[6주차]통계적 사고와 이해

데이터테크전공 20173204 곽명빈

aX

- 상관계수 계산식:

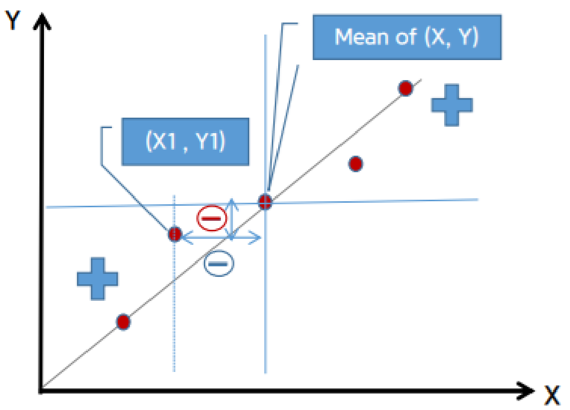
$$\rho_{X,Y} = \mathrm{corr}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

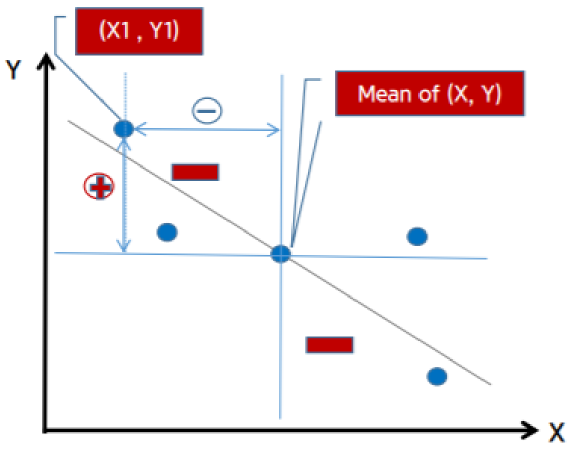
Tip

④ $\cancel{C}$ X $=\{c_1,c_2,\cdots,c_N\}$로 표시되는 유한 모집단의 평균을 μ, 표준편차를 σ라 할 때, 모집단 $\cancel{D}$ Y $=\{d_1,d_2,\cdots,d_N\}$ (단, $d_i=ac_i+b$, $i=1,\cdots,N$, $a,c=$상수)의 평균과 표준편차를 구하라.

$$E[d_i] = E(ac_i+b) = aE(c_i)+b$$
$$= a\mu + b$$
$$Var(d_i) = Var(ac_i+b) = a^2 Var(c_i)$$
$$= a^2\sigma^2$$
$$\mathrm{sd}(d_i) = \sqrt{Var(d_i)} = \sqrt{a^2\sigma^2} = |a|\sigma$$

- 공분산의 해석 : Cov(X, Y) ?





