기초 통계 II

- ADsP -

조상구
빅데이터과 경복대학교

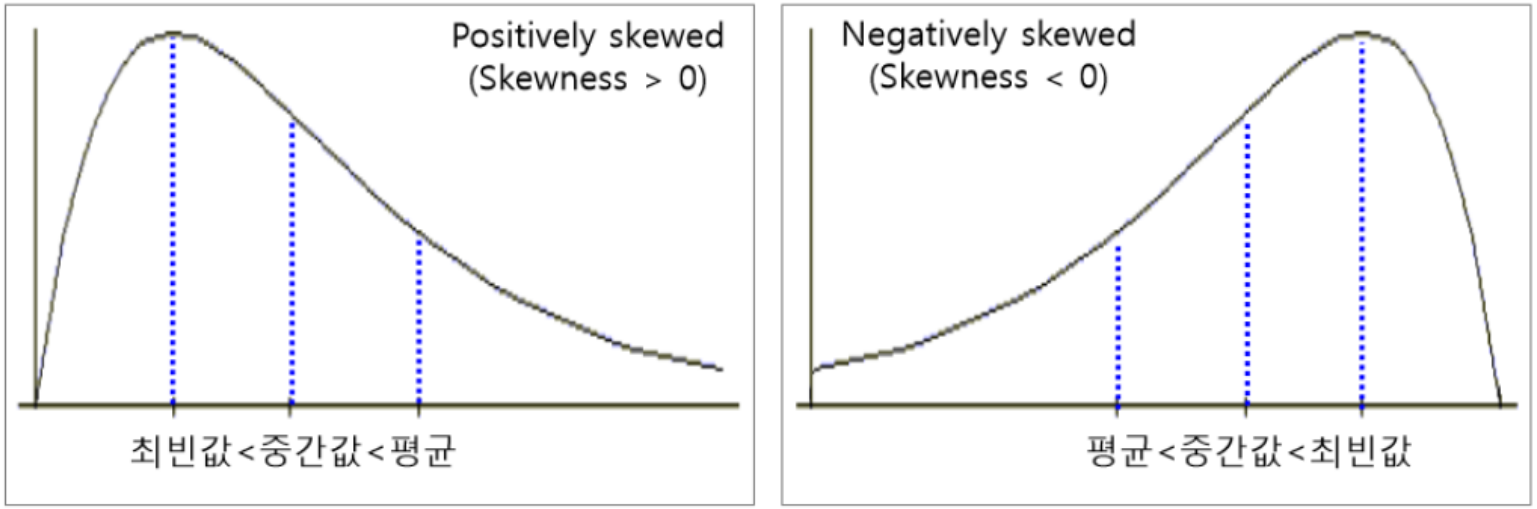
기술 통계분석(Descriptive Statistics Analyssis)

- 모집단으로부터 수집된 데이터를 잘 요약하여 모집단에 대하여 유용한 정보를 생산한다.
- 기술통계 분석에서는 데이터를 그래프, 표 또는 대표값을 나타내는 숫자(통계량)로 요약한다.
- 우리는 실생활에서 많은 기술통계 분석 자료를 접할 수 있다.

