6-2 공분산 및 상관계수

6-2-1. 공분산

공분산(covariance)

두 확률변수의 결합분포를 알고 있는 경우에 구할 수 있는 모수

✓주로 두 변수 사이의 관계의 밀접도를 측정하는 상관계수를 구하는 과정에서 계산되는 경우가 많다.

✓공분산의 부호가 **양수이면 두 변수들이 서로 같은 방향으로** 변하고, **음수이면 두 변수들이 서로 반대 방향**으로 변한다.

✓공분산은 **두 변수의 측정단위**에 따라 달라진다.

두 확률변수 ***X*** 와 ***Y*** 의 공분산(covariance)은 ***Cov*** **(*X*, *Y*)** 또는 $\boldsymbol{\sigma_{XY}}$ 로 표현하며 다음과 같이 계산한다.

$$\sigma_{xy} = E[(X-\mu_x)(Y-\mu_y)]$$
$$= E(XY) - \mu_x \times \mu_y$$

(예: 키(X)와 몸무게(Y)의 공분산)

-측정단위($\boldsymbol{X_1}$ **(m), *Y*(kg))**: ***Cov*** **(*X₁*, *Y*)** $= \sigma_{x1,y} = \mathrm{E}[(X_1 - \mu_{x1})(Y - \mu_y)]$

-측정단위($\boldsymbol{X_2}$ **(cm), *Y*(kg))**: ***Cov*** **(X_2, *Y*)** $= \sigma_{x2,y} = \mathrm{E}[(X_2 - \mu_{x2})(Y - \mu_y)]$

$$= \mathrm{E}[(100X_1 - 100\mu_{x1})(Y - \mu_y)]$$

$$= 100\sigma_{x1,y}$$

6-2-2. 상관계수

상관계수(correlation coefficient)

두 변수 사이의 관계의 밀접도를 측정하는 통계량

✓상관관계수의 부호가 양수이면 두 변수들이 서로 같은 방향으로 변하고, 음수이면 두 변수들이 서로 반대 방향으로 변한다.

✓상관계수는 두 변수의 측정단위에 영향을 받지 않는다.

두 확률변수 X 와 Y 에 대하여 σ_X^2, σ_Y^2 을 각각의 분산이라 하고,

σ_{XY} 를 X 와 Y 의 공분산이라고 할 때, X 와 Y 의 상관계수는 ρ 로 표현한다.

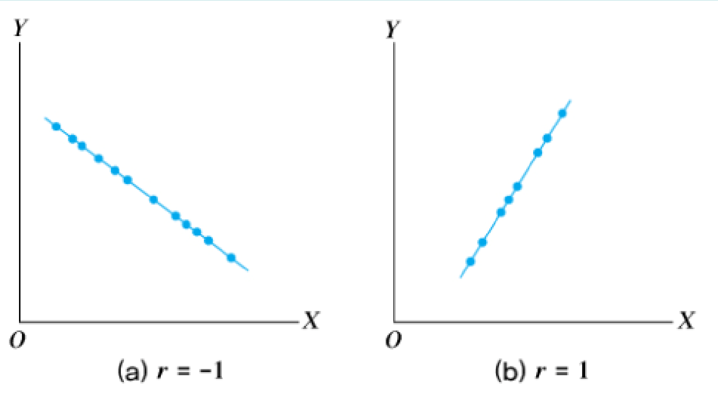
$$\rho = \frac{\sigma_{xy}}{\sqrt{\sigma_X{}^2 \cdot \sigma_Y{}^2}} = \frac{\sigma_{xy}}{\sigma_X \cdot \sigma_Y}$$

$$S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$$

$$S_X = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

$$S_Y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2}$$

① 완전 선형인 경우



왼쪽 그림과 같이 (a), (b)와 같이
(X , Y) 의 관계가 완전선형
Y=a+bX (b ≠0) 인 경우에는
상관계수가 r=1(b>0)
또는 r=-1(b<0)으로 나타난다.

