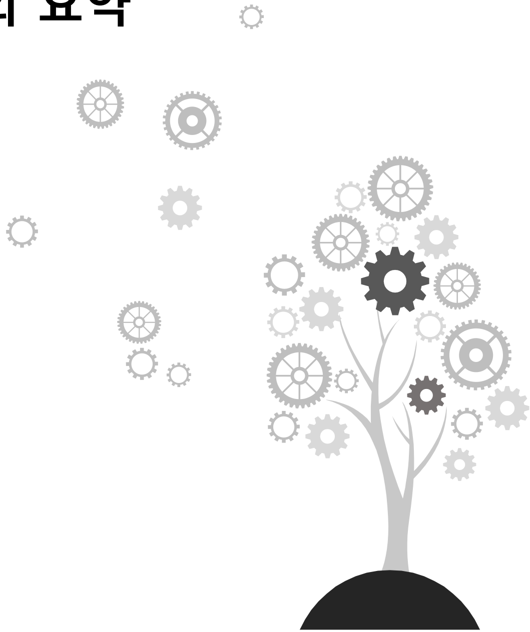
Contents



Chapter 3 수치를 통한 연속형 자료의 요약

- 그림을 통한 요약은 일관성과 객관성이 떨어지며, 통계적 추론에서 요구되는 이론적 근거를 제시하기 어렵다.

- 몇 개의 의미 있는 수치로 요약하여 자료의 대략적인 분포상태를 파악하고자 한다.
(중심위치의 측도, 퍼진 정도의 측도, 상자그림)

01 중심 위치의 측도 (Measure of center)

- 연속형 자료가 어떤 값을 중심으로 분포되어 있는지를 나타낸다.

1 표본평균 (sample mean)
① 극단값에 크게 영향 받는다.

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

모든 관측 값을 다 더해서 표본의 크기로 나눈다.

Chapter 3 수치를 통한 연속형 자료의 요약

'n'은 '표본의 크기' 이다. (=표본의 개수)

② 중앙값 (median)

① 전체 관측값을 크기 순서로 배열 했을 때 가운데 위치하는 값
② 극단값에 크게 영향 받지 않는다.
③ 관측값의 변화에 민감하지 않다.

핵심!!!

$$
\text{median} = \begin{cases} x_{\left(\frac{n+1}{2}\right)}, & \text{if n is odd} \\ \dfrac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}, & \text{if n is even} \end{cases}
$$

이 때문에,

평균은 계급구간의 중앙값×상대도수

③ 최빈값 (mode)

① 관측값 중에서 가장 자주 나오는 값
② 이산형, 범주형에서 사용가능하다. → 이산형, 순위형, 명목형.
③ 단봉형 분포를 갖는 자료에서 유용하다.
④ 연속형 자료에서는 도수분포표에서 최대도수를 갖는 계급구간의 중앙값으로 최빈값을 정한다.
그러나 계급구간의 폭에 영향을 받는다.

핵심!!!

수치를 통한 연속형 자료의 요약

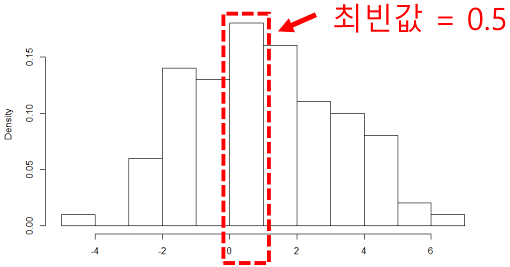
- 최빈값의 예시

① 범주형 자료

범주	도수
A	22
B	20
AB	7
O	11
합	60

② 연속형 자료

계급	계급구간	도수
1	-4 ~ -3	1
2	-3 ~ -2	3
3	-2 ~ -1	7
4	-1 ~ 0	14
5	0 ~ 1	21
6	1 ~ 2	20
7	2 ~ 3	19
8	3 ~ 4	9
9	4 ~ 5	4
10	5 ~ 6	2
합		100



Chapter 3 수치를 통한 연속형 자료의 요약

예제 1 & 2. 어떤 과목에서 6명의 학생의 점수가 89, 74, 91, 88, 72, 84일 때, 표본평균과 중앙값을 구해라.

① 표본평균

② 중앙값

예제 3. 콩의 개수에 관한 자료가 아래와 같을 때 최빈값을 구해라.

