7.4 다변수 확률변수

하나의 값을 가질 수 있는 이산확률변수가 두 개 이상 있는 경우에는
각각의 확률변수에 대한 확률분포 이외에도 확률분포 상의 개개의 복합적인 확률분포를
살펴보아야 한다. 이 절에서는 여러 개의 다변수 확률변수의 확률분포를 표현하기 위한
결합확률분포함수를 알아본다.

- 두 확률변수 값의 쌍이 어떤 확률분포를 갖는지 안다면, 둘 중 하나의 확률분포의 값을 알고 있을 때 다른 확률분포가 어떻게 되는지도 알 수 있다.
- 이러한 정보를 나타내는 조건부확률분포에 대해서도 공부한다.

결합확률질량함수

주사위처럼 1부터 6까지의 값을 가지는 카테고리 분포 확률변수 X와 Y를 생각하자.
확률변수 각각의 확률적 특성은 확률질량함수 $P_X(x)$, $P_Y(y)$로 나타낼 수 있다.

(공정한 주사위)

$$P_X(1) = \frac{1}{6}, \cdots, P_X(6) = \frac{1}{6}$$

$$P_Y(1) = \frac{1}{6}, \cdots, P_Y(6) = \frac{1}{6}$$

특정한 숫자 쌍 (pair)이 나타나는 경우
… 하나하나의 숫자쌍에 대해 확률을 알려주는 PMF (확률질량함수)만 있으면
전체 확률분포를 알 수 있다.

→ 결합확률질량함수 (joint probability mass function)

$$P_{XY}(x, y)$$

(공정한 주사위 X, Y)

$$P_{XY}(1,1)=\frac{1}{36},\ P_{XY}(1,2)=\frac{1}{36},\ \cdots,\ P_{XY}(6,6)=\frac{1}{36}$$

주변확률질량함수 "marginal pmf"

"주변확률질량함수"(marginal probability mass function)는 두 확률변수 중 하나의 확률변수 값에 대해서만 확률분포를 표시한 함수이다. 즉, 다변수가 되기 이전의 단변수 확률질량함수를 말한다.

결합확률질량함수에서 주변확률질량함수를 구하려면 전체 확률의 법칙에 의해 "다른 변수"가 가질 수 있는 모든 값의 결합확률질량함수를 합한 확률이 된다.

$$P_X(x)=\sum_{y_i} P_{XY}(x,y_i)$$

$$P_Y(y)=\sum_{x_i} P_{XY}(x_i,y)$$

위에서 예로 든 이산확률변수의 경우에 관심 X만 관심이 있다면, 결합확률질량함수 $P_{XY}(x,y)$로부터 X에 대한 주변확률질량함수 $P_X(x)$를 구해야 한다.

주변확률질량함수를 계산한 값은 다음과 같다.

$$P_X(A)=P_{XY}(A,A)+P_{XY}(A,B)+P_{XY}(A,C)$$

(ex) y가 주어진 경우의 결합확률질량함수의 단면.

$P(x, y=A)$

$P(x, y=B)$

$P(x, y=C)$

$P(x, y=D)$

$P(x, y=E)$

$P(x, y=F)$

조건부확률질량함수 $P_{X|Y}(x|y) = \frac{P_{XY}(x,y)}{P_Y(y)}$, $P_{Y|X}(y|x) = \frac{P_{XY}(x,y)}{P_X(x)}$

결합질량함수 $P_{XY}(x,y)$에서 y값이 고정된 함수, 즉 결합질량함수의 단면과 같아진다. (다만 조건부확률질량함수의 합은 1이 된다.)

y = A일 때의 결합확률질량함수 단면 $P(x, y=A)$

y=A일 때의 조건부확률질량함수 $p(x|y=A)$

결합누적확률분포함수

두 연속확률변수 X, Y에 대한 결합누적확률분포함수 $F_{XY}(x,y)$는 다음과 같이 정의된다.

$$F_{XY}(x,y) = P(\{X<x\} \cap \{Y<y\}) = P(\{X<x, Y<y\})$$

결합누적확률분포함수 $F_{XY}(x,y)$는 다음과 같은 특성을 갖는다.

$$F_{XY}(\infty,\infty) = 1$$

$$F_{XY}(-\infty, y) = F_{XY}(x, -\infty) = 0$$

결합확률밀도함수

결합누적확률밀도함수를 D분하여 결합확률밀도함수(joint probability density function)를 정의할 수 있다. 즉, 변수 2개에 대해 각각에 대해 편미분(partial diff.)해야 한다.

