제 13 장 상관분석

제1절 모집단의 경우

1. 공분산(Covariance)

$$Cov(X,Y) = E\left[(X-\mu_X)(Y-\mu_y)\right]$$

Note: $V(X) = E\left[(X-\mu_x)(X-\mu_x)\right] = E(X-\mu_x)^2$

$V(Y) = E\left[(Y-\mu_y)(Y-\mu_y)\right] = E(Y-\mu_y)^2$

예제 1. 카드를 두 번 뽑는 게임; X: 첫 번째 카드의 수, Y: 두 번째 카드의 수
주어진 결합확률표를 사용하여 두 변수의 기대값, 분산, 공분산을 구하시오.

Y \ X	100	200	300
0	0.12	0.60	0.08
100	0.08	0.10	0.02

풀이 한계확률 $p(X)$와 $p(Y)$

Y \ X	100	200	300	$p(Y)$
0	0.12	0.60	0.08	0.80
100	0.08	0.10	0.02	0.20
$p(X)$	0.20	0.70	0.10	1.00

$$E(X) = \sum_{j=1}^{3} X_j \cdot p(X_j) = 100(0.20) + 200(0.70) + 300(0.10) = 190$$

$$E(Y) = \sum_{i=1}^{2} Y_i \cdot p(Y_i) = 0(0.80) + 100(0.20) = 20$$

$$V(X) = \sigma_X^2 = \sum_{j=1}^{3} \left(X_j - E(X)\right)^2 \cdot p(X_j)$$

$$= (100-190)^2(0.20) + (200-190)^2(0.70) + (300-190)^2(0.10) = 2{,}900$$

$$V(Y) = \sigma_Y^2 = \sum_{i=1}^{2} \left(Y_i - E(Y)\right)^2 \cdot p(Y_i) = (0-20)^2(0.80) + (100-20)^2(0.20) = 1{,}600$$

$$Cov(X,Y) = \sigma_{XY} = \sum_{i=1}^{2}\sum_{j=1}^{3} \left((X_j - E(X)(Y_i - E(Y)\right) \cdot p(X_j, Y_i)$$

= (100-190)(0-20) · 0.12 + (200-190)(0-20) · 0.60 + (300-190)(0-20) · 0.08

+ (100-190)(100-20) · 0.08 + (200-190)(100-20) · 0.10 + (300-190)(100-20) · 0.02

= -400

연습문제 1. 아래 결합확률표를 사용하여 두 변수의 기대값, 분산, 공분산을 구하시오.

문제 1.

Y \ X	100	200
300	0.30	0.40
400	0.10	0.20

문제 2.

Y \ X	20	80
40	0.40	0.10
70	0.20	0.30

두 확률변수가 서로 독립적인 경우의 $E(XY)$, $Cov(X, Y)$, $V(X \pm Y)$

Y \ X	100	200	$P(Y)$
300	0.42	0.28	0.70
400	0.18	0.12	0.30
$P(X)$	0.60	0.40	1.00

$E(X) = 100(.6) + 200(.4) = 140, \quad E(Y) = 300(.7) + 400(.3) = 330$

독립적이므로 다음이 성립한다.

$P(X=100, Y=300) = 0.42$인데, 독립적이므로 $P(X=100) \times P(Y=300) = 0.6 \times 0.7$과 같다.

$P(X=100, Y=400) = 0.18$인데, 독립적이므로 $P(X=100) \times P(Y=400) = 0.6 \times 0.3$과 같다.

$P(X=200, Y=300) = 0.28$인데, 독립적이므로 $P(X=200) \times P(Y=300) = 0.4 \times 0.7$과 같다.

