

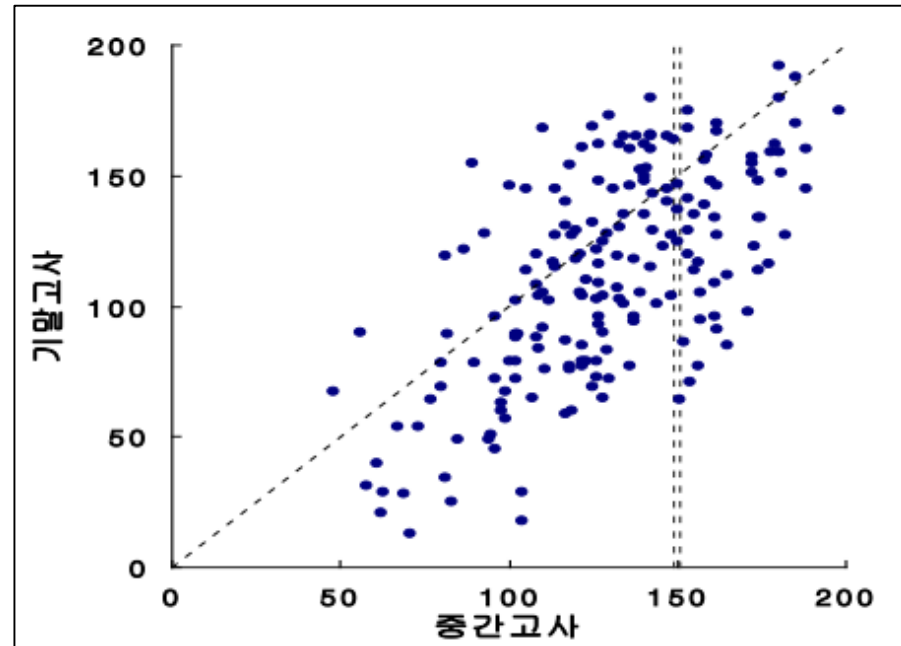
상관 분석

들어가기

- 과학적 실험이나자연 또는 사회현상의 관찰결과에 대한 분석의 중요한 목적 중 하나는 연구대상을 구성하는 변수들 사이의 관계를 규명하는 일이다.
 - 화학실험에서 실험온도(X)와 화학반응 속도(Y) 사이의 관계
 - 수학능력시험성적(X)과 대학평점(Y)과의 관계
 - 부모의 키(X)와 자녀의 키(Y) 사이의 관계
- 두 변수 사이의 선형관계가 유의한지, 어느 정도의 선형관계가 존재하는지 등을 상관계수를 통해 알아볼 수 있다. 이러한 상관계수의 추정과 검정을 상관분석(correlation analysis)이라고 한다.

