3.1 상관관계

1) 산포도와 상관관계
2) 상관계수 구하기, 상관계수의 특징
3) 상관계수와 선형관계

1. 산포도와 상관관계

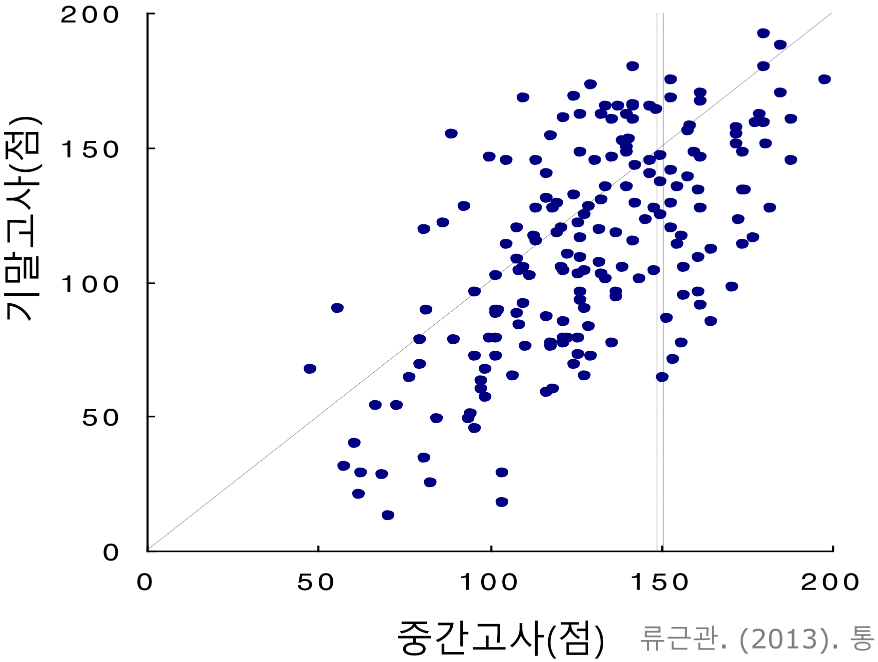
결합분포

- 이제까지는 한 번에 하나의 변수만을 다루는 방법에 대해 살펴보았다.
- 이제부터는 두 변수 사이의 상호관계를 분석하기 위한 방법을 살펴본다.
- 남녀간의 관계처럼 많은 경우 둘간의 관계가 중요하다.
  - 예: 교육과 임금
  - 예: 통화증가율과 물가상승률
  - 예: 학급 규모와 학생 성적
- 결합분포(joint distribution): 두 변수간의 관계 전모를 보여줌

1. 산포도와 상관관계

산포도 (scatter plot)