$P(X=200, Y=400) = 0.12$인데, 독립적이므로 $P(X=200) \times P(Y=400) = 0.4 \times 0.3$과 같다.

$$E(XY) = \sum_{i=1}^{2}\sum_{j=1}^{2} X_i Y_j \cdot P(X_i Y_j) \qquad \{\text{개념에 따라 전개한 경우}\}$$

= 100 · 300(0.42) + 100 · 400(0.18) + 200 · 300(0.28) + 200 · 400(0.12)

= 12,600 + 7,200 + 16,800 + 9,600 = 46,200

위의 식은 X와 Y가 서로 독립적이면 아래와 같이 정리된다.

$$E(XY) = \sum_{i=1}^{2}\sum_{j=1}^{2} X_i Y_j \cdot P(X_i Y_j)$$
$$= \sum_{i=1}^{2}\sum_{j=1}^{2} X_i Y_j \cdot P(X_i)P(Y_j) \qquad \text{서로 독립적이면}$$
$$= \sum_{i=1}^{2}\sum_{j=1}^{2} X_i \cdot P(X_i) \cdot Y_j \cdot P(Y_j)$$
$$= \sum_{i=1}^{2} X_i \cdot P(X_i) \cdot \sum_{j=1}^{2} Y_j \cdot P(Y_j)$$
$$= E(X) \cdot E(Y)$$

$$E(XY) = \sum_{i=1}^{2}\sum_{j=1}^{2} X_i Y_j \cdot P(X_i Y_j)$$
$$= 100 \cdot 300(.6\times.7) + 100 \cdot 400(.6\times.3) + 200 \cdot 300(.4\times.7) + 200 \cdot 400(.4\times.3)$$
$$= 100(.6) \cdot 300(.7) + 100(.6) \cdot 400(.3) + 200(.4) \cdot 300(.7) + 200(.4) \cdot 400(.3)$$
$$= 100(.6) \cdot [300(.7)+400(.3)] + 200(.4) \cdot [300(.7)+400(.3)]$$
$$= [100(.6)+200(.4)] \cdot [300(.7)+400(.3)]$$
$$= E(X) \cdot E(Y)$$

두 확률변수 X, Y가 독립적이면 아래 식들이 성립한다.

(1) $E(XY) = E(X)E(Y)$ [1]
(2) $Cov(X, Y) = 0$ [2]
(3) $V(X \pm Y) = V(X) + V(Y)$ [3]

---

1) $E(XY) = \sum_i \sum_j X_i Y_j \cdot P(X_i Y_j)$ $E(XY)$의 정의
$= \sum_i \sum_j X_i Y_j \cdot P(X_i)P(Y_j)$ 서로 독립적이므로 $P(X_i Y_j) = P(X_i)P(Y_j)$
$= \sum_i \sum_j X_i \cdot P(X_i) \cdot Y_j \cdot P(Y_j)$
$= \sum_i X_i \cdot P(X_i) \cdot \sum_j Y_j \cdot P(Y_j)$
$= E(X) \cdot E(Y)$

2) $Cov(X, Y) = E((X-\mu_X)(Y-\mu_y))$ $Cov(X, Y)$의 정의
$= E(XY - X\mu_y - \mu_X Y + \mu_X \mu_Y)$
$= E(XY) - \mu_Y E(X) - \mu_X E(Y) + \mu_X \mu_Y$
$= 0$ 서로 독립적이면 $E(XY) = \mu_X \mu_Y$

연습문제 2. 아래 결합확률표를 사용하여 두 변수의 기대값, 분산, 공분산을 구하시오.

문제 1. 두 확률변수가 서로 독립적인 경우

Y \ X	20	80
40	0.42	0.28
70	0.18	0.12

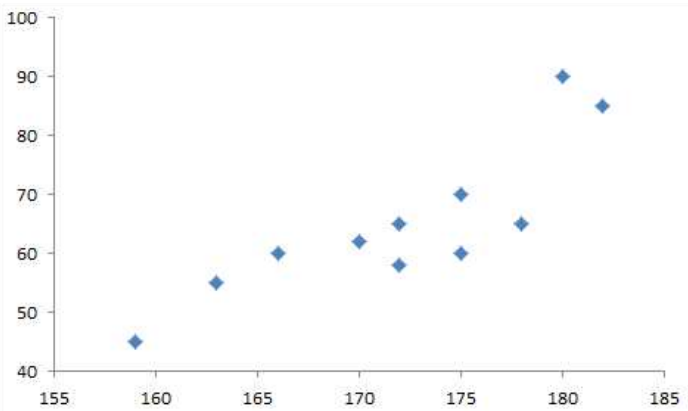
예제 2. 키와 몸무게 자료(모집단)

