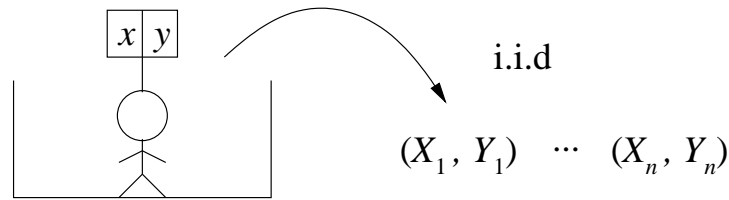


Lecture 11: Joint Distribution of (X, Y) and Related Quantities

We have seen the prediction of one variable based on the other, using a data set of pairs. If we regard (X_i, Y_i) as i.i.d. samples from a joint distribution (X, Y) there is a corresponding story of predicting one r.v. from another r.v.

e.g.



The population box determines the distribution of the pair (X, Y) or the joint distribution of (X, Y)

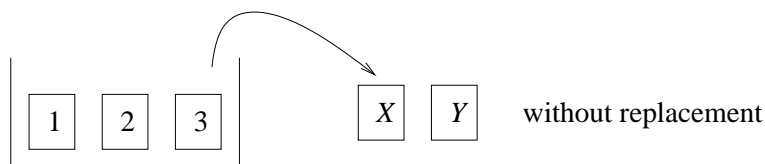
Recall a joint distribution of discrete r.v's X and Y is described by a joint probability table. For example,

		Y			
		y_1	\dots	y_j	↓
X	x_1	$p_{ij} = (X = x_i, Y = y_j)$			
	\vdots				
	x_l				
					←

marginal distribution of X = row sums

marginal distribution of Y = column sums

e.g.



		Y			
		1	2	3	
X	1	0	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{3}$
	2	$\frac{1}{6}$	0	$\frac{1}{6}$	$\frac{1}{3}$
	3	$\frac{1}{6}$	$\frac{1}{6}$	0	$\frac{1}{3}$
		$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	

Properties of $\text{Var}(\cdot)$, $\text{E}(\cdot)$

1. $\text{E}(X + Y) = \text{E}X + \text{E}Y$
2. $\text{Var}(X + Y) = \text{Var}(X) + 2\text{Cov}(X, Y) + \text{Var}(Y)$

where $\text{Cov}(X, Y) = \text{E}[(X - \text{E}X)(Y - \text{E}Y)]$

Computation formula: $\text{Cov}(X, Y) = \text{E}[XY] - \text{E}X\text{E}Y$

3. $\text{Cov}(\sum a_i X_i, \sum b_j Y_j) = \sum_i \sum_j a_i b_j \text{Cov}(X_i, Y_j)$

e.g.

$$\begin{aligned}
 \text{E}(XY) &= \sum_{x_i} \sum_{y_j} x_i y_j \text{P}(X = x_i, Y = y_j) \\
 &= [1 \times 2 + 1 \times 3 + 2 \times 1 + 2 \times 3 + 3 \times 1 + 3 \times 2] \times \frac{1}{6} \\
 &= 22 \times \frac{1}{6} = \frac{11}{3}
 \end{aligned}$$

$$\begin{aligned}
 \text{E}X = 2 = \text{E}Y \implies \text{Cov}(X, Y) &= \frac{11}{3} - 4 = -\frac{1}{3} \\
 \text{Var}(X) &= \frac{1}{3}[(-1)^2 + 0^2 + 1^2] = \frac{2}{3}
 \end{aligned}$$

$$\text{Recall: } \text{Cov}(X, Y) = -\frac{1}{N-1}\sigma^2 = -\frac{1}{2} \times \frac{2}{3} = -\frac{1}{3}$$

If (X, Y) are continuous, then the joint distribution is described by a joint density function $f(x, y)$ and integration replaces summation in the discrete case:

$$\begin{aligned}
 \text{marginal } f_X(x) &= \int_{-\infty}^{\infty} f(x, y) dy \\
 \text{marginal } f_Y(y) &= \int_{-\infty}^{\infty} f(x, y) dx
 \end{aligned}$$

$$\begin{aligned}\text{Cov}(X, Y) &= E[XY] - EXEY \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xyf(x, y)dx dy - \int_{-\infty}^{\infty} x f_X(x)dx \int_{-\infty}^{\infty} y f_Y(y)dy\end{aligned}$$

e.g. Bivariate Normal distribution: standard.

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right) \quad -1 \leq \rho \leq 1$$

$$f(x, y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp \left\{ -\frac{1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2) \right\}$$

$$EX = 0, EY = 0, \text{Var}(X) = 1, \text{Var}(Y) = 1, \text{Cov}(X, Y) = \rho.$$

Marginal distribution of X and Y are $N(0,1)$.

General bivariate Normal:

$$\begin{aligned}\begin{pmatrix} X \\ Y \end{pmatrix} &\sim N \left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right) \\ \begin{pmatrix} \frac{X-\mu_1}{\sigma_1} \\ \frac{Y-\mu_2}{\sigma_2} \end{pmatrix} &\sim N \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right)\end{aligned}$$

Prediction of Y in terms of mean squared error (MSE)

1. by a constant

Find a to minimize $E[Y - a]^2$. Let $g(a) = E[Y - a]^2 = E(Y^2) - 2aEY + a^2$.

$$g'(a) = -2EY + 2a = 0 \implies a = EY$$

i.e. the best constant prediction of Y is EY . The data-version, $a = \bar{y}$ minimizes $g(a) = \frac{1}{n} \sum_{i=1}^n (y_i - a)^2$.

