2장. 모집단과 표본

박명현

서울대학교 학부대학

기술통계(descriptive statistics)

- 기술통계: 자료를 정리, 요약하여 기술하는 것
  - 자료의 대략적인 모습을 보여주고(summary), 섬세한 분석을 위한 기초단계로 활용된다.

1. 그래프나 표를 이용한 기술통계
   - 표본을 도표화함으로써 모집단 분포의 개형을 파악
   - 도수분포표, 분할표, 히스토그램, 산점도 등
2. 수치에 의한 기술통계
   - 중심위치 척도: 평균, 중앙값 사분위수 등
   - 변동성 척도: 분산, 범위 등
   - 연관성 척도: 공분산, 상관계수

자료의 구분(사칙연산 가능 여부에 따른)

- 질적자료(qualitative)
  - 사칙연산 불가능
  - 관측결과가 몇개의 범주 또는 항목의 형태로 나타남
  - 예) 직업, 종교, 성별 등
  - 범주형 자료(categorical data)라 부르기도 함
- 양적자료(quantitative)
  - 사칙연산 가능
  - 관측된 값이 수치로 측정됨
  - 합계, 평균, 분산 등으로 자료를 요약 정리할 수 있으며 비교도 가능
  - 예) 몸무게, 용돈, 가격 등
  - 숫자형 자료(numerical data)라 부르기도 함

자료의 구분(측정척도에 따른)

- 질적자료
  - 명목형 자료(nominal data)
    - 이름이나 문자로 나타내어지는 자료.
    - 범주간 순서가 없음
      예: 혈액형(A,B,O,AB), 성별(남, 여)
      예: 직업종류(전문직, 회사원, 공무원)
  - 순서형 자료(ordinal data)
    - 이름이나 문자로 나타내어지는 자료이나 범주들 간에 순서가 있는 자료
    - 사칙연산을 할 수 없음
      예: 선호도(아주좋아함, 좋아함, 보통, 싫어함, 아주싫어함)
      예: 학점(A, B, C, D, F), 옷사이즈(S,M,L,XL)

자료의 구분(측정척도에 따른)

- 양적자료
  - 연속형 자료
    - 연속적인 값을 가지는 자료들
      예: 키, 몸무게
    - 많은 경우에 정확한 수치보다는 반올림 등을 통해 나타냄
      예: 키170cm 는 169.5에서 170.5사이의 값을 의미함
  - 이산형 자료
    - 관측가능한 값이 셀수 있음
      예: 주사위 결과, 교통사고 건수

자료 종류의 예

자료	종류
전문직업분류(변호사, 의사, 회계사등)	질적자료, 범주형자료
몸무게	양적자료, 숫자형자료
학력(초, 중, 고, 대)	질적자료, 순서형자료, 범주형자료
학력(교육연한)	양적자료, 숫자형자료

- 자료의 계층에 따라 선택할 수 있는 통계분석 도구가 달라지게 됨
  - 나이는 연수로 표시되면 숫자형자료. 평균과 분산을 구할 수 있고, 히스토그램 등의 그래프로 표시가능.
  - 나이를 10년 단위로 그룹화 하여 도수분포표와 바차트 등으로 표시할 수 있음.
  - 나이를 소년, 청년, 중년, 장년 등으로 순서형 자료로 표시하였다면 도수분포표와 바차트는 가능. 그러나 평균이나 분산을 구할 수 없음.

그래프나 표를 이용한 기술통계
(graphical and tabular description)

도수분포표(frequency table)

- 표본자료를 하나의 표로 요약한 것.
- 서로 다른 특성값에 대한 자료의 개수, 즉 도수(frequency)나 상대도수(relative frequency)를 구하여 특성값과 함께 나열한 것
- 상대도수=도수/전체자료의 갯수

표 1–1 어느 시의 시장 선거 여론조사 결과

후보명	지지율
John Smith	42%
Emily Johnson	37%
Michael Lee	15%
Sarah Davis	6%

- 수치형 자료도 전체 범위를 몇 개의 계급(구간)으로 나누어 도수분포표 작성 가능

도수분포표 예제

Example

예제 2.6) 다음은 어떤 집단에서 40명을 뽑아 키를 측정한 결과이다. 이 표본자료를 이용하여 도수분포표를 작성하여라(단위 : cm)

183 168 165 158 180 176 168 160 160 150
180 183 170 170 185 170 165 170 168 154
152 173 173 165 163 163 160 178 138 171
163 168 165 165 178 170 168 183 145 174

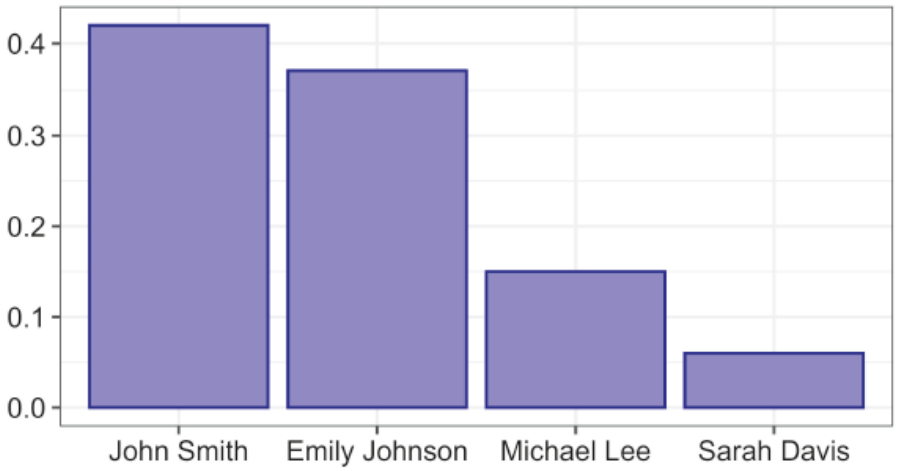
sol) 먼저 최대값=185, 최소값=138이므로 범위=47이다.
계급의 개수를 6으로 하면 $(47/6 = 7.83)$이므로 폭은 8로 한다.
시작점= 138-1/2× 1=137.5

표 2-6 [예제 2-6]의 도수분포표

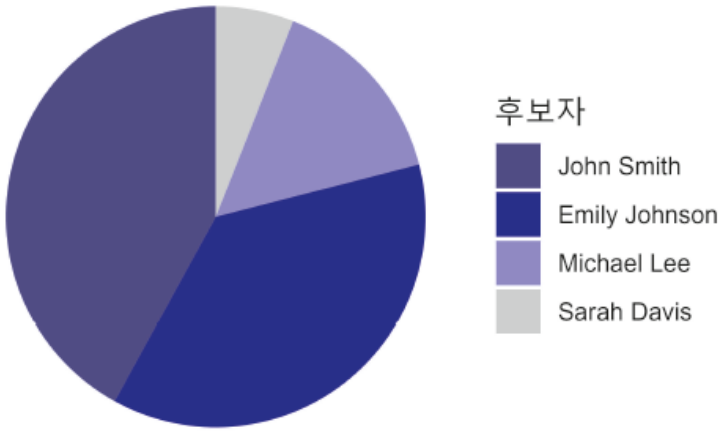
계급(cm)	도수	상대도수
$137.5^{\text{이상}} \sim 145.5^{\text{미만}}$	2	0.050
$145.5^{\text{이상}} \sim 153.5^{\text{미만}}$	2	0.050
$153.5^{\text{이상}} \sim 161.5^{\text{미만}}$	5	0.125
$161.5^{\text{이상}} \sim 169.5^{\text{미만}}$	13	0.325
$169.5^{\text{이상}} \sim 177.5^{\text{미만}}$	10	0.250
$177.5^{\text{이상}} \sim 185.5^{\text{미만}}$	8	0.200
합계	40	1

막대그래프와 원형그래프

- 막대그래프(bar graph): 각 특성값의 막대의 높이가 상대도수에 비례하게 그린 것
- 원형그래프(pie graph): 원을 부채꼴 모양으로 나누는데, 각 부채꼴의 넓이(또는 중심각의 크기)가 상대도수에 비례하게 나눈 것



(a) 막대그래프



(b) 원형그래프

그림 2–11 [예제 2–5]의 막대그래프와 원형그래프

히스토그램(histogram)