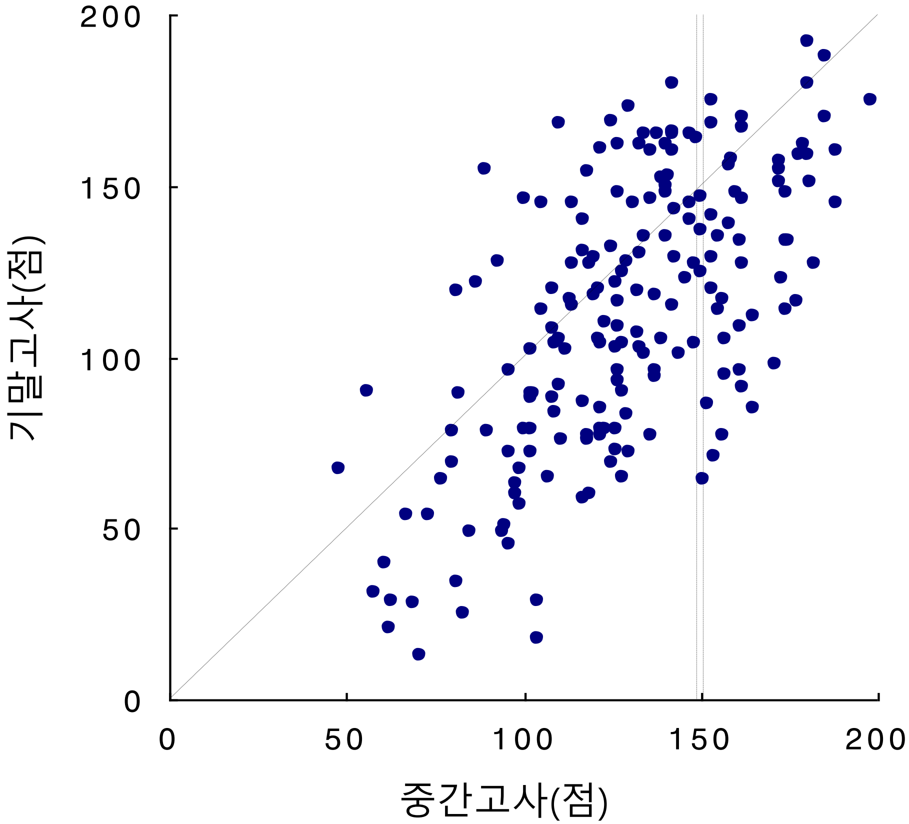
- 두 변수 사이의 관계를 살펴보기 위해 산포도를 이용한다.
- 설명변수는 x로 표기하고 가로축에 표시
- 피설명변수는 y로 표기하고 세로축에 표시



