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

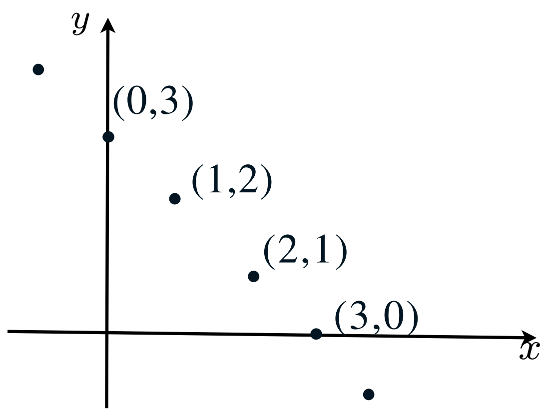
The distribution of $X+Y$: the discrete case

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

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

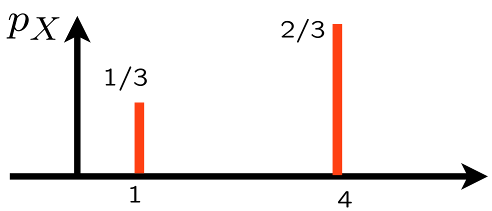
  known PMFs

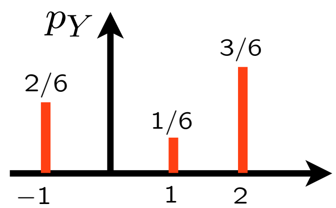
$p_Z(3) =$



Discrete convolution mechanics

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

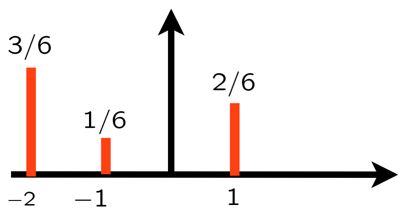




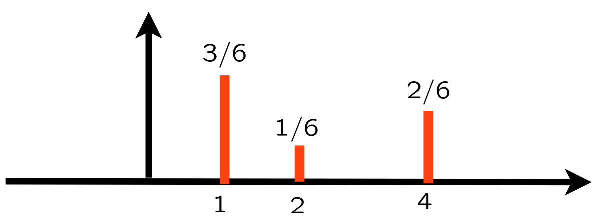
- To find $p_Z(3)$:

  – 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

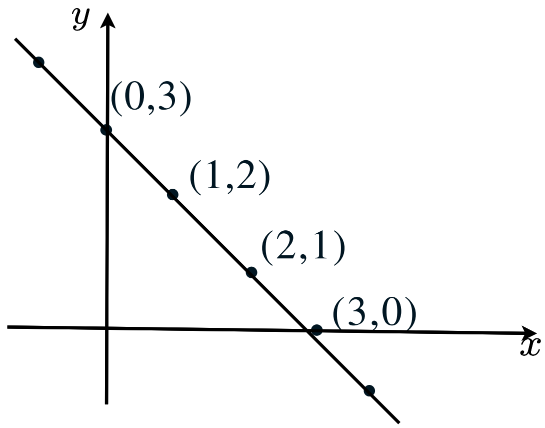
$$p_Z(z) = \sum_x p_X(x)\, p_Y(z-x)$$

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

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

  known PDFs

Conditional on $X = x$:



Joint PDF of Z and 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

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

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

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

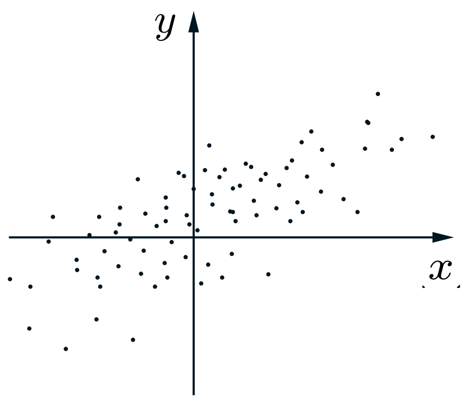
$$\begin{aligned}
f_Z(z) &= \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)\, dx \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\,\sigma_x} \exp\left\{-\frac{(x-\mu_x)^2}{2\sigma_x^2}\right\} \frac{1}{\sqrt{2\pi}\,\sigma_y} \exp\left\{-\frac{(z-x-\mu_y)^2}{2\sigma_y^2}\right\} dx \\
\text{(algebra)} &= \frac{1}{\sqrt{2\pi(\sigma_x^2+\sigma_y^2)}} \exp\left\{-\frac{(z-\mu_x-\mu_y)^2}{2(\sigma_x^2+\sigma_y^2)}\right\}
\end{aligned}$$

The sum of finitely many independent normals is normal

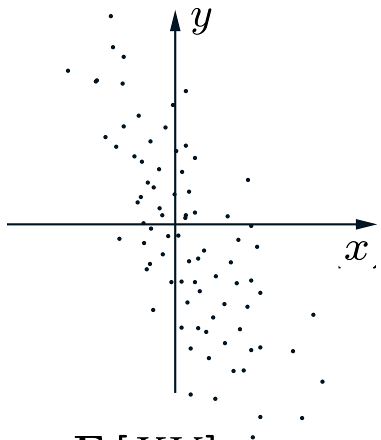
Covariance

- Zero-mean, discrete X and Y

– if independent: $\mathbf{E}[XY] =$



$\mathbf{E}[XY]$

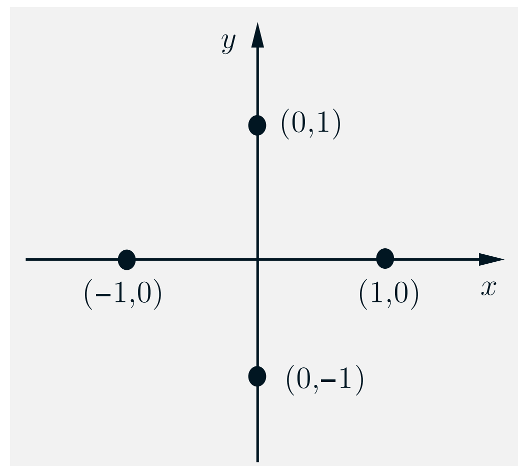


$\mathbf{E}[XY]$

Definition for general case:

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

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



Covariance properties

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

$\text{cov}(X,X) =$

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

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

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

The variance of a sum of random variables

$\mathsf{var}(X_1 + X_2) =$

The variance of a sum of random variables

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

$$\mathrm{var}(X_1 + \cdots + X_n) =$$

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

The Correlation coefficient

- Dimensionless version of covariance:

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

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

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

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

- $\rho(X,X) =$

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

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

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 $\rho(X,Y) = \mathbf{E}[XY]$

$$\mathbf{E}\Big[(X-\rho Y)^2\Big] =$$

If $|\rho| = 1$, then

Interpreting the correlation coefficient

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

- 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

    $$X = Z + V \qquad\qquad Y = Z + W$$

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

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).

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

- If the X_i are uncorrelated, then:

  $\text{var}(X_1 + \cdots + X_{10}) =$

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

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

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