금융인을 위한

통계와 데이터 분석 입문

기술통계량 이해하기
: 왜도, 첨도, 상관계수

한국금융연수원
KOREA BANKING INSTITUTE

학습 내용

1. 연속형/수치형 자료를 요약하는 기술통계량 (descriptive statistics)의 종류
   - 자료의 비대칭성 : 왜도(skewness)
   - 자료의 분포 모양 : 첨도(kurtosis)
   - 두 변수 간의 관계 : 상관계수(correlation coefficient)

2. (실습) 파이썬을 활용한 연속형/수치형 자료의 왜도, 첨도, 상관계수 분석

왜도(skewness)

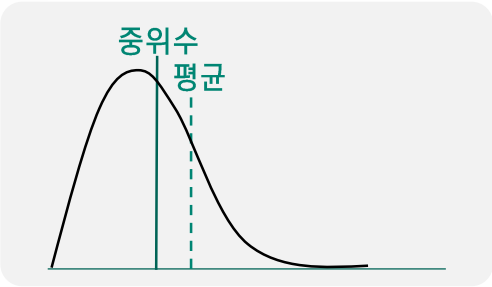
왜도(skewness)

자료의 비대칭성을 알아보는 기술 통계량

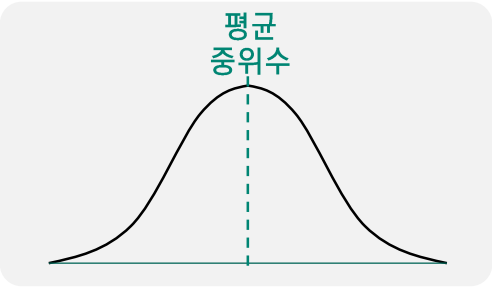
- 왜도 계산식은 관측치들과 평균과의 거리의 세 제곱의 합을 포함함
- 계산된 왜도 값
  - 0이면 분포가 **좌우대칭**(symmetric)
  - 0 **이상**이면 오른쪽으로 긴 **꼬리**를 가진 분포(positive skew)
  - 0 **이하**이면 왼쪽으로 긴 **꼬리**를 가진 분포(negative skew)

왜도(skewness)

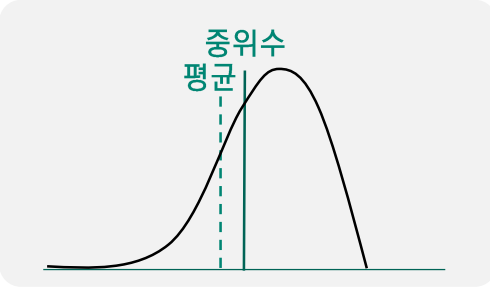
- 왜도의 절대값이 커질수록 비대칭 정도가 심하다는 것을 의미



positive
skew



symmetrical
distribution



negative
skew

첨도(kurtosis)

첨도(kurtosis)

자료의 분포모양 중 **꼬리부분의 두꺼운 정도**를
나타내는 기술통계량

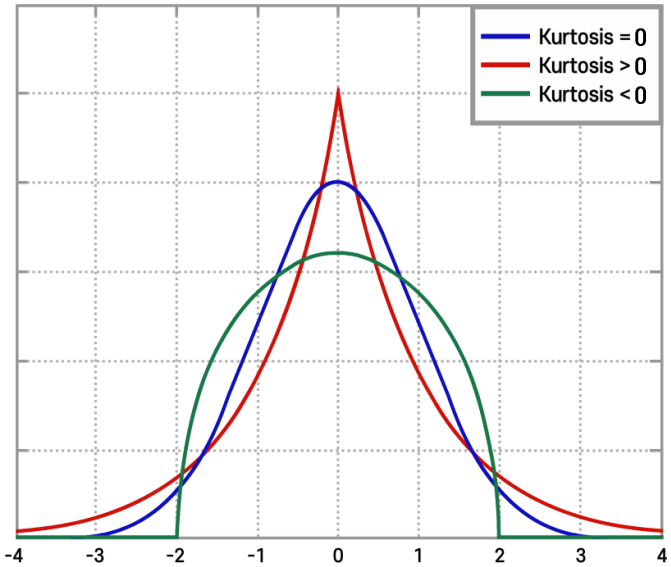
- 첨도 계산식은 관측치들과 평균과의 거리의 네 제곱의 합을 포함함
- 정규분포를 기준으로 함 : 정규분포의 첨도는 0
  - 0 이상이면 정규분포보다 두꺼운 꼬리를 가진 분포
  - 0 이하이면 정규분포보다 얇은 꼬리를 가진 분포
- 자료의 산포도와는 다른 개념

예

분산이 다른 정규분포들의 첨도는 모두 0

첨도(kurtosis)

자료의 분포모양 중 **꼬리부분의 두꺼운 정도**를
나타내는 기술통계량



상관계수(correlation coefficient)

피어슨 상관계수

두 연속형/수치형 변수간의 **선형 관계**의
강도와 **방향**을 나타내는 기초통계량

- 두 변수 X,Y의 상관계수는 공분산(covariance)를 각각의 표준편차의 곱으로 나눠준 값
  - $(x_i, y_i), i = 1, \cdots, n$ 들이 표본이라 할 때

$$r = \frac{\sum_1^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_1^n (x_i - \bar{x})^2}\sqrt{\sum_1^n (y_i - \bar{y})^2}}$$