Descriptive Statistics: 왜도

- **왜도 (Skewness)**
  - 자료의 분포에 대한 비대칭의 정도
  - 비대칭이 없는 경우 왜도 = 0

$$skewness = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{s}\right)^3$$

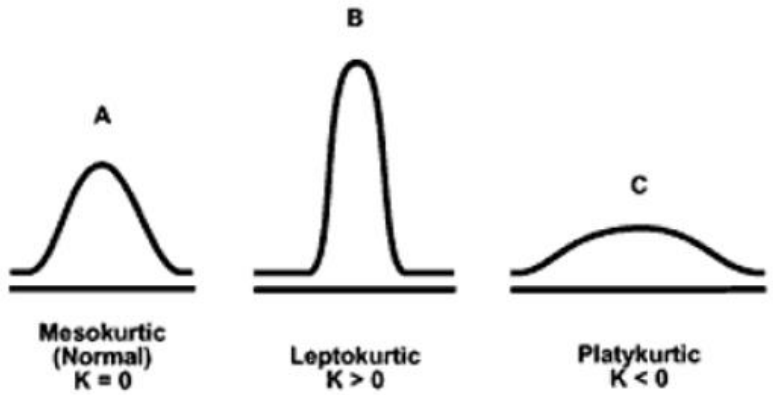


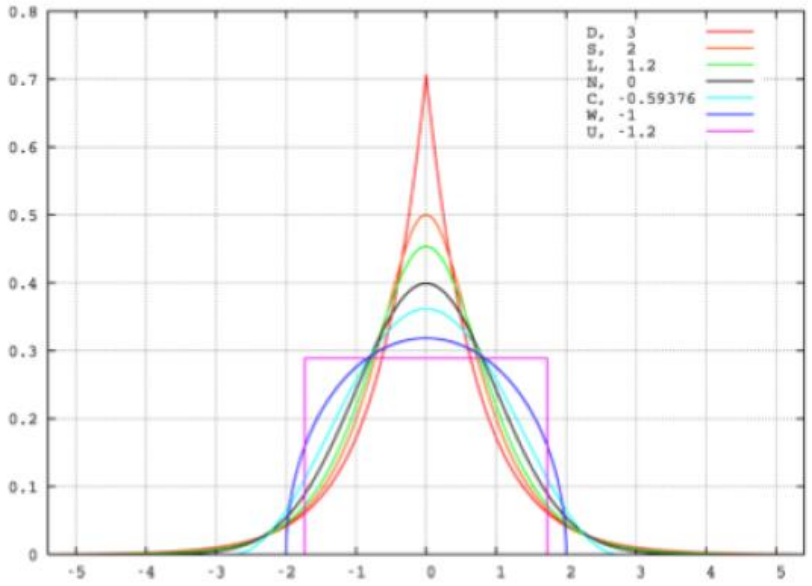
- Ex) 500, 489, 495, 493, 505, 248 → *skewness* = -2.5

Descriptive Statistics: 첨도

- **첨도 (Kurtosis)**
  - Data 분포의 뾰족한 정도
  - 표준정규분포의 경우 첨도 = 0

$$Kurtosis = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{s}\right)^4 - 3$$