- 도수분포표를 이용하여 표본자료의 분포를 그래프로 나타낸 것

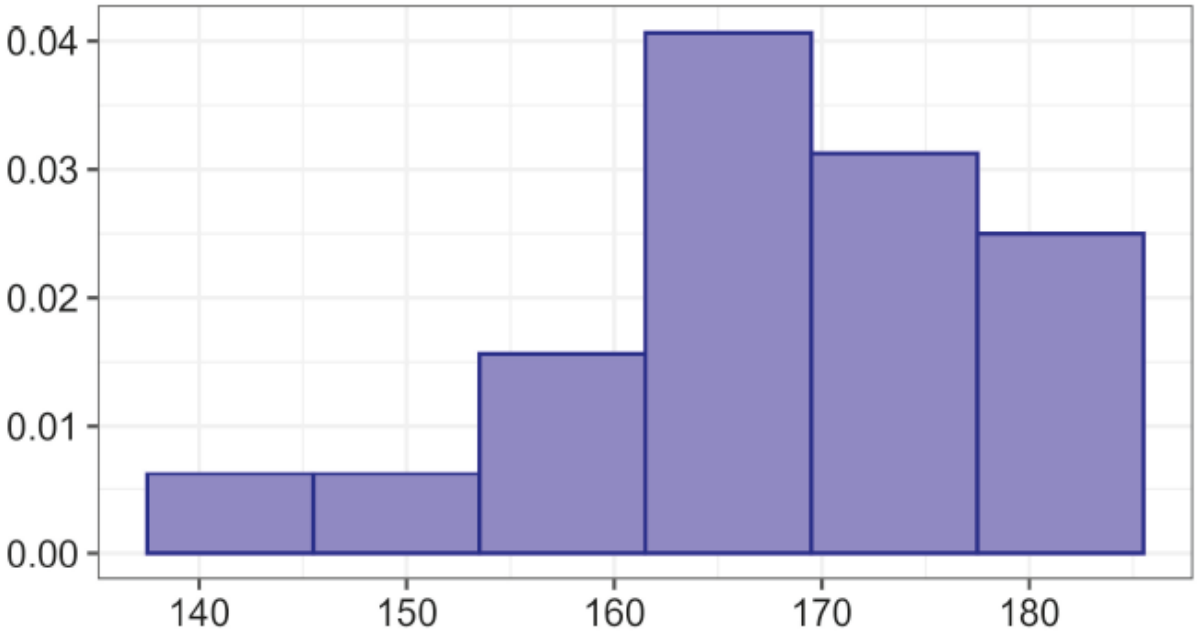


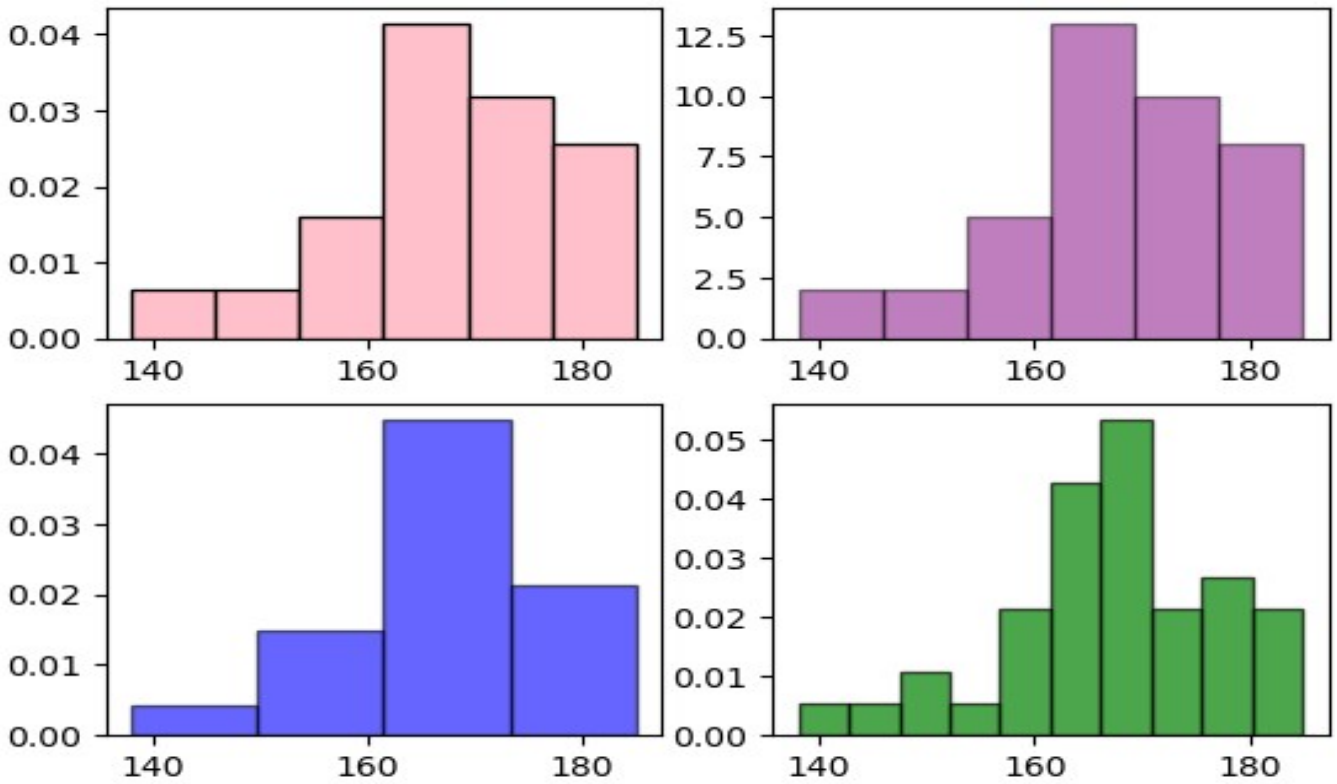
그림 2-12 [예제 2-6]의 히스토그램 : 모든 계급의 폭이 일정한 경우

히스토그램(histogram)

- 수평축위에 계급구간을 표시하고 그 위로 각 계급의 상대도수에 비례하는 넓이의 직사각형을 그린다.
- 직사각형의 넓이가 상대도수에 비례
- 직사각형의 높이=상대도수/계급의 폭
- 각 계급의 폭은 일정
- 전체 직사각형 넓이의 합은 1이 되어야 함
- 폭의 너비에 따라 분포의 형태가 달라짐

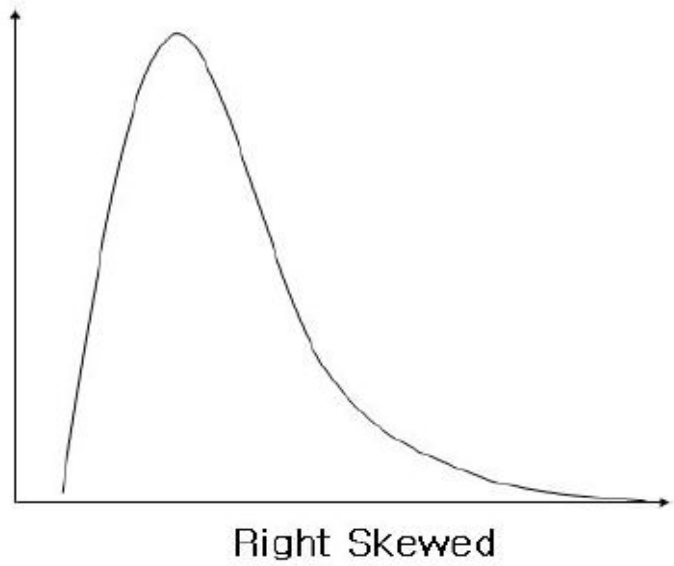
히스토그램(histogram)

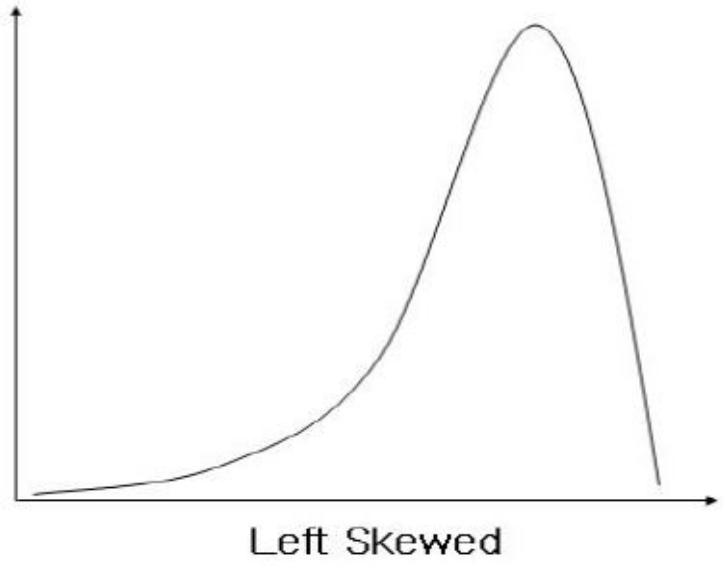
- 높이를 도수, 계급의 폭을 4개, 10개로 바꾸어 그린 히스토그램
  - hist(data,bins=6,color ='purple',edgecolor='black',alpha = 0.5,density=True)



히스토그램(histogram)

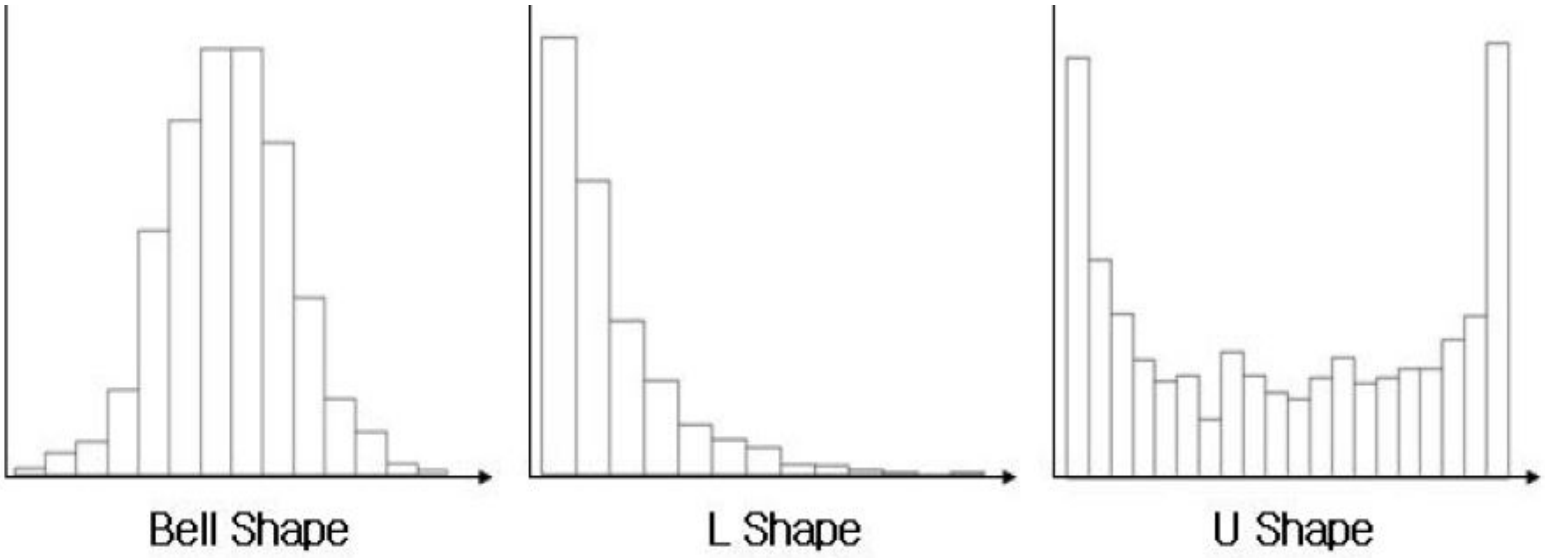
- 자료에 대한 정보를 특징적 형태로 제공함
- 대칭(Symmetry)
- 왜도(Skewness): 좌우로 늘어진 정도
  - 양의 왜도(Right skewed) : 오른쪽으로 길게 늘어짐
  - 음의 왜도(Left skewed) : 왼쪽으로 길게 늘어짐





히스토그램(histogram)

- 모양(Shape): 종 모양(Bell Shape), L자 모양, U자 모양



- 봉우리 개수: 단봉(Unimodal), 2봉(Bimodal)
  - 상이한 집단의 자료들이 섞여 있을때
  - 남녀 구별하지 않은 몸무게 자료
  - 서울 강남과 강원도 삼척의 Apartment가격 자료

줄기-잎 그림(stem-and-leaf display)

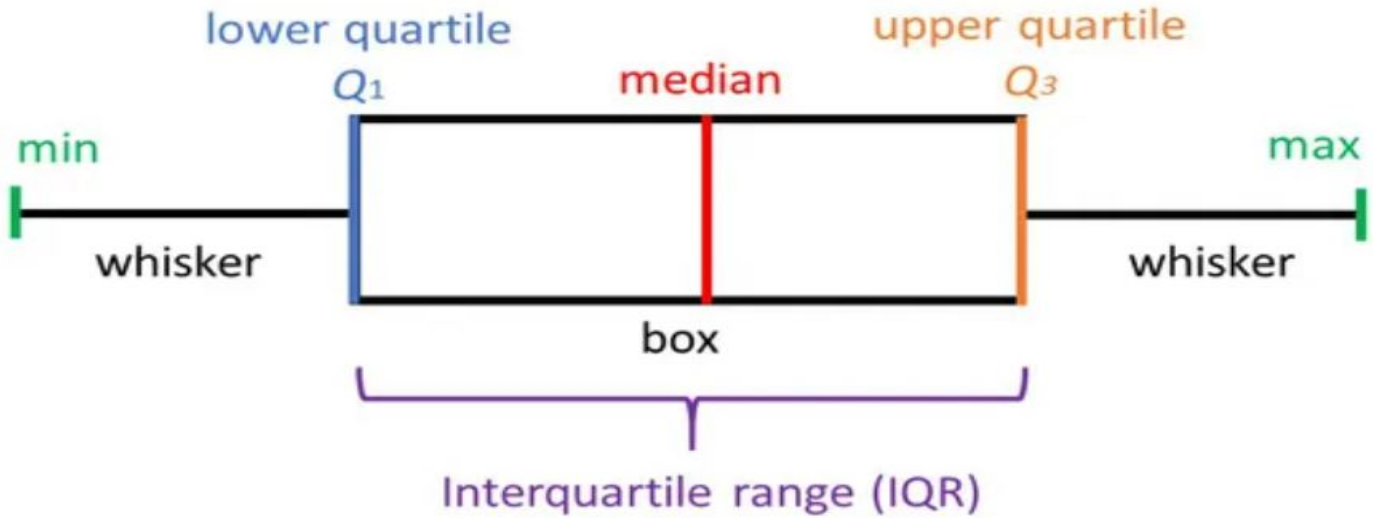
- 히스토그램을 옆으로 돌려놓은 것과 같은 모양
- 세로 열을 줄기(stem), 오른쪽으로 나열된 가로 행을 잎(leaf)
- 자료의 값을 그대로 유지하고 있다는 장점
- 방대한 자료의 경우에는 그리기 어려움

