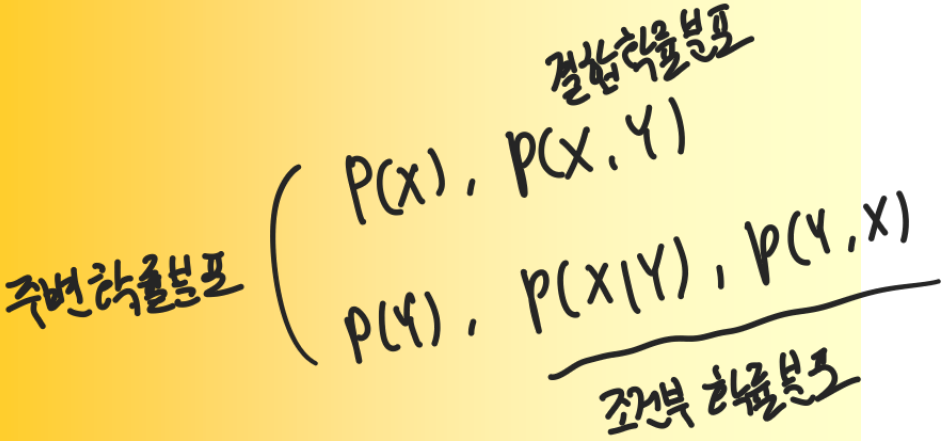
6. 확률변수의 결합분포



담당교수 : 김 덕 기

toby123@cbnu.ac.kr

이산결합확률분포(discrete joint probability distribution)

- 결합분포: 두 확률변수 X와 Y에 대하여 X가 가지는 값과 Y가 가지는 값의 각 쌍에 대응하는 확률을 나타낸 것

- 이산형 결합분포
  - X가 가지는 값: $x_1, x_2, \ldots, x_m$
  - Y가 가지는 값: $y_1, y_2, \ldots, y_n$
  - $f(x_i, y_j) = P(X = x_i, Y = y_j)$: 결합확률질량함수(joint p.m.f.)

이산: Jpmf
연속: Jpdf

		Y			
		y_1	y_2	…	y_n
X	x_1	$f(x_1, y_1)$	$f(x_1, y_2)$	…	$f(x_1, y_n)$
	x_2	$f(x_2, y_1)$	$f(x_2, y_2)$	…	$f(x_2, y_n)$
	…			…	
	x_m	$f(x_m, y_1)$	$f(x_m, y_2)$	…	$f(x_m, y_n)$

이산결합확률분포(discrete joint probability distribution)

결합확률질량함수(Joint probability mass function)

이산확률변수 X와 Y의 결합확률질량함수 $f(x,y) = P(X = x, Y = y)$는 다음 조건을 만족한다.

➔ 임의의 집합 $A = \{(x,y): i \le x \le j, k \le y \le l\}$ 에 대하여

$$P[(X,Y) \in A] = \sum_{x=i}^{j} \sum_{y=k}^{l} P(x,y)$$

(1) 모든 (x,y)에서 $0 \le P(x,y) \le 1$

(2) $\sum_x \sum_y P(x,y) = 1$

두 이산 확률변수의 결합확률분포

(1) 이산형 확률변수의 결합 확률분포

실험 A와 B 두 종류의 곤충에 대한 공존상태를 연구하는 실험

확률변수 X, Y : 각각, 한 나무에 서식하는 곤충 A와 B의 수

$A \rightarrow X$, $B \rightarrow Y$

① X와 Y의 결합확률분포

$$P(X,Y) = \sum_{x=1}^{4}\sum_{y=1}^{3} P(x,y) = 1$$

Y \ X	1	2	3	4	계
0	0	0.05	0.05	0.10	0.20
1	0.08	0.15	0.10	0.10	0.43
2	0.20	0.12	0.05	0	0.37
계	0.28	0.32	0.20	0.20	1

$* \quad P(X) = \sum_{y=0}^{2} P(X,Y)$, $P(Y) = \sum_{x=1}^{4} P(X,Y)$ $\Rightarrow$ 주변확률분포

두 확률변수의 주변확률분포

② 주변 확률분포

x	1	2	3	4	계
$P(X=x)$	0.28	0.32	0.20	0.20	1

[확률변수 X의 주변 확률분포]

y	0	1	2	계
$P(Y=y)$	0.20	0.43	0.37	1

[확률변수 Y의 주변 확률분포]

Question 한 나무에 서식하는 곤충 B의 수가 곤충 A의 수보다 많을 확률은?

