Week 11. 통계, 기댓값, 분산, 공분산, 상관계수, 공분산 행렬

11.1 기댓값, 분산, 표준편차

- 확률변수의 기댓값 (expectation)
  확률적 사건에 대한 평균값, 사건의 일어나는 값과 그 사건이 일어날 확률을 곱한 것을 모든 사건에 대해 합한 값
  → "확률적 사건에 대한 평균의 의미"
- 확률변수의 분산 (variance)
  확률변수가 기댓값으로부터 얼마나 떨어진 곳에 분포하는지를 가늠하는 수.
- 확률변수의 표준편차 (standard deviation)
  분산의 양의 제곱근

(1) 기댓값 $E(X) = \mu = \sum_i x_i f(x_i)$

(2) 분산 $V(X) = \sigma^2 = \sum_i (x_i - \mu)^2 f(x_i) = E(X^2) - \mu^2$

(3) 표준편차 $S(X) = \sigma = \sqrt{V(X)} = \sqrt{\sum_i (x_i - \mu)^2 f(x_i)}$

※ 연속확률변수 X의 기댓값과 분산, 표준편차 계산

(1) 기댓값 $E(X) = \mu = \int_{-\infty}^{\infty} x f(x)\, dx$

(2) 분산 $V(X) = \sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx = E(X^2) -$

(3) 표준편차 $S(X) = \sigma = \sqrt{V(X)} = \sqrt{\int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx}$

기댓값, 분산에 대해서 다음 성질이 만족한다.

① $E(X+b) = E(X)+b$, $E(aX) = aE(X)$, $E(ax+b) = aE(X)+b$

② $V(X+b) = V(X)$, $V(aX) = a^2V(X)$, $V(aX+b) = a^2V(X)$

③ 확률변수 X에 대하여 새로운 확률변수를 $Z = \frac{X-\mu}{\sigma}$로 정의하면,
위의 성질 ①, ②에 의해 평균과 분산이 항상 0과 1이다.
따라서 이 확률변수 Z를 확률변수 X의 **표준화 확률변수** 라 한다.
Standardized random variable

4.2 결합 확률분포

확률변수가 두 개 이상 있는 경우에는 각각의 확률변수에 대한 확률분포 이외에도 확률분포 쌍이 가지는 복합적인 확률분포를 살펴보아야 한다.

두 확률변수의 값의 쌍이 어떤 확률분포를 가지는지 안다면, 둘 중 하나의 확률분포의 값을 알고있을때 다른 확률분포가 어떻게 되는지도 알 수 있다.

이를 위하여 **결합 확률분포** (=결합분포)에 대한 개념이 필요하다.

(1) X와 Y가 이산확률변수이면,
X와 Y의 결합 확률함수 (joint prob. func)

$P_{ij} = p(x_i, y_j) = P(X=x_i, Y=y_j)$

$(x_i, y_j \in \mathbb{R}, i=1,\cdots s, j=1,\cdots t)$

(2) X와 Y의 가능한 모든 값에 대하여
$p(x_i, y_j)$의 값을 나타낸 것을
결합 확률분포라 한다 ⟹

X \ Y	y_1	y_2	⋯	y_t	
x_1	p_{11}	p_{12}	⋯	p_{1t}	→ $P(X=x_1)$
x_2	p_{21}	p_{22}	⋯	p_{2t}	→ $P(X=x_2)$
⋮	⋮	⋮	⋱	⋮	⋮
x_s	p_{s1}	p_{s2}	⋯	p_{st}	→ $P(X=x_s)$
	↓	↓		↓	
Sum	$P(Y=y_1)$	$P(Y=y_2)$	⋯	$P(Y=y_t)$	→ 1

(X에 관한 주변확률분포: Sum 열)

(Y에 관한 주변확률분포: Sum 행)

(3) X, Y의 결합분포가 주어져 있을 때 주변확률분포는 다음과 같이 정의

$$P(X=x_i) = \sum_{j=1}^{t} P(X=x_i, Y=y_j) = \sum_{j=1}^{t} p_{ij}$$

$$P(Y=y_j) = \sum_{i=1}^{s} P(X=x_i, Y=y_j) = \sum_{i=1}^{s} p_{ij}$$

● 주변분포란,
결합확률분포에서 하나의 확률변수만 고려한 확률분포를 뜻한다.

결합확률분포에 관하여 다음이 성립한다.

