통계학
- 제 2 장 자료의 구조(2)

0011 0010 1010 1101 0001 0100 1011

12

수치에 의한 자료의 요약 – 중심위치의 척도

- 대표값 (measure of centrality)
  - 표본평균 (sample mean)
    - 대표값 중에서 실제적으로 가장 많이 사용됨
    - 통계적 추론 과정에서 활용
    - 일반적으로 산술평균(arithmetic mean)을 이용하여 계산
      - n개의 관측값을 각각 $x_1, x_2, \cdots, x_n$이라 하면

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

    - 산술평균은 예외적으로 크거나 작은 관측값에 영향을 많이 받음
      ⇒ 절사평균(trimmed mean)을 이용

- 중앙값 (median)
  - 관측값들을 크기 순으로 정렬하였을 때 중앙에 위치하는 값
    - 관측값의 개수(n)가 홀수일 때

$$\frac{(n+1)}{2} \text{ 번째 관측값}$$

    - 관측값의 개수(n)가 짝수일 때

$$\frac{n}{2} \text{ 번째 관측값과 } \frac{n}{2}+1 \text{ 번째 관측값의 평균}$$

  - 중앙값은 예외적으로 크거나 작은 관측값에 영향을 받지 않음

- 예 – [사무직 직원의 월급여액]
  - 사무직 직원 10명에 대한 월급여액을 조사(단위 : 만원)

    85 92 103 102 83 97 119 109 320 115

  - 월급여액의 표본평균
    - (85+92+103+102+83+97+119+320+115)/10 = 122.5 (만원)
  - 월급여액의 중앙값
    - 크기 순 정렬 : 83 85 92 97 102 103 109 115 119 320
    - 관측값의 개수(n) = 10
    - 중앙값 : 5번째 관측값과 6번째 관측값의 평균
    - (102+103)/2 = 102.5 (만원)
  - 어느 것을 중심위치를 나타내는 통계량으로 쓰는 것이 바람직한가?

- 최빈값 (mode)
  - 전체 관측값들 중에서 가장 빈도가 많은 관측값
  - 자료의 범위가 비교적 작고 서로 겹치는 관측값들이 많은 경우 유용
  - 주로 이산형 자료 및 범주형 자료의 대표값으로 활용
  - 두 개 이상의 최빈값이 존재할 경우 대표값으로의 의미 약화됨
- 예 – [통계학과 신입생의 키]
  - 전체 학생을 대상으로 한 경우
    - 159.5cm 이상 164.5cm 미만인 학생수 14명
    - 169.5cm 이상 174.5cm 미만인 학생수 12명
  - 최빈값으로는 자료의 중심을 단언하기 힘듦

수치에 의한 자료의 요약 – 퍼진 정도의 척도

- 산포도 (measure of dispersion)
  - 분산 (variance)
    - 관측값들이 표본평균을 중심으로 흩어진 정도
    - 편차(deviation)
      - n개의 관측값을 각각 $x_1, x_2, \cdots, x_n$이라 하고, 이들의 평균을 $\bar{x}$라 하면

$$\text{편차} = (x_i - \bar{x}) \qquad i = 1, \cdots, n$$

      - 각각의 관측값에 대한 편차의 합은 항상 0이 됨.
    - 표본분산
      - 편차의 제곱합

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

      - 관측값 측정단위의 제곱

- 표준편차 (standard deviation, S.D.)
  - 관측값들이 표본평균을 중심으로 흩어진 정도
  - 표본표준편차
    - 표본분산의 양의 제곱근

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

    - 관측값의 측정단위와 일치
- 변동계수(coefficient of variation, CV)
  - 표본변동계수
    - 표본평균을 중심으로 상대적인 산포 비율(%)
    - 측정 단위가 다르거나 중심위치가 매우 다른 경우 비교 가능

$$CV = \frac{s}{\bar{x}} \times 100 \ (\%)$$

- 예 – [통계학 시험 성적]
  - 통계학 수업을 듣는 A분반과 B분반 학생 각각 10명씩의 시험 성적

    A분반 : 50 60 65 70 75 75 80 85 90 100

    B분반 : 50 70 70 70 75 75 85 85 85 100

  - 두 분반의 평균 시험 성적
    - A분반 : 75 (점)
    - B분반 : 75 (점)
  - 두 분반 시험 성적의 표준편차
    - A분반 : 약 14.72 (점)
    - B분반 : 약 12.47 (점