0.20 (Y=2, X=1)

이러한 확률은 확률변수 X, Y의 각각의 확률분포로부터 구할 수 없다. 따라서 결합확률분포의 정보가 각각의 확률분포의 정보보다 더 많은 정보를 제공해준다는 사실을 알 수 있다.

결합확률정보=
각각의 확률변수의 주변확률 정보+두 변수의 Correlation(상관정도) 정보

두 확률변수의 기대 값과 분산계산

확률변수 X의 주변확률분포

x	1	2	3	4	계
$P(X=x)$	0.28	0.32	0.20	0.20	1

$$E(X)=\sum_{all\,x} xf(x)=1\times 0.28+2\times 0.32+3\times 0.20+4\times 0.20=2.32$$

$$E(X^2)=\sum_{all x} x^2 f(X)=1^2\times 0.28+2^2\times 0.32+3^2\times 0.20+4\times 0.20=6.56$$

$$Var(X)=E(X^2)-[E(X)^2]=6.56-(2.32)^2=1.1776$$

확률변수 Y의 주변확률분포

y	0	1	2	계
$P(Y=y)$	0.20	0.43	0.37	1

$$E(Y)=\sum_{all\,y} yf(y)=0\times 0.20+1\times 0.43+2\times 0.37=1.17$$

$$E(Y^2)=\sum_{all y} x^2 f(y)=0^2\times 0.20+1^2\times 0.43+2\times 0.37=1.91$$

$$Var(Y)=E(Y^2)-[E(Y)]^2=1.91-(1.17)^2=0.5411$$

두 확률변수의 상관계수추정

|r| ≈ 1 매우 강한
|r| > 2/3 강한
|r| ≈ 1/2 보통
|r| ≈ 1/3 약한
|r| ≈ 0 무상관

결합 확률분포로부터 :

$$E(XY) = \sum_x \sum_y xy f(xy)$$

P(x, y)

$$= 1\times 0\times 0 + 1\times 1\times 0.08 + 1\times 2\times 0.20 + \cdots + 4\times 1\times 0.10 + 4\times 2\times 0$$

$$= 2.26$$

↳ 단위동일

↳ E[(X-Mx)(Y-My)]

$$\therefore\ Cov(X,Y) = E(XY) - E(X)E(Y)$$

공분산

model $y = f(x) + e$

$$= 2.26 - 2.32\times 1.17 = -0.4544$$

(A↑ B↓) or (A↓ B↑)

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

상관계수

$$-1 \le corr(x,y) \le 1$$

$$= \frac{-0.4544}{\sqrt{1.1776}\sqrt{0.5411}} = -0.5692 \ge \frac{1}{2}$$

보통이상.

$r = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$

r ≒ 0 ~ 선형성 X

독립관계 → 상관관계 X

// ⇐ //

두 이산 확률변수의 결합확률분포-example

- [예제 1] 10대의 차 중에서 5대는 좋은 상태(G), 2대는 기아변속기에 문제(DT), 3대는 엔진에 문제(DE)가 있다. 임의로 2대를 선택할 때
  - X = 기어변속기에 문제가 있는 차량의 수
  - Y = 엔진에 문제가 있는 차량의 수
  - 결합확률질량함수:

$$f(0,0) = P(X=0, Y=0) = \frac{\binom{5}{2}}{\binom{10}{2}} = \frac{10}{45}$$

$$f(1,0) = P(X=1, Y=0) = \frac{\binom{2}{1}\binom{5}{1}}{\binom{10}{2}} = \frac{10}{45}$$