산포도

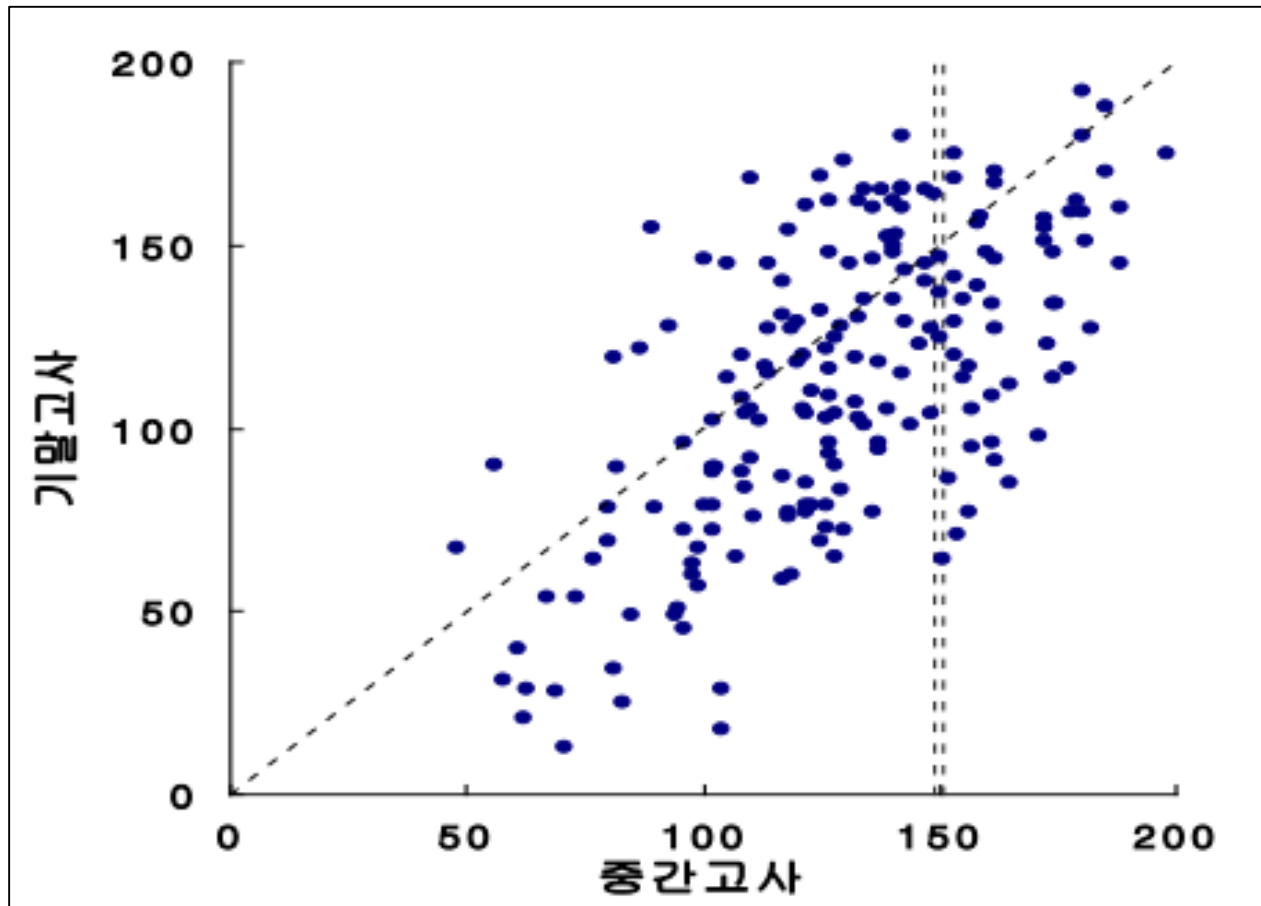
- 두 변수 사이의 관계를 살펴보기 위하여 산포도(scatter plot)를 이용한다.
 - 설명(독립)변수는 x 로 표기하고 가로축에 표시
 - 종속변수는 y 로 표기하고 세로축에 표시