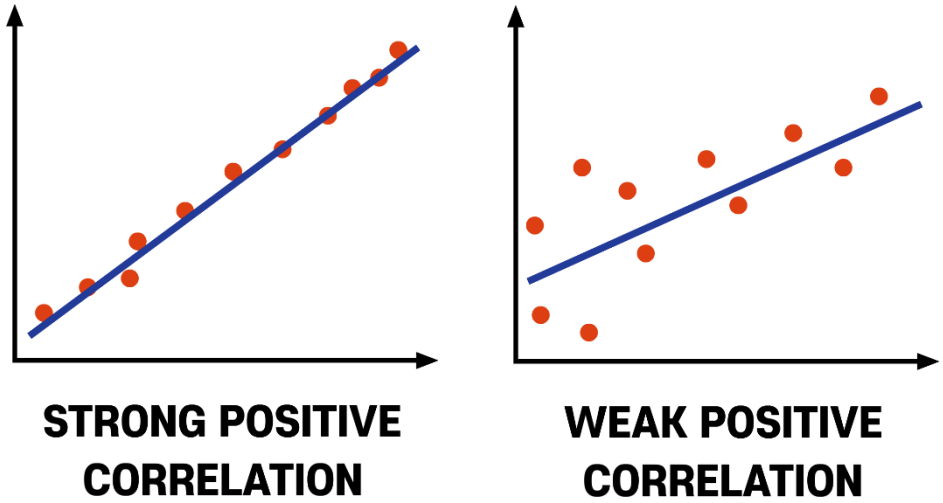
- $-1 \le r \le 1$ 또는 $-100\% \le r \le 100\%$ (단위 없음)

참고 두 변수의 값이 아닌 순위(rank)들 간의 관계에 관심 있다면 스피어만 순위상관계수, 켄달 타우 등을 고려함

상관계수(correlation coefficient)

양의 상관관계($r > 0$)

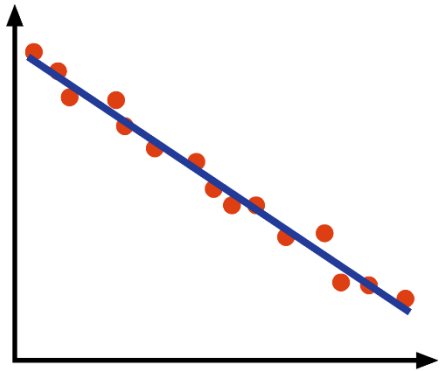
- 두 변수의 움직임이 같은 방향인 경우
- 두 변수가 양(positive)의 방향으로 상관관계가 강할수록 상관계수는 1에 가깝게 나타남



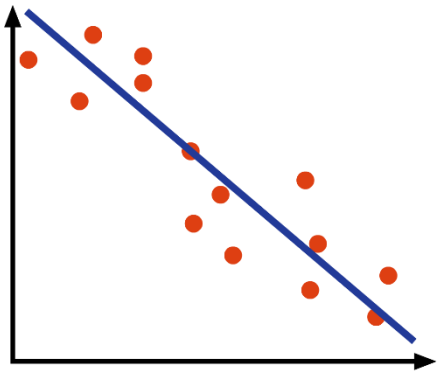
상관계수(correlation coefficient)

음의 상관관계($r < 0$)

- 두 변수의 움직임이 반대 방향인 경우
- 두 변수가 음(negative)의 방향으로 상관관계가 강할수록 상관계수는 -1에 가깝게 나타남



STRONG NEGATIVE CORRELATION

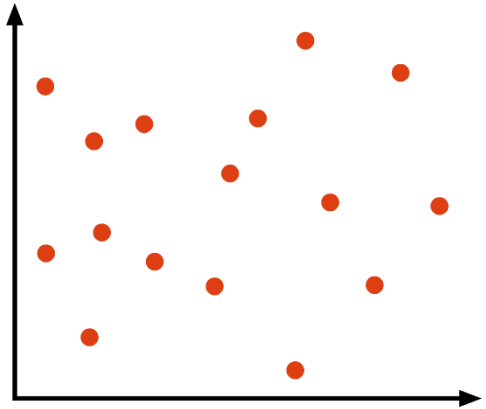


WEAK NEGATIVE CORRELATION

상관계수(correlation coefficient)

두 변수 사이에 선형 관계가 없는 경우($r \approx 0$)

- 상관계수가 0에 가깝게 나타남



NO CORRELATION

- 상관계수는 두 변수 간의 선형관계의 정도와 방향을 나타낼 뿐 인과관계를 나타내는 것은 아님

학습정리

- **왜도(skewness)**는 수치형 변수의 **비대칭성**을 나타내는 기술통계량임
- **첨도(kurtosis)**는 수치형 변수의 분포 모양 중 **꼬리부분의 두꺼운 정도**를 **정규분포**를 기준으로 나타내는 기술통계량임
- **상관계수(correlation coefficient)**는 두 수치형 변수 간의 **선형 관계의 정도와 방향**을 나타내는 기술통계량으로 인과관계로 확대해석하지 않도록 주의해야 함
- 파이썬을 활용하여 연속형/수치형 자료의 **왜도, 첨도, 상관계수**를 분석할 수 있음