$\cdots$

		Y 0	1	2
	0	10/45	15/45	3/45
X	1	10/45	6/45	0
	2	1/45	0	0

- $P(X=Y) = P(X=0, Y=0) + P(X=1, Y=1) = \frac{10}{45} + \frac{6}{45} = \frac{16}{45}$
- $P(X=1) = \sum_{y=0}^{2} P(X=1, Y=y) = \frac{10}{45} + \frac{6}{45} = \frac{16}{45}$
- $P(Y \geq 1) = P(Y=1) + P(Y=2) = \frac{10}{45} + \frac{6}{45} + \frac{3}{45} = \frac{19}{45}$

주변확률분포(marginal distribution)

- 두 확률변수 중에서 어느 한 확률변수의 확률분포를 주변분포라고 한다.
  - X의 주변분포:
    $P_X(x_i) = \sum_j P(X = x_i, Y = y_j) = \sum_j P(x_i, y_j)$
  - Y의 주변분포:
    $P_Y(y_j) = \sum_i P(X = x_i, Y = y_j) = \sum_i P(x_i, y_j)$

		Y				합계
		y_1	y_2	…	y_n	
X	x_1	$p(x_1, y_1)$	$p(x_1, y_2)$	…	$p(x_1, y_n)$	$P_X(x_1)$
	x_2	$p(x_2, y_1)$	$p(x_2, y_2)$	…	$p(x_2, y_n)$	$P_X(x_2)$
	…			…		…
	x_m	$p(x_m, y_1)$	$p(x_m, y_2)$	…	$p(x_m, y_n)$	$P_X(x_m)$
합계		$P_Y(y_1)$	$P_Y(y_2)$	…	$P_Y(y_n)$	1

주변확률분포의 기댓값, 분산

- 두 확률변수 X와 Y의 결합분포가 주어져 있을 때, X의 기댓값이나 분산 등을 X의 주변확률분포를 이용하여 이들을 구할 수 있다.

- 예제 1에서
  - $P_X(0) = \frac{10}{45} + \frac{15}{45} + \frac{3}{45} = \frac{28}{45}$
  - $P_X(1) = \frac{10}{45} + \frac{6}{45} + 0 = \frac{16}{45}$
  - $P_X(2) = \frac{1}{45} + 0 + 0 = \frac{1}{45}$
  - $P_Y(0) = \frac{10}{45} + \frac{10}{45} + \frac{1}{45} = \frac{21}{45}$
  - $P_Y(1) = \frac{15}{45} + \frac{6}{45} + 0 = \frac{21}{45}$
  - $P_Y(2) = \frac{3}{45} + 0 + 0 = \frac{3}{45}$

		Y 0	1	2	합계
X	0	10/45	15/45	3/45	28/45
	1	10/45	6/45	0	16/45
	2	1/45	0	0	1/45
합계		21/45	21/45	3/45	1

- $E(X) = \sum_i x_i P_X(x_i) = 0 \times \frac{28}{45} + 1 \times \frac{16}{45} + 2 \times \frac{1}{45} = \frac{18}{45} \equiv \mu_X$

  $E(X^2) = \sum_i x_i^2 P_X(x_i) = 0^2 \times \frac{28}{45} + 1^2 \times \frac{16}{45} + 2^2 \times \frac{1}{45} = \frac{20}{45}$

  $\sigma_X{}^2 = \frac{20}{45} - \left(\frac{18}{45}\right)^2 = \frac{64}{225}$

➔ E(Y), V(Y)
➔ Cov(X,Y)
➔ Corr(X,Y)

연속결합확률분포(continuous joint probability distribution)

결합확률밀도함수(joint probability density function)

연속확률변수 X와 Y의 결합확률밀도함수 $f(x,y)$는 다음 조건을 만족한다.