4	3	4	3	3	5	5	6	4	4	4	3
3	4	3	3	6	4	5	3	6	3	2	1
4	4	4	4	4	5	3	4	3	1	2	2
5	2	4	3	5	5	3	3	3	3	5	5
3	3	6	4	3	5	6	4	4	3	3	4

콩의 개수	도수	상대도수
1		
2		
3		
4		
5		
6		
합		

- 최빈값 =

Chapter 3 수치를 통한 연속형 자료의 요약

- 표본평균, 중앙값, 최빈값의 비교

① 전체의 경향을 볼 때 극단적인 영향을 배제해야 하는 경우에는 중앙값을 사용하는 것이 적절하다.

② 전체의 경향을 볼 때 전체 관측값을 모두 포함하고자 하는 경우에는 평균을 사용하는 것인 바람직하다.

※ 중심 위치 측도
표본 평균
중앙값
최빈값

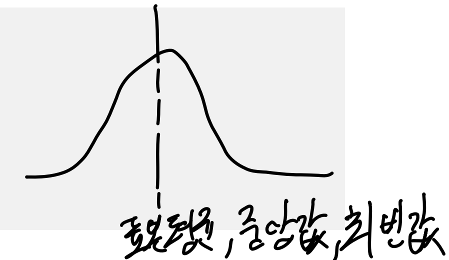
예제 4. 예제 1의 자료에서 74점이 50점으로 바뀌었다고 하자. 표본평균과 중앙값을 구하라.

① 표본평균

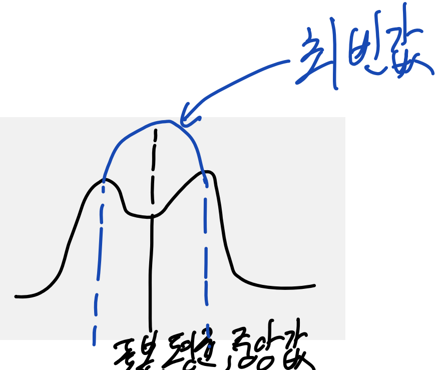
② 중앙값

Chapter 3 수치를 통한 연속형 자료의 요약

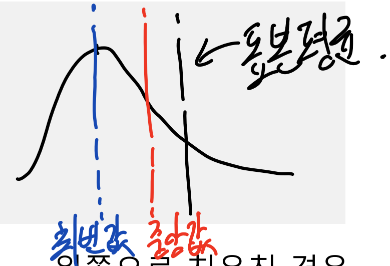
- 분포의 모양에 따른 표본평균, 중앙값, 최빈값 위치



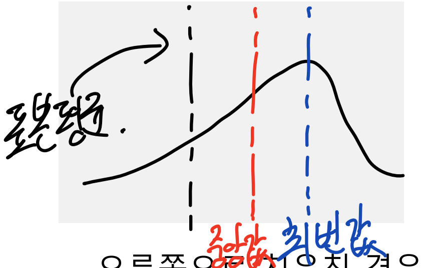
단봉형, 대칭분포



이봉형, 대칭분포



왼쪽으로 치우친 경우



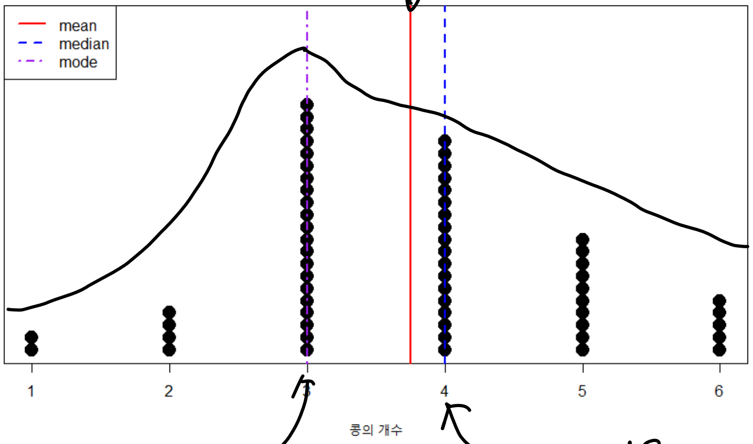
오른쪽으로 치우친 경우

수치를 통한 연속형 자료의 요약

- 표본평균, 중앙값, 최빈값 위치 비교 예시

예제 3의 콩의 개수에 관한 자료

중앙값



최빈값

표본평균

02 퍼진 정도의 측도 (Measure of dispersion)