	1	2	3	4	5	6	7	8	9	10	평균	분산
키	170	159	180	172	175	163	166	182	178	175	172	50.8
몸무게	62	45	90	58	70	55	60	85	65	60	65	165.8

공분산 = 78.6

[주의] 모집단 자료이므로 분산을 구할 때 분산과 공분산을 구할 때 분모는 10이다.

산포도(scatter diagram) (주의: Excel에서는 분산형이라 한다.)



---

3) $V(X+Y) = E\big((X+Y)-(\mu_X+\mu_Y)\big)^2$

$= E\big((X-\mu_X)+(Y-\mu_Y)\big)^2$

$= E\big((X-\mu_X)^2+2(X-\mu_X)(Y-\mu_Y)+(Y-\mu_Y)^2\big)$

$= E(X-\mu_X)^2 + 2E\big((X-\mu_X)(Y-\mu_Y)\big) + E(Y-\mu_Y)^2$

$= V(X) + 2\,Cov(X, Y) + V(Y)$

$= V(X) + V(Y)$ 서로 독립적이면 $Cov(X, Y) = 0$

연습문제 3. 평균, 분산 및 공분산을 구하고 산포도를 작성하시오. (모집단)

	1	2	3	4	5	6
X	10	20	30	40	50	60
Y	50	60	30	40	10	20

2. 상관계수 (Correlation Coefficient)

모집단상관계수 $\rho = \frac{\sigma_{xy}}{\sqrt{\sigma_x^2 \cdot \sigma_y^2}} = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y}$

$-1 \leq \rho \leq 1$

연습문제 4. ρ의 하한값과 상한값

문제 1. ρ가 취할 수 있는 가장 작은 값
문제 2. ρ가 취할 수 있는 가장 큰 값

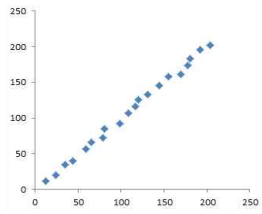
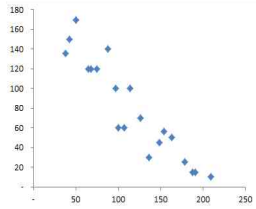
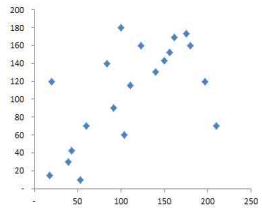
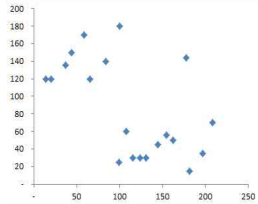
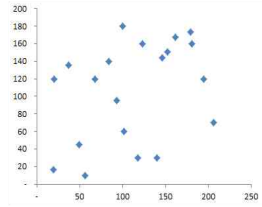
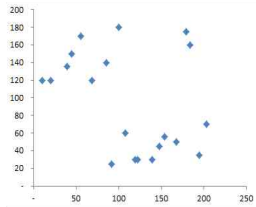
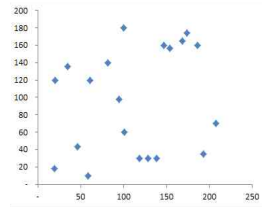
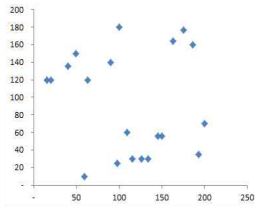
연습문제 5. (모집단) 상관계수를 구하시오.