(1) 모든 실수 (x,y)에서 $f(x,y) \geq 0$

(2) $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dx\,dy = 1$ 결합확률질량함수 $\text{Jpmf} = \sum_{x}\sum_{y} x \cdot y \cdot P(X,Y)$

➔ 임의의 2차원 공간의 집합 A에 대하여 $P[(X,Y) \in A] = \int\int_{A} f(x,y)dxdy$

주변확률밀도함수(marginal probability density function)

연속확률변수 X와 Y의 결합확률밀도함수 $f(x,y)$로 주어지면

(1) X의 주변확률밀도함수 $f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy,\ -\infty < x < \infty$

(2) Y의 주변확률밀도함수 $f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx,\ -\infty < y < \infty$

$f(x) = \int_{R_y} f(x,y)\,dy$, $f(y) = \int_{R_x} f(x,y)\,dx$

주변확률밀도함수(marginal probability density function)

- [예제 6] $f(x,y) = x + y,\ 0 < x, y < 1$ 에서
  - $f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy = \int_0^1 (x+y)dy = x + \frac{1}{2}, 0 < x < 1$
  - $f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx = \int_0^1 (x+y)dx = y + \frac{1}{2}, 0 < y < 1$

$f(x,y) \neq f(x) \cdot f(y)$ ∴ X와 Y는 종속관계

- [예제 7] 주변분포의 기댓값과 분산은
  - $E(X) = \int_0^1 x\, f_X(x)dx = \int_0^1 x(x + \frac{1}{2})dx = \frac{7}{12}$
  - $E(X^2) = \int_0^1 x^2\, f_X(x)dx = \int_0^1 x^2(x + \frac{1}{2})dx = \frac{5}{12}$
  - $Var(X) = E(X^2) - E(X)^2 = \frac{5}{12} = \left(\frac{7}{12}\right)^2 = \frac{11}{144}$.

$E(X) = \sum x p(x)$ / $\int x f(x)$

➔ E(Y), V(Y) , Cov(X,Y) , Corr(X,Y)

$E(X,Y) = \sum_x \sum_y xy\, p(x,y)$
$= \int\int xy f(x,y)$

주변확률밀도함수-example

- [예제 8] $f(x,y) = cxy,\ 0 < x, y < 1, x + y < 1$
  - $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dxdy = 1$ 을 풀면
  - $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dxdy = c\int_0^1\int_0^{1-y} xydx\,dy = \frac{c}{24} = 1$ 에서 $c = 24.$
  - 주변확률밀도함수 $f_X(x) = \int_0^{1-x} 24xy\,dy = 12x(1-x)^2,\ 0 < x < 1\,.$
  - $P(X \le 1/2) = \int_0^{1/2} f_X(x)dx = \int_0^{1/2} 12x(1-x)^2\,dx$
    $= [6x^2 - 8x^3 + 3x^4]_0^{\frac{1}{2}} = \frac{11}{16}$
  - $E(X) = \int_0^1 x\,f_X(x)dx = \int_0^1 12x^2(1-x)^2dx = [4x^3 - 6x^4 + \frac{12}{5}x^5]_0^1 = \frac{2}{5}$

$P(X,Y) = P(X) \cdot P(Y)$, $E(XY) = E(X) \cdot E(Y)$
면, X와 Y는 독립이다. $Cov(X,Y) = 0$
상관관계 ×.

확률변수의 독립

두 확률변수 X와 Y의 결합확률밀도(질량)함수가 $f(x,y)$이며 각 주변확률밀도(질량)함수가 $f_X(x)$, $f_Y(y)$ 일 때

$$f(x,y) = f_X(x) \cdot f_Y(y) \quad \forall x, y$$

이면 X와 Y는 서로 독립이며 성립하지 않으면 서로 종속이다.