1단계	2단계	3단계
18	18 | 3, 0, 0, 3, 5, 3	18 | 0, 0, 3, 3, 3, 5
17	17 | 6, 0, 0, 0, 0, 3, 3, 8, 1, 8, 0, 4	17 | 0, 0, 0, 0, 0, 1, 3, 3, 4, 6, 8, 8
16	16 | 8, 5, 8, 0, 0, 5, 8, 5, 3, 3, 0, 3, 8, 5, 5, 8	16 | 0, 0, 0, 3, 3, 3, 5, 5, 5, 5, 5, 8, 8, 8, 8, 8
15	15 | 8, 0, 4, 2	15 | 0, 2, 4, 8
14	14 | 5	14 | 5
13	13 | 8	13 | 8

그림 2-15 [예제 2-6]의 줄기-잎 그림을 그리는 과정

상자그림(boxplot)

- 다섯숫자요약(five number summary)을 그래프로 표현한 것
  - 최솟값, Q_1(first quartile), 중위수, Q_3(third quartile), 최댓값
- 상자(box): Q_1, 중위수, Q_3를 사용하여 만든다.
- 상자의 길이: 사분위범위(IQR)
- 상자로부터 최솟값, 최댓값까지 선을 그어 수염(whiskers)을 만든다.

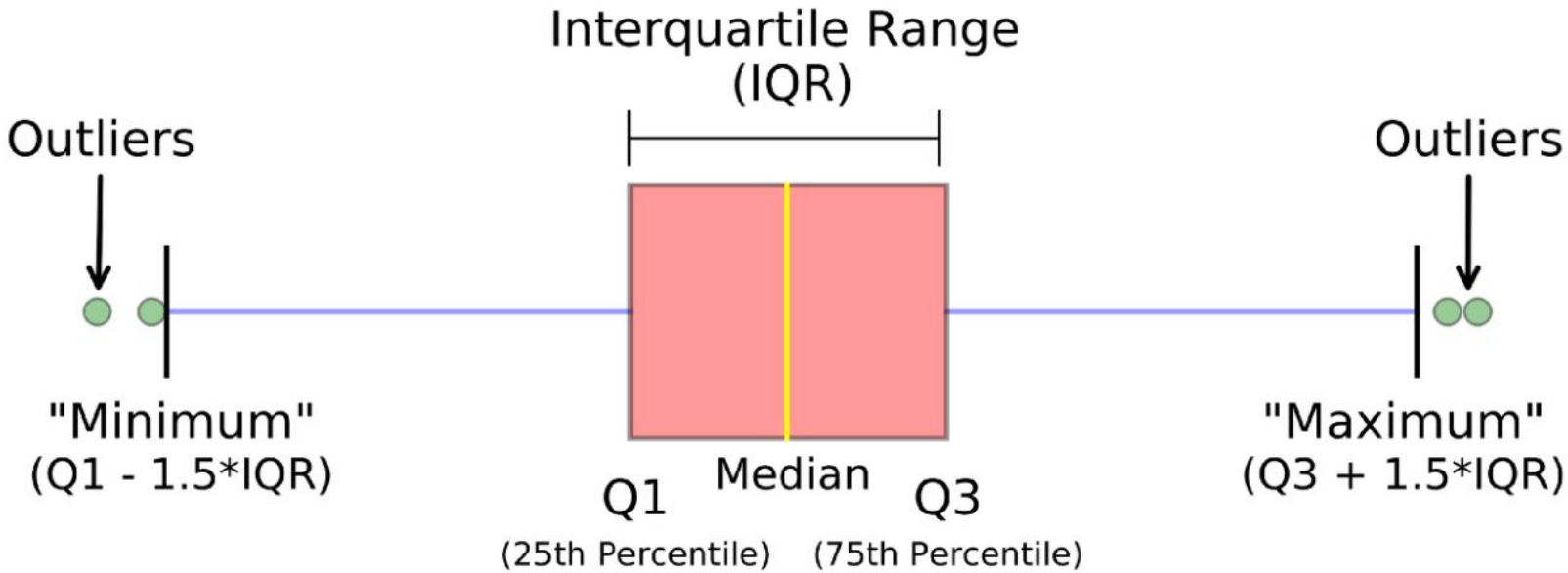


- 이상치(outlier)가 있으면 수염이 과도하게 길어진다.

상자그림(boxplot)_투키(John W. Tukey)

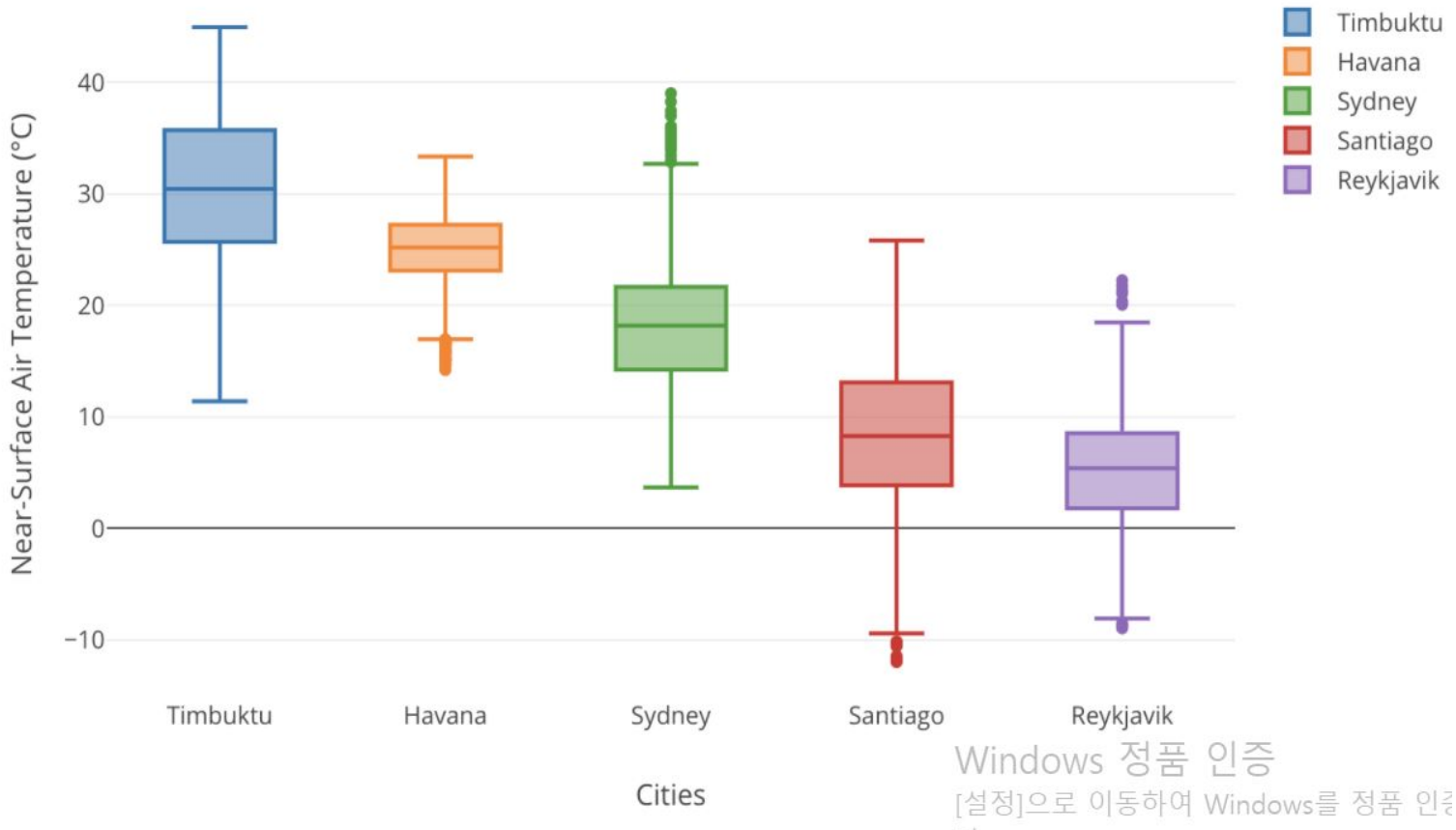
- 투키가 제안한 방법:

- 상자(box)를 만든다.
- 박스의 각 모서리 (Q1, Q3)로 부터 IQR의 1.5배 내에 있는 가장 멀리 떨어진 데이터 점까지 수염을 그린다.
- 수염보다 바깥쪽에 데이터가 존재한다면 이상치(outlier)로 간주 (*등의 기호로 표시)



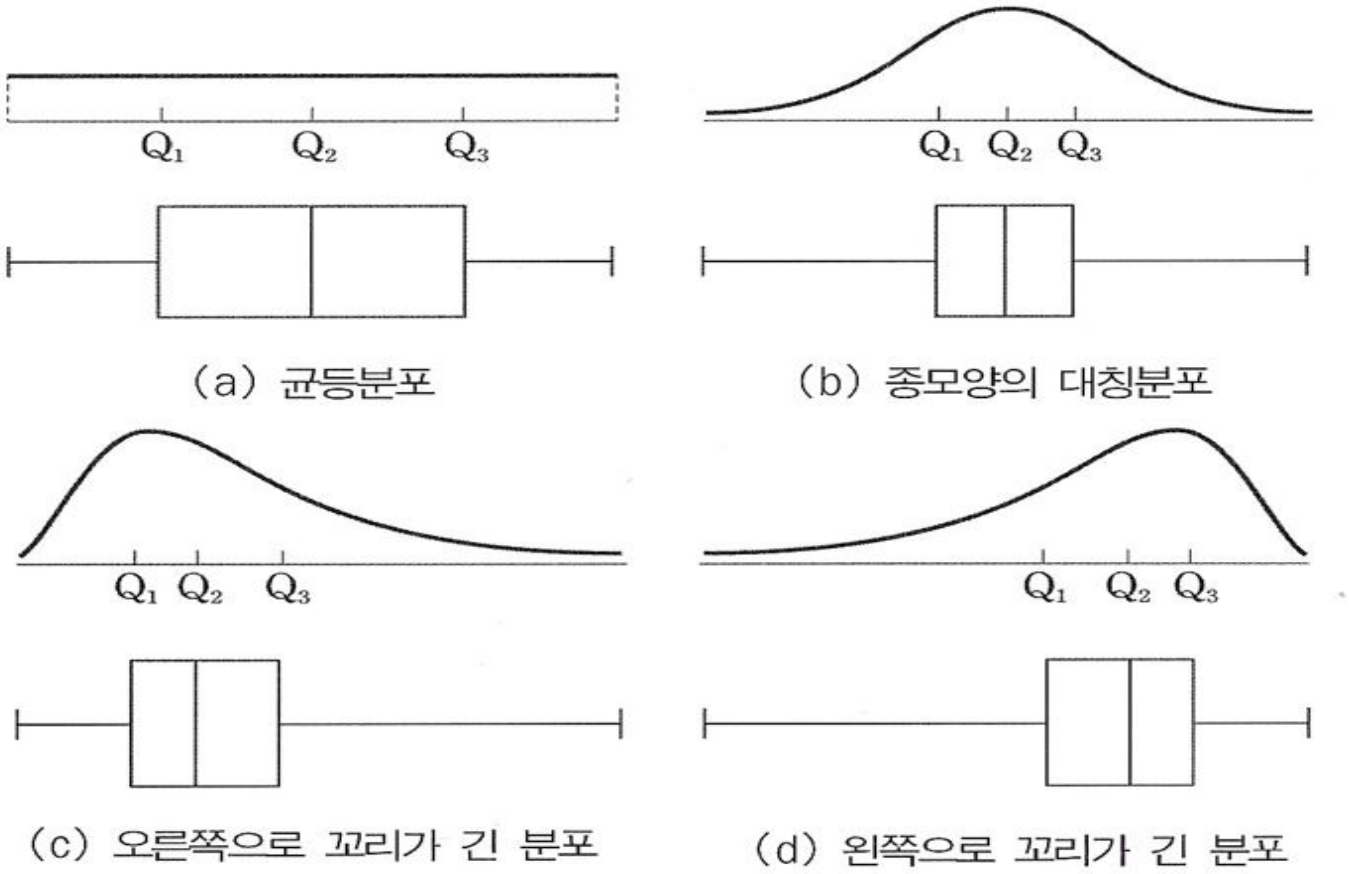
상자그림(boxplot)

