2. 수치를 이용한 자료의 정리

【참고】 자료를 요약하고 그 특성을 파악하기 위하여 자료를 일련의 수치 즉,

① 대푯값(=중심경향값)

② 산포도

③ 상대적 위치

등을 이용하여 나타낸다.

2-1. 중심경향값

① 중심경향(central tendency)이란? 자료에서 얻어진 모든 값들이 어떤 값을 중심으로 몰리는 경향

② 중심경향값(measure of central tendency)이란? 자료를 대표하는 값

〖예〗 평균, 중앙값, 최빈값 등

2-1-1. 평균

① 관측된 n개의 자료의 총합을 자료의 개수 n으로 나눈 산술평균

② 가장 보편적으로 사용

③ 모집단 전체의 평균을 모평균(population mean)이라 하며 보통 그리스 문자 μ(뮤)로 표기

④ 모집단에서 추출된 표본의 평균을 표본평균(sample mean)이라 하며 보통 $\overline{X}$나 $\overline{Y}$ 등으로 표기

〖예〗 6개의 수집된 자료 $15, 5\ 10, 20, 17, 7$의 평균은 $\frac{15+5+10+20+17+7}{6}=12.3$이다.

2-1-2. 중앙값

① 자료를 크기 순서로 나열하였을 때 중앙의 위치에 해당하는 값, $\widehat{M}$, M_e, M_d로 표기

② 백분위수의 개념에 비추면 제50백분위 점수

③ 자료의 개수가 홀수개; $\frac{n+1}{2}$번째 값

④ 자료의 개수가 짝수개; $\frac{n}{2}$번째와 $\frac{n}{2}+1$번째 값의 평균

〖예〗 $15, 5\ 10, 20, 17, 7$; 6개

㉠ 크기 순으로 배열 $5, 7, 10, 15, 17, 20$

㉡ 3번째 자료와 4번째 자료의 평균인 $\frac{10+15}{2}$=12.5

2-1-3. 최빈값

① 전체 자료 중 가장 빈도가 높은 값, M_o로 표기

② 모든 사례가 각기 다른 값을 가지면 최빈값은 존재하지 않으며 도수가 많은 값이 여러 개인 경우 최빈값이 여러 개 존재할 수 있음

〖예〗 10문항을 10명의 대학생에게 질문, 맞힌 문항 수

8, 7, 9, 4, 8, 10, 9, 9, 3, 5

도수가 가장 많은 3명의 점수인 9점

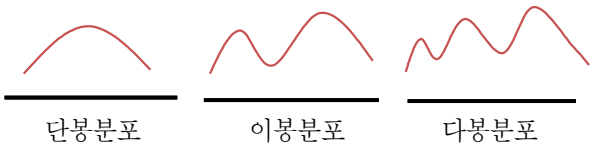
③ 도수가 가장 많은 수의 값을 알고자 할 때만 사용

〖예〗 구두 제작에서 발 크기의 평균보다는 가장 흔한 발 크기에 맞는 구두를 제작함

④ 최빈치가 1개일 때, 단봉분포(unimodal distribution)

2개일 때, 이봉분포(bimodal distribution)

3개 이상일 때, 다봉분포(multi-modal distribution)

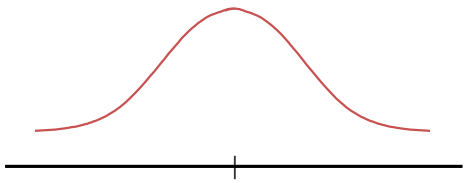


단봉분포 이봉분포 다봉분포

【참고】 평균, 중앙값, 최빈값의 관계

① 좌우대칭 분포일 경우 : 평균=중앙값=최빈값

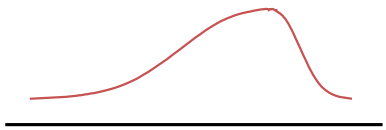
〖예〗 정규분포; 좌우대칭, 점근선적



평균=중앙값=최빈값

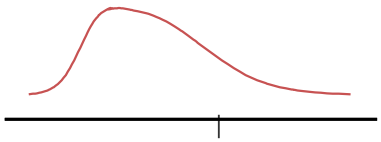
② 부적편포(negatively skewed distribution) : 평균<중앙값<최빈값

〖예〗 완전학습



③ 정적편포(positive skewed distribution) : 최빈값<중앙값<평균

〖예〗 정신지체아의 지능



2-2. 산포도(분산도)

① 분포의 흩어진 분포를 알면 자료의 특성 파악 용이
② 중심경향값이 같더라도 넓은 범위에 흩어진 분포가 있을 수 있고 좁은 범위에 흩어진 분포가 있을 수 있어 그 흩어진 정도를 고려
③ 산포도란 개개의 관찰값이 대푯값 주위에 어떻게 분포되어 있는가를 계량하는 척도
④ 산포도가 클수록 그 분포의 흩어진 폭이 넓고, 산포도가 작을수록 분포의 흩어진 폭은 좁음
⑤ 범위, 분산, 표준편차 등

2-2-1. 범위

① 자료의 최댓값과 최솟값의 차이
〖예〗 신생아 체중 자료의 관측값이 5.6에서 9.4 사이
; 범위 $R = 9.4 - 5.6 = 3.8$
② 계산하기 쉽고 설명하기 쉬워 적은 자료의 변동에 적당
③ 범위는 자료의 두 극단적인 값의 차이만을 나타내기 때문에 이상점(outlier)이 있을 경우 올바른 산포의 측도가 되지 못한다는 단점이 있음. 범위의 이러한 단점을 일부 보완한 것이 사분위 범위(interquartile range)임

