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

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

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

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

$g(x, y)$ known PMFs

$p_Z(3) = \cdots + p(X = 0, Y = 3) + p(X = 1, Y = 2) + \cdots$

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

$P_z(2) = \sum_x P(X = x, Y = z - x)$

$P_z(2) = \sum_x P_X(x) P_Y(z - x)$
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

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

Known PDFs

- Conditional on $X = x$: $Z = x + Y$
  - $x = 3$, $z = y + 3$
  - $f_{Z|X}(z|x) = f_Y(z - x)$
  - $f_{Z|X}(z|x) = f_Y(z - x)$

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

- \( X \sim N(\mu_x, \sigma_x^2), \ Y \sim N(\mu_y, \sigma_y^2), \) independent

\[
Z = X + Y
\]

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

\[
f_X(x) = \frac{1}{\sqrt{2\pi \sigma_x^2}} e^{-\frac{(x-\mu_x)^2}{2\sigma_x^2}} \quad f_Y(y) = \frac{1}{\sqrt{2\pi \sigma_y^2}} e^{-\frac{(y-\mu_y)^2}{2\sigma_y^2}}
\]

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

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

\[
\text{(algebra)} = \frac{1}{\sqrt{2\pi(\sigma_x^2 + \sigma_y^2)}} \exp \left\{ -\frac{(z-\mu_x-\mu_y)^2}{2(\sigma_x^2 + \sigma_y^2)} \right\} \quad N\left(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2\right)
\]

The sum of finitely many independent normals is normal.
Covariance

- Zero-mean, discrete $X$ and $Y$
  - if independent: $E[XY] = E[X]E[Y] = 0$

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

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

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

\[
\text{cov}(aX + b, Y) = E[(aX + b)Y] = aE[XY] + bE[Y] = a \cdot \text{cov}(X, Y)
\]

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

\[
\text{cov}(X, Y) = E[XY] - E[X]E[Y]
\]
The variance of a sum of random variables

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

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

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

\[ = \text{var}(X_1) + \text{var}(X_2) + 2 \text{cov}(X_1, X_2). \]
The variance of a sum of random variables

\[ \text{var}(X_1 + X_2) = \text{var}(X_1) + \text{var}(X_2) + 2 \text{cov}(X_1, X_2) \]

\[ \text{var}(X_1 + \cdots + X_n) = \sum_{i=1}^{n} \text{var}(X_i) + \sum_{i \neq j} \text{cov}(X_i, X_j) \]

(assume 0 means)

\[ = E \left[ \left( \sum_{i=1}^{n} X_i \right)^2 \right] = E \left[ \sum_{i=1}^{n} X_i^2 + \sum_{i=1}^{n} \sum_{j=1}^{n} X_i X_j \right] \]

\[ \{ \text{n}\text{-n terms} \} \]

\[ = \sum_{i} \text{var}(X_i) + \sum_{i \neq j} \text{cov}(X_i, X_j) \]
The Correlation coefficient

- Dimensionless version of covariance:

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

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

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

- Measure of the degree of “association” between $X$ and $Y$

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

- $|\rho| = 1 \iff (X - \mathbb{E}[X]) = c(Y - \mathbb{E}[Y])$ (linearly related)

- $\text{cov}(aX + b, Y) = a \cdot \text{cov}(X, Y) \Rightarrow \rho(aX + b, Y) = \frac{a \text{cov}(x, y)}{1a |\sigma_x \sigma_y} = \text{sign}(a) \cdot \rho(x, y)$

$-1 \leq \rho \leq 1$
Proof of key properties of the correlation coefficient

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

$$-1 \leq \rho \leq 1$$

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

$$\begin{align*}
E[(X - \rho Y)^2] &= E[X^2] - 2 \rho E[XY] + \rho^2 E[Y^2] \\
0 &\leq 1 - 2\rho^2 + \rho^2 = 1 - \rho^2 \\
1 - \rho^2 &\geq 0 \Rightarrow \rho^2 \leq 1
\end{align*}$$

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
    \[ X = Z + V \quad \quad Y = Z + W \]
  - Assume, for simplicity, that \( Z, V, W \) have zero means, unit variances
    \[ \text{var}(x) = \text{var}(Z) + \text{var}(V) = 2 \Rightarrow \sigma_x = \sqrt{2} \quad \sigma_y = \sqrt{2} \]
    \[ = 1 + 0 + 0 + 0 = \frac{1}{2} \]
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:

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

$$\sigma(X_1 + \cdots + X_{10}) = 4.1$$

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

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

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

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