- 자료를 개괄적으로 알아보기 위함
- 여러 개의 분포를 한눈에 비교할 때 유용



상자그림(boxplot)

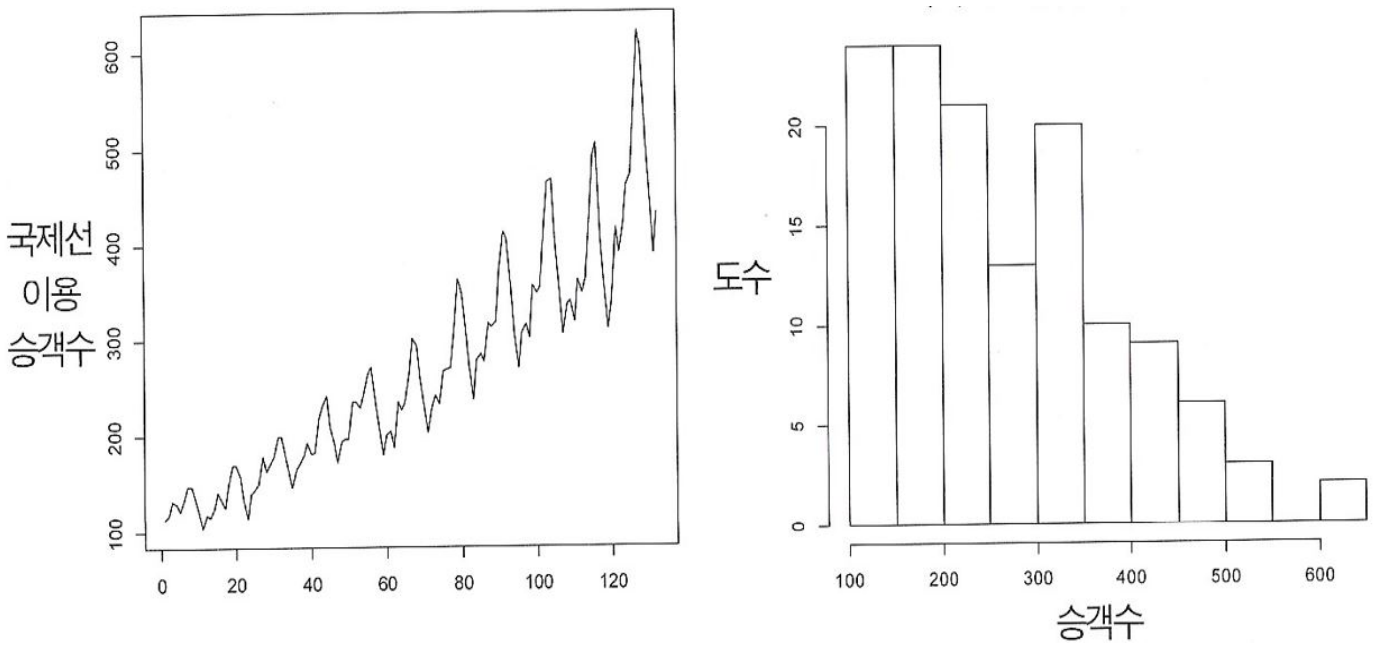
- 분포에 따른 상자그림의 형태를 나타낸 것



시계열그림(time series plot)

- 시간의 변화에 따라 얻게 되는 시계열 자료를 그래프로 표현
- X축: 시간, Y축: 해당자료
- 추세(trend)와 계절변동(seasonal variation) 등을 찾을 수 있음

Figure: 12년간의 월별 국제선 항공 승객수의 시계열그림과 히스토그램



이차원 자료의 도표화

1. 분할표(contingency table)

- 이차원 표본자료를 하나의 표로 요약한 것
- 각 특성의 값과 함께 각 이차원 특성값의 상대도수(또는 도수)를 이차원 표에 나열 (일차원인 경우의 도수분포표표)

예) 다음의 표는 400명의 학생에 대하여 국어, 영어, 수학 세 과목의 선호도를 조사한 결과를 분할표로 나타낸 것이다.

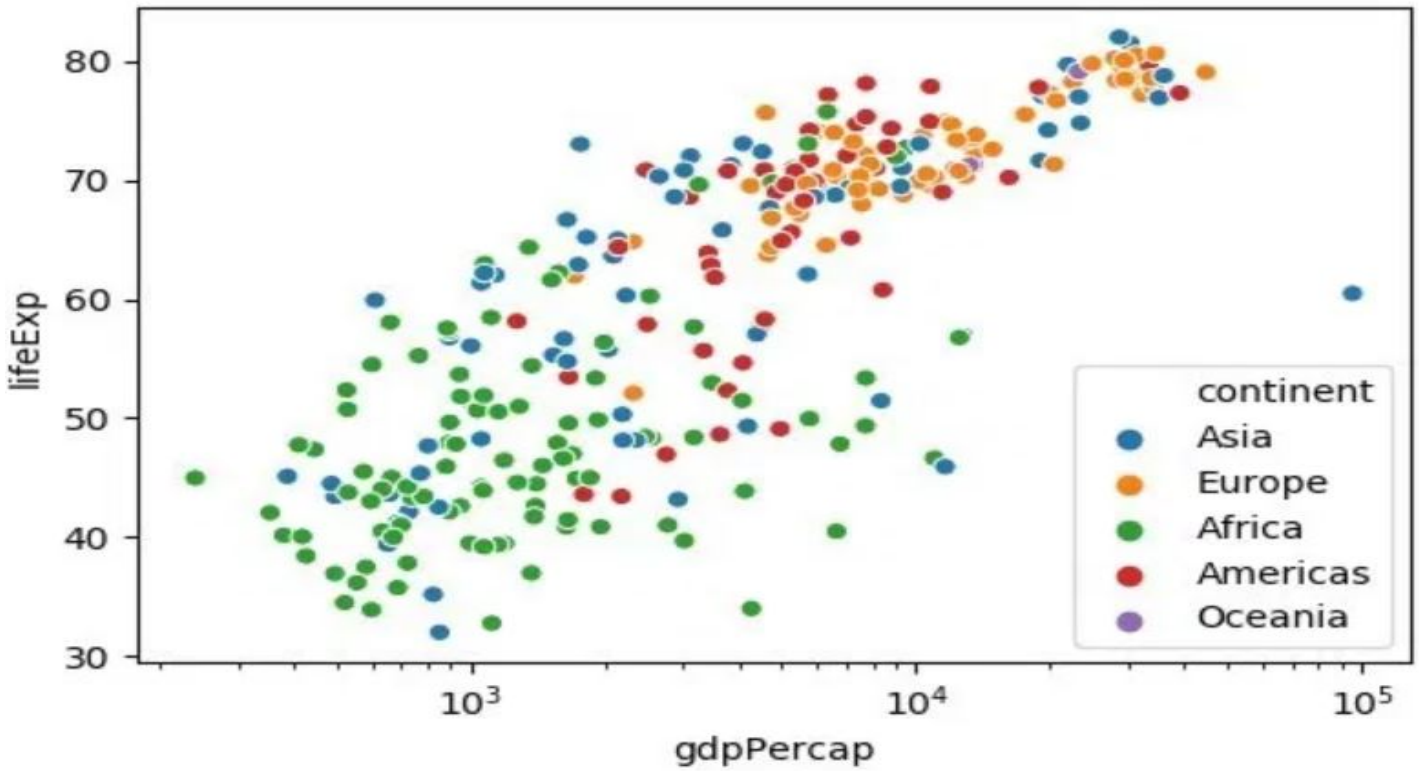
표 2-7 이차원 표본자료의 분할표

	국어	영어	수학	합계
남자	0.18	0.24	0.18	0.60
여자	0.14	0.16	0.10	0.40
합계	0.32	0.40	0.28	1.00

- 수치형 자료도 범위를 몇 개의 계급으로 나누어 분할표 작성 가능

이차원 자료의 도표화

2. 산점도(scatter plot): 각 이차원 자료값의 (x축:특성1의 값, y축:특성2의 값)의 점을 좌표평면 위에 찍은 것.
   - 두 변수의 연관성과 변화 관계를 쉽게 알아볼 수 있음



이차원 자료

Example

예제 2.7) 다음의 첫번째 값은 키(cm)이고, 두번째 값은 몸무게(kg)이다. 이 표본자료를 이용하여 분할표와 이차원 히스토그램, 산점도를 작성하여라.

(183,82) (168,52) (165,52) (158,55) (180,93)
(176,57) (168,60) (160,48) (160,50) (150,49)
(180,79) (183,73) (170,60) (170,66) (185,74)
(170,82) (165,57) (170,54) (168,57) (154,55)
(152,50) (173,63) (173,60) (165,49) (163,63)
(163,57) (160,57) (178,77) (138,54) (171,69)
(163,50) (168,57) (165,54) (165,59) (178,79)
(170,54) (168,79) (183,70) (145,37) (174,65)

sol) 키의 계급의 갯수는 6개이고(예제 2.6), 몸무게도 같은 방법으로 구해보면

- 최대값=93, 최소값=37, 범위=56이고, 계급의 개수를 6으로 하면 $56/6 = 9.33$이므로 폭은 10이다.