$$P_{XY} = \frac{\partial^2 F_{XY}(x,y)}{\partial x \partial y}$$

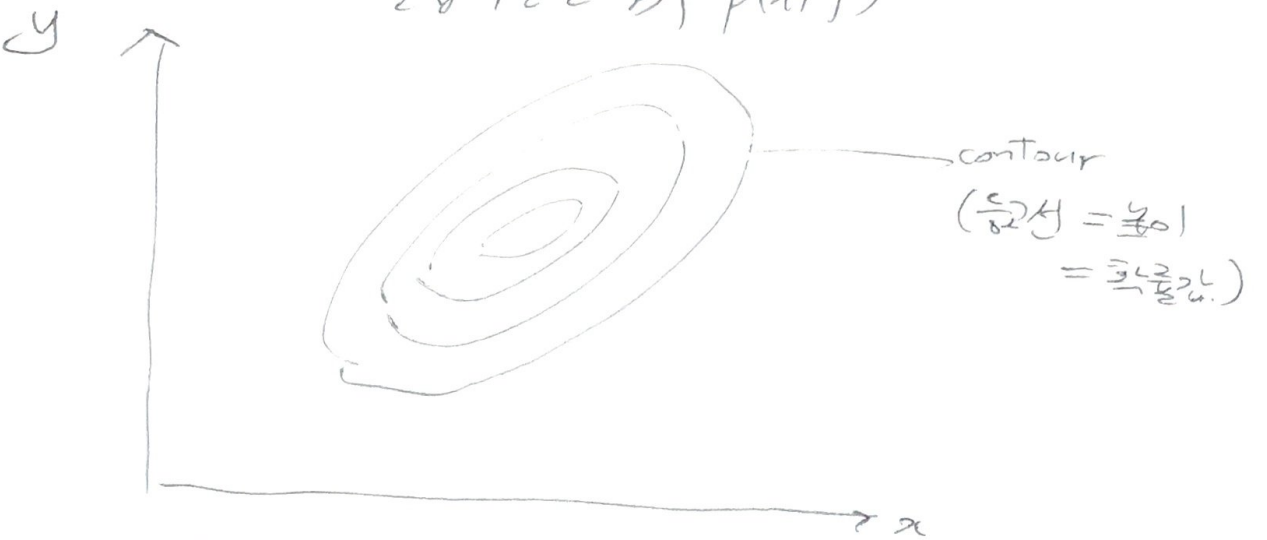
결합확률밀도함수를 특정구간에 대해 적분하면, 해당 구간에 대한 확률이 된다.

$$\int_{x_1}^{x_2}\int_{y_1}^{y_2} P_{XY}(x,y)\,dxdy = P(\{x_1 \le X \le x_2, y_1 \le Y \le y_2\})$$

따라서 결합확률밀도함수를 모든 변수에 대해 $-\infty$에서 ∞까지 적분하면 1이 된다.