2. by a linear function of X .

Find (a, b) that minimizes:

$$\begin{aligned}g(a, b) &= E[Y - (a + bX)]^2 \\ &= E(Y^2) - 2aEY - 2bE(XY) + a^2 + 2abEX + b^2E(X^2)\end{aligned}$$

Differentiate with respect to a and b

$$\begin{aligned}\frac{\partial g}{\partial a} &= -2EY + 2a + 2bEX = 0 \\ \frac{\partial g}{\partial b} &= -2E[XY] + 2aEX + 2bE(X^2) = 0\end{aligned}$$

This leads to

$$EY = a + bEX \quad (1)$$

$$E(XY) = aEX + bE(X^2) \quad (2)$$

Plug (1) into (2): $a = EY - bEX$

$$E[XY] = (EY - bEX)EX + bE[X^2]$$

$$\begin{aligned} \implies b &= \frac{E(XY) - EXEY}{E(X^2) - (EX)^2} = \frac{\text{Cov}(X, Y)}{\text{Var}(X)} \\ &= \frac{\text{SD}(Y)}{\text{SD}(X)} \frac{\text{Cov}(X, Y)}{\text{SD}(X)\text{SD}(Y)} = \frac{\text{SD}(Y)}{\text{SD}(X)} \rho_{XY} \end{aligned}$$

$$\rho_{XY} = \frac{\text{Cov}(X, Y)}{\text{SD}(X)\text{SD}(Y)} = \text{correlation coefficient of } (X, Y)$$

We have seen the data version: the minimizer of $\frac{1}{n} \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2$ is:

$$\alpha = \bar{y} - \beta \bar{x} \quad \text{and} \quad \beta = \frac{\text{SD}(y)}{\text{SD}(x)} r_{xy}.$$

The regression line $y = \alpha + \beta x$ can be viewed as an estimate of the population best linear predictor:

$$\begin{aligned} y &= a + bx \\ a &= EY - bEX \\ b &= \frac{\text{SD}(Y)}{\text{SD}(X)} \rho_{XY} \end{aligned}$$

Conditional Distribution & Independence

- For events A, B

$$P(A \cap B) = P(B)P(A|B) = P(A)P(B|A)$$

- For discrete r.v. (X, Y)

$$P(X = x_i, Y = y_j) = P(Y = y_j)P(X = x_i|Y = y_j) = P(X = x_i)P(Y = y_j|X = x_i)$$

- Fix $Y = y$,

— $P(X = x_i|Y = y_j)$, $i = 1, \dots, I$ gives the conditional distribution of X given $Y = y$.

— $\sum_{i=1}^I P(X = x_i|Y = y_j) = 1$. $P_{X|Y=y}$ denotes this distribution.

- Fix $X = x$,
 - $P(Y = y_j | X = x)$, $j = 1, \dots, J$ gives the conditional distribution of Y given $X = x$.
 - $P_{X|Y=y}$ denotes this distribution.
- $P(x, y) = P_{Y|X=x}(y)P_X(x) = P_{X|Y=y}P_Y(y)$
- X and Y are independent:
 - iff $P(X = x, Y = y) = P(X = x)P(Y = y)$ for all x, y
 - iff $P_{X|Y=y}(x) = P_X(x)$ for all x, y
 - iff $P_{Y|X=x}(y) = P_Y(y)$ for all x, y

- Facts

1. If (X, Y) are independent,

$$E[g(X)h(Y)] = E[g(X)]E[h(Y)]$$

and in particular,

$$E(XY) = EXEY \implies \text{Cov}(X, Y) = 0$$

2. If X, Y are bivariate normal and $\text{Cov}(X, Y) = 0$, then they are **independent**.

- In the continuous case, the above hold too.

$$\begin{aligned}
 & f(x, y) = f_X(x)f_Y(y) \\
 \iff & \text{independent} \\
 \iff & f_{Y|X=x}(y) = f_Y(y) \\
 \iff & f_{X|Y=y}(x) = f_X(x)
 \end{aligned}$$

- If X and Y are independent then $X + Y$ is also called the **convolution** of X and Y and its density is

$$f(z) = \int f(x)h(z - x)dx$$

where $f(x)$ is the density of X and $h(y)$ is the density of Y .

Central Limit Theorem (CLT)

If X_1, \dots, X_n i.i.d with $EX = \mu$, $\text{Var}(X) = \sigma^2$, then for n large.

$$\frac{X_1 + \dots + X_n - n\mu}{\sqrt{n}\sigma} \sim N(0, 1)$$

Equivalently for large n ,

$$\begin{aligned} \sum X_i &\sim N(n\mu, n\sigma^2) \\ \text{and } \bar{X} = \frac{1}{n} \sum X_i &\sim N\left(\mu, \frac{\sigma^2}{n}\right) \end{aligned}$$