중간고사와 기말고사와의 관계를 나타내는 산포도

산포도 예제

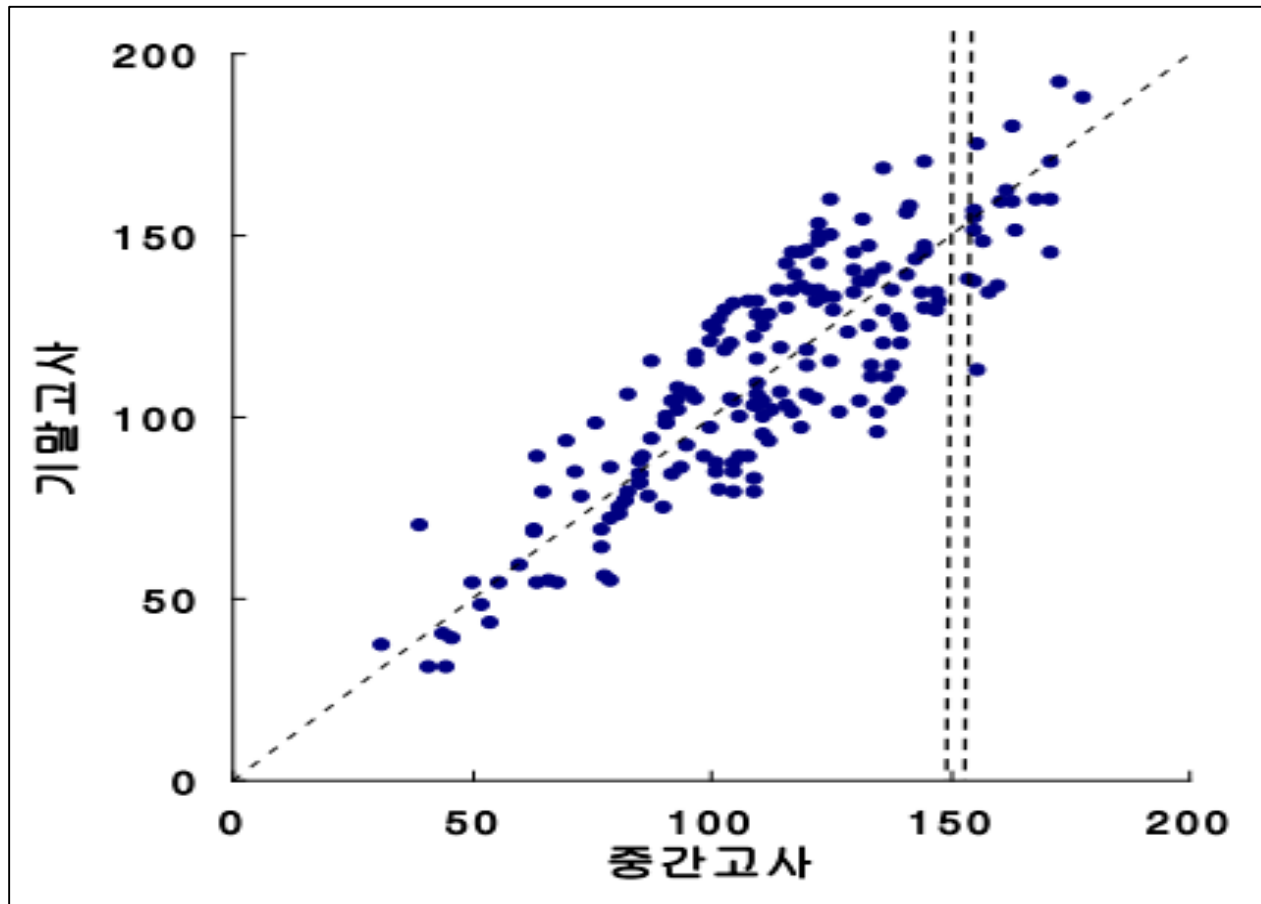
중간고사 / 기말고사 성적간 관계가 약한 경우



- ① 변수 사이의 관계가 약하면 한 변수 값이 다른 변수 값을 예측하는 데 큰 도움이 안된다.
- ② 중간고사에서 150점을 받은 학생들의 기말고사 성적은 55점에서 175점 사이에 분포한다.

산포도 예제

중간고사 / 기말고사 성적간 관계가 강한 경우



- ① 변수 사이의 관계가 강하면 한 변수 값이 다른 변수 값을 예측하는 데 큰 도움이 된다.
- ② 중간고사에서 150점을 받은 학생들의 기말고사 성적은 105점에서 175점 사이에 분포한다.

