sqrt{2}} = \frac{1}{2}$$

  $$X = \underline{\underline{Z}} + V \qquad Y = \underline{\underline{Z}} + W$$

  Assume, for simplicity, that Z, V, W have zero means, unit variances

$$\text{var}(x) = \text{var}(z) + \text{var}(v) = 2 \Rightarrow \sigma_x = \sqrt{2} \qquad \sigma_y = \sqrt{2}$$

$$\text{cov}(x,y) = E[(z+v)(z+w)] = E[z^2] + E[vz] + E[zw] + E[vw]$$

$$= 1 + 0 + 0 + 0$$

Correlations matter...

- A real-estate investment company invests \$10M in each of 10 states. At each state i, the return on its investment is a random variable X_i, with mean 1 and standard deviation 1.3 (in millions).

$$\text{var}(X_1 + \cdots + X_{10}) = \sum_{i=1}^{10} \text{var}(X_i) + \sum_{\{(i,j):\, i \neq j\}} \text{cov}(X_i, X_j)$$

$E[X_1 + \ldots + X_{10}] = 10$

- If the X_i are uncorrelated, then:

$\text{var}(X_1 + \cdots + X_{10}) = 10 \cdot (1.3)^2 = 16.9$ $\qquad \sigma(X_1 + \cdots + X_{10}) = 4.1$

- If for $i \neq j$, $\rho(X_i, X_j) = 0.9$:

$\text{cov}(X_i, X_j) = \rho\, \sigma_{X_i} \sigma_{X_j} = 0.9 \times 1.3 \times 1.3 = 1.52$

$\text{var}(X_1 + \quad + X_{10}) = 10 \cdot (1.3)^2 + 90 \cdot 1.52 = 154$

$\sigma(X_1 + \cdots + X_{10}) = 12.4$