류근관. (2013). 통계학, 제 3 판. 서울: 법문사. P. 104

1. 산포도와 상관관계

중간고사 기말고사 성적간 관계가 약한 경우

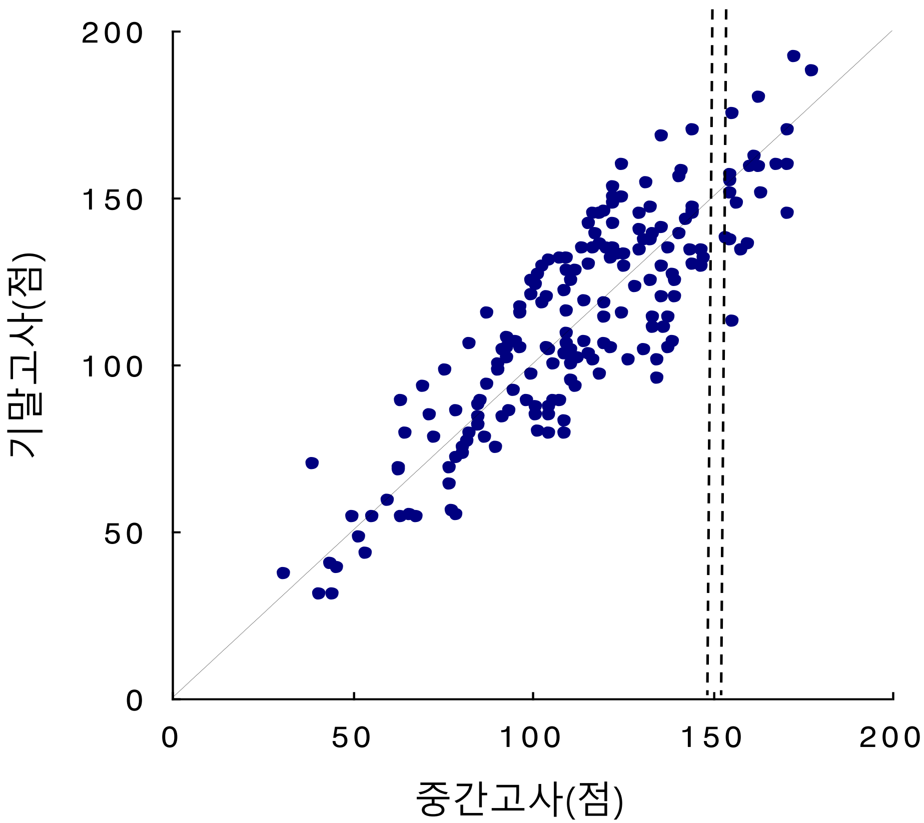


류근관. (2013). 통계학, 제 3 판. 서울: 법문사. P. 100

- 변수 사이의 관계가 약하면 한 변수 값이 다른 변수 값을 예측하는 데 큰 도움 안됨
- 중간고사에서 150점 받은 학생들의 기말고사 성적은 55점에서 175점 사이에 분포

1. 산포도와 상관관계

중간고사 기말고사 성적간 관계가 강한 경우



- 변수 사이의 관계가 강하면 한 변수 값이 다른 변수 값을 예측하는 데 크게 도움됨
- 중간고사에서 150점 받은 학생들의 기말고사 성적은 105점에서 175점 사이에 분포

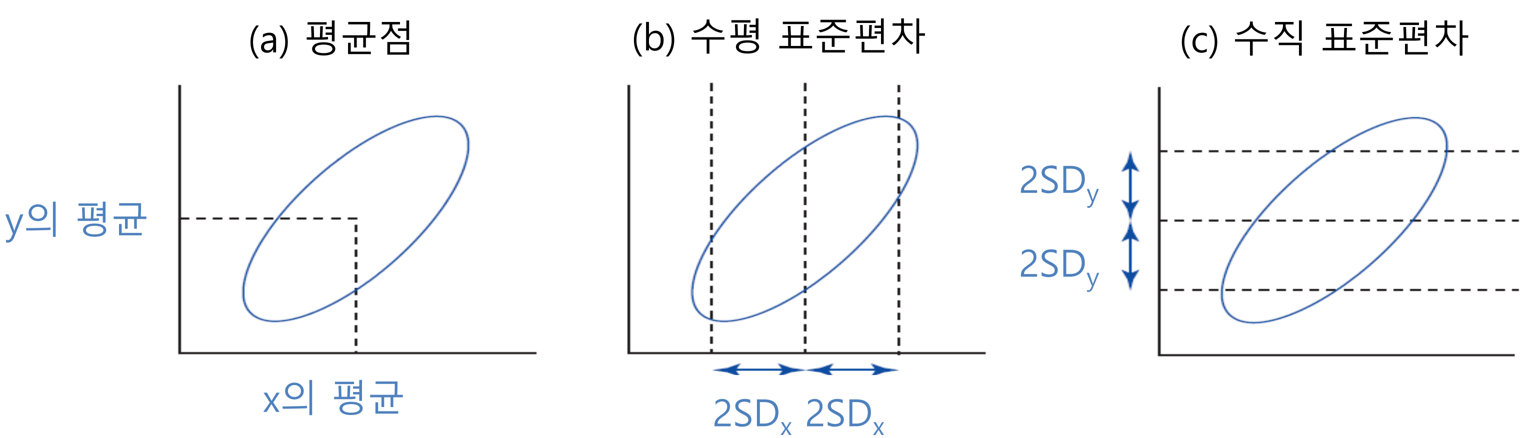
류근관. (2013). 통계학, 제 3 판. 서울: 법문사. P. 100

1. 산포도와 상관관계

산포도의 요약

- 가로로 보면 대략 95%의 점들이 x평균점을 기준으로 $\pm 2SD_x$ 이내에 위치함
- 세로로 보면 대략 95%의 점들이 y평균점을 기준으로 $\pm 2SD_y$ 이내에 위치함
- x의 평균과 표준편차, y의 평균과 표준편차는 x와 y의 분포를 따로따로 요약