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} P_{XY}(x,y)\,dxdy = 1$$

* 결합확률밀도함수 $p(x,y)$



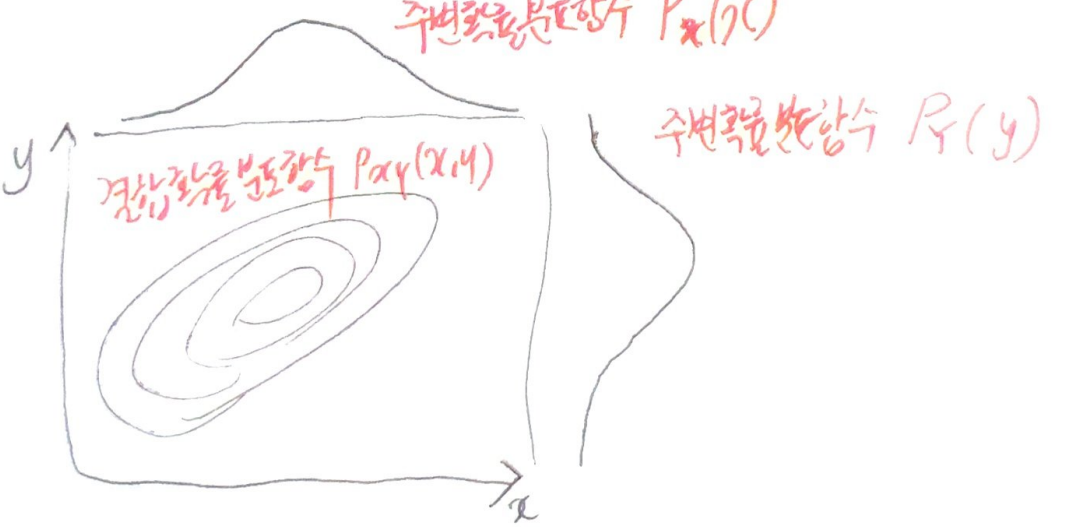
주변확률밀도함수

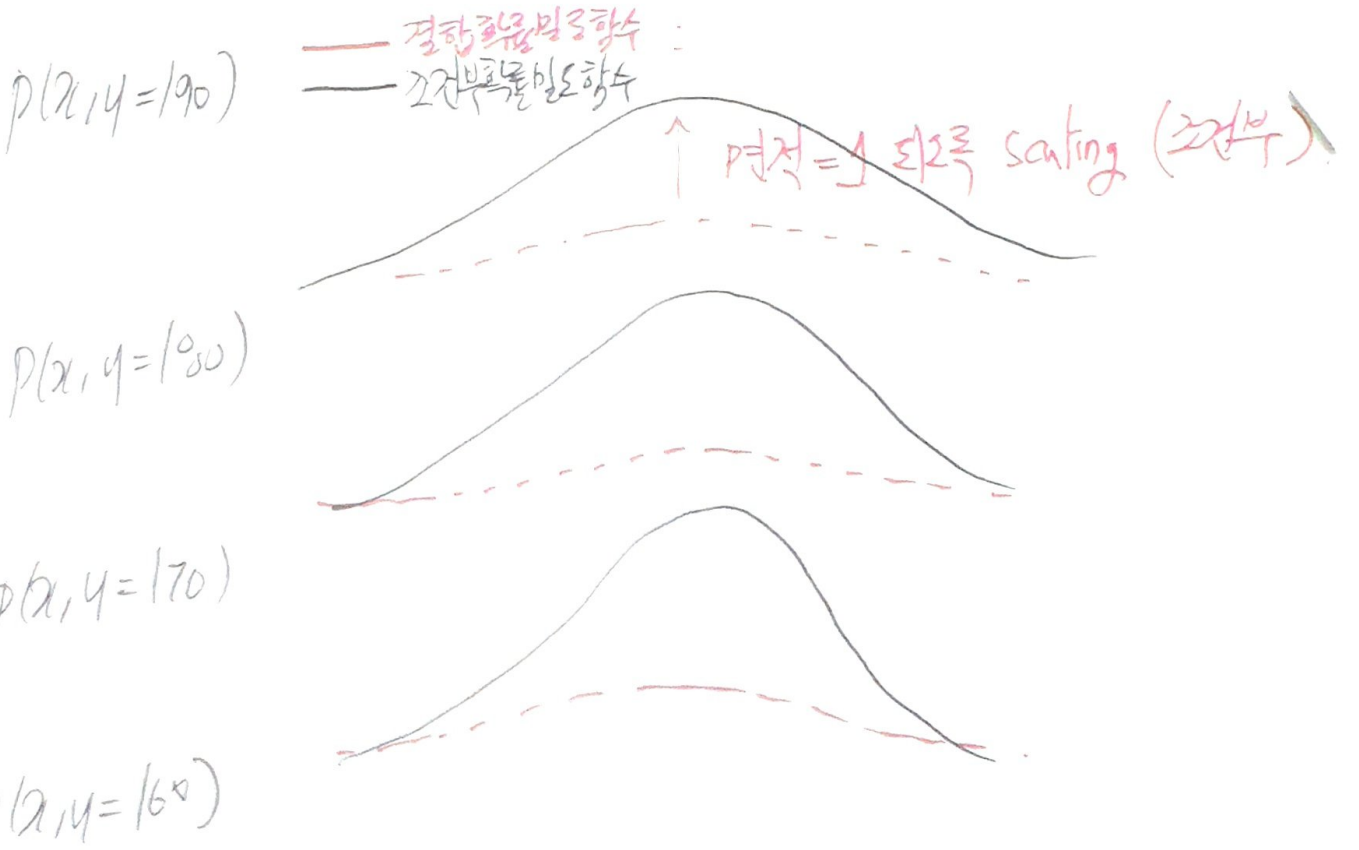
주변확률밀도함수(marginal probability density function)