- 분산과 표준편차 (Variance and standard deviation)

① 편차(deviation) : 관측값과 평균의 차이, $x_i - \bar{x}, \quad i = 1, \dots, n$

d_i

편차의 합 : $\sum_{i=1}^{n}(x_i - \bar{x}) = 0,$ 편차제곱합 : $\sum_{i=1}^{n}(x_i - \bar{x})^2 \geq 0$

② 표본분산(sample variance) : 편차 제곱합을 자유도(= $n-1$)로 나눈 값으로 정의

s^2

↑ 도출되는 모든 편차 제곱을 더한 후, n-1로 나눈다.

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

③ 표본표준편차(sample standard deviation) :표본분산의 제곱근

$$\text{표본표준편차} = s$$

s

↳ 표준 편차 : 편차들의 기댓값 (=평균)

※ 편차의 합은 무조건 '0'이다.
⇒ 그래서, 편차의 합을 활용해선, 표준편차를 구하기 어렵다!!
⇒ 이 때문에, '분산'을 통해 표준편차를 구한다!!

표준편차 : 분산의 제곱근(양수) 표준편차를 구한다..

Chapter 3 수치를 통한 연속형 자료의 요약

④ 표본분산(sample variance)의 간편식

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

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

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

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

Chapter 3 수치를 통한 연속형 자료의 요약

<=> 자료의 범위, 관측값의 범위.

- 범위(Range)

① 범위 = 최대값 – 최소값

- 백분위수

① 제 $100 \times p$ 백분위수 (the $100 \times p -$ th percentile)

: 자료의 수가 n일 때, 제 $100 \times p$ 백분위수는 그 값보다 작거나 같은 관측값의 개수가 np개 이상이고 그 값보다 크거나 같은 관측값이 $n(1-p)$개 이상인 값

② 중앙값의 일반화 => 백분위수는 중앙값과 관련있다.

- 제 $100 \times p$ 백분위수 구하는 방법

① 관측값을 작은 순서로 배열한다. <- sort 먼저함!!!

② 관측값의 개수(n)에 p를 곱한다.

1) 만약 $n \times p$가 정수이면, $n \times p$번째로 작은 관측값과 $n \times p + 1$번째로 작은 관측값의 평균을 제 $100 \times p$ 백분위수로 한다.

2) 만약 $n \times p$가 정수가 아니면, $n \times p$에서 정수부분에서 1을 더한 값 m을 구한 후, m번째로 작은 관측값을 제 $100 \times p$ 백분위수로 한다.

<애가 상대적으로 구하기 쉽다.>

핵심!!!

- 사분위수 범위

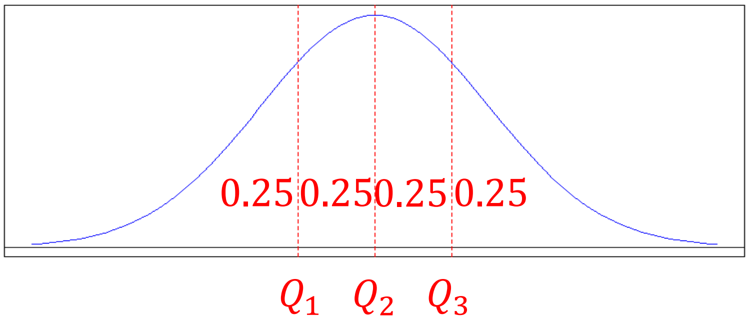
① 사분위수 (quartile)

제1사분위수: Q_1=제25백분위수
제2사분위수: Q_2=제50백분위수=중앙값
제3사분위수: Q_3=제75백분위수

② 사분위수 범위(interquartile range)

: 사분위수범위는 중앙값을 중심으로 전체 자료의 50% 들어오는 범위이며, 극단값에 영향을 덜 받는다.

$$\mathrm{IQR} = Q_3 - Q_1$$



- 표준편차, 범위, 사분위수 범위 비교
  ① 표준편차 : 표본평균에 대응하는 측도 (전체 관측값의 퍼진 정도를 모두 반영)
  ② 사분위수 범위 : 중앙값에 대응하는 측도 (극단적인 관측값이 배제되어 있어 극단값에 크게 영향 받지 않음)
  ③ 범위 : 극단값에 민감하게 반응하고, 관측값을 골고루 반영하지 못함

