상관분석과 회귀분석

1. 산포도

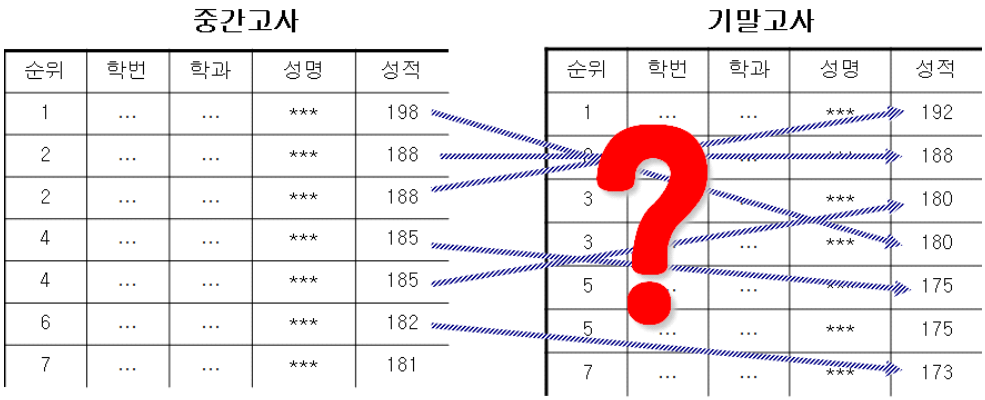
A. 예) 중간고사와 기말고사의 성적

중간고사

순위	학번	학과	성명	성적
1	…	…	***	198
2	…	…	***	188
2	…	…	***	188
4	…	…	***	185
4	…	…	***	185
6	…	…	***	182
7	…	…	***	181

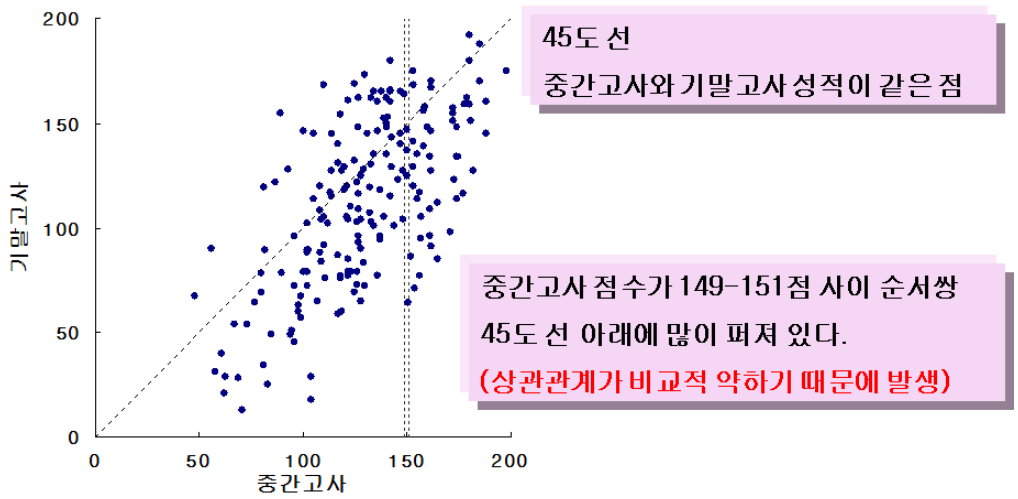
기말고사

순위	학번	학과	성명	성적
1	…	…	***	192
2	…	…	***	188
3	…	…	***	180
3	…	…	***	180
5	…	…	***	175
5	…	…	***	175
7	…	…	***	173