- Ex) 500, 489, 495, 493, 505, 248 → *Kurtosis* = 5.93

Descriptive Statistics: 실습

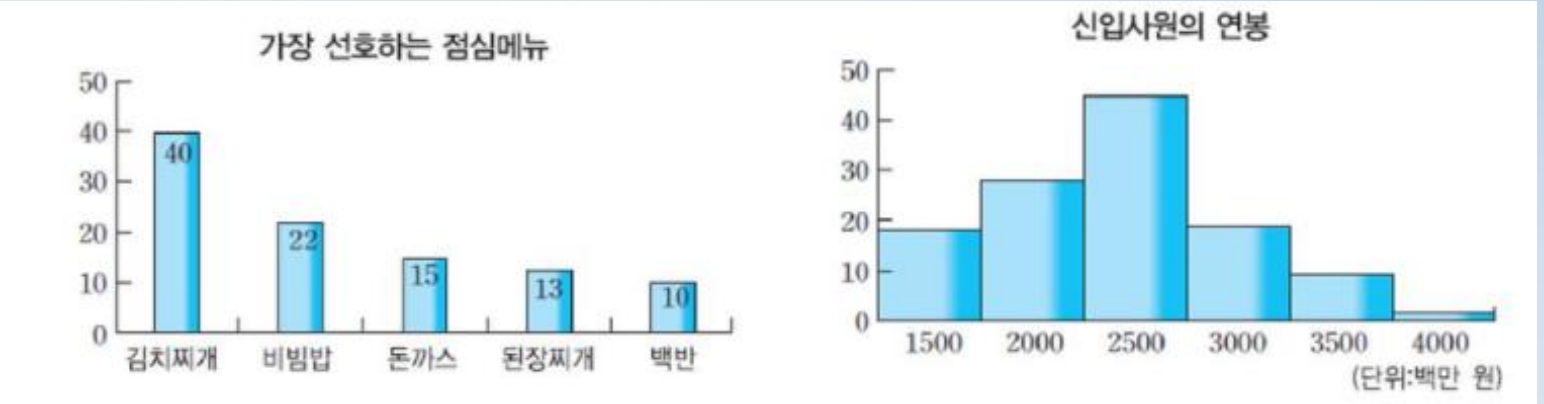
- **Example) 여대생 신장 자료**

170	151	154	160	158	154	171	156	160
157	160	157	148	165	158	159	155	151
152	161	156	164	156	163	174	153	170
149	166	154	166	160	160	161	154	163
164	160	148	162	167	165	158	158	176

  - 대표값
    - 평균 =
    - 중앙값 =
  - 산포
    - 범위=
    - (Q1, Q2, Q3) =
    - 사분위범위 =
    - 90%백분위수 =
    - (분산, 표준편차) =
  - 왜도 =
  - 첨도 =

※ Tip) 관련 Excel 함수
: average, median, quartile, percentile, var, stdev, skew, kurt

막대그래프 vs. 히스토그램

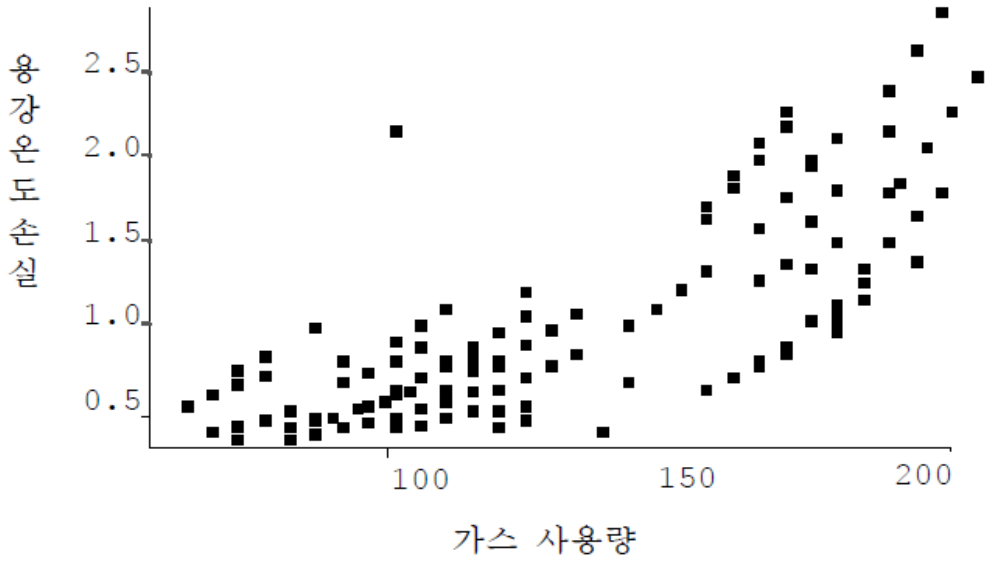


구분	막대그래프	히스토그램
데이터 유형	• 범주적 (종교, 직업 등)	• 연속적 (몸무게, 성적 등)
막대 순서	• 임의적으로 변경 가능	• 변경 불가능
막대 간격	• 일정한 간격 유지	• 간격 없이 표현
용도	• 범주의 비교, 순위 표시 등	• 데이터 분포 표현

상관 분석

산점도

- **산점도(scatter plot)**
  - (x_1,y_1), (x_2,y_2), …, (x_n, y_n)평면좌표에 표시한 도표
  - 시각적으로 두 변수의 관계를 쉽게 알 수 있게 해줌
  - 변수간 관계의 방향, 형태, 관계의 강도를 알 수 있음