Figure: 예제 2.7의 분할표

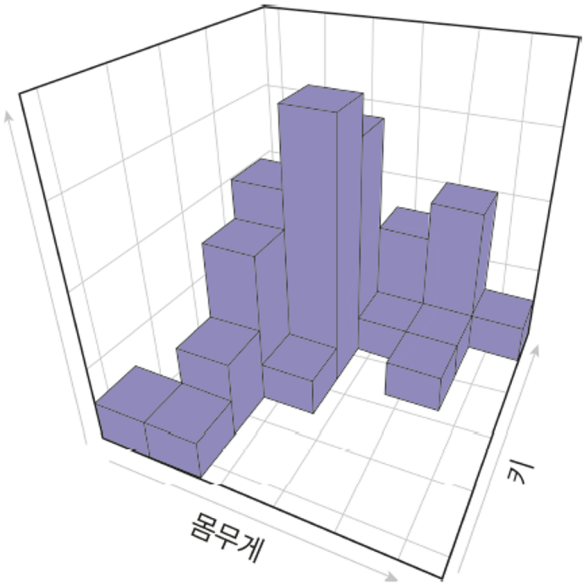
표 2-8 [예제 2-7]의 분할표

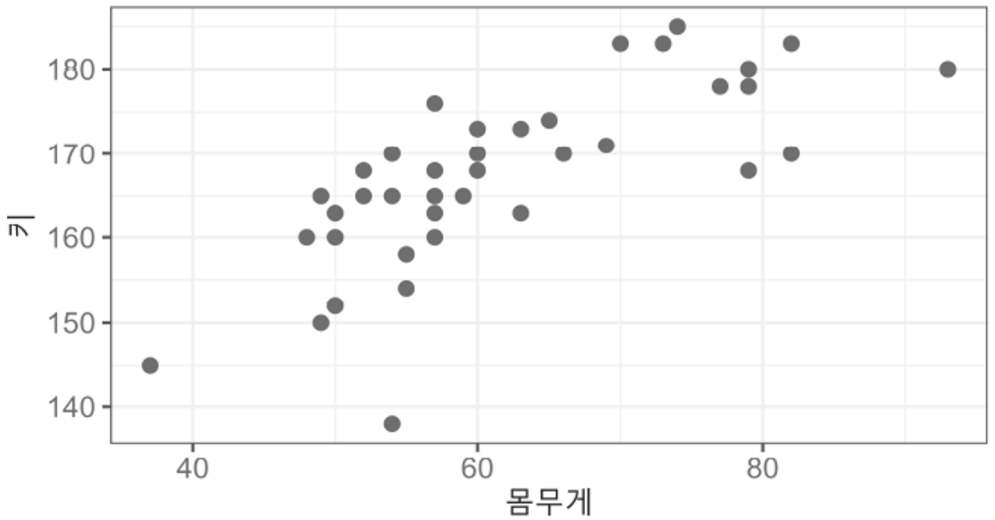
키 \ 몸무게	36.5 이상~ 46.5 미만	46.5 이상~ 56.5 미만	56.5 이상~ 66.5 미만	66.5 이상~ 76.5 미만	76.5 이상~ 86.5 미만	86.5 이상~ 96.5 미만	합계
137.5 이상~145.5 미만	0.025	0.025	0.000	0.000	0.000	0.000	0.050
145.5 이상~153.5 미만	0.000	0.050	0.000	0.000	0.000	0.000	0.050
153.5 이상~161.5 미만	0.000	0.100	0.025	0.000	0.000	0.000	0.125
161.5 이상~169.5 미만	0.000	0.125	0.175	0.000	0.025	0.000	0.325
169.5 이상~177.5 미만	0.000	0.050	0.150	0.025	0.025	0.000	0.025
177.5 이상~185.5 미만	0.000	0.000	0.000	0.075	0.100	0.025	0.200
합계	0.025	0.350	0.350	0.100	0.150	0.025	1.000

이차원 자료의 도표화

3. 히스토그램이차원 계급을 평면 위에 표시하고 그 위에 각 이차원 계급의 상대도수를 부피로 갖는 직육면체를 그린 것.

- 예제 2.7의 이차원 히스토그램과 산점도





수치에 의한 기술통계
(numerical description)

용어정리

모집단

- 모수(parameter): 모집단의 특징을 나타내는 대표값
- 모집단의 대표값: 모평균, 모분산, 모표준편차 등

예) 2021년도 대학교 신입생들의 한달 용돈에 관심. 모평균$= \mu$

표본

- 표본의 대표값: 표본평균, 표본분산, 표본표준편차 등

예) 전체 대학교 신입생 중 100명을 랜덤하게 추출하여 한달 용돈의 평균값을 조사. 표본평균$= \bar{x}$

- 이러한 표본의 대표값들은 모수를 추론하는 데 사용됨. $(\bar{x} = \hat{\mu})$

중심위치 척도

- 중심위치 척도: 자료들이 대략 어떠한 값을 갖는지를 알아보는 것 어느 위치를 중심으로 자료들이 모여 있는지를 나타내는 척도
- 위치를 나타내는 대표값:
  1. 평균
  2. 절사평균
  3. 중앙값
  4. 최빈값
  5. 백분위수
  6. 사분위수
- 모집단 $\{c_1, c_2, \cdots, c_N\}$에서 c_i는 i번째 추출단위의 특성값을 의미
- 표본의 특성값: $\{x_1, x_2, \cdots, x_n\}$

중심위치 척도

1. 평균(mean): 자료를 모두 더하여 자료의 수만큼으로 나누어준 값. 산술평균을 의미

- 모평균(population mean):

$$\mu = \frac{1}{N}\sum_{i=1}^{N} c_i$$

- 특성값 중에서 서로 다른 것들을 $c_1^*, c_2^*, \cdots, c_k^*$, 각각의 도수를 $f_1, f_2, \cdots, f_k$라고 하면,

$$\mu = \sum_{i=1}^{k} c_i^* \frac{f_i}{N}$$

- 표본평균(sample mean): μ을 추측하는 데 사용됨: $\bar{x} = \hat{\mu}$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

중심위치 척도

Example (2.1)

다음은 2020년 기준 우리나라 15세 이상 기혼 여성 16,549천 명의 출생아 수를 조사한 결과이다. 이 유한모집단의 모평균을 구하라.

특성값(출생아 수)	0	1	2	3	4	5*
상대도수	0.083	0.189	0.477	0.150	0.054	0.047

단, 표의 5*는 출생아 수가 5명 초과이면, 5명으로 간주했음을 의미한다.

sol) $\mu = 0 \times .083 + 1 \times .189 + \cdots + 5 \times .047 = 2.044$

- 평균은 물리적으로 무게중심의 의미를 가짐.

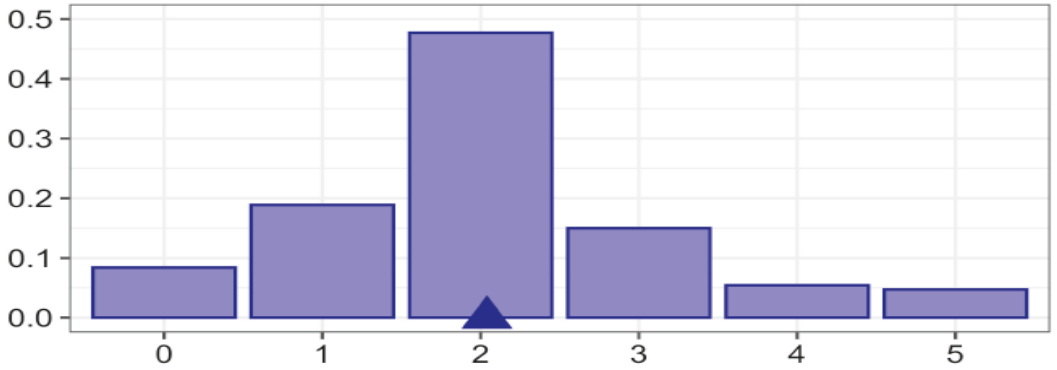


그림 2–7 균형점으로서의 평균(유한모집단)

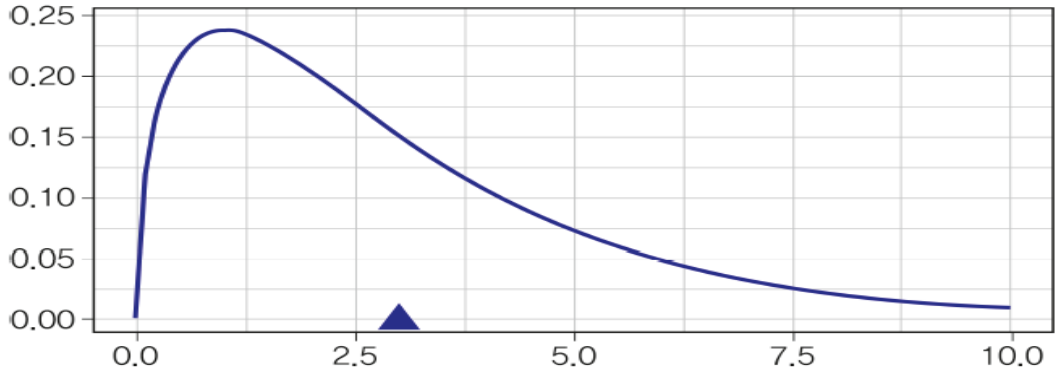


그림 2–8 균형점으로서의 평균(무한모집단)

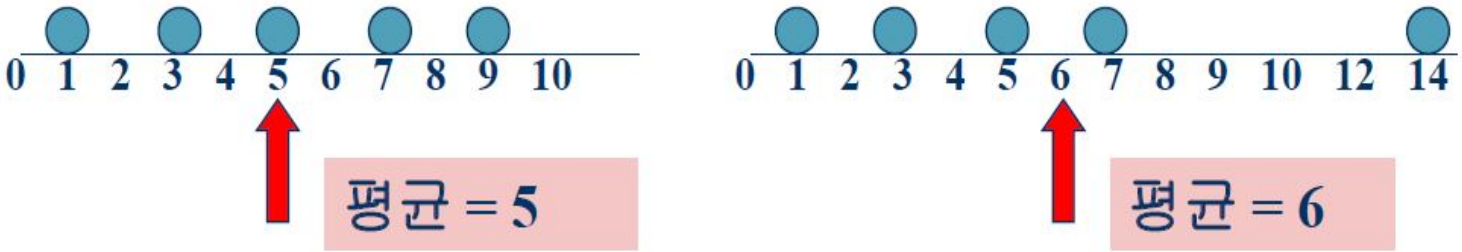
- 막대 위의 각 특성값 위치에 상대도수에 비례하는 질량을 놓았을 때의 균형점 위치
- 즉, 분포의 균형점으로서의 중심위치를 나타낸다.