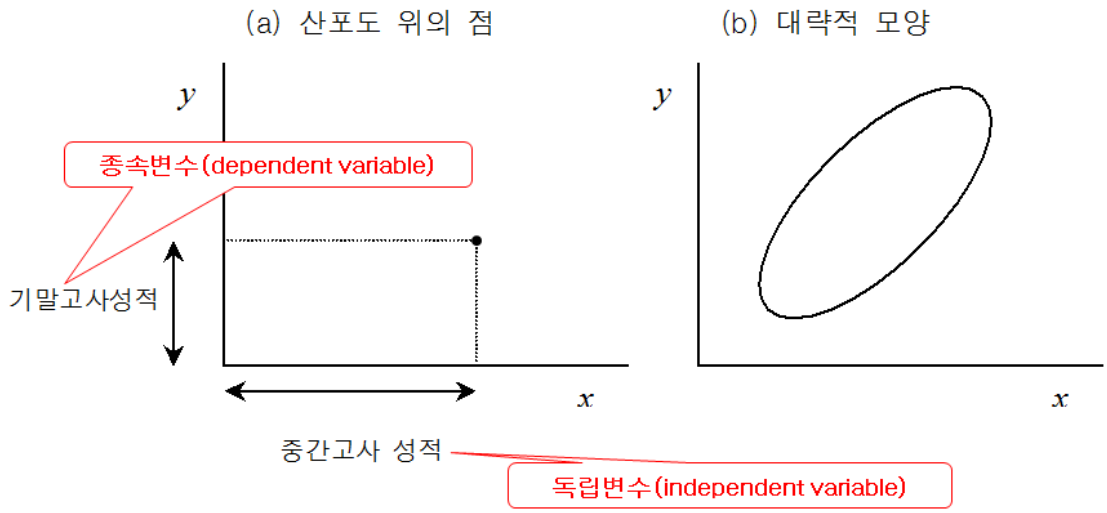


두 변수들 사이의 관계를 파악하려면 결합분포를 보아야 함.

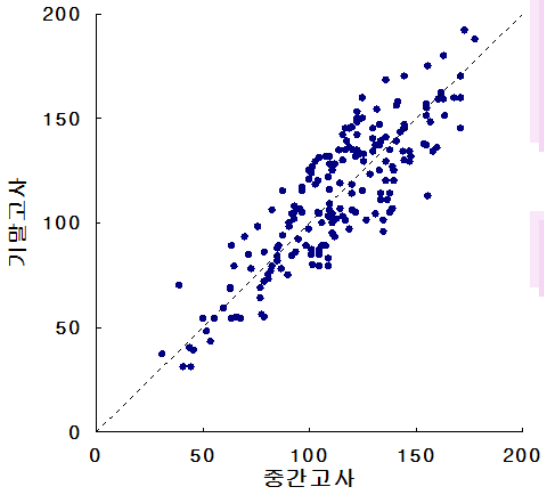
B. 예) 중간고사와 기말고사의 산포도



C. 산포도에 점을 표시하는 방법



D. 중간고사 성적과 기말고사 성적이 비슷하다면 ?

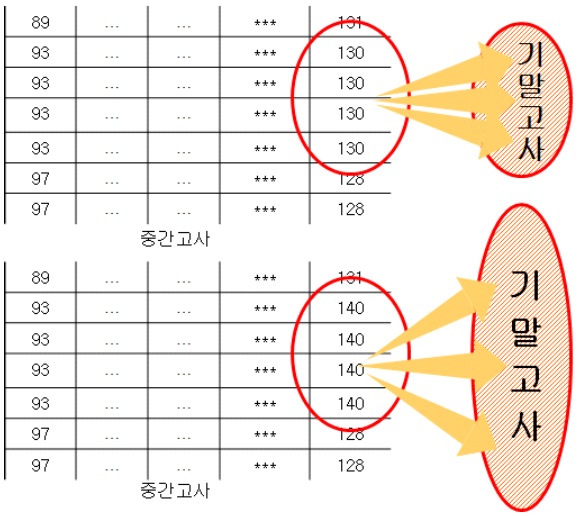


45도 선

(중간고사 점수, 기말고사 점수)로 이루어진 순서쌍들이 45도선 주위에 분포

중간고사 점수가 기말고사 점수를 예측하는 데 큰 도움이 된다.

E. 기말고사 성적을 예측하려 한다면 ?



중간고사 점수로 본 특정집단의 기말고사 성적이 엇비슷하면, 중간고사 성적이 기말고사 성적 예측에 도움을 준다

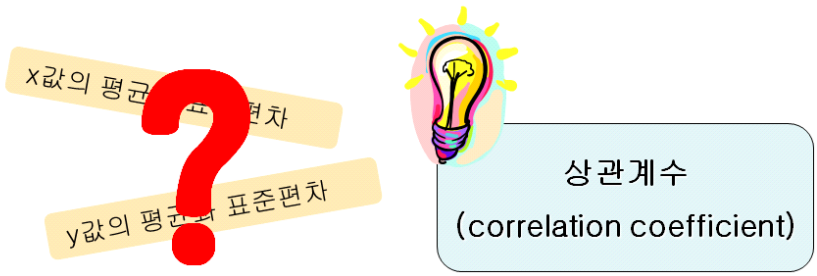
중간고사 점수로 본 특정집단의 기말고사 성적이 흩어져 있으면, 중간고사 성적이 기말고사 성적 예측에 별 도움을 주지 못 한다