② 어느 정도 상관관계가 있는 경우

산포도의 분포 폭이 중심축으로부터 커질수록 ㅣrㅣ 값의 크기는 0에 가까워진다.

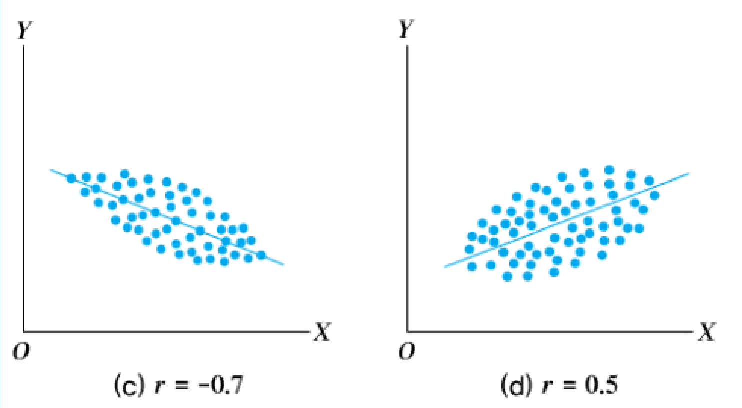


그림 (c) 와 (d) 를 비교할 때ㅣrㅣ=0.5인 경우가 ㅣrㅣ=0.7인 경우보다
분포의 폭이 더 큰 것을 알 수 있다.

③ 상관관계가 없는 경우

r=0인 경우는 두 변수 사이에 선형관계가 없음을 의미한다.

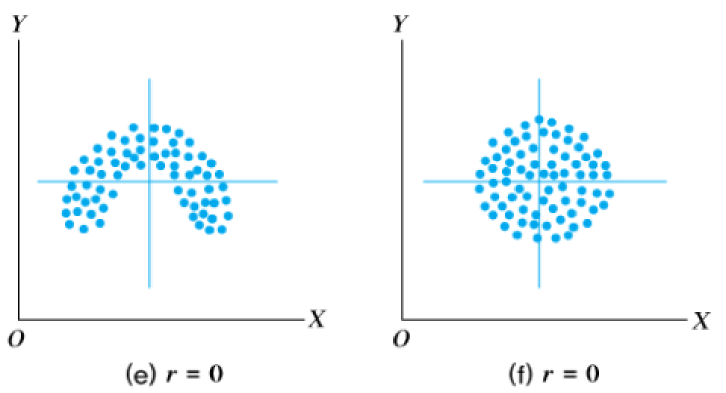


그림 (e), (f)와 같이 산포도가 하나의 특정한 중심축을 그릴 수 없는 경우와 관계식 Y=a+bX에서 b=0인 경우가 여기에 해당된다.

예 이산형 확률변수의 공분산 및 상관계수 계산

예 : 포트폴리오의 구성

포트폴리오를 구성하고 있는 두 증권 X와 Y의 수익률에 대한 분포가 다음의 표와 같고 투자금액을 X증권에 60%, Y증권에 40%투자할 때 포트폴리오의 기대수익률과 분산 그리고 X,Y의 공분산 및 상관계수를 구하라.

X \ Y	-1	6	2	20	f(x)
5	0.1	0	0	0	0.1
7	0	0.4	0	0	0.4
-4	0	0	0.3	0	0.3
15	0	0	0	0.2	0.2
f(y)	0.1	0.4	0.3	0.2	1.0