- 결합확률밀도함수를 특정한 해석의 변수에 대해 가중합한 값.
- 따라서 결합확률밀도함수를 해석의 확률변수에 대해서만 적분하여 구한다.
- 가중합인 (적분)으로 차원이 1개 줄어들기 때문에, 2차원 확률변수의 주변확률밀도함수는 1차원 함수가 된다.

$$P_X(x) = \int_{-\infty}^{\infty} P_{XY}(x,y)\,dy$$

$$P_Y(y) = \int_{-\infty}^{\infty} P_{XY}(x,y)\,dx$$



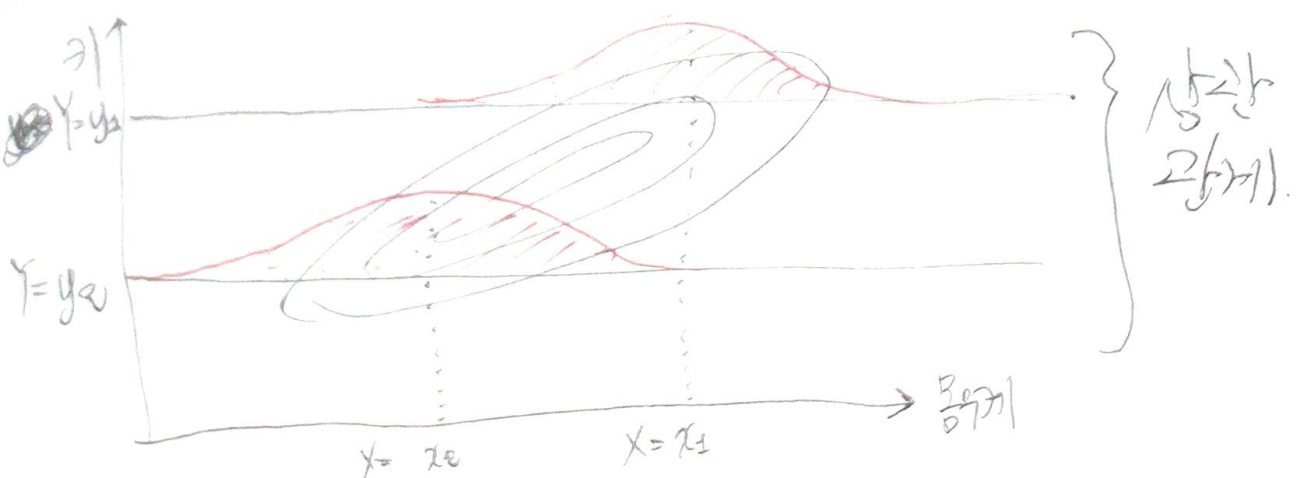


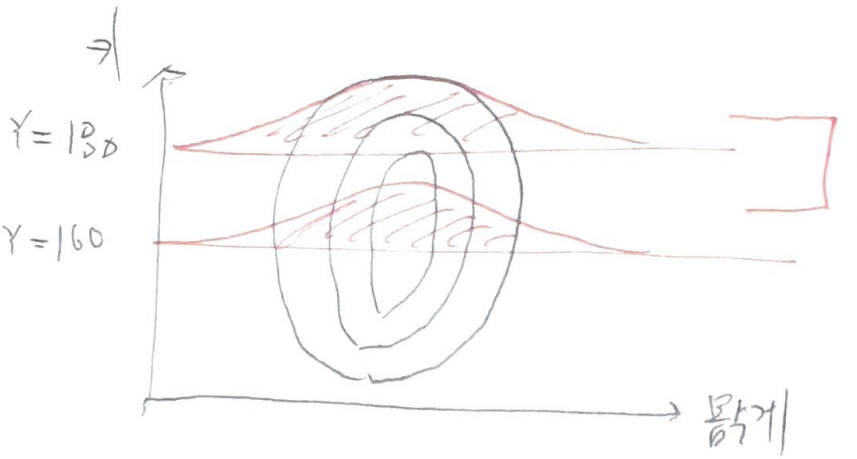
조건부확률밀도함수

$$P(x \mid Y=y_1) = \frac{P_{XY}(x, y_1)}{P_Y(y_1)}$$

독립과 상관

두 확률변수가 있을 때, 한 확률변수의 표본값이 달라져도 다른 확률변수의 조건부 분포가 달라질 때 서로 상관관계가 있다고 한다. 반대로 상관관계가 없으면 독립이라 한다.