- 예 – [A분반 학생의 통계학과 미적분학 시험 성적]
  - A분반 학생 10명의 통계학과 미적분학 시험 성적
    - 통계학 : 100점 만점, 미적분학 : 200점 만점

통 계 학	:	50	60	65	70	75	75	80	85	90	100
미적분학	:	115	120	125	125	130	135	150	175	185	200

  - 두 과목의 평균 시험 성적
    - 통계학 : 75 (점), 미적분학 : 146 (점)
  - 두 과목 시험 성적의 표준편차
    - 통계학 : 14.72, 미적분학 : 30.17
  - 두 과목 시험 성적의 변동계수(CV)
    - 통계학 : (14.72/75) x 100 = 19.63
    - 미적분학 : (30.17/146) x 100 = 20.66

- 사분위수(quartile)와 백분위수(percentile)
  - 사분위수
    - 관측값들을 크기 순으로 정렬하였을 때 4등분되는 지점의 값
  - 백분위수
    - 관측값들을 크기 순으로 정렬하였을 때 100등분되는 지점의 값
    - 사분위수와 백분위수의 관계

제1사분위수(Q_1) = 제25백분위수

제2사분위수(Q_2) = 제50백분위수 = 중앙값

제3사분위수(Q_3) = 제75백분위수

- 제 $100 \times p$ 백분위수를 구하는 방법 $(0 \le p \le 1)$
  - 관측값을 작은 순서로 배열한다.
  - 관측값의 개수(n)에 p 를 곱한다.
  - 만약 $n \times p$ 가 정수이면, 제 $100 \times p$ 백분위수는

    $n \times p$번째 작은 관측값과 $n \times p + 1$번째 작은 관측값의 평균

  - 만약 $n \times p$ 가 정수가 아니면, 제 $100 \times p$ 백분위수는

    ($n \times p$ 의 정수부분 + 1) 번째 작은 관측값

- 범위와 사분위범위
  - 범위 (range)
    - 관측값들을 크기 순으로 정렬하였을 때 양 끝점의 차이
    - 관측값들의 흩어진 총 길이

범위(range) = 최대값(Max) – 최소값(Min)

  - 사분위범위(interquartile range)
    - 관측값들을 크기 순으로 정렬하였을 때
      중앙값을 중심으로 전체 관측값들 중 50%가 흩어진 범위

$$IQR = Q_3 - Q_1$$

- 예 – [교통소음의 정도]
  - 특정 교차로에서 발생하는 교통소음을 측정 (단위 : dB)

55.9	63.8	57.2	59.8	65.7	62.7	60.8	51.3	61.8	56.0
66.9	56.8	66.2	64.6	59.5	63.1	60.6	62.0	59.4	67.2
63.6	60.5	66.8	61.8	64.8	55.8	55.7	77.1	62.1	61.0
58.9	60.0	66.9	61.7	60.3	51.5	67.0	60.2	56.2	59.4
67.9	64.9	55.7	61.4	62.6	56.4	56.4	69.4	57.6	63.8