2-2-2. 분산과 표준편차

① 모든 자료를 각각 고려하여 분포의 흩어진 정도를 나타낸 것
② 표준편차란 각 점수가 평균으로부터 떨어진 정도인 편차(deviation)들의 평균

$$(\text{분산}) = \frac{(\text{편차})^2\text{의 총합}}{(\text{도수})\text{의 총합}} \quad (\text{표준편차}) = \sqrt{(\text{분산})}$$

〖예〗 학생 8명의 충치 수

(단위: 개)

2	5	3	1	3	4	0	6

㉠ $(\text{평균}) = \dfrac{2+5+3+1+3+4+0+6}{8} = 3$ (개)

㉡ $(\text{분산}) = \dfrac{1}{8}\{(2-3)^2 + (5-3)^2 + (3-3)^2 + (1-3)^2$
$+ (3-3)^2 + (4-3)^2 + (0-3)^2 + (6-3)^2\}$
$= \dfrac{7}{2} = 3.5$

㉢ $(\text{표준편차}) = \sqrt{3.5} \fallingdotseq 1.87$ (개)

③ 분산이 크면 분포가 흩어져 있으므로 그 분포를 구성하는 요소들은 이질적(heterogeneous)이며, 분산이 작으면 작을수록 그 분포를 구성하는 요소는 동질적(homogeneous)이라 할 수 있다.

2-3. 상대적 위치

① 많은 량의 자료들에 대해 특정 자료값이 전체 자료에서 어떤 위치에 있는가를 알 수 있는 방법
② 사분위수, 백분위수 그리고 Z-값 등

2-3-1. 사분위수

① 크기 순서에 따라 늘어놓은 자료를 사등분할 때 각각 사등분되는 위치의 값

㉠ 제1사분위수(first quartile, Q_1); 전체 자료의 1/4 값보다 작거나 같게 되는 값
㉡ 제2사분위수(second quartile, Q_2); 자료의 중앙값
㉢ 제3사분위수(third quartile, Q_3); 전체 자료의 3/4 값보다 작거나 같게 되는 값
㉣ 제4사분위수(fourth quartile, Q_4); 자료의 최대값

② 제1사분위수 Q_1을 Lower Quartile, 제3사분위수 Q_3을 Upper Quartile 이라고 부르며,
$Q_3 - Q_1$을 사분위수 범위(Inter Quartile Range ; IQR)이라고 함

2-3-2. 백분위수

① 분위수의 개념을 확대하여 크기 순서에 따라 자료를 100 등분하는 값
② $p\%$ 백분위수(p-th percentile)는 전체 자료의 $p\%$가 이 값보다 작거나 같게 되는 값을 의미
㉠ 제1사분위수 Q_1(25% 백분위수) = $0.25\times(n+1)$번째 자료 값
㉡ 제2사분위수 Q_1(50% 백분위수) = $0.50\times(n+1)$번째 자료 값
㉢ 제3사분위수 Q_1(75% 백분위수) = $0.75\times(n+1)$번째 자료 값
③ Q_1, Q_3가 정수가 아니면 사분위수는 인접한 두 개 값을 사용하여 보간법으로 구한다.
④ 중앙값과 사분위수는 자료를 4등분한다. Q_1과 Q_3는 가운데 50% 자료값의 하한과 상한이다.
제3사분위수와 제1사분위수와의 차이를 자료의 사분위수범위(IQR)라 부른다.
⑤ 자료 “16, 25, 4, 18, 11, 13, 20, 8, 11, 9”에 대하여 제1사분위수와 제3사분위수를 구하시오.

[풀이]

㉠ 관측값을 크기순서로 정렬한다. ; 4, 8, 9, 11, 11, 13, 16, 18, 20, 25
㉡ Q_1의 위치 = $0.25(n+1) = 0.25(10+1) = 2.75$
㉢ Q_3의 위치 = $0.75(n+1) = 0.75(10+1) = 8.25$
㉣ 정수가 아니므로 보간법을 사용하면
- $Q_1 = 8 + 0.75(9-8) = 8 + 0.75 = 8.75$
- $Q_3 = 18 + 0.25(20-18) = 18 + 0.5 = 18.5$

㉤ 따라서 $IQR = Q_3 - Q_1 = 18.5 - 8.57 = 9.75$이다.

2-3-3. $Z-$값

① 개별 자료값들이 그들의 평균으로부터 표준편차의 몇 배 만큼 떨어져 있는가를 나타내는 수치

$$Z\text{-값}=\frac{X-\mu}{\sigma}\quad(\mu=\text{평균},\ \sigma=\text{표준편차})$$

〖예〗 다음 자료의 사분위수, 백분위수 그리고 $Z-$값

85, 89, 107, 109, 110, 129, 144, 161, 187, 193, 194, 196, 202, 203, 224

순위	1	2	3	4	5	6	7	8	9
점수	85	89	107	109	110	129	144	161	187
z-값	-1.487	-1.403	-1.023	-0.981	-0.960	-0.559	-0.243	0.116	0.665

순위	10	11	12	13	14	15
점수	193	194	196	202	203	224
z-값	0.791	0.812	0.854	0.981	1.002	1.445

[풀이]

㉠ 제1사분위수 Q_1 = 25% 백분위수 ; $0.25(n+1)=0.25(15+1)=4$번째 109

㉡ 제3사분위수 Q_3 = 75% 백분위수 ; $0.75(n+1)=0.75(15+1)=12$번째 196

㉢ 사분위수 범위 $IQR=Q_3-Q_1=196-109=87$이다.