면적이 그이 되도록 scaling 해주면 동일한 분포 → 독립

두 확률변수 X, Y의 결합확률밀도함수 (joint pdf) 가 각각의 주변확률밀도함수 (marginal pdf)의 곱과 같으면 서로 독립 (independent) 이다.

$$P_{XY}(x,y) = P_X(x)\,P_Y(y)$$

$$P_{XYZ}(x,y,z) = P_X(x)\,P_Y(y)\,P_Z(z)$$

⋮ ⋮

↓ 이때 X, Y, Z 중 아무 확률변수를 골라도 서로 독립

$$P_{XY}(x,y) = \sum_{z\in\Omega} P_{XYZ}(x,y,z)$$ 전체 확률의 합칙

$$= \sum_{z\in\Omega} P_X(x)\,P_Y(y)\,P_Z(z)$$

$$= P_X(x)\,P_Y(y)\underbrace{\sum_{z\in\Omega} P_Z(z)}_{=1}$$

$$= P_X(x)\,P_Y(y)$$

반복시행

같은 확률변수에서 복수의 표본 데이터를 취하는 것.
이 표본들은 서로 독립인 확률변수들에서 나온 표본으로 볼 수 있다.

- 확률밀도함수 $f(x)$
- 표본데이터 $\{x_1, x_2, \cdots, x_N\}$
- 데이터 $(x_1, x_2, \cdots, x_N)$가 나올 확률은 다음과 같다.

$$p(x_1, x_2, \cdots, x_N) = \prod_{i=1}^{N} p(x_i)$$

조건부 확률분포

독립인 두 확률변수 X, Y의 조건부확률밀도함수는 주변확률밀도함수와 같다.

$$P_{X|Y}(x|y) = \frac{P_{XY}(x,y)}{P_Y(y)} = \frac{P_X(x)P_Y(y)}{P_Y(y)} = P_X(x) \quad y\text{랑 상관} \times$$

$$P_{Y|X}(y|x) = \frac{P_{XY}(x,y)}{P_X(x)} = \frac{P_X(x)P_Y(y)}{P_X(x)} = P_Y(y) \quad x\text{랑 상관} \times$$

<상관관계인 두 확률변수의 결합확률질량함수>

	0.1	0.2	0.4	0.2	0.1
0.1	0	0	0	0.05	0.05
0.2	0	0.05	0.05	0.05	0.05
0.4	0	0.05	0.3	0.05	0
0.2	0.05	0.05	0.05	0.05	0
0.1	0.05	0.05	0	0	0