① 모든 x_i, y_j에 대하여 $p(x_i, y_j) \geq 0$ $i, j = 1, 2, \cdots$

② 모든 x_i, y_j에 대하여 $p(x_i, y_j)$의 합은 1이다.
즉 $\sum_{i=1}^{s} \sum_{j=1}^{t} p(x_i, y_j) = 1$

③ 모든 x_i, y_j에 대하여 $P(a<X<b, c<Y<d) = \sum_{a<x_i<b} \sum_{c<y_j<d} p(x_i, y_j)$

예제 크기같은 파란공 3개, 붉은공 2개, 녹색공 3개가 한 주머니 안에 들어있다.
이 주머니에서 임의로 2개 공을 꺼낸다. 꺼낸 공 중 파란공 개수 = X,
빨간공 개수 = Y라 할때 다음물음에 답하시오.

① X와 Y의 결합확률함수를 구하여라.

$$p(x_i, y_j) = \frac{{}_3C_x \; {}_2C_y \; {}_3C_{2-x-y}}{{}_8C_2} \quad (x=0,1,2;\ y=0,1,2;\ 0 \leq x+y \leq 2)$$

② 결합 확률분포를 작성하여라.

$(X,Y) = \{(0,0), (0,1), (0,2), (1,0), (1,1), (1,2), (2,0)\}$

$$P(X=0, Y=0) = \frac{{}_3C_0\ {}_2C_0\ {}_3C_2}{{}_8C_2} = \frac{3}{28}$$

$$P(X=0, Y=1) = \frac{{}_3C_0\ {}_2C_1\ {}_3C_1}{{}_8C_2} = \frac{3}{14}$$

$$P(X=1, Y=0) = \frac{{}_3C_1\ {}_2C_0\ {}_3C_1}{{}_8C_2} = \frac{9}{28}$$

⋮ ⋮ ⋮

X \ Y	0	1	2	sum	
0	$\frac{3}{28}$	$\frac{3}{14}$	$\frac{1}{28}$	$\frac{5}{14}$	→ $P(X=0)$
1	$\frac{9}{28}$	$\frac{3}{14}$	0	$\frac{15}{28}$	→ $P(X=1)$
2	$\frac{3}{28}$	0	0	$\frac{3}{28}$	→ $P(X=2)$
sum	$\frac{15}{28}$	$\frac{3}{7}$	$\frac{1}{28}$	①	
	↓	↓	↓		
	$P(Y=0)$	$P(Y=1)$	$P(Y=2)$		

③ $P(X+Y \le 1)$을 구하여라.

$p(0,0), p(1,0), p(0,1)$ for $p(X,Y) \rightarrow \frac{3}{28} + \frac{3}{14} + \frac{9}{28} = \frac{9}{14}$

④ X의 주변분포를 구하여라.

$$P(X=0) = p(0,0) + p(0,1) + p(0,2) = \frac{5}{14}$$
$$P(X=1) = p(1,0) + p(1,1) = \frac{15}{28}$$
$$P(X=2) = p(2,0) = \frac{3}{28}$$

⑤ Y의 주변분포를 구하여라.

$$P(Y=0) = p(0,0) + p(1,0) + p(2,0) = \frac{15}{28}$$
$$P(Y=1) = p(0,1) + p(1,1) = \frac{3}{7}$$
$$P(Y=2) = p(0,2) = \frac{1}{28}$$

■ 여러 개의 연속확률변수에 대하여는 다음과 같은 결합밀도함수를 이용하여 결합 확률분포를 나타낸다. (2중적분 필요)
연속확률변수 X와 Y의 결합밀도함수(joint density function) $f(x,y)$는 다음과 같이 정의된다.

① 모든 x,y에 대하여 $f(x,y) \geq 0$ 이다.

② 모든 x,y에 대하여 $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dxdy = 1$

③ 모든 x,y에 대하여 $P(a<X<b,\ c<Y<d) = \int_c^d \int_a^b f(x,y)\,dxdy$

④ (X,Y)가 xy평면상의 임의의 영역 A에 들어갈 확률은

$$P\{(X,Y)\in A\} = \iint_A f(x,y)\,dxdy$$

⑤ X와 Y의 주변확률밀도함수(marginal probability density function)는 각각 다음과 같이 정의된다.

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,dy,\quad f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\,dx$$

11.3 공분산, 상관계수

1) 하나의 확률변수 X가 갖는 분포를 이해하기 위해서 첫번째로 사용하는 것은 (평균)이다.
— 평균을 이용하면 분포에 관한 정보를 하나의 숫자(분포중간부분)로 표시가능

2) 두 번째로 사용하는 개념은 분산 — "분포가 평균으로부터 얼마나 퍼져 있는가?"