문제 1. $\sigma_x^2 = 3{,}200,\ \sigma_y^2 = 2{,}800,\ \sigma_{xy} = 2{,}500$
문제 2. $\sigma_x^2 = 3{,}200,\ \sigma_y^2 = 2{,}800,\ \sigma_{xy} = 1{,}500$
문제 3. $\sigma_x^2 = 3{,}200,\ \sigma_y^2 = 2{,}800,\ \sigma_{xy} = -1{,}000$
문제 4. $\sigma_x^2 = 3{,}200,\ \sigma_y^2 = 2{,}800,\ \sigma_{xy} = -2{,}600$

상관계수와 변수의 관련성

0.0~0.2	(-0.2~0.0)	관련이 없음
0.2~0.4	(-0.4~-0.2)	약간의 관련성
0.4~0.6	(-0.6~-0.4)	상당한 관련성
0.7~1.0	(-1.0~-0.7)	매우 강한 관련성

상관계수와 산포도

σ_x^2 = 3,288 σ_y^2 = 3,387 σ_{xy} = 3,333 ρ = 0.998		σ_x^2 = 2,621 σ_y^2 = 2,390 σ_{xy} = -2,334 ρ = -0.932	
σ_x^2 = 3,355 σ_y^2 = 2,890 σ_{xy} = 1,854 ρ = 0.595		σ_x^2 = 3,312 σ_y^2 = 2,891 σ_{xy} = -1,788 ρ = -0.578	
σ_x^2 = 3,202 σ_y^2 = 3,088 σ_{xy} = 1,117 ρ = 0.355		σ_x^2 = 3,350 σ_y^2 = 3.,057 σ_{xy} = -1,179 ρ = -0.368	
σ_x^2 = 3,330 σ_y^2 = 3,476 σ_{xy} = 719 ρ = 0.211		σ_x^2 = 3,135 σ_y^2 = 3,235 σ_{xy} = -388 ρ = -0.122	

제2절 표본의 경우

표본공분산 $\quad S_{xy} = \dfrac{\sum(x-\bar{x})(y-\bar{y})}{n-1}$

표본상관계수 $\quad r_{xy} = \dfrac{S_{xy}}{\sqrt{S_x^2 \cdot S_y^2}} = \dfrac{S_{xy}}{S_x \cdot S_y}, \qquad -1 \le r_{xy} \le 1$

제3절 상관계수의 가설검정

① H_0: $\rho = \hat{\theta}$

H_A: $\rho \neq \hat{\theta}$ (또는 $\rho < \hat{\theta}$, $\rho > \hat{\theta}$)

② Test Statistic: $\quad t_{n-2} = \dfrac{r-\rho}{s_r} = \dfrac{r-\rho}{\sqrt{\dfrac{1-r^2}{n-2}}} \quad \leftarrow s_r = \sqrt{\dfrac{1-r^2}{n-2}}$

③ For α, rejection Region:

(1) H_A: $\rho \neq \hat{\theta}$ $\rightarrow$ $t < -t_{n-2, \frac{\alpha}{2}}$ or $t > t_{n-2, \frac{\alpha}{2}}$

(2) H_A: $\rho < \hat{\theta}$ $\rightarrow$ $t < -t_{n-2, \alpha}$

(3) H_A: $\rho > \hat{\theta}$ $\rightarrow$ $t > t_{n-2, \alpha}$

연습문제 6. ρ에 대한 가설검정을 실시하시오.

문제 1. H_0: $\rho = 0$, H_A: $\rho \neq 0$; $r = -0.54$, $n = 8$, $\alpha = 0.05$

문제 2. H_0: $\rho = 0$, H_A: $\rho > 0$; $r = 0.3250$, $n = 10$, $\alpha = 0.10$

연습문제 정답

1. (1) $p(X=100) = p(X=100,\ Y=300) + p(X=100,\ Y=400) = 0.30 + 0.10 = 0.40$
$p(X=200) = p(X=200,\ Y=300) + p(X=200,\ Y=400) = 0.40 + 0.20 = 0.60$
$p(Y=300) = p(X=100,\ Y=300) + p(X=200,\ Y=300) = 0.30 + 0.40 = 0.70$
$p(Y=400) = p(X=100,\ Y=400) + p(X=200,\ Y=400) = 0.10 + 0.20 = 0.30$