  - 오름차순 정렬

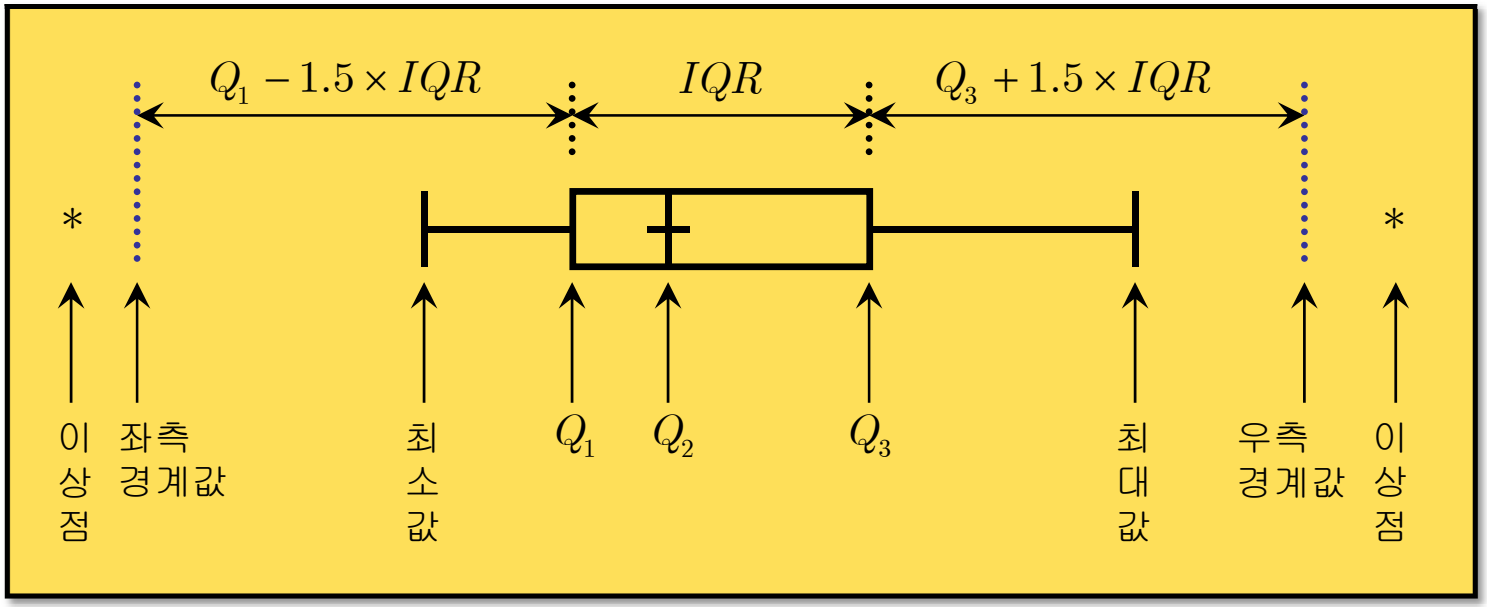
51.3	51.5	55.7	55.7	55.8	55.9	56.0	56.2	56.4	56.4
56.8	57.2	57.6	58.9	59.4	59.4	59.5	59.8	60.0	60.2
60.3	60.5	60.6	60.8	61.0	61.4	61.7	61.8	61.8	62.0
62.1	62.6	62.7	63.1	63.6	63.8	63.8	64.6	64.8	64.9
65.7	66.2	66.8	66.9	66.9	67.0	67.2	67.9	69.4	77.1

- 예 – [교통소음의 정도]
  - 관측값의 개수(n) = 50
  - 제2사분위수(Q_2) = 제50백분위수 = 중앙값 : p = 0.5
    - $n \times p$ = 50×0.5 = 25
    - 25번째 관측값과 26번째 관측값의 평균 : (61.0+61.4)/2 = 61.2
  - 제1사분위수(Q_1) = 제25백분위수 : p= 0.25
    - $n \times p$ = 50×0.25 = 12.5
    - 13번째 관측값 : 57.6
  - 제3사분위수(Q_3) = 제75백분위수 : p = 0.75
    - $n \times p$ = 50×0.75 = 37.5
    - 38번째 관측값 : 64.6
  - 범위(range)
    - 최대값(Max) – 최소값(Min) = 77.1–51.3 = 25.8
  - 사분위범위
    - $IQR = Q_3 - Q_1$ = 64.6–57.6 = 7