(?) 그렇다면 확률변수가 2개일 때, 이 확률분포들이 어떤 방향으로 되어 있는지 어떻게 알 수 있는가?

1. X의 평균과 Y의 평균
2. 분산을 이용하여 X, Y가 퍼진 정도를 확인

* 그러나 확률변수 간의 상관관계를 알기 위해서는 공분산(covariance) 개념 필요!

확률변수 X와 Y의 공분산은 다음과 같이 정의된다.

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

즉, 공분산은 X의 편차와 Y의 편차를 곱한 것의 평균이다.
— 그런데 공분산에도 X, Y 단위의 크기에 영향을 받는다는 문제점이 있다.
— 이것을 보완하기 위해 상관계수(correlation)를 사용한다.
▰ 확률변수의 절대크기에 영향을 받지 않도록, 각 확률변수의 표준편차로 나누어 표준화시킨 것

확률변수 X, Y 사이의 상관계수 correlation

$$Corr(X,Y) = \rho = \frac{Cov(X,Y)}{S(X)S(Y)} = \frac{E[(X-\mu_x)(Y-\mu_y)]}{\sqrt{E(X-\mu_x)^2 \cdot E(Y-\mu_y)^2}}$$

11.4 공분산 행렬 (covariance matrix)

행렬을 이용하면 여러 개의 확률변수가 서로 어떤 관계를 갖는지를 쉽게 표현할 수 있다.
특히, 각 데이터의 분산과 공분산을 이용해 만든 공분산 행렬이 이에 해당한다.

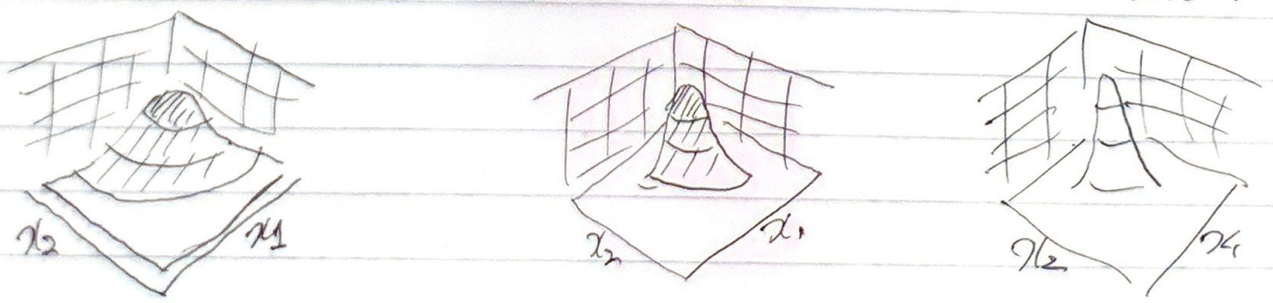
p개의 확률변수 $\{X_1, \cdots, X_p\}$에 대한 공분산행렬 (covariance matrix)은 (i,j) 행렬성분이 $i \neq j$일 때는 i번째 확률변수 x_i와 j번째 확률변수 x_j 사이의 공분산 σ_{ij}으로, $i=j$일 때는 i번째 확률변수의 분산 $\sigma_{ii} = \sigma_i^2$으로하는 $p \times p$ 행렬로 정의하고 Σ로 표기한다.

$$\Sigma = \begin{bmatrix} Var(X_1) & Cov(X_1, X_2) & \cdots & Cov(X_1, X_p) \\ Cov(X_2, X_1) & Var(X_2) & \cdots & Cov(X_2, X_p) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(X_p, X_1) & Cov(X_p, X_2) & \cdots & Var(X_p) \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_p^2 \end{bmatrix}_{p \times p}$$

쉽게 말하면, 정사각행렬의 성분을 각 변수의 분산 (주대각선)과 공분산으로 채운 것이 바로 공분산 행렬이다. 공분산 행렬은 아래처럼 데이터의 분포를 나타낸다고 볼 수 있다.

$p = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ $\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ $p = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ $\Sigma = \begin{bmatrix} 1 & .5 \\ .5 & 1 \end{bmatrix}$ $p = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ $\Sigma = \begin{bmatrix} 1 & .8 \\ .8 & 1 \end{bmatrix}$



* 공분산 행렬은 고차원 데이터의 분포를 최대한 유지하면서 차원을 효과적으로 줄이는 차원축소 (dimension reduction) 에서 중요한 역할을 한다.
대표적 기법 — PCA (principal component analysis) // 주성분 계산 시 특이값분해 (SVD) 사용