$E(X) = 100(0.40) + 200(0.60) = 160$
$E(Y) = 300(0.70) + 400(0.30) = 330$

$V(X) = (100\text{-}160)^2(0.40) + (200\text{-}160)^2(0.60) = 2{,}400$
$V(Y) = (300\text{-}330)^2(0.70) + (400\text{-}330)^2(0.30) = 2{,}100$

$Cov(X, Y) = (100\text{-}160)(300\text{-}330)(0.30) + (200\text{-}160)(300\text{-}330)(0.40)$
$+ (100\text{-}160)(400\text{-}330)(0.10) + (200\text{-}160)(400\text{-}330)(0.20)$
$= 540 - 480 - 420 + 560 = 200$

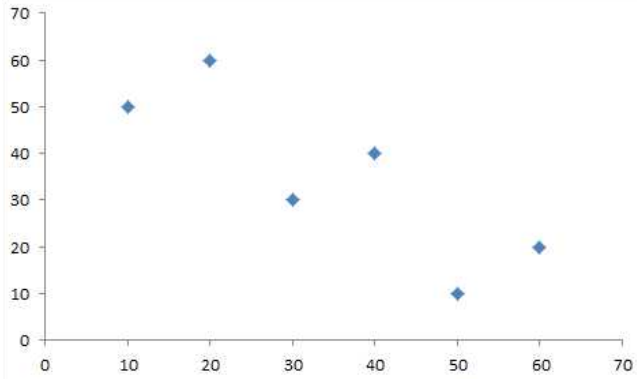
(2) $p(X=20) = 0.60,\ p(X=80) = 0.40,\ p(Y=40) = 0.50,\ p(Y=70) = 0.50$
$E(X) = 44.0,\ E(Y) = 55.0$
$V(X) = 864,\ V(Y) = 225,\ Cov(X, Y) = 180$

2. (1) $p(X=20) = 0.60,\ p(X=80) = 0.40,\ p(Y=40) = 0.70,\ p(Y=70) = 0.30$
$E(X) = 44.0,\ E(Y) = 49.0$
$V(X) = 864,\ V(Y) = 189,\ Cov(X, Y) = 0$

3. (1)

	1	2	3	4	5	6	평균	분산
X	10	20	30	40	50	60	35.0	291.667
Y	50	60	30	40	10	20	35.0	291.667

공분산 = -241.667



4. (1) -1 (2) 1

5. (1) 0.835 (2) 0.501 (3) -0.334 (4) -0.869

6. (1) ① H_0: $\rho = 0$, H_A: $\rho \neq 0$

② Test Statistic: $t_6 = \dfrac{r-\rho}{s_r} = \dfrac{r-\rho}{\sqrt{\dfrac{1-r^2}{n-2}}}$

③ For $\alpha = 0.05$, Rejection Region: $t < -t_{6,\,0.025} = -2.4469$ or $t > t_{6,\,0.025} = 2.4469$

④ Value of the Test Statistic: $t = \dfrac{-0.54-0}{\sqrt{\dfrac{1-(-0.54)^2}{8-2}}} = -1.5716$

⑤ Conclusion: Do not reject H_0.

(2) ① H_0: $\rho = 0$, H_A: $\rho > 0$

② Test Statistic: $t_8 = \dfrac{r-\rho}{s_r} = \dfrac{r-\rho}{\sqrt{\dfrac{1-r^2}{n-2}}}$

③ For α=0.10, Rejection Region: $t > t_{8,\,0.10} = 1.3968$

④ Value of the Test Statistic: $t = \dfrac{0.3250-0}{\sqrt{\dfrac{1-0.3250^2}{10-2}}} = 0.9720$

⑤ Conclusion: Do not reject H_0.