중심위치 척도

- 자료에 특이하게 작거나 큰 값, 즉 특잇값(outlier)이 있으면 표본평균은 이들 값에 크게 영향을 받음

  예) 1, 3, 5, 7, 9 -> 평균=25/5=5

  1, 3, 5, 7, 14 -> 평균=30/5=6



중심위치 척도

2. $100p\%$ 절사평균(trimmed mean):

   순서대로 나열된 표본의 특성값 중에서 양쪽 $100p\%$를 버린 후 가운데 $100(1-2p)\%$ 특성값의 평균으로 정의.

$$\bar{x}_p$$

- 특이하게 작거나 큰 값이 자료에 있어도 영향을 받지 않게 됨

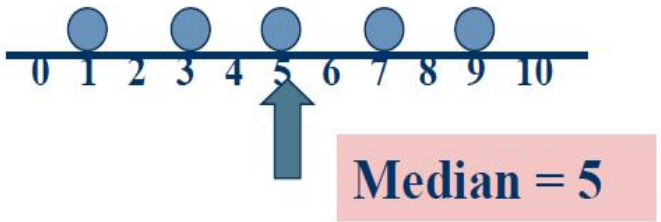
중심위치 척도

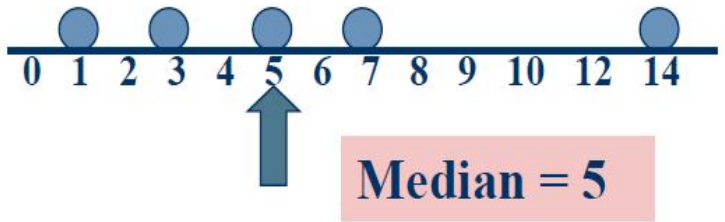
3. 중앙값(median): 특성값을 크기 순으로 늘어 놓았을 때, 가운데에 해당하는 값. 연속적인 경우에는 밀도곡선의 전체 넓이를 이등분하는 점

- 자료가 홀수인 경우: 가운데 값, $\frac{n+1}{2}$번째 자료
- 자료가 짝수인 경우: 가운데 두 개($\frac{n}{2}$, $\frac{n}{2}+1$번째) 자료의 평균
- 특이하게 작거나 큰 값이 자료에 있어도 영향을 받지 않게 됨

예) 1, 3, 5, 7, 9 -> 중앙값= 5

1, 3, 5, 7, 14 -> 중앙값= 5





중심위치 척도

4. 최빈값(mode): 자료 중 그 빈도수가 최대인 값

- 명목자료인 경우 평균과 중앙값은 의미가 없으므로 최빈값을 사용
- 자료가 적은 경우 최빈값은 무의미
- 최빈값은 여러 개가 나올 수 있음

예) 올해의 유행 옷 색깔

예) 한 의류매장에서 어느날 판매된 남성복 바지들의 허리Size는 다음과 같다. 최빈값은34이고 중앙값은 33.5이다. 이 경우 자료의 중심위치가 33.5 Inch라고 하는 것보다 34 Inch라고 하는 것이 더 상식적이다.

(31, 34, 36, 33, 28, 34, 30, 34, 32, 40)

중심위치 척도

5. 제p백분위수(pth percentile):

   특성값을 작은 것부터 순서대로 나열하였을 때 p%의 특성값이 그 값보다 작고, 또한 $(100-p)$%의 특성값이 그 값보다 크게 되는 값.

   - 연속형 자료의 경우
   - 이산형 자료의 경우

$$(n+1)\frac{p}{100}, \quad or \quad 1+(n-1)\frac{p}{100}$$

중심위치 척도

6. 사분위수:

- 자료를 크기 순으로 나열한 후 똑같은 크기의 네 덩어리로 만들 때 그 경계에 해당하는 값
- 자료의 $\frac{1}{4}$, $\frac{2}{4}$, $\frac{3}{4}$, 혹은 25%,50%, 75%에 해당하는 값

- 제1사분위수(first quartile): 제25백분위수, Q_1

$$Q_1 = (n+1)\frac{25}{100}\text{번째 순위 값}$$

- 중앙값(median): 제50백분위수, Q_2
- 제3사분위수(third quartile): 제75백분위수, Q_3

$$Q_3 = (n+1)\frac{75}{100}\text{번째 순위 값}$$

중심위치 척도

- 예) 다음과 같은 자료를 얻었다고하자.

  1, 0, 7, 5, 3, 2, 0, 1, 8, 4

  선형보간법을 이용하여 4분위수를 구하여라.

sol) 자료를 크기순으로 나열 -> 0,0,1,1,2,3,4,5,7,8

Q_1 : (10 + 1) × 25 / 100 = 2.75 번째에해당

따라서 0 + (1 –0) × 0.75 = 0.75.

Q_2 : (10 + 1) × 50 / 100 = 5.5 번째에해당

따라서 2 + (3 –2) × 0.5 = 2.5.

Q_3 : (10 + 1) × 75 / 100 = 8.25 번째에해당

따라서5 + (7 –5) × 0.25 = 5.5.

중심위치 척도 예제

Example

예제 2.6) 다음은 어떤 집단에서 40명을 뽑아 키를 측정한 결과이다. 이 표본자료를 이용하여 평균, 사분위수, 10% 절사평균을 구하여라.

183 168 165 158 180 176 168 160 160 150
180 183 170 170 185 170 165 170 168 154
152 173 173 165 163 163 160 178 138 171
163 168 165 165 178 170 168 183 145 174

sol) $\bar{x} = \frac{1}{40}(183 + 168 + \cdots + 174) = 167.45$

$Q_1 = (40+1) \times \frac{25}{100}$ 번째 값=10.25번째 값=163

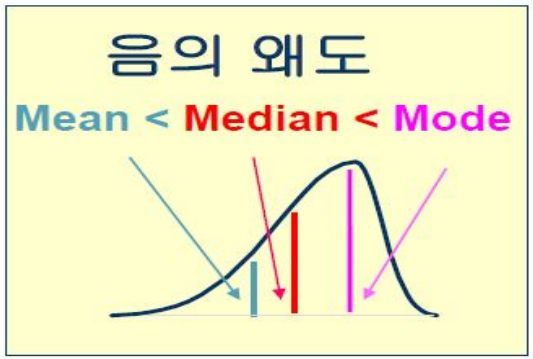
$Q_2 = (40+1) \times \frac{50}{100}$ 번째 값=20.5번째 값=168

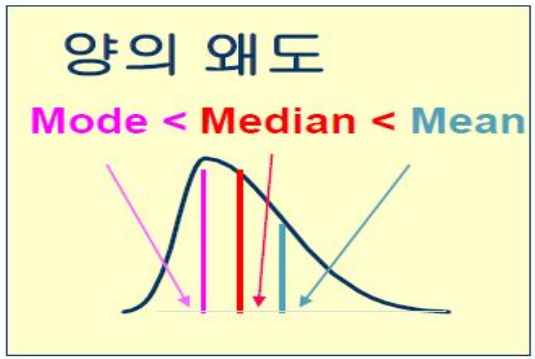
$Q_3 = (40+1) \times \frac{75}{100}$ 번째 값=30.75번째 값=173.75

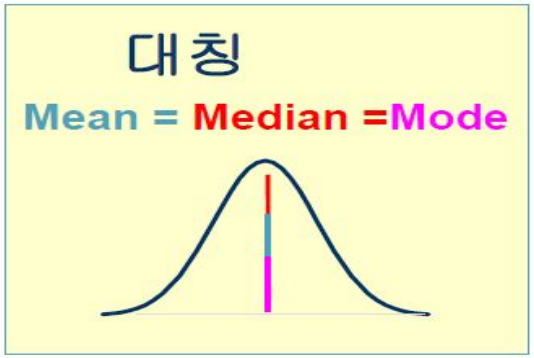
$\bar{x}_{0.1} = \frac{1}{32}(180 + 180 + \cdots + 154) = 168.09$

(양끝 4개의 값을 버린 후 가운데32개 값의 평균)

중심위치 척도의 비교







- 대칭: mean=median=mode
- 양의 왜도: mode< median<mean
- 음의 왜도: mean< median< mode

변동성 척도

- 변동성 척도: 자료들이 얼마나 변동하거나 퍼져있는지를 표시.
- 분포의 산포(특성값이 흩어져 있는 상태)를 나타내는 대표값:
  1. 평균절대편차
  2. 범위
  3. 사분위수범위
  4. 분산, 표준편차

1. 평균절대편차(mean absolute deviation(MAD)):
   각 특성값이 중앙값으로부터 떨어진 평균거리를 나타냄

$$MAD = \frac{1}{n}\sum_{i=1}^{n}|x_i - Q_2|$$

변동성 척도

2. 범위 (Range) : 최댓값과 최솟값의 거리
   - 쉽고 빠르게 구할 수 있음
   - 특이하게 크거나 작은 값이 있을 경우 자료의 범위가 왜곡됨
   - 자료의 개수와 상관없이 같게 나올 수 있음

   ⇒ 자료의 변동성을 대표하지 못하는 경우가 많음

3. 사분위수범위(interquartile range(IQR)):
   제3사분위수와 제1사분위수의 차로 정의

$$IQR = Q_3 - Q_1$$

   - 가운데 50%특성값의 범위를 나타냄
   - 양쪽 극단 값에서 자료의 25%씩 안쪽으로 들어와 있는 값의 거리 므로 특잇값의 영향을 거의 받지 않음

변동성 척도

4. 분산(variance), 표준편차(standard deviation):

- 분산: 자료 각각이 그 평균으로부터 떨어져 있는 거리를 자승한 것의 평균값.
- 표준편차: 항목들의 자승효과를 없애기 위해 분산에 자승근을 한 것
- 분포의 산포를 나타내는 대표값.

- 모분산과 모표준편차: σ^2, σ 로 표시

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(c_i - \mu)^2, \quad \sigma^2 = \sum_{i=1}^{k}(c_i^* - \mu)^2 \frac{f_i}{N}$$

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(c_i - \mu)^2}$$

변동성 척도

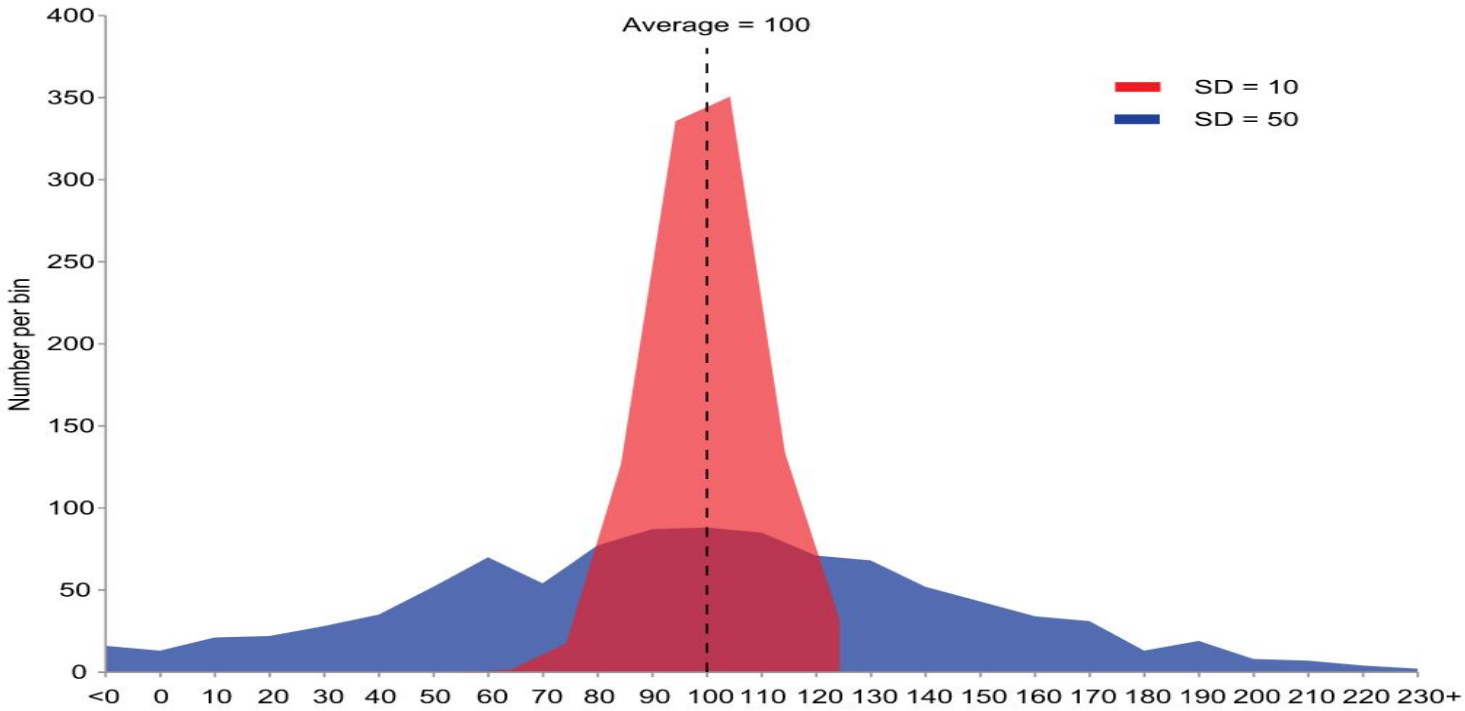
- 표본분산(sample variance)과 표본표준편차(sample standard deviation): s^2, s로 표시

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2, \qquad s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- 자료 하나하나의 값이 전부 고려되어 구해진 변동성 척도
- 값이 클수록 변동성이 큼, 넓게퍼짐
- 음수가 될 수 없음

변동성 척도

Figure: 평균이 같은 두 모집단의 분산 비교



- The red population has mean 100 and variance 100 (SD=10)
- while the blue population has mean 100 and variance 2500 (SD=50).