그림 5–4 산포도의 요약

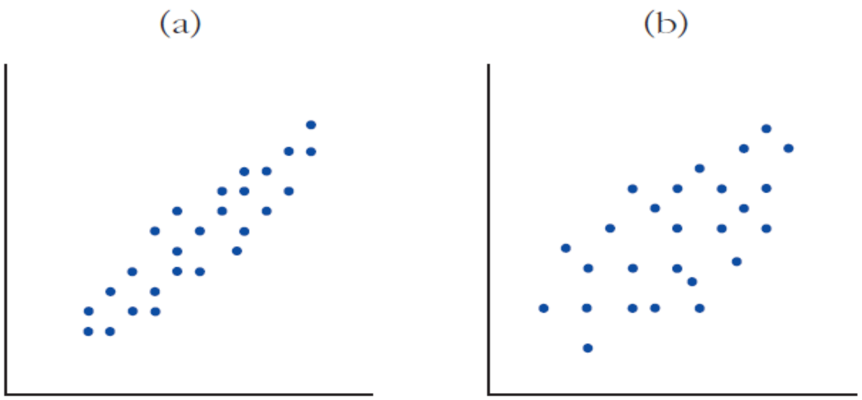


류근관. (2013). 통계학, 제 3 판. 서울: 법문사. P. 102

2. 상관계수 구하기, 상관계수의 특징

상관계수의 필요성

- 가로든 세로든 평균과 표준편차가 동일해도 두 변수의 관계는 상이



류근관. (2013). 통계학, 제 3 판. 서울: 법문사. P. 103

- 위의 두 산포도는 가로든 세로든 중심과 퍼진 정도가 동일하지만 (a)가 (b)보다 더 강한 선형관계를 보임
- 두 변수간 선형관계의 방향과 강도가 얼마나 되는지 측정할 필요성 대두
- 상관계수는 두 변수간 선형관계의 방향과 강도 측정