공분산과 표본공분산

- **공분산(Covariance)**
  - 두 확률변수 X와 Y의 선형적 상관관계를 측정할 수 있는 척도
  - X와 Y가 각각의 평균을 중심으로 하여 같은 방향으로 변화하는 정도를 나타냄
  - $\sigma_{xy}=Cov[X, Y] = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$
  - 공분산 값의 크기가 단위 등에 따라 달라짐

- **표본 공분산 (Sample Covariance)**
  - σ_{xy} 의 추정량으로 S_{xy} 로 표현

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \frac{1}{n-1}(\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y})$$

상관계수와 표본상관계수 (1/3)

- **상관계수**
  - 공분산을 단위와 무관하게 표준화시킴
  - -1 에서 1까지의 값을 가짐

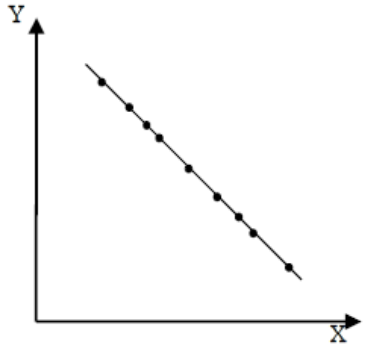
$$\rho_{XY} = \frac{Cov[X,Y]}{\sqrt{Var[X]}\sqrt{Var[Y]}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

- **표본상관계수**
  - 상관계수의 추정량으로 r_{xy} 로 표현

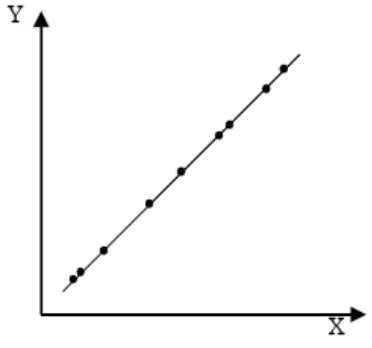
$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

상관계수와 표본상관계수 (2/3)

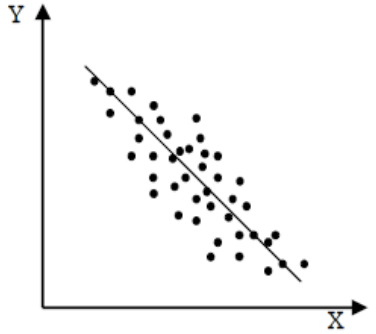
- 표본상관계수에 따른 산점도



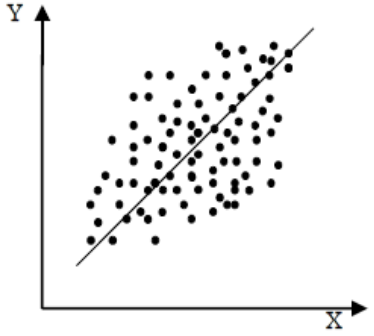
(a) r = -1



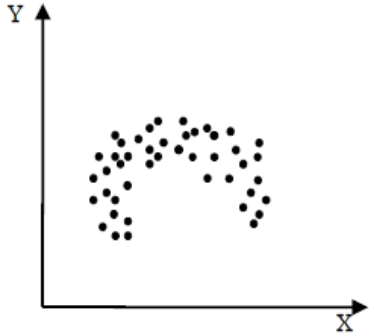
(b) r = 1



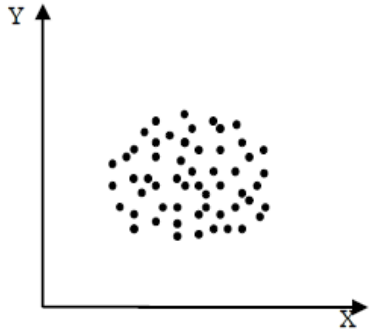
(c) r = -0.7



(d) r = 0.5



(e) r = 0



(f) r = 0

autodesk

- https://www.research.autodesk.com/publications/same-stats-different-graphs/