$$R = 0.6X + 0.4Y$$

$$E(X) = \sum xf(x) = 5(0.1) + 7(0.4) + (-4)(0.3) + 15(0.2) = 5.1$$

$$E(Y) = \sum yf(y) = (-1)(0.1) + 6(0.4) + 2(0.3) + 20(0.2) = 6.9$$

$$Var(X) = E(X^2) - [E(X)]^2 = 45.89$$

$$Var(Y) = E(Y^2) - [E(Y)]^2 = 48.09$$

$$E(XY) = \sum \sum xyf(x, y) = 73.9$$

$$Cov(X, Y) = E(XY) - E(X)E(Y) = 73.9 - (5.1)(6.9) = 38.71$$

$$\rho_{X,Y} = \frac{Cov(X, Y)}{\sigma_X \sigma_Y} = \frac{38.71}{\sqrt{45.89}\sqrt{48.09}} = 0.824$$

$$E(R) = 0.6(5.1) + 0.4((6.9) = 5.82$$

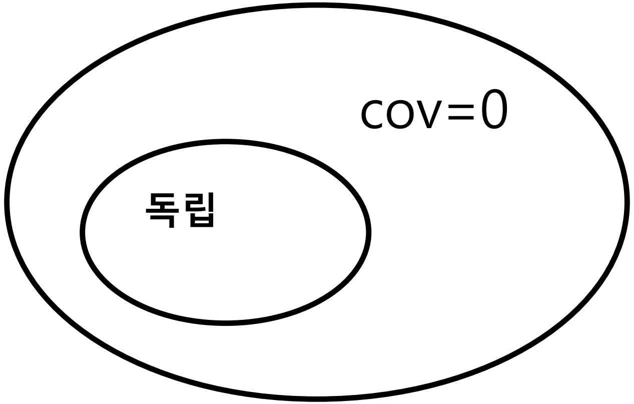
$$Var(R) = 0.6^2(45.89) + 0.4^2(48.09) + 2(0.6)(0.4)(38.71) = 42.8$$

6-2-3. 독립성과 공분산

두 확률변수 X와 Y가 서로 독립이면 cov(X,Y)=0이지만 cov(X,Y)=0이라고 해서 반드시 X와 Y가 서로 독립인 것은 아니다.

즉, cov(X,Y)=0은 독립이기 위한 필요조건이지 충분조건은 아니다.

$$\text{독립} \xrightarrow[\leftarrow \times]{\bigcirc \rightarrow} Cov=0$$



예 다음의 결합확률분포를 이용해 위의 내용을 보여라.

Y/X	-1	0	1	Y의 주변확률
-1	1/6	1/3	1/6	2/3
0	0	0	0	0
1	1/6	0	1/6	1/3
X의 주변확률	1/3	1/3	1/3	1.0

$$E(X) = \sum xf(x) = (-1)\frac{1}{3} + 0\frac{1}{3} + 1\frac{1}{3} = 0$$

$$E(Y) = \sum yf(y) = (-1)\frac{2}{3} + 00 + 1\frac{1}{3} =- \frac{1}{3}$$

$$E(XY) = \sum \sum xyf(x,y) = (-1)(-1)\frac{1}{6} + (-1)(0)(0) + (-1)(1)(\frac{1}{6}) +$$

$$\dots + (1)(1)(\frac{1}{6}) = 0$$

$Cov(X, Y) = E(XY) - E(X)E(Y) = 0$

$f(x,y) = f(x)f(y)$인가?

예를 들어,

$$f(X=-1, Y=-1) = \frac{1}{6} \neq f(X=-1)(Y=-1) = \frac{1}{3}\frac{2}{3} = \frac{2}{9}$$

따라서, Cov(X,Y)=0이지만 X,Y는 서로 독립이 아니다.

독립성

만약에 X와 Y가 서로 독립이면 다음의 관계가 성립한다.

ⅰ) $cov(X, Y) = E(XY) - E(X)E(Y) = 0$ <- $E(XY) = E(X)E(Y)$ 만약에 X, Y가 독립이면)

ⅱ) $var(X \pm Y) = var(X) + var(Y)$

ⅲ) $\rho_{XY} = \dfrac{cov(X, Y)}{\sigma_X \sigma_Y} = 0$