[예제 9]

- $f(1,1) = \frac{6}{45}$
- $f_X(1)f_Y(1) = \frac{16}{45}\frac{21}{45}$
- $f(1,1) \neq f_X(1)f_Y(1)$
- X와 Y는 서로 독립아님

종속

		Y 0	1	2	합계
X	0	10/45	15/45	3/45	28/45
	1	10/45	6/45	0	16/45
	2	1/45	0	0	1/45
합계		21/45	21/45	3/45	1

[예제 10] $f(x,y) = x + y,\ 0 < x, y < 1$ 에서 ➔ Cov(X,Y) , Corr(X,Y)

- $f(x,y) \neq f_X(x)f_Y(y) = (x + \frac{1}{2})(y + \frac{1}{2})$ 이므로 종속

조건부 확률분포, 조건부 기대값

두 확률변수 X와 Y의 결합확률밀도(질량)함수가 $f(x,y)$이며 각 주변확률밀도(질량)함수가 $f_X(x)$, $f_Y(y)$ 라 하면, $X=x$로 주어졌을 때 Y의 조건부 확률질량(밀도)함수는 다음과 같다.

$$f_{Y|X}(y|x)=\frac{f(x,y)}{f_X(x)} \text{ (단, } f_X(x)>0)$$

E[Y|X]

$y=f(x)+e$

target feature

$X=x$로 주어졌을 때 Y의 조건부 기댓값은

$$E(Y|X=x)=\begin{cases}\sum_y y\, f_{Y|X}(y|x) & \text{이산}\\ \int_y y\, f_{Y|X}(y|x)\,dy & \text{연속}\end{cases}$$

조건부 확률분포와 기대값-example

[예제 15] (연속) 예제 5에서 $X = x$로 주어졌을 때 Y의 조건부 확률밀도함수와 $P(Y \le 0.5|X = 0.5)$ 를 구하라.

$f(x) = \int f(x,y)\,dy = x + \frac{1}{2}$

[풀이] $f(x,y) = x + y$에서 $f_X(x) = x + 0.5$ 이므로

- $f_{Y|X}(y|x) = \frac{x+y}{x+0.5}, 0 < y < 1$
- $P(Y \le 0.5|X = 0.5) = \int_0^{0.5} f_{Y|X}(y|0.5)dy = \frac{3}{8}$

$f(x) = \lambda e^{-\lambda x}$

$X \sim Exp(\lambda)$

[예제 16] $f(x,y) = 2e^{-(x+y)},\ 0 < x < y < \infty$ 에서 $X = x$ 로 주어졌을 때 Y 의 조건부 확률밀도함수와 조건부 기댓값

[풀이] $f_X(x) = \int_x^{\infty} 2e^{-(x+y)}dy = 2e^{-2x},\ x > 0$ 즉 $X \sim \text{Exp}(2)$.

따라서 $f_{Y|X}(y|x) = \frac{2e^{-(x+y)}}{2e^{-2x}} = e^{-(y-x)},\ y > x.$

$E(Y|X) = \int y \cdot f(y|x)\,dy$

$$E(Y|X = x) = \int_x^{\infty} ye^{-(y-x)}dy = x + 1$$

공분산과 상관계수

X와 Y의 공분산(covariance)

$$Cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)]$$

여기서 $\mu_X = E(X), \mu_Y = E(Y)$

- 공분산은 결합분포를 가지는 두 확률변수의 관계를 나타내는 척도

[정리] 결합 확률변수 (X,Y) 함수 $h(X,Y)$ 의 기댓값

$$E[h(X,Y)] = \begin{cases} \sum_x \sum_y h(x,y)\, f(x,y) & \text{이산} \\ \int_y \int_x h(x,y)\, f(x,y)\,dxdy & \text{연속} \end{cases}$$

- 함수 $h(X,Y)$의 간단한 예 : $X, Y, XY, (X-\mu_X)(Y-\mu_Y)$ 등

공분산과 상관계수-example

- 공분산의 간편 계산식

$$Cov(X,Y) = E[X-\mu_X)(Y-\mu_Y)] = E(XY) - \mu_X\,\mu_y$$

[예제 18] 예제5의 결합확률밀도함수 $f(x,y) = x+y$ 이므로

- $E(X+Y) = \int_0^1\int_0^1 (x+y)f(x,y)dxdy = \int_0^1\int_0^1 (x+y)^2 dxdy = \frac{7}{6}$

$E(X) = \int_0^1\int_0^1 xf(x,y)dxdy = \int_0^1\int_0^1 x(x+y)dxdy = \frac{7}{12}$ 이고 $E(Y) = \frac{7}{12}$.