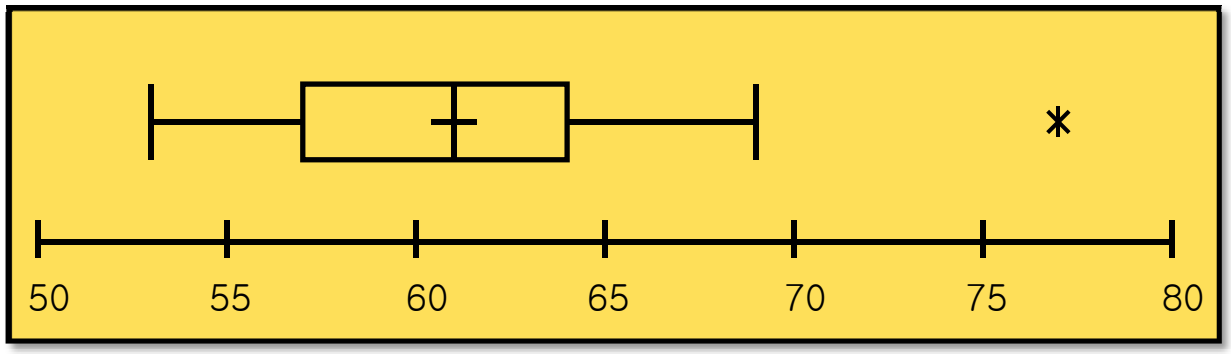
수치 정보를 나타내는 그림

- 상자그림(box plot)
  - 최소값, 사분위수, 최대값을 활용한 그림
  - 자료의 중심위치, 산포도, 대칭성, 극단점 등을 파악 가능
    - 상자그림의 작성 과정
      - 사분위수(Q_1, Q_2, Q_3)를 결정한다.
      - 제1사분위수(Q_1)와 제3사분위수(Q_3)를 네모 상자로 연결한다.
      - 중앙값(median, Q_2)의 위치에 선을 긋는다.
      - 사분위범위(IQR)를 계산한다.
      - 양측 경계값 $Q_1 - 1.5 \times IQR$ 과 $Q_3 + 1.5 \times IQR$ 을 계산한다.
      - 경계 범위내의 최소값과 최대값을 찾아 선을 긋는다.
      - 경계 범위내의 최소값과 최대값을 상자와 연결한다.
      - 양측 경계를 벗어나는 관측값(이상점)들을 찾아 *로 표시한다.

- 상자그림의 정보


$Q_1 - 1.5 \times IQR$
IQR
$Q_3 + 1.5 \times IQR$
*
*
이상점
좌측경계값
최소값
Q_1
Q_2
Q_3
최대값
우측경계값
이상점

- 예 – [교통소음의 정도]
  - 상자그림을 위한 정보
    - 제1사분위수(Q_1) = 57.6, 제2사분위수(Q_2) = 61.2, 제3사분위수(Q_3) = 64.6, 사분위범위(IQR) = 7
    - 경계값 :
      $Q_1 - 1.5 \times IQR = 57.6 - 1.5 \times 7 = 47.1$
      $Q_3 + 1.5 \times IQR = 64.6 + 1.5 \times 7 = 75.1$
    - 47.1 < 51.3 ⇒ 경계 범위내의 최소값 : 51.3
    - 75.1 < 77.1 ⇒ 경계 범위내의 최대값 : 69.4
  - 상자그림



두 범주형 변수의 요약

- 분할표 (contingency table)
  - 두 개의 범주형 변수에 대한 관측값 패턴을 파악하기 위한 표
  - 작성요령
    - 두 변수의 범주를 행과 열에 배치
    - 각각의 범주들이 교차하는 칸(cell)마다 도수 표시
    - 분석 목적에 따라 행상대도수 / 열상대도수 / 전체상대도수 표시
  - 예 – [시험문제의 수준에 대한 의견 조사]
    - 임의의 400명을 대상 (남 : 176명, 여 : 224명)
    - 시험문제의 수준에 대한 의견 조사 (어렵다-보통이다-쉽다)

