4. 확률 변수와 확률 분포

숫자의 종류

상수 : 상(항상 똑같은) 수(숫자)

변수 : 변(변하는) 수(숫자)

확률 변수 (Random Variable)

특정사건에서 여러 가지 결과들이 시현되는데(변수), 모든 가능한 결과들을 표현하는 방식

즉, 확률이라는 규칙을 가지면서 변하는 사건을 숫자로 표현한 것을 의미

ex) 동전을 한번 던질 때 앞면이 나오는 경우의 수를 확률변수 X라 하면, X = 0, 1

확률 변수의 종류

이산형 확률변수	확률 변수가 가질 수 있는 값들을 셀 수 있는 경우 ex) 동전 던지기, 주사위 던지기 등
연속형 확률변수	확률 변수가 가질 수 있는 값들을 셀 수 없는 경우 ex) 키, 몸무게 등

확률 분포 함수

확률 변수가 가지는 규칙, 즉 확률을 수식으로 표현한 것을 확률 분포 함수라 함

확률 분포 함수의 종류

종류	정의	조건
이산형 확률분포함수	확률 변수가 가질 수 있는 값들을 셀 수 있는 경우, 이를 함수식으로 표현한 것을 의미	1) $0 \leq f(x_i) \leq 1, i = 1, 2, 3, \ldots$ 2) $\sum_{i=1} f(x_i) = 1$
연속형 확률분포함수	확률 변수가 가질 수 있는 값들을 셀 수 없는 경우, 이를 함수식으로 표현한 것을 의미	1) $f(x) \geq 0$ 2) $\int_{-\infty}^{\infty} f(x)dx = 1$

결합 확률 분포

(Joint Probability Distribution)

두 개 이상의 변수에 대한 확률 분포에 대해 정의하는 것

이산형 결합 확률 분포

	y_1	$\cdots$	y_j	$\cdots$	$f(x)$
x_1	p_{11}		p_{1j}		$p_{1.}$
$\cdots$					$\cdots$
x_i	p_{i1}		p_{ij}		$p_{i.}$
$\cdots$					$\cdots$
$g(y)$	$p_{.1}$	$\cdots$	$p_{.j}$	$\cdots$	1

1) p_{ij}는 확률변수 $X = x_i$이고 $Y = y_j$ 에 해당하는 교집합의 확률을 의미

2) 확률의 전체 합은 1

$$p_{i.} = \sum_{j=1} p_{ij} = P(X = x_i) = f(x_i)$$

$$p_{.j} = \sum_{j=1} p_{ij} = P(Y = y_j) = g(y_j)$$

연속형 결합 확률 분포

이산형과 동일하게 다른 변수에 대한 확률의 합으로 구할 수 있으며, 연속형에서 이러한 확률의 합은 적분을 통해 이뤄짐

$f(x) = \int f(x, y)\, dy$	각각의 y값에 대응되는 하나의 x값의 합
$g(x) = \int f(x, y)\, dx$	각각의 x값에 대응되는 하나의 y값의 합
$\int\int f(x, y) dxdy = 1$	전체 합 = 1

기댓값 (Expectation)

기댓값은 확률 변수의 값들이 가질 수 있는 확률을 가중치로 부여하여 계산한 가중평균의 개념

확률 변수가 어느 값을 가질 것으로 기대되는가를 구하는 것

분포의 중심위치를 계산하는 것으로 이해

Cf. 평균 : 각각이 가지는 변수의 가중치가 동일 (기댓값은 다를 수 있음)

기댓값 (Expectation)

이산형 확률변수의 기댓값

$$E(X) = \sum xf(x)$$

연속형 확률변수의 기댓값

$$E(X) = \int xf(x)dx$$

분산 (Variance)

확률 변수 값들이 평균으로부터 얼마나 퍼져있는지를 나타내는 통계량

$$V(X) = \sigma^2$$

분산 (Variance)