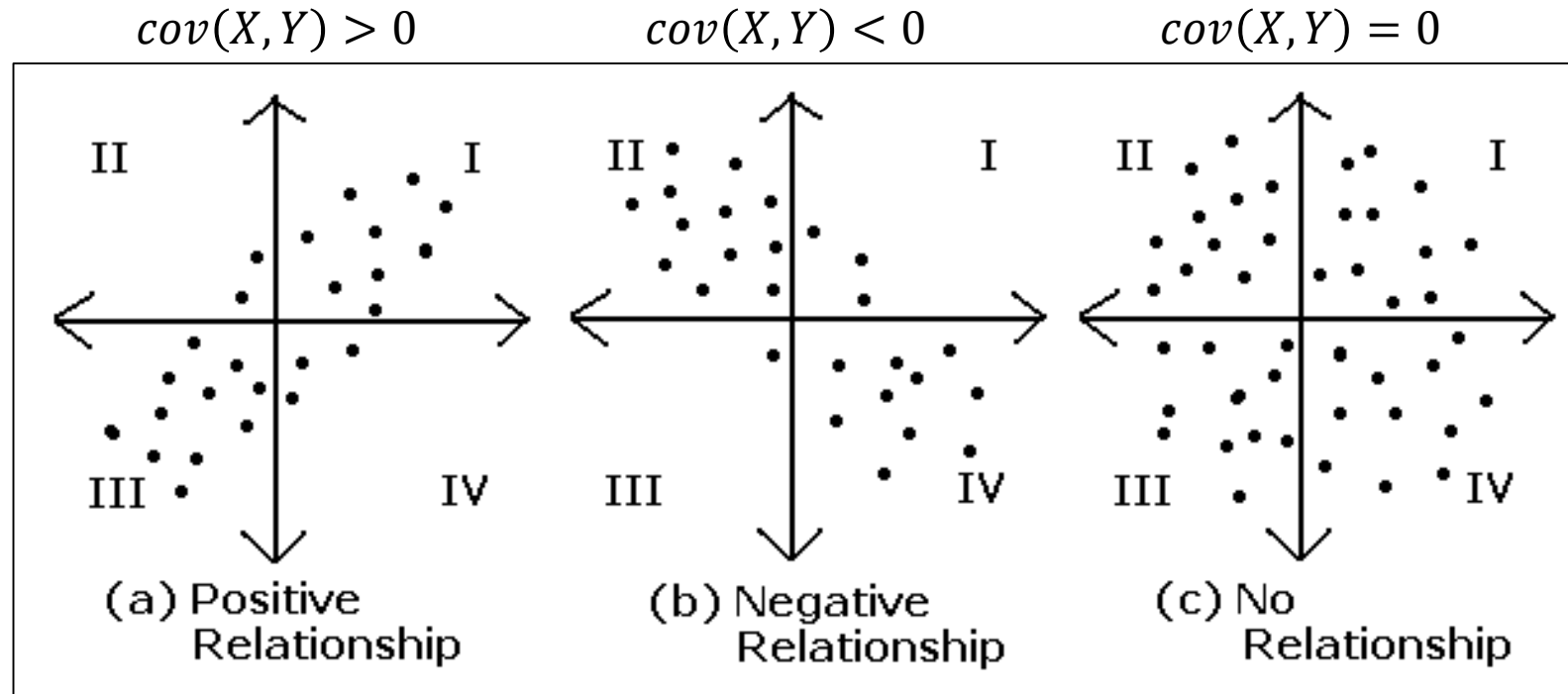
변수간 상관관계

- 두 변수간 선형관계의 방향과 강도가 얼마나 되는지 측정할 필요성이 있다.
 - 공분산
 - 상관계수

공분산

- 두 확률변수 X 와 Y 사이의 흠어진 관계를 측정하는 통계량으로 공분산을 사용한다.
 - ✓ 확률변수 X 의 기대값 : $E[X]$
 - ✓ 확률변수 Y 의 기대값 : $E[Y]$
- 공분산 : $cov(X, Y) = E[(X - E[X])(Y - E[Y])]$

공분산



$cov(X, Y) > 0$: X 가 증가할 때 Y 도 증가한다.

$cov(X, Y) < 0$: X 가 증가할 때 Y 도 감소한다.

$cov(X, Y) = 0$: X 와 Y 는 독립이다.

공분산의 단점

- 공분산은 변수의 측정단위(measurement unit)에 의존한다.
- 변수 X 와 변수 Y 의 측정단위를 바꾸면 변수 사이의 관계가 변하는 것이 아닌데, 공분산의 값이 변한다.

상관계수

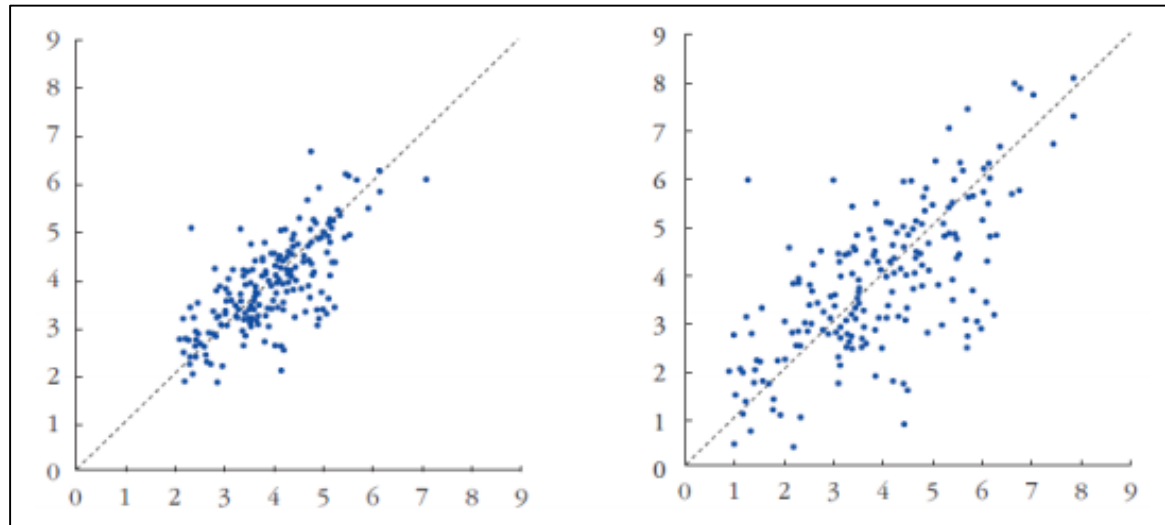
- 단위에 의존하지 않도록, 공분산을 두 변수의 표준편차의 곱으로 나누는 것을 상관계수 (correlation coefficient)라 정의하고, 이것을 두 변수 사이의 선형관계를 측정하는 단위로 주로 사용한다.