로 $E(X+Y) = E(X) + E(Y)$ 이고 공분산을 정의에 따라서 계산해보면

- $Cov(X,Y) = \int_0^1\int_0^1 (x-\frac{7}{12})(y-\frac{7}{12})(x+y)\,dxdy = -\frac{1}{144}$

$\int x f(x)\,dx =$

$$E(XY) = \int_0^1\int_0^1 xy(x+y)dxdy = \frac{1}{3}$$

$$E(X) = \mu_X = \frac{7}{12},\ E(Y) = \mu_Y = \frac{7}{12}$$

$$Cov(X,Y) = E(XY) - \mu_X\,\mu_y = \frac{1}{3} - \left(\frac{7}{12}\right)^2 = -\frac{1}{144}$$

$f(x) = \int_0^1 x+y\ dy$

$f(Y) = \int x+y\ dx$

공분산과 상관계수–특징

임의의 상수 a, b, c, d 에서

- $Cov(aX + b, cY + d) = ac\ Cov(X, Y)$
- $Cov(X, Y) = Cov(Y, X)$
- $Cov(X, X) = Var(X)$

↳ 독립 X

$|r| \approx 0$, 비선형관계

$|r| = 0$, 선형성

독립 → $|r| = 0$

$|r| = 0$ ✗ 독립

- $Cov(aX, cY) = ac\ Cov(X, Y)$
  - X를 체중, Y를 키라 할 때
  - 측정단위를 kg과 m 로한 공분산 $Cov(X, Y)$
  - 측정단위를 g과 cm로한 공분산 $Cov(aX, cY)$ $(a = 1000, b = 100)$
- 측정 단위와 변동성에 영향 받지 않는 연관성 측도인 상관계수

확률변수 X와 Y의 상관계수 (correlation coefficient)

$$Corr(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}} = \rho$$

상관계수의 해석

- 상관계수의 성질
  - $-1 \le \rho \le 1$
  - $Y = aX + b$ 이면 $\rho = 1(a > 0), \rho = -1(a < 0)$
  - $\rho = 0$ 이면 선형관계가 없다.(uncorrelated)
  - $Corr(aX, bY) = \frac{ab}{|ab|} Corr(X, Y)$

[예제 21] 예제 5의 결합분포에서 $Cov(X, Y) = -\frac{1}{144}$ 주변확률분포에서 $\sigma_X = \sqrt{\frac{11}{144}}, \sigma_Y = \sqrt{\frac{11}{144}}$ 이므로

$$\rho = \frac{Cov(X, Y)}{\sigma_X \cdot \sigma_Y} = -\frac{1}{11}$$

- 선형관계의 강도
  - $|\rho| \approx 1$: 매우 강한 선형 관계, $|\rho| \ge 2/3$: 강한 선형 관계
  - $|\rho| \approx 1/2$: 중간 정도 선형 관계
  - $|\rho| \approx 1/3$: 약한 선형 관계, $|\rho| \approx 0$: 선형 관계가 없다(무 상관)

두 확률변수가 서로 독립일 경우

- 상관계수는 두 확률변수 사이의 선형관계를 측정
  - $\rho = 0$: 선형관계가 없다. 그러나 비선형 관계는 있을 수 있다.
  - 두 확률변수의 독립 : 모든 관계가 없다.
  - 독립이면 $\rho = 0$ 이지만 $\rho = 0$ 이라고 해서 독립이라고 할 수는 없다.

임의의 상수 a, b 에대하여

- $E(aX \pm bY) = aE(X) \pm bE(Y)$
- $Var(aX \pm bY) = a^2Var(X) + b^2Var(Y) \pm 2Cov(X,Y)$

$f(x)$ $f(y)$
$E(x)$ $E(y)$
$V(x)$ $V(y)$

- 두 확률변수가 서로 독립이면
  - $Var(X \pm Y) = Var(X) + Var(Y)$

$f(x,y) = x+y$
$0 \le x$, $y \le 1$
$Cov(x,y) = E(x)E(y) - M_x M_y$