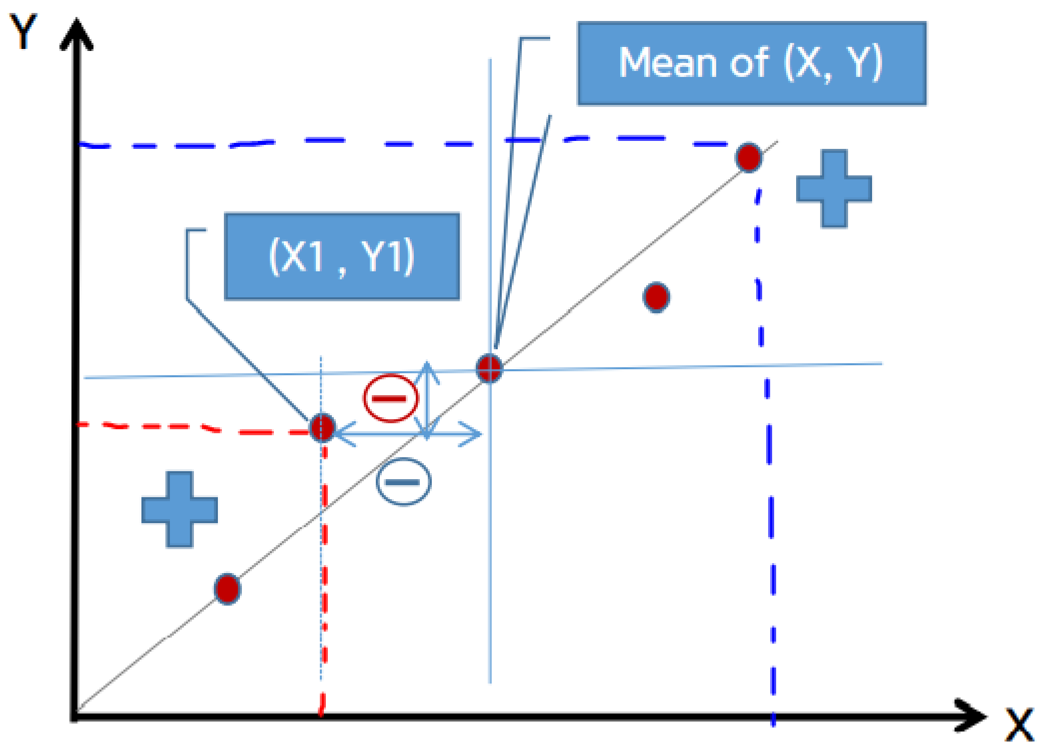
문제 1

p8의 공분산 그래프를 해석하시오.

공분산의 식은 다음과 같습니다. $cov(X,Y) = E[(X-\mu_x)(Y-\mu_Y)]$

식을 있는 그대로 해석해보면, 기대값(X변수의 편차)(Y변수의 편차)입니다.

양의 상관



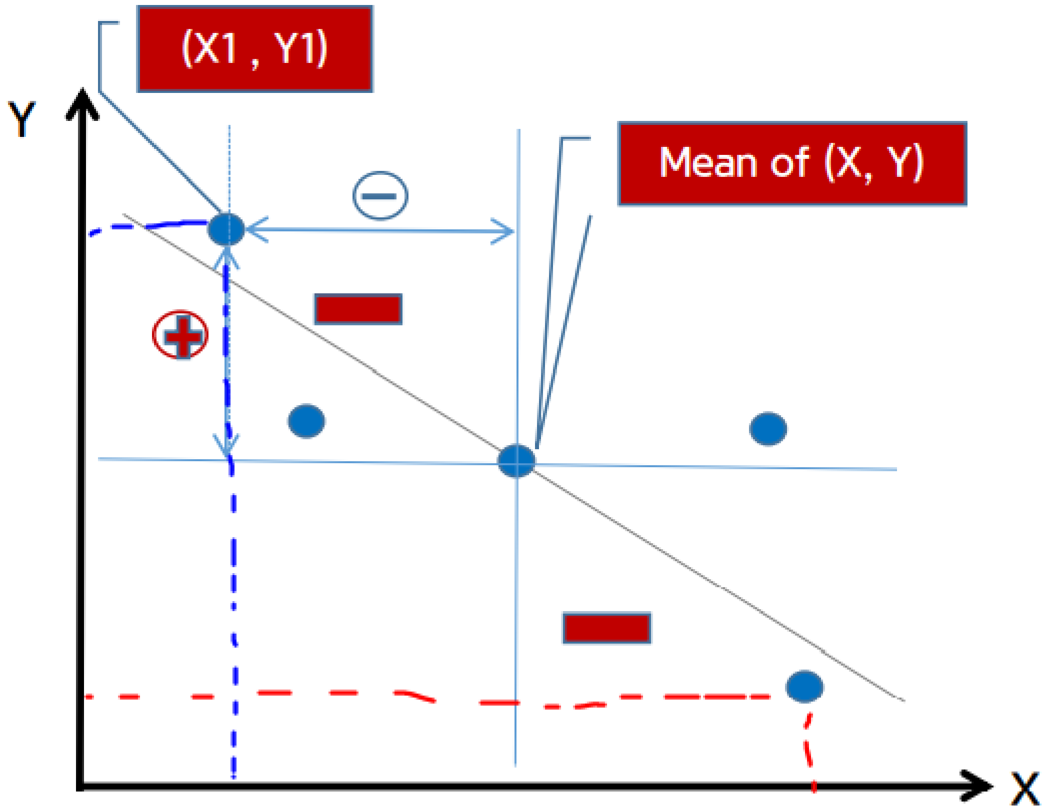
양의 상관의 그래프를 보면 1,3사분면에 데이터가 많은것을 확인할 수 있습니다.