2. 상관계수 구하기, 상관계수의 특징

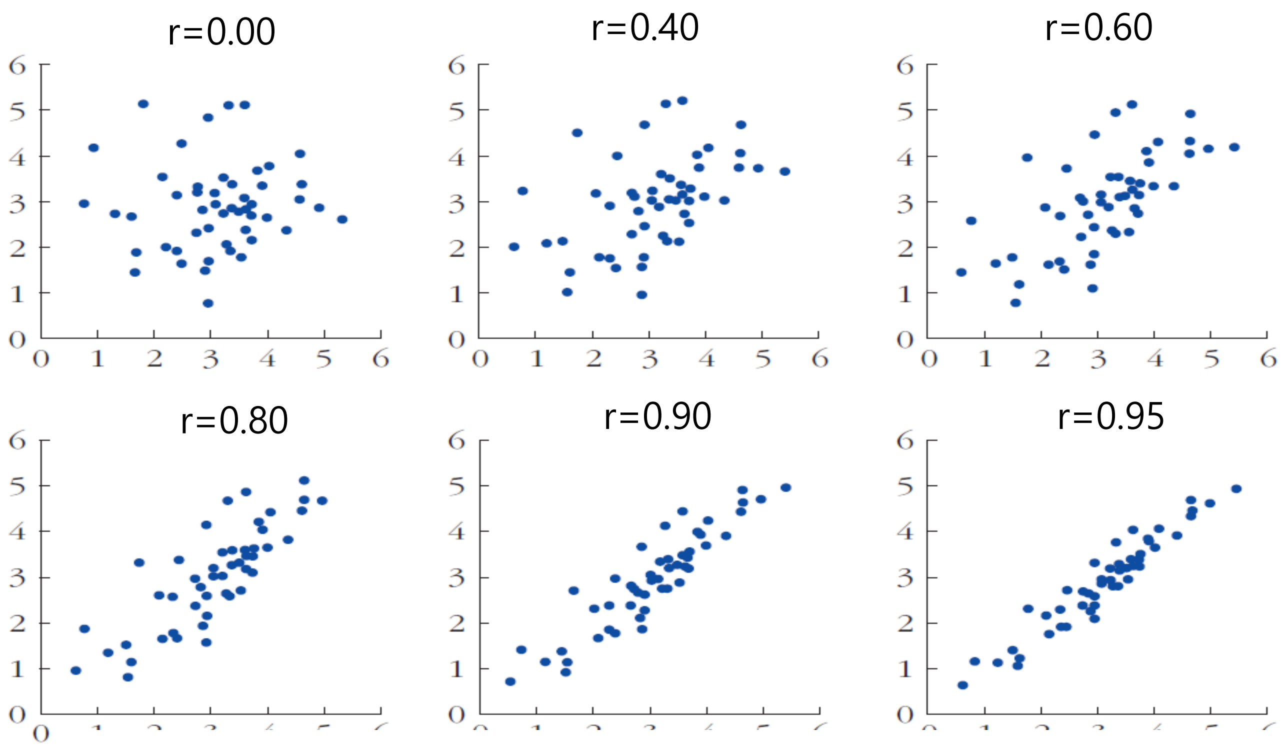
두 변수 사이의 관계

- 이변량 자료의 요약 통계량
  - x의 평균과 표준편차
  - y의 평균과 표준편차
  - x와 y간 상관계수 (correlation coefficient) : r로 표기

2. 상관계수 구하기, 상관계수의 특징

상관계수가 다른 여섯 개의 산포도

그림 5-6 양(+)의 상관계수 값을 갖는 여섯 개의 산포도



류근관. (2013). 통계학, 제 3 판. 서울: 법문사. P. 104

주: 각각의 산포도는 가로, 세로 모두 평균 3, 표준편차 1의 동일한 값을 갖는다.

각각의 산포도에는 50개씩의 점이 찍혀 있다.

2. 상관계수 구하기, 상관계수의 특징

상관계수의 범위, 부호

- 범위: $-1 \le r \le 1$

- 상관계수 = 1 또는 -1 이면 완전상관(perfect correlation)
  - 모든 점들이 정확히 하나의 선 위에 위치

- 양의 상관관계이면 점의 분포가 우상향
- 음의 상관관계이면 점의 분포가 우하향

- 두 변수의 표준편차가 모두 0이면 상관계수를 정의할 수 없음
- 두 변수 중 어느 한 변수만의 표준편차가 0이면 상관계수는 0

2. 상관계수 구하기, 상관계수의 특징

상관계수 구하는 절차 1

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

1) 각 변수를 평균으로부터의 편차로 바꾼다.

2) 두 편차를 서로 곱한 뒤 합친다.

3) 각 편차를 제곱하여 합치고, 다시 제곱근을 취한다. 두 제곱근을 곱한다.

4) 위 2)에서 얻은 값을 위 3)에서 얻은 값으로 나눈다.

2. 상관계수 구하기, 상관계수의 특징

상관계수 구하는 절차 2

변형된 공식

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

1) 각 변수를 평균으로부터의 편차로 변환한다.

2) 두 편차를 서로 곱하여 합친 뒤 자유도 $n-1$로 나누어 공분산을 구한다.

3) 두 표준편차를 곱한다.

4) 위 2)에서 구한 값을 위 3)에서 구한 값으로 나눈다.

2. 상관계수 구하기, 상관계수의 특징

상관계수의 특징

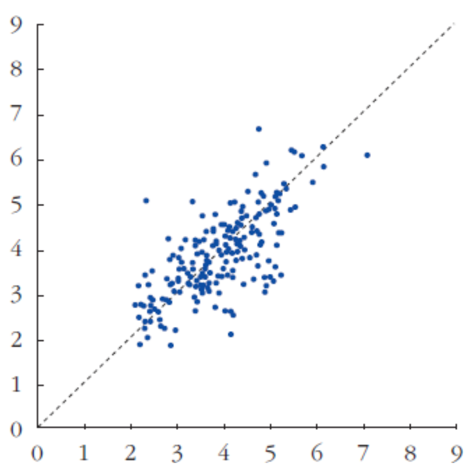
- 상관계수는 단위를 갖지 않음. 즉, 측정단위와 독립적으로 정의됨
  - 하나의 변수가 취하는 모든 값에 상수를 더하거나 빼는 변환을 해도 상관계수는 변하지 않음
  - 하나의 변수가 취하는 모든 값에 양의 상수를 곱하거나 양의 상수로 나누는 변환을 해도 상관계수는 변하지 않음