		의견			합계
		어렵다	보통이다	쉽다	
성별	남자	112	36	28	176
	여자	84	68	72	224
합계		196	104	100	400

- 예 – [시험문제의 수준에 대한 의견 조사]
  - 전체 의견분포 확인 : 전체상대도수 이용

	어렵다	보통이다	쉽다	합계
남자	112 (0.28)	36 (0.09)	28 (0.07)	176 (0.44)
여자	84 (0.21)	68 (0.17)	72 (0.18)	224 (0.56)
합계	196 (0.49)	104 (0.26)	100 (0.25)	400 (1.00)

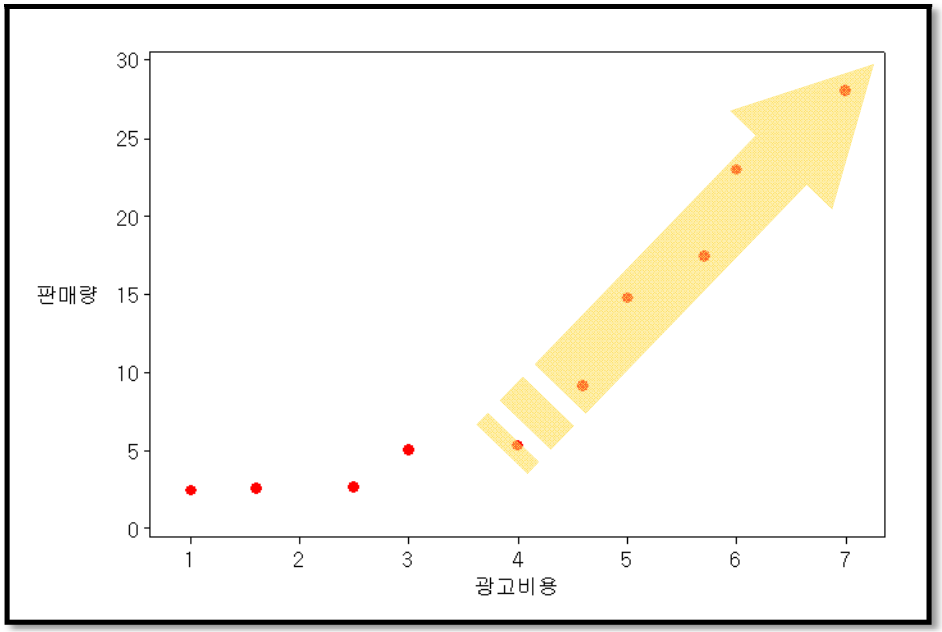
  - 자료 전반적인 개요 파악 용이

그림을 통한 두 연속형 변수의 요약

- 산점도 (scatter plot)
  - 두 개의 연속형 변수에 대한 관측값 패턴을 파악하기 위한 그림
  - 작성요령
    - 두 개의 변수를 각각 수평축과 수직축에 배치
    - 각각의 관측값 짝을 좌표 위에 표시
  - 예 – [광고비용과 판매량]
    - 유사 제품을 판매하는 10개 회사 대상
    - 광고비용(단위 : 억 원)과 판매량(단위 : 백만 개)을 조사

회사	광고비용	판매량	회사	광고비용	판매량
A	1.0	2.5	F	4.6	9.1
B	1.6	2.6	G	5.0	14.8
C	2.5	2.7	H	5.7	17.5
D	3.0	5.0	I	6.0	23.0
E	4.0	5.3	J	7.0	28.0

- 예 – [광고비용과 판매량]
  - 산점도



- 광고비용이 1에서 3사이에는 판매량 증가폭 낮음
- 광고비용이 5 이상인 경우 판매량의 증가폭이 높아짐

수치를 통한 두 연속형 변수의 요약

- 상관계수 (correlation coefficient)
  - 두 변수간의 선형성의 정도
  - 표본상관계수 (sample correlation coefficient, r)
    - 두 변수 X, Y 에 대하여

      n 개의 관측값 짝 (x_1, y_1), (x_2, y_2), $\cdots$, (x_n, y_n) 이 주어질 때,
      - 두 변수의 표본평균

