

확률과 통계

- 평균 구하기 -

참고서: Head First Statistics

(2장 평균구하기)

작성자 : 이계식

한경대학교, 컴퓨터공학과



평균 구하기

1. 평균(Average)의 종류

- ① 평균값 (Mean)
- ② 중앙값 (Median)
- ③ 최빈값 (Mode)

p.88

2. 평균값(Mean)

예제 : 파워워크다웃 고실 참가자들의 나이

19, 20, 20, 20, 21

질문: 참가자들 나이의 평균값은?

답: $\frac{19+20+20+20+21}{5}$ ← 나이의 총합
← 참가자수

p.90

일반화: 평균값 (μ) = $\frac{\sum X}{n}$

μ: 평균값의 기호

드수의 합
(조사대상 총 개체수)

$\sum X = x_1 + x_2 + \dots + x_n$

1그다지
(합을 나타내는 기호)

$x_i = i$ 번째 대상과
관련된 숫자

(예: 나이, 체중, 키, 가격, 예산 등)

p.91

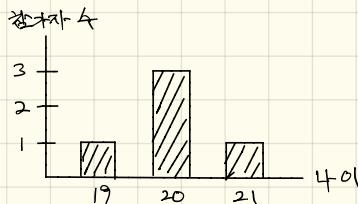
예제: 파워워크아웃교실 참가자들의 나이를 담은 표 다시 살펴보기

나이	19	20	21
도수	1	3	1

참가자 수

$$\Rightarrow \text{평균값} = \frac{19 \cdot 1 + 20 \cdot 3 + 21 \cdot 1}{5}$$

주의: 막대그래프는 평균값에 대한 정보를 주지 않는다.



3. 이상치 (Outliers)

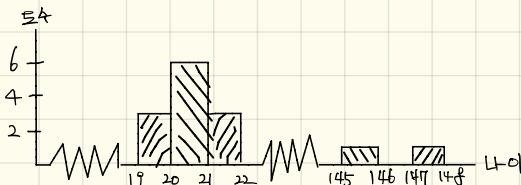
p.96

평균값의 한계

예제: 컴퓨터교실 참가자들 나이의 평균연령 구하기

나이	19	20	21	145	147
도수	3	6	3	1	1

히스토그램으로 연령분포를 확인해 보자.



$$\text{평균연령}(\mu) = \frac{19\cdot3 + 20\cdot6 + 21\cdot3 + 145 + 147}{14}$$

$$= 38$$

P. 97-98

문제점: 평균연령이 38세이지만 거의 대부분은 20대 초반임.

나이가 아주 많은 두 명은 무질의 고수.

50대 초반의 창가 희망자가 쿵푸수업을

할 따라갈 수 있을까?

창가자들 나의 평균값(평균연령)만을 보고

창가신청하면 매우 고생할 수 있음.

극단적인 값

원인: 두 개의 이상치 존재!

즉, 145, 147 두 개의 이상치가 평균값을 왜곡시킴.

결론: 이상치를 포함한 편향된 데이터는 조심해서 다루어야 함.

또한 이상치를 고려하여 평균을 정해야 함.

위 예제의 경우 평균값보다 중앙값이 창가들의 나이를
보다 잘 대표함.

4. 중앙값 (Median)

P. 101

정의: 가운데에 위치한 값

예제 1: 19 19 20 20 20 21 21 100 102

예제 2: 19 20 20 20 21 21 10 102

두 수의 평균값 (20.5)

P. 102

중앙값 구하기 3단계

① 주어진 N 개의 데이터를 크기순으로 정렬

② N 이 홀수인 경우 :

$$\text{중앙값} = \frac{N+1}{2} \text{ 번째 위치한 데이터}$$

③ N 이 짝수인 경우 :

$$\text{중앙값} = \frac{N}{2} \text{ 번째 데이터와}$$

$$\frac{N}{2} + 1 \text{ 번째 데이터의 평균값}$$

P. 104

편향된 데이터의 평균값과 중앙값 비교

① 오른쪽으로 편향된 데이터

예제)

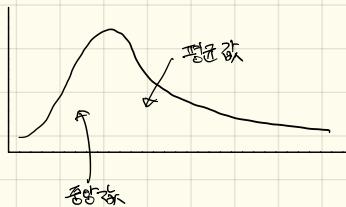
값	1	2	3	4	5	6	7	8
도수	4	6	4	4	3	2	1	1

$$\text{중앙값} = 3$$

$$\begin{aligned}\text{평균값} &= \frac{4+2 \cdot 6+3 \cdot 4+4 \cdot 4+5 \cdot 3+6 \cdot 2+7+8}{25} \\ &= 3.44\end{aligned}$$

결론 : 중앙값 < 평균값

데이터의 그래프는 아래 모양을 갖는다.