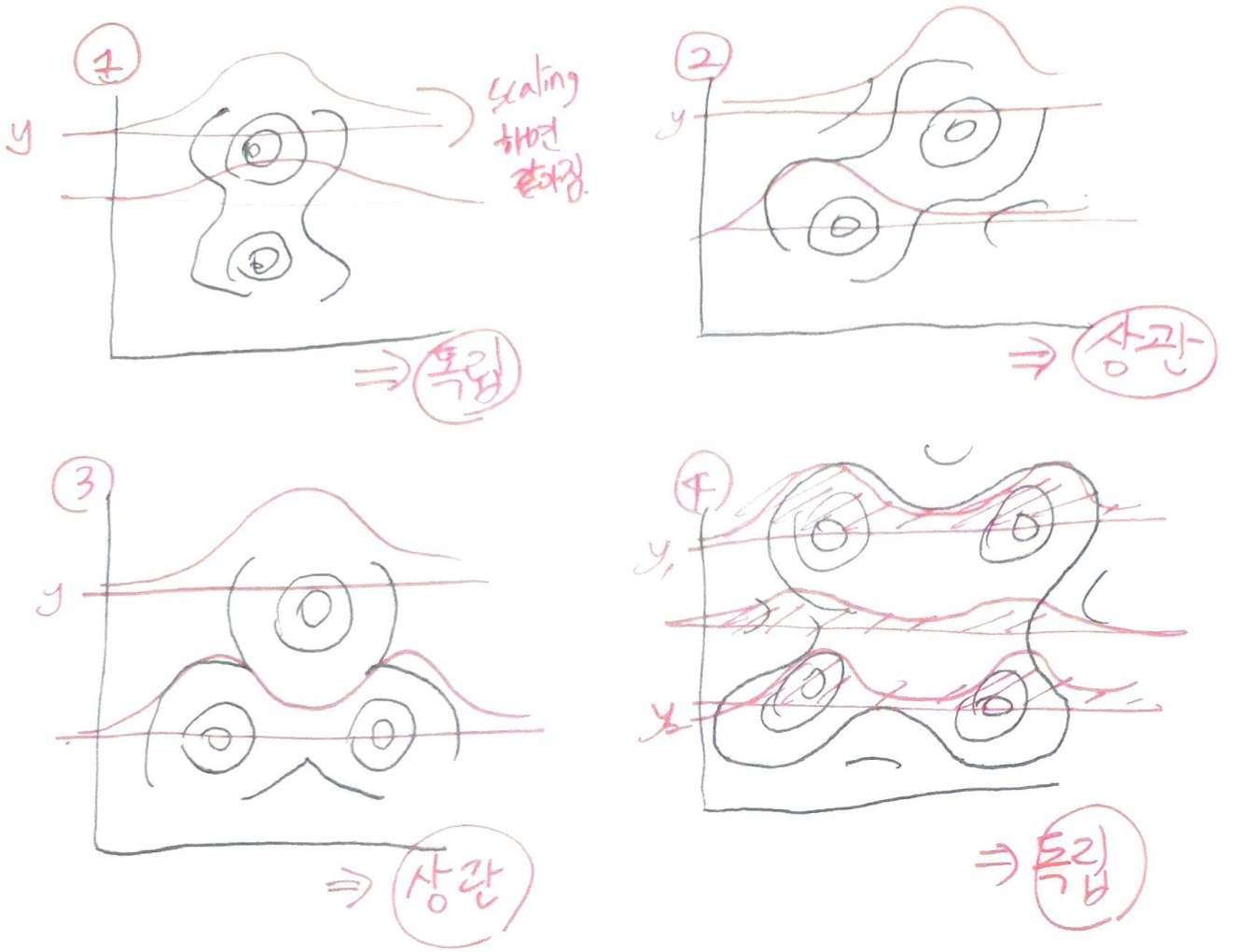
≠

<두 확률변수의 주변확률질량함수의 곱>

0.01	0.02	0.04	0.02	0.01
0.02	0.04	0.08	0.04	0.02
0.04	0.08	0.16	0.08	0.04
0.02	0.04	0.08	0.04	0.02
0.01	0.02	0.04	0.02	0.01

=

Joint PDF 보고 독립/상관 판단하기.



독립확률변수의 기댓값

독립인 두 확률변수 X, Y의 기댓값은 다음 성질을 만족한다.

$$E[XY] = E[X]E[Y] \quad \cdots ①$$

$$E[(X-\mu_X)(Y-\mu_Y)] = 0 \quad \cdots ②$$

(증명)

$$E[XY] = \iint xy\, P_{XY}(x,y)\,dxdy$$

$$= \iint xy\, P_X(x)\, P_Y(y)\,dxdy$$

"다중적분의 값은 적분을 연속하여 한 값과 같다는 푸비니(Fubini) 정리에 의해 다음처럼 증명할 수 있다.

$$E[XY] = \int \left(\int xy P_X(x) P_Y(y) dx \right) dy$$

$$= \int \left(y P_Y(y) \left(\int x P_X(x) dx \right) \right) dy$$

$$= \left(\int x P_X(x) dx \right) \left(\int y P_Y(y) dy \right)$$

$$= E[X] E[Y] \quad \cdots \text{①}$$

∴ 결과를 이용하여 두번째 등식도 다음처럼 증명한다.

$$E[(X-\mu_X)(Y-\mu_Y)] = E[XY - \mu_X Y - \mu_Y X + \mu_X \mu_Y]$$

$$= E[XY] - \mu_X E[Y] - \mu_Y E[X] + \mu_X \mu_Y$$

$$= E[XY] - \mu_X \mu_Y$$

$$= E[XY] - E[X]E[Y] = 0 \quad \cdots \text{②}$$

독립 확률변수의 분산

독립인 두 확률변수 X, Y의 분산은 다음 성질을 만족한다.
(바로 윗줄에서 설명한 내용으로 증명 생략)

$$Var[X+Y] = Var[X] + Var[Y]$$