예제 7. 전철역 사이의 시간 (단위: 분)

-※ 퍼진 정도를 측정하는 측도.
- 표준편차
- 사분위수 범위
- 범위

표본 : 42, 40, 38, 37, 43, 49, 78, 38, 45, 44, 40, 38, 41, 35, 31, 44
정렬된 표본 : 31, 35, 37, 38, 38, 38, 39, 40, 40, 41, 42, 43, 44, 44, 45, 78

제 20백분위수 : 16 * 0.2 = 3.2 이므로 4번째로 작은 값인 38
제 50백분위수 : 16 * 0.5 = 8 이므로 8번째 및 9번째로 작은 값의 평균인 40
(중앙값의 정의와 일치)

※ 제 0 백분위수 = 최소값 : 31, 제 100 백분위수 = 최대값 : 78

① 단위가 다르거나 ② 중심 위치가 매우 다른

- 변동계수(Coefficient of variation)

두 개 이상의 분포를 비교할 때 사용함.

① 상대적으로 퍼진 정도를 나타내는 수치
② 서로 다른 단위를 갖는 자료들을 비교할 때 유용함
③ 표본평균에 대한 상대적인 퍼진 정도, 즉 표본표준편차를 백분율로 나타낸 것

기준값

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

④ 상대적인 수치이므로 단위가 존재하지 않음

예제 10

비율에선 '분모' 자리에 들어감.

$$비율 = \frac{비교값\ (\text{'}\sim의\text{'})}{기준값\ (\text{'}\sim에\ 대한\text{'})}$$

03 상자 그림(Box plot)

① 최소값, Q_1, Q_2, Q_3, 최대값을 가지고 그린 그림으로 요약된 수치와 자료의 전체적인 모양을 함께 제공.

② 중심 위치, 퍼진 정도의 수치뿐만 아니라 분포의 대칭성, 분포의 집중 정도, 이상점(극단값) 파악가능

③ 단봉형 자료 분석에 적절하며, 다봉형에서는 효과적인 분석이 어려움

- 상자그림의 작성과정

① 사분위수(Q_1, Q_2, Q_3)를 결정한다

② Q_1과 Q_3을 네모난 상자로 연결하고 중앙값(Q_2)의 위치에 수직선을 긋는다

③ $IQR = Q_3 - Q_1$을 계산한다

④ 상자 양끝에서 $1.5 \times IQR$크기의 범위를 경계로 하여, 이 범위에 포함되는 최솟값과 최댓값을 Q_1과 Q_3으로 부터 각각 선으로 연결한다

⑤ 양 경계를 벗어나는 자료값들을 *로 표시하고, 이 점들을 이상점이라고 한다

핵심이다!!

절대, $Q_1 - 1.5 \times IQR$을 Q_1과

연결하면 안된다.

04 도수분포표에서의 자료의 요약

① 연속형 자료가 도수분포표로 요약되고 원자료가 주어지지 않았을 때를 사용
② 각 계급구간의 중간값을 선택하여 그 계급구간의 모든 관측값이 그 값을 갖는 것처럼 평균과 분산 등을 계산한다.
③ 원자료에서의 표본평균과 표본분산과는 값이 다를 수 있다.

1) 도수분포표에서의 표본평균

각 계급의 중간값

$$\bar{x}_g = \frac{1}{n}\sum_{i=1}^{k} m_i \cdot f_i = \sum_{i=1}^{k} m_i \cdot \left(\frac{f_i}{n}\right) = \sum(\text{중간값} \times \text{상대도수})$$

2) 도수분포표에서의 표본분산

$$s_g^2 = \frac{1}{n-1}\sum_{i=1}^{k}(m_i - \bar{x}_g)^2 \cdot f_i = \frac{1}{n-1}\left(\sum_{i=1}^{k} m_i^2 f_i - n\bar{x}_g^2\right)$$

3) 도수분포표에서의 표본표준편차

$$s_g = \sqrt{s_g^2}$$

k: 계급의 개수, f_i: i번째 계급의 도수, m_i: i번째 계급구간의 중앙값, n: 자료의 개수($= \sum_{i=1}^{k} f_i$)

※ 어떤 자료의 표본 평균과 중앙값의 차이가 심하다면, 중앙값이 중심 위치 측도로서 더 신뢰할 수 있는 것이다.