2. 상관계수

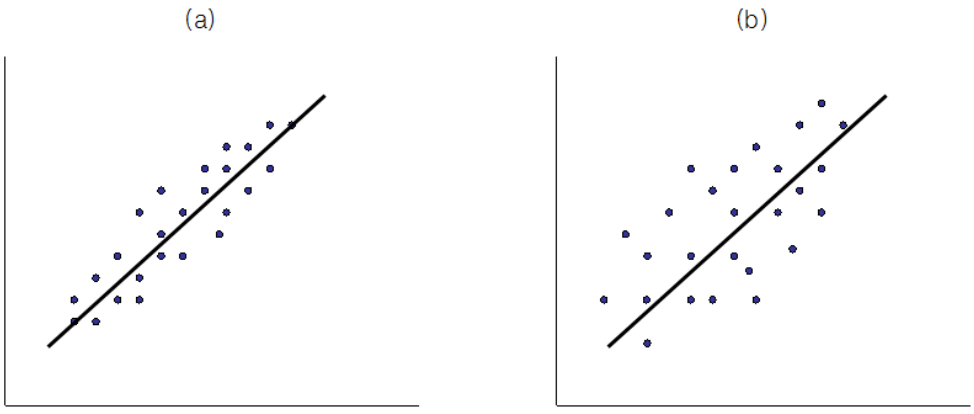
A. 목적

1) 두 변수 사이에 관계가 어느정도 강한가 ?