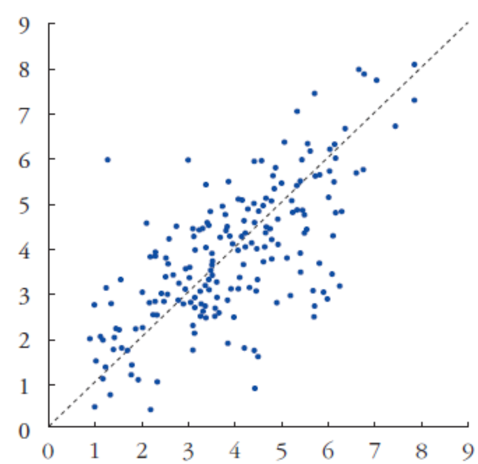
- 상관계수는 방향성을 갖지 않음. 즉 x와 y의 상관계수는 y와 x의 상관계수와 같음

3. 상관계수와 선형관계

상관계수의 해석

- 상관계수의 의미
  - '상관계수=0.8'은 산포도 상에서 80%의 점들이 하나의 선 주위에 빽빽하게 밀집해 있다는 것을 의미하지 않는다.
  - '상관계수=0.8'은 상관계수가 0.4일 때보다 선형관계의 강도가 강하기는 하지만 정확히 두 배로 강하다는 것을 의미하지도 않는다.
- 산포도상에서 표준편차를 변화시킬 때의 시각적 효과





사실상 두 그림에서
상관계수는 0.7로 같음

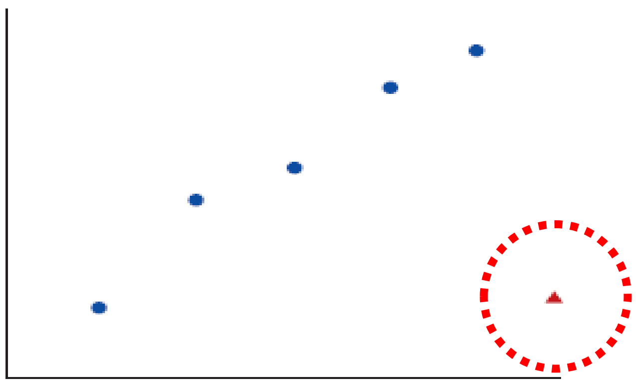
류근관. (2013). 통계학, 제 3 판. 서울: 법문사. P. 110

3. 상관계수와 선형관계

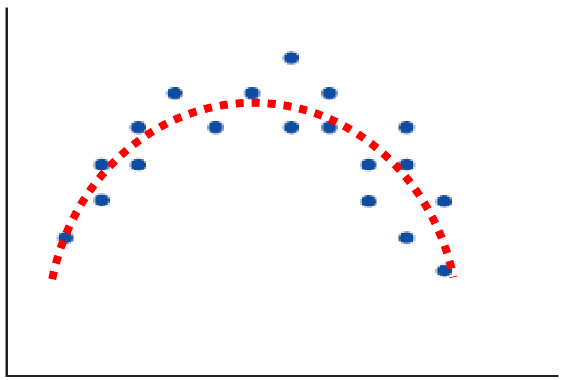
상관계수가 유용하지 않은 경우

- 이탈값(outlier)이 존재하는 경우
- 두 변수간 관계가 비선형인 경우

(a) 이탈값



(b) 비선형관계



류근관. (2013). 통계학, 제 3 판. 서울: 법문사. P. 111