이산형 확률변수의 분산

$$V(X) = E[(X-\mu)^2] = \sum (x-\mu)^2 f(x)$$

연속형 확률변수의 분산

$$V(X) = E[(X-\mu)^2] = \int (x-\mu)^2 f(x)dx$$

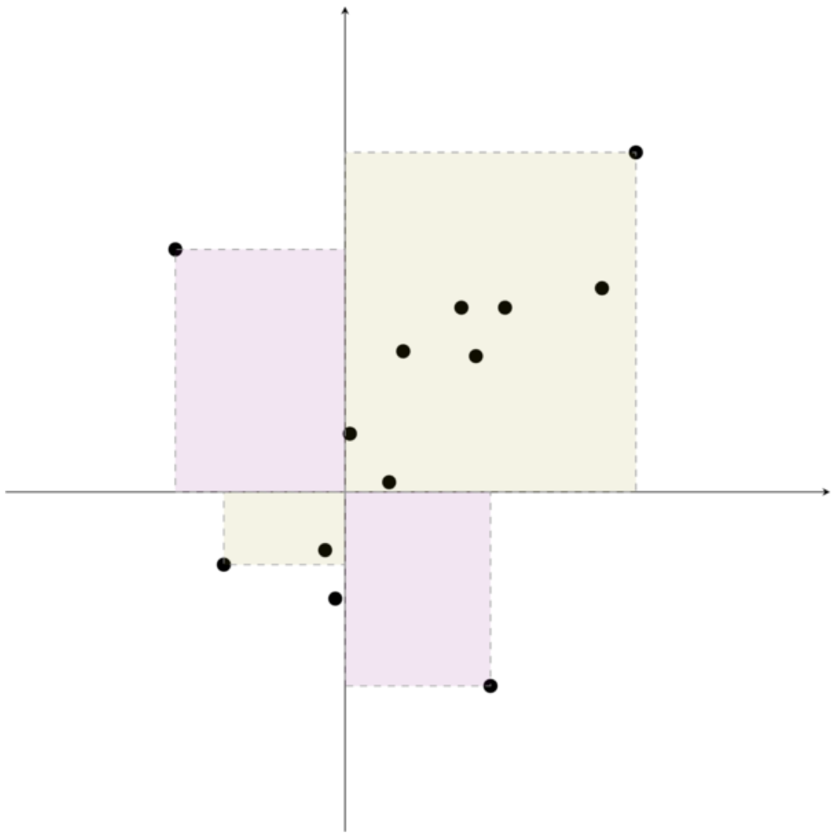
두 변수의 관계 요약

예를 들어 키와 몸무게는 어떤 관계를 가지고 있는지, 소득 수준과 소비 성향은 어떤 관계를 가지고 있는지를 설명해야 하는 경우

이러한 관계를 설명하고자 하는 통계량이 공분산, 상관계수

공분산 (Covariance)

두 확률변수의 값이 평균값으로부터 떨어져 있는 면적들의 기댓값을 의미



$$\mathrm{COV}(X,Y) = E\left[(X-\mu_x)(Y-\mu_y)\right] = E(XY) - \mu_x\mu_y$$

이산형 : $E(XY) = \sum_x \sum_y xyf(x,y)$

연속형 : $E(XY) = \int\int xyf(x,y)dxdy$

면적의 넓이가 1,3분면이 크면 양의 관계

면적의 넓이가 2,4분면이 크면 음의 관계

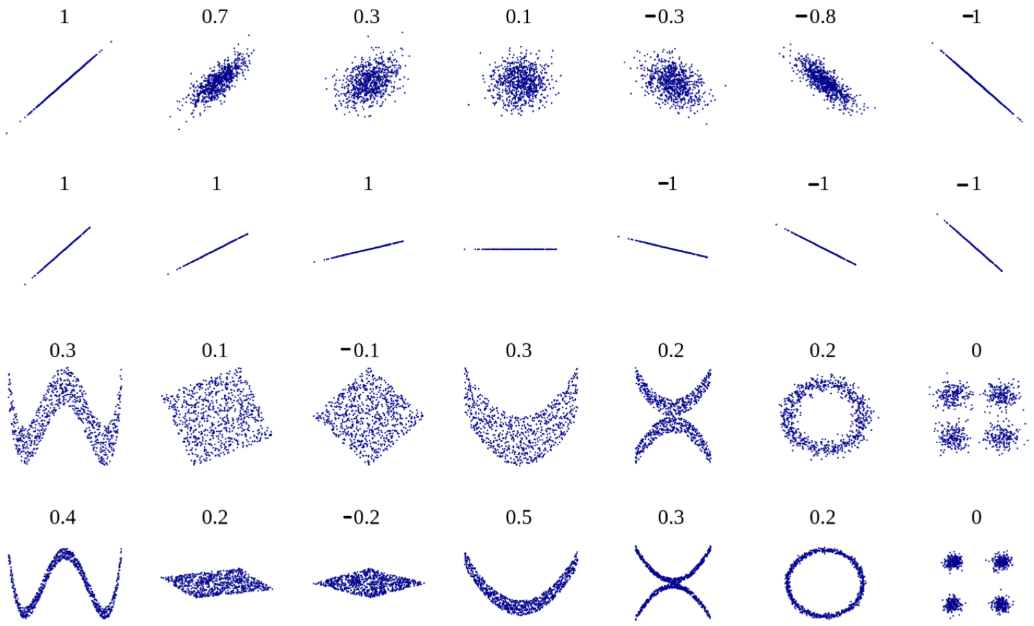
관계의 방향은 설명할 수 있으나, 관계의 정도를 측정하지 못함

ex) 단위에 따른 절대값이 다름

상관계수 (Correlation)

공분산의 문제점 (변수가 가지는 단위의 값에 따라 크기가 결정)을 해결하기 위해 만들어진 개념

공분산에 표준편차를 나눠 값의 표준화로 만듦



$$p_{xy} = \mathrm{Corr}(X, Y) = E\left[\frac{(X-\mu_x)(Y-\mu_y)}{\sigma_x\sigma_y}\right] = \frac{Cov(X, Y)}{\sigma_x\sigma_y}$$

값의 표준화를 통해 상관계수를 $-1 \le p_{xy} \le 1$ 범위에 있음

1에 가까울수록 양의 관계가 큼

-1에 가까울수록 음의 관계가 큼

0에 가까울수록 무상관일 가능성이 큼

두 변수가 독립적일 때 의미

하나의 변수가 다른 변수가 가지는 값에 대하여 영향을 미치지 못한다는 것

두 변수가 독립이라면 공분산과 상관계수는 0

$$\mathrm{COV}(X, Y) = E\left[(X - \mu_x)(Y - \mu_y)\right] = E(XY) - \mu_x\mu_y = \mu_x\mu_y - \mu_x\mu_y = 0$$

$$p_{xy} = \mathrm{Corr}(X, Y) = E\left[\frac{(X - \mu_x)(Y - \mu_y)}{\sigma_x\sigma_y}\right] = \frac{Cov(X, Y)}{\sigma_x\sigma_y} = \frac{0}{\sigma_x\sigma_y} = 0$$