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \ , \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

      - 두 변수의 분산

$$s_{xx} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \ , \qquad s_{yy} = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

- 두 변수의 표본공분산(sample covariance)

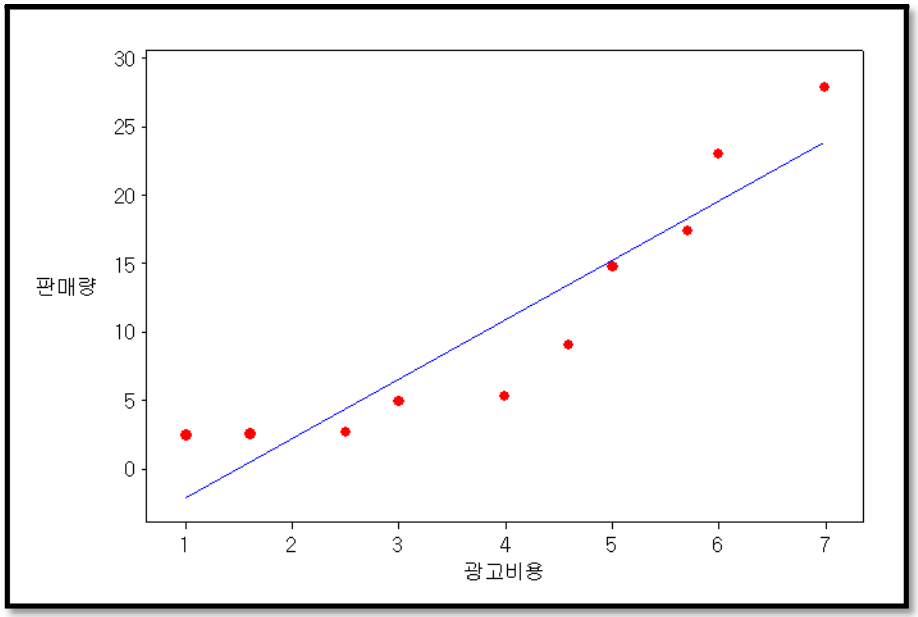
$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

- 두 변수의 표본상관계수

$$r = \frac{s_{xy}}{\sqrt{s_{xx}}\sqrt{s_{yy}}}$$

  - 상관계수는 단위가 상쇄되어 없음
  - 단위가 다른 여러 쌍의 변수들 간 직선관계의 정도 비교 가능

- 예 – [광고비용과 판매량]
  - 표본평균 – 광고비용 : 4.04 , 판매량 : 11.05
  - 표본분산 – 광고비용 : 3.92 , 판매량 : 85.94
  - 표본공분산 :
    - $\{(1.0-4.04)\times(2.5-11.05)+ \cdots +(7.0-4.04)\times(28.0-11.05)\}/9$
      $= 16.98$
  - 표본표준편차
    - 광고비용 : 1.98
    - 판매량 : 9.27
  - 표본상관계수
    - $r = \dfrac{16.98}{1.98\times 9.27}$
      $= 0.925$



- 상관계수의 특징
  - 상관계수는 항상 -1과 1사이의 값($-1 \le r \le 1$)을 가짐
  - 상관계수의 부호는 직선관계의 방향을 나타냄
    - 상관계수가 양수($r > 0$)이면
      한 변수의 값이 작으면 다른 변수의 값도 작음
      한 변수의 값이 크면 다른 변수의 값도 큼
    - 상관계수가 음수($r < 0$)이면
      한 변수의 값이 작으면 다른 변수의 값은 큼
      한 변수의 값이 크면 다른 변수의 값은 작음
  - 상관계수의 절대값 크기는 직선관계의 정도를 나타냄
    - 상관계수의 절대값이 1에 가까울수록
      관측값들이 직선 주위에 몰려 있으며 강한 상관관계를 가짐
    - 상관계수의 절대값이 0에 가까울수록
      약한 상관관계를 가지며 기울기가 없는 직선의 관계를 보임

- 산점도와 상관계수