간편 계산식

- 간편 계산식을 통해서 다음과 같이 계산할 수 있다.

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum x_i^2 - n\bar{x}^2$$

$$pf) \quad \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum x_i^2 - 2\bar{x}\sum x_i + n\bar{x}^2$$

$$= \sum x_i^2 - 2\bar{x}^2 n + n\bar{x}^2 = \sum x_i^2 - n\bar{x}^2$$

- 따라서 표본분산을 다시 표현하면 다음과 같다.

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)$$

변동성 척도

Example (2.3)

example 2.2의 모집단에 대하여 모분산과 모표준편차를 구하여라.

sol) $\sigma^2 = (0-1.9)^2 \times 1/10 + (1-1.9)^2 \times 3/10 + \cdots + (4-1.9)^2 \times 1/10 = 1.29$

$\sigma = \sqrt{1.29} \simeq 1.14$

Example

예제 2.6의 표본자료를 이용하여 표본사분위수범위와 표본표준편차를 구하여라.

sol) $Q_1 = 163,\ Q_3 = 173.75$이므로 $IQR = 173.75 - 163 = 10.75$

$s^2 = \frac{1}{40-1}\{(183 - 167.45)^2 + \cdots + (174 - 167.45)^2\} = 108.61$

$s = 10.42$

연관성 척도

- 연관성(association): 이차원 또는 그 이상의 항목에 대한 자료들이 상호 관련되어 있는 성질
- 이차원 특성값의 분포에서 선형관계(linear relationship)의 연관성을 나타내는 대표값들:
  1. 상관계수
  2. 공분산
- (특성1, 특성2)의 이차원 모집단 $(c_{11}, c_{21}), (c_{12}, c_{22}), \cdots, (c_{1N}, c_{2N})$가 있을 때, 특성1과 2의 모평균, 표준편차를 $\mu_1, \sigma_1, \mu_2, \sigma_2$라고 하자.
- 이차원 표본자료 $(x_1, y_1), \cdots, (x_n, y_n)$가 있을때, 특성1과 2의 표본평균, 표본표준편차를 $\bar{x}$, s_1, $\bar{y}$, s_2라 하자.

두 이차원 모집단의 결합분포

모집단 A				
	특성1			
특성2	0	1	2	합
0	.1	.2	.1	.4
1	.0	.2	.1	.3
2	.0	.0	.3	.3
합	.1	.4	.5	1.0

모집단 B				
	특성1			
특성2	0	1	2	합
0	.0	.0	.4	.4
1	.0	.2	.1	.3
2	.1	.2	.0	.3
합	.1	.4	.5	1.0

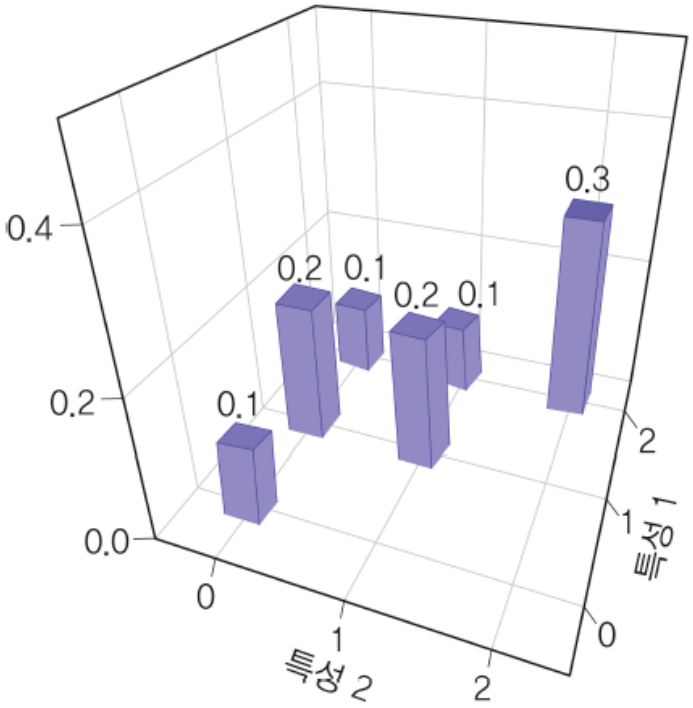
- 두 모집단 A, B의 특성별 모평균과 모표준편차는 같다.

모집단 A:

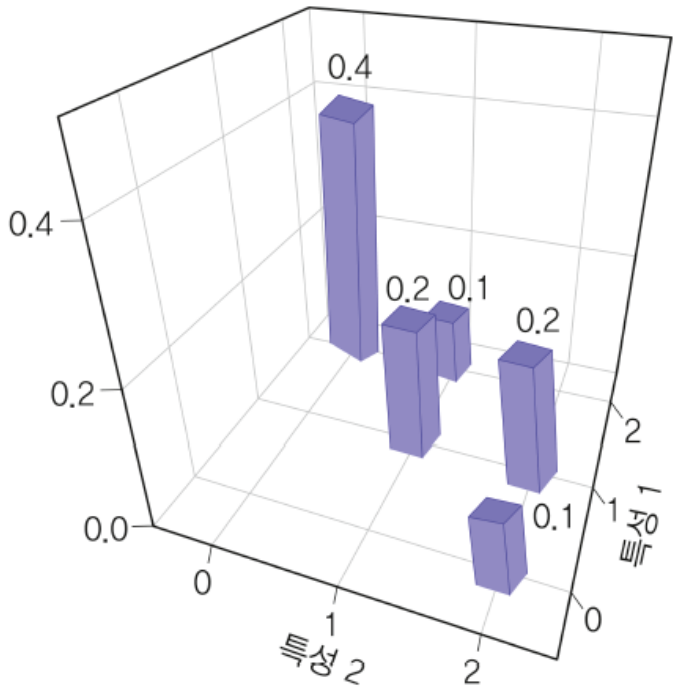
- $\mu_1 = 1.4$, $\sigma_1 = .6633$
- $\mu_2 = .9$, $\sigma_2 = .8307$

모집단 B:

- $\mu_1 = 1.4$, $\sigma_1 = .6633$
- $\mu_2 = .9$, $\sigma_2 = .8307$



(a) 모집단 A



(b) 모집단 B

그림 2-10 두 이차원 모집단의 결합분포

- 그러나 모집단 A의 분포는 좌표평면에서 양의 방향, 모집단 B는 음의 방향으로 놓여있다.
- 두 특성의 변화 관계를 나타내는 추가적인 대표값이 필요.

연관성 척도

1. 공분산(covariance): 두 변수 사이의 선형관계의 연관성을 측정
   - 두 변수가 각기 평균으로부터 떨어진 값을 서로 곱한 후, 더하여 표본에서는 $n-1$, 모집단에서는 N으로 나눈 값.

- 모집단 공분산(population covariance):

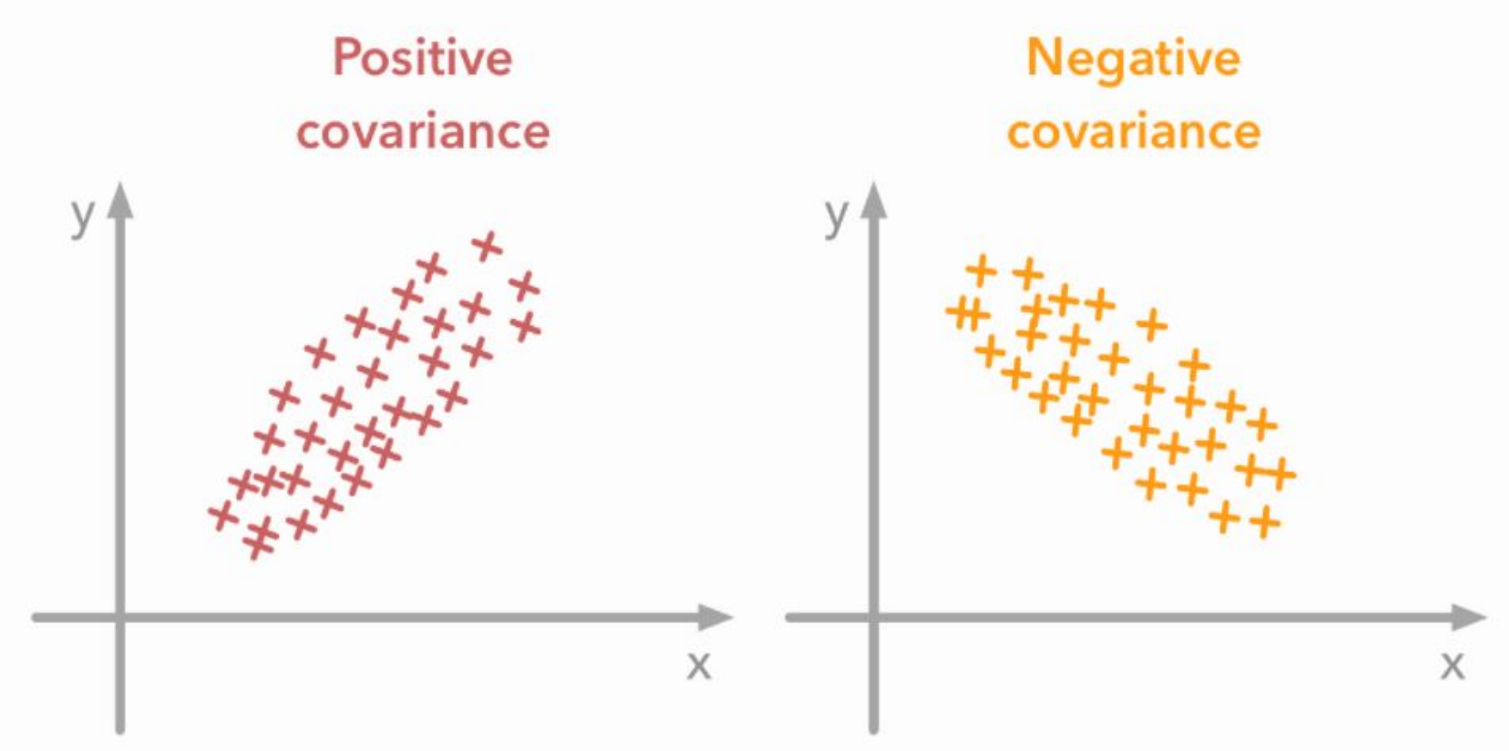
$$\rho\sigma_1\sigma_2 = \frac{1}{N}\sum_{i=1}^{N}(c_{1i}-\mu_1)(c_{2i}-\mu_2)$$

- 표본공분산(sample covariance):

$$rs_1s_2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

연관성 척도

- 공분산의 부호
  - 양의 관계(Positive Relationship)
    : x와 y가 평균에 대하여 서로 같은 값을 가질 때
  - 음의 관계(Negative Relationship)
    : x와 y가 평균에 대하여 서로 반대 값을 가질 때



연관성 척도

- 공분산의 성질: (rs_1s_2에서도 동일하게 성립)

1. $0 < \rho\sigma_1\sigma_2$: x와 y는 양의 선형관계를 의미한다.
2. $\rho\sigma_1\sigma_2 < 0$: x와 y는 음의 선형관계를 의미한다.
3. $\rho\sigma_1\sigma_2 = 0$ x와 y는 선형 관계가 존재하지 않는다.