$$\rho(X, Y) = \frac{Cov(X, Y)}{SD(X) \times SD(Y)}$$

- 표준화된 두 변수 사이의 공분산과 동일하다.
 - ✓ 상관계수의 절대값은 1을 넘을 수 없다. $-1 \leq \rho \leq 1$
 - ✓ 확률변수 X, Y 가 독립이라면 상관계수는 0이다.
 - ✓ X, Y 가 완전한 선형적 관계라면 상관계수는 1 혹은 -1이다.

상관계수의 유의점

- 상관계수의 의미
 - 상관계수 0.8은 80%의 점들이 선형관계를 이루며 뽀뽀하게 밀집해 있는 것을 의미하지 않는다.
 - 상관계수 0.8은 상관계수 0.4보다 선형관계의 강도가 강하기는 하지만, 정확히 2배가 강하다는 것을 의미하지 않는다.



둘의 상관관계수는 같다.

상관계수의 유의점

- 상관계수의 유용성
 - 이탈값(outlier)가 존재하면 상관계수의 값은 유용하지 않다.
 - 두 변수간 관계가 비선형인 경우 상관계수의 값은 유용하지 않다.



이탈값 존재

비선형 관계

표본상관계수

- Sample Correlation Coefficient
- 크기가 n 인 이변량 랜덤포본 (X_1, Y_1) , (X_2, Y_2) , ..., (X_n, Y_n) 에 대한 상관계수 R

$$R = \frac{\sum_{i=1}^n \left\{ \frac{(X_i - \bar{X})(Y_i - \bar{Y})}{n} \right\}}{\sqrt{\sum_{i=1}^n \frac{(X_i - \bar{X}_n)^2}{n}} \sqrt{\sum_{i=1}^n \frac{(Y_i - \bar{Y}_n)^2}{n}}}$$
$$= \frac{\sum_{i=1}^n \frac{X_i Y_i}{n} - \bar{X} \bar{Y}}{\sqrt{\sum_{i=1}^n \frac{(X_i - \bar{X}_n)^2}{n}} \sqrt{\sum_{i=1}^n \frac{(Y_i - \bar{Y}_n)^2}{n}}}$$

표본상관계수의 분포이론

- 랜덤표본 (X_1, Y_1) , (X_2, Y_2) , ..., (X_n, Y_n) 의 모분포를 모수벡터가 $(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$ 인 이변량 정규분포라고 하자. 이 때 $\rho(X, Y) = 0$ 이면 표본상관계수 R 의 함수인 통계량 T 는 자유도가 $(n - 2)$ 인 t 분포를 따른다.

$$T = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}}$$

상관관계의 가설검정 ($H_0 : \rho = 0$)

- 표본상관계수 R 의 함수인 통계량 T 는 이변량 정규성하에서 모상관계수에 대한 가설검정에 사용된다.
 - $H_0 : \rho = 0$
 - $H_1 : \rho \neq 0$
 - $|T| \geq t_{\frac{\alpha}{2}}(n - 2)$ 가 관측되었을 때 귀무가설을 유의수준 α 에서 기각

상관관계의 가설검정 ($H_0 : \rho = \rho_0$)

- $\rho(X, Y) \neq 0$ 인 경우 표본상관계수 R 의 분포
 - $H_0 : \rho = \rho_0 (\neq 0)$ 의 귀무가설을 검정할 경우는 근사적인 분포이론을 이용해야 한다.
 - $Z = \frac{1}{2} \log \frac{1+R}{1-R}$ 통계량은 정규분포를 근사적으로 따른다.
 - ✓ $E(Z) = \frac{1}{2} \log \frac{1+\rho}{1-\rho}$
 - ✓ $Var(Z) = \frac{1}{n-3}$
 - 이것에 근거하여 모상관계수 ρ 에 관한 근사검정법과 근사신뢰구간을 구축한다.