2) 평균과 표준편차만으로는 알 수 없는 한계



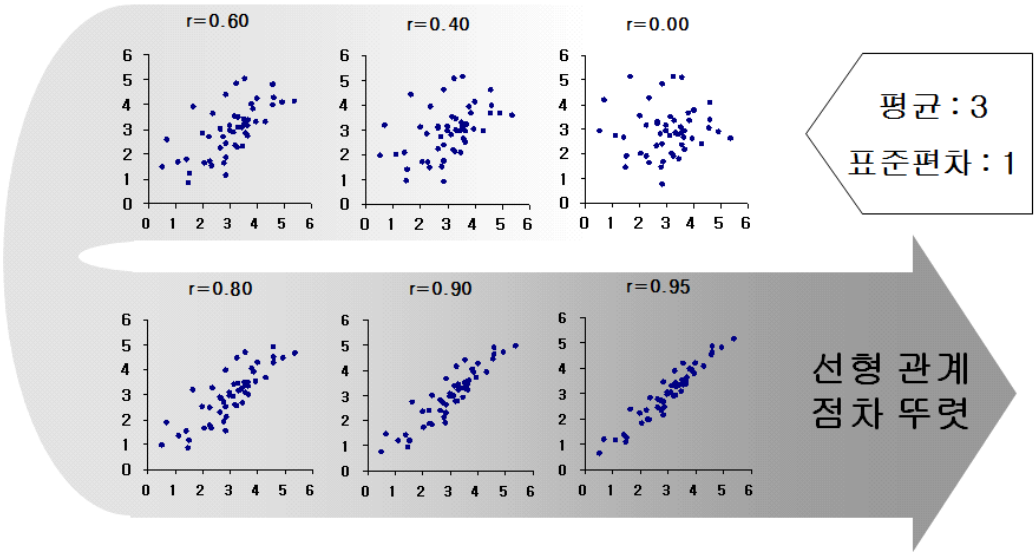
B. 산포도와 상관계수

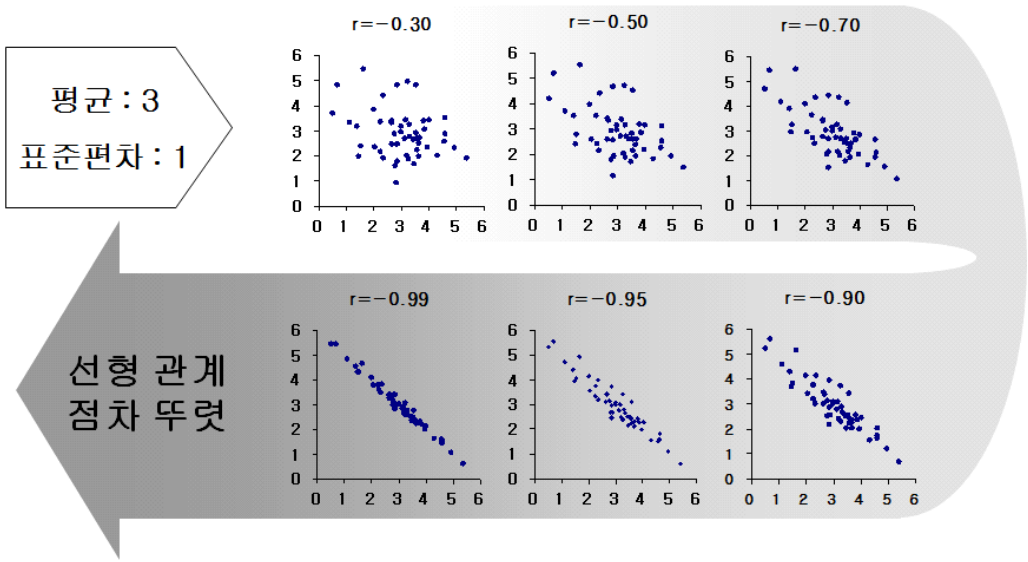


1에 가까운 상관계수는 점들이
선 주위에 몰려 있음을 의미

0에 가까운 상관계수는 점들이
선 주위에 퍼져 있음을 의미

B. 양의 상관관계와 음의 상관관계





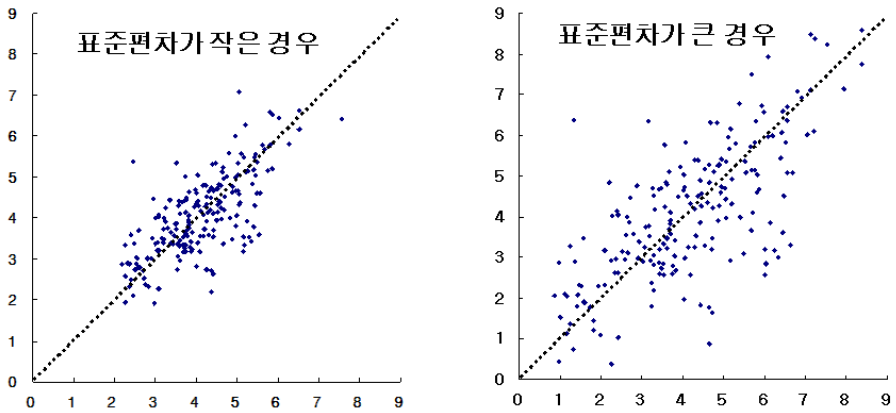
C. 상관계수 (Correlation Coefficient)

$$R = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

$$= \frac{n\sum_{i=1}^{n}x_iy_i - \sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{\sqrt{n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2}\sqrt{n\sum_{i=1}^{n}y_i^2-\left(\sum_{i=1}^{n}y_i\right)^2}}$$

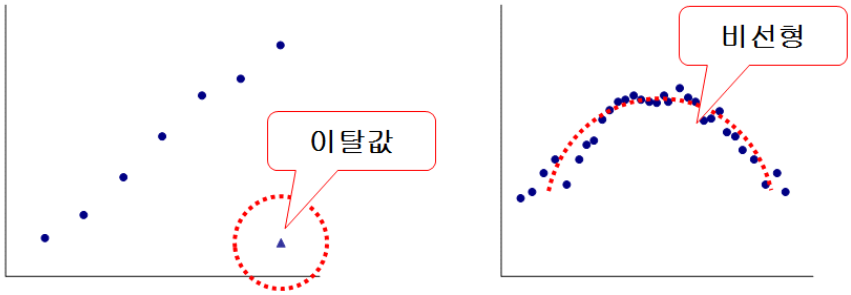
C. 상관계수의 특징

1) 얼마만큼 빽빽하게 밀집되어 있는가를 의미하지 않는다.



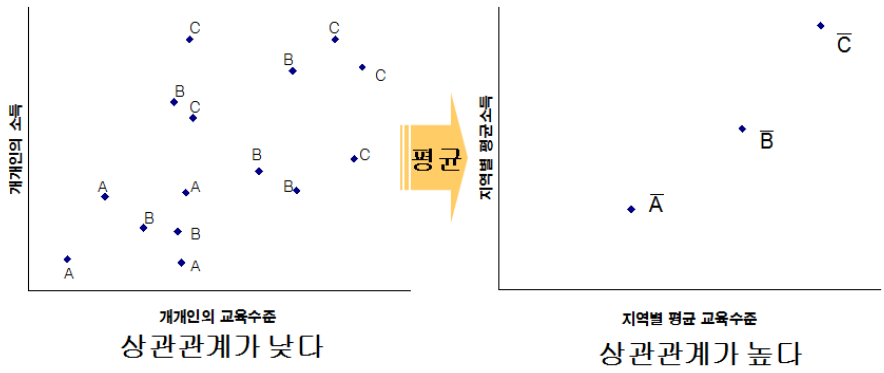
상관계수가 같아도 표준편차가 작으면
산포도가 더 빽빽하게 밀집된 것처럼 보인다.

2) 유용하지 않은 경우가 존재한다.



상관계수는 이탈값이 존재하거나 분포가 비선형일 때 유용성이 떨어진다.

3) 실제의 관계를 과장할 수 있다.



상관관계가 낮다 상관관계가 높다

집단별 비율이나 평균에 기초하여 구한 상관계수는 실제의 관계를 과장한다.