1사분면 $\Rightarrow X-\mu_x>0, Y-\mu_y>0$ 이기 때문에 $(X-\mu_x)(Y-\mu_Y)>0$라고 할 수 있습니다.

3사분면 $\Rightarrow X-\mu_x<0, Y-\mu_y<0$ 인데, 음수끼리의 곱은 양수이기 때문에 $(X-\mu_x)(Y-\mu_Y)>0$ 입니다.

그러므로 공분산의 값이 0보다 커지기 때문에 양의 상관이라고 할 수 있습니다.

음의 상관



음의 상관의 그래프를 보면 2,4사분면에 데이터가 많은것을 확인할 수 있습니다.

2사분면 $\Rightarrow X-\mu_x<0, Y-\mu_y>0$ 이기 때문에 $(X-\mu_x)(Y-\mu_Y)<0$라고 할 수 있습니다.

4사분면 $\Rightarrow X-\mu_x>0, Y-\mu_y<0$ 이기 때문에

$(X-\mu_x)(Y-\mu_Y)<0$ 입니다.

그러므로 공분산의 값이 0보다 작아지기 때문에 음의 상관이라고 할 수 있습니다.

문제 2

p8의 상관계수 계산식을 이용하여, Y = aX (a는 상수)일 때 상관계수가 1또는 -1 임을 확인하시오.

$$Corr(X,aX)=\frac{Cov(X,aX)}{\sigma_X\sigma_{aX}}$$

공분산의 성질

$Cov(X,aX)=a\ \mathrm{Cov}(X,X)$

$$corr(X, aX) = \frac{a\ E((X-\mu_x)(X-\mu_x))}{\sigma_X \sigma_{aX}}$$

$$\sigma_{XY} = E[(X-\mu_x)(Y-\mu_Y)]$$

$Y = aX$일때

$$\begin{aligned} &= E[(X-\mu_x)(Y-\mu_Y)] \\ &= E[(X-\mu_x)(aX-a\mu_X)] \\ &= aE[(X-\mu_x)(X-\mu_X)] \\ &= aE[(X-\mu_x)^2] \end{aligned}$$

$$\sigma_x^2 = E[(X-\mu_x)^2]$$

σ는 아래와 같이 표현 가능

$$\sigma_x = \sqrt{E(X-\mu_x)^2}$$

$$a\sigma_x = |a|\sqrt{E(X-\mu_x)^2}$$

그러므로

a>0일때

$$corr(X, aX) = \frac{a\ E\{(X-\mu_x)(X-\mu_x)\}}{|a|\sqrt{E(X-\mu_x)^2}\sqrt{E(X-\mu_x)^2}}$$

$$corr(X, aX) = \frac{a\ E(\{(X-\mu_x)(X-\mu_x)\}}{|a|\ E\{(X-\mu_x)(X-\mu_x)\}} = 1$$

$$\therefore\ corr(X, aX)\frac{a\sigma_x^2}{|a|\sigma_x\sigma_x} = 1$$

a<0일때

$$corr(X, aX) = \frac{E\{(X-\mu_x)(-aX+a\mu_x)\}}{|a|\sqrt{E(X-\mu_x)^2}\sqrt{E(X-\mu_x)^2}}$$

$$corr(X, aX) = \frac{-a\ E\{(X-\mu_x)(X-\mu_x)\}}{|a|\ E\{(X-\mu_x)(X-\mu_x)\}} = -1$$

$$\therefore\ corr(X, aX) = \frac{-a\sigma_x^2}{|a|\sigma_x\sigma_x} = -1$$