- 각 변수간의 다른 척도 기준이 그대로 값에 반영되어 공분산 값의 크기에 대한 기준이 없음

예) (키,몸무게)

연관성 척도

2. 상관계수(correlation coefficient):
   공분산을 각자의 표준편차로 나눈 값, 공분산을 표준화 한 것.

- 모집단의 모상관계수: ρ로 표시.

$$\rho = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{c_{1i}-\mu_1}{\sigma_1}\right)\left(\frac{c_{2i}-\mu_2}{\sigma_2}\right)$$

- 표본상관계수(sample correlation coefficient): r로 표시

$$\begin{aligned} r &= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_1}\right)\left(\frac{y_i-\bar{y}}{s_2}\right) \\ &= \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \end{aligned}$$

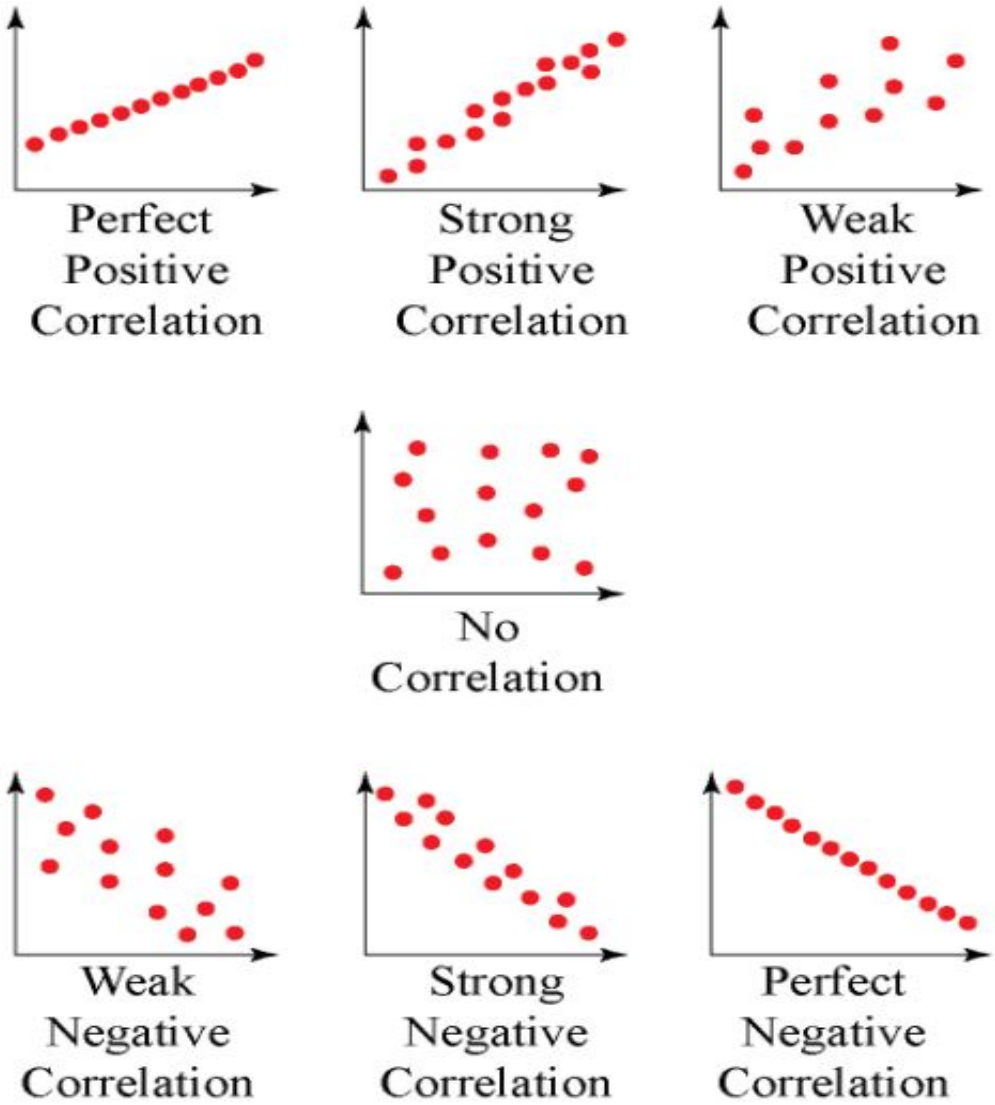
연관성 척도

- 두 변수가 어느 방향으로 얼마나 변동하는지 또는 퍼져있는지를 나타냄(부호(sign)와 크기(magnitude))

- 상관계수의 성질:

1. 항상 $-1 \leq \rho \leq 1$ (CH3, pf)
2. $0 < \rho \leq 1$: x와 y는 양의 선형관계를 의미
3. $-1 \leq \rho < 0$: x와 y는 음의 선형관계를 의미
4. $\rho = 0$: 선형 관계가 존재하지 않음
5. 기울기가 한 직선 주위에 모집단의 분포가 집중될수록 상관계수의 값은 그 직선의 방향에 따라 양 극단값-1, +1에 가까워진다

연관성 척도


Perfect
Positive
Correlation
Strong
Positive
Correlation
Weak
Positive
Correlation
No
Correlation
Weak
Negative
Correlation
Strong
Negative
Correlation
Perfect
Negative
Correlation

간편 계산식

- $\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum x_i^2 - n\bar{x}^2$, $\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum y_i^2 - n\bar{y}^2$
- $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - n\bar{x}\bar{y}$

$$
\begin{aligned}
pf) \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) &= \sum (x_i - \bar{x}) y_i - \bar{y} \sum (x_i - \bar{x}) \\
&= \sum x_i y_i - \bar{x} \sum y_i \\
&= \sum x_i y_i - n\bar{x}\bar{y}
\end{aligned}
$$

- $$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_1} \right) \left(\frac{y_i - \bar{y}}{s_2} \right) = \frac{1}{n-1} \frac{\sum x_i y_i - n\bar{x}\bar{y}}{s_1 s_2}$$

연관성 척도 예제

Example

다음 표의 광고비와 매출액의 선형적 연관성을 측정하여라.

	광고비(x_i)	매출액(y_i)	$x_i - \overline{x}$	$y_i - \overline{y}$	$(x_i - \overline{x})(y_i - \overline{y})$
	5	16	-5	-9	45
	6	19	-4	-6	24
	7	18	-3	-7	21
	8	20	-2	-5	10
	9	24	-1	-1	1
	11	26	1	1	1
	12	30	2	5	10
	13	32	3	7	21
	14	31	4	6	24
	15	34	5	9	45
합	100	250	0	0	202

sol) $rs_1s_2 = \frac{202}{10-1} = 22.44,\ r = \frac{22.44}{(3.496)(6.532)} = 0.98$

연관성 척도 예제

Example

예제 2.7의 표본자료를 이용하여 표본상관계수를 구하여라.

sol) $\bar{x} = 167.45$, $s_1 = 10.42$, $\bar{y} = 61.45$, $s_2 = 11.94$이므로

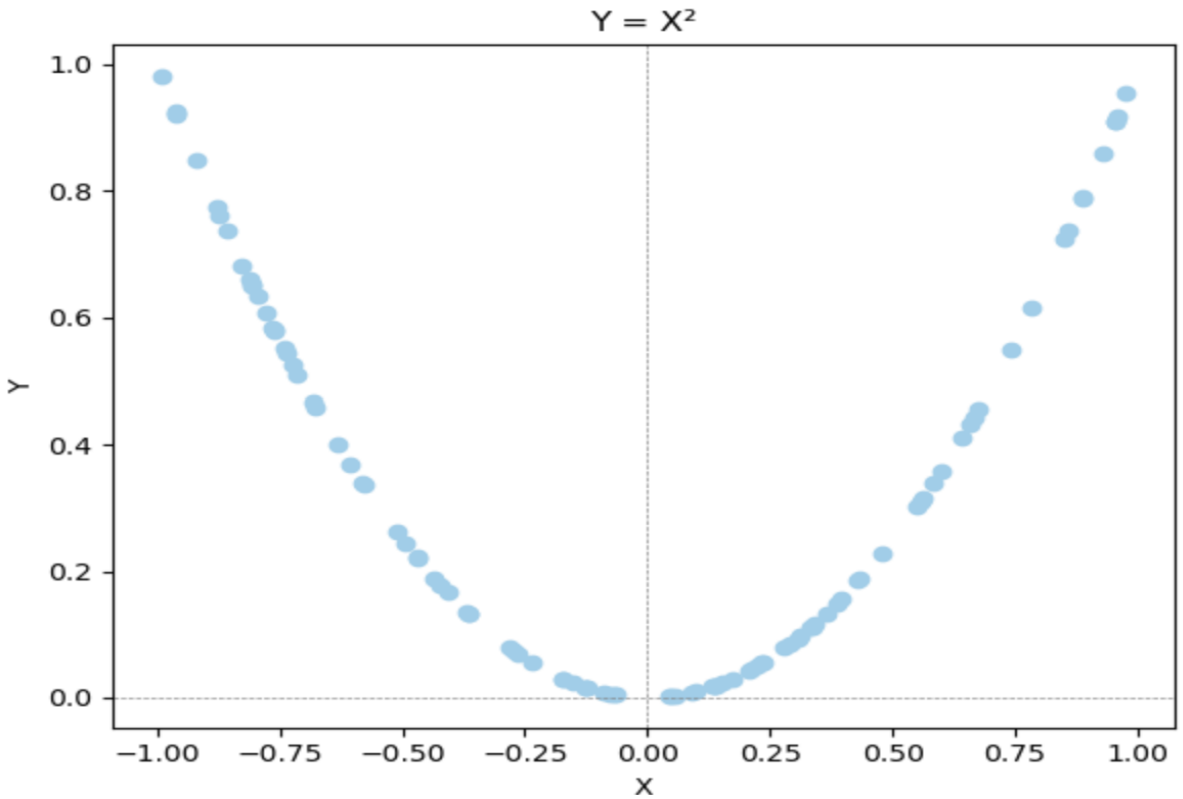
$$\begin{aligned} rs_1s_2 &= \frac{1}{40-1}\{(183-167.45)(82-61.45) \\ &\quad + \cdots + (174-167.45)(65-61.45)\} \\ &= 90.57 \end{aligned}$$

- 따라서 표본상관계수는 다음과 같다.

$$r = \frac{90.57}{10.42 \times 11.94} = 0.728$$

연관성 척도

Figure: 비선형 관계인 경우. $\rho = 0$



연관성 척도

Figure: $\rho = 0.9$ 값을 갖는 여러 경우