(2) 원쪽으로 편향된 데이터

예제)

값	1	4	6	8	9	10	11	12
도수	1	1	2	3	4	4	5	5

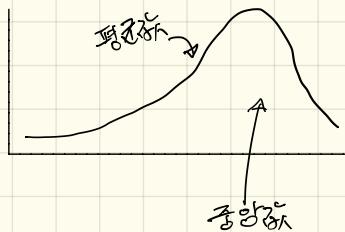
$$\text{중앙값} = 10$$

$$\text{평균값} = \frac{1+4+6 \cdot 2 + 8 \cdot 3 + 9 \cdot 4 + 10 \cdot 4 + 11 \cdot 5 + 12 \cdot 5}{25}$$

$$= 9.28$$

결론: 중앙값 > 평균값

데이터의 그래프는 아래 모양을 갖는다.

5. 최빈값 (Mode)평균값과 중앙값의 합계

예제: 부모와 함께하는 아이 수영교실

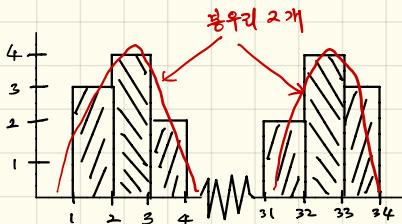
나이	1	2	3	31	32	33
도수	3	4	2	2	4	3

$$\text{평균값} = \frac{1 \cdot 3 + 2 \cdot 4 + 3 \cdot 2 + 31 \cdot 2 + 32 \cdot 4 + 33 \cdot 3}{18} = 17$$

$$\text{중앙값} = 17 \quad (\text{총 } 18\text{ 명, 따라서 } 9\text{ 번째 값 } 3\text{ 라 } 10\text{ 번째 값인 } 3\text{의 평균값})$$

} 두 경우 모두
최빈값들을 선택
대표하지 못함.

데이터의 히스토그램은 아래 모양을 갖는다.



결론: 한 개의 나이가 아니라 두 개의 나이가 대표성을 가짐
이런 경우 최빈값을 활용하여 평균(Average)을 구한다.

p. 113

최빈값 = 가장 큰 도수를 갖는 값 또는 범주

부모와 할머니라는 어린이 수영교실의 경우 2세와 32세 모두 최빈값이 된다.

p. 114

연습

①

값	1	2	3	4	5	6	7	8
도수	4	6	4	4	3	2	1	1

$$\text{평균값} = \frac{1 \cdot 4 + 2 \cdot 6 + 3 \cdot 4 + 4 \cdot 4 + 5 \cdot 3 + 6 \cdot 2 + 7 \cdot 1 + 8 \cdot 1}{25} = 3.44$$

중앙값 = 3 ← 대표성이 가장 높.

최빈값 = 2

오른쪽으로 편향됨

②

별주	파랑	빨강	초록	분홍	노랑
도수	4	5	8	1	3

평균값 = (의미 없음)

숫자와 관련된 데이터에서만 의미 있음.

중앙값 = (의미 없음)

최빈값 = 초록 ← 대표성이 가장 높.

(2)

값	1	2	3	4	5
도수	2	3	3	3	3

$$\text{평균값} = \frac{1 \cdot 2 + 2 \cdot 3 + 3 \cdot 3 + 4 \cdot 3 + 5 \cdot 3}{14} = 3.14 \quad \leftarrow \text{데이터가 편향되지 않은 값을 경우}$$

$$\text{중앙값} = 3 \quad \leftarrow \text{데이터가 편향되지 않은 값을 경우 평균값을 우선시함.}$$

$$\text{최빈값} = 2, 3, 4, 5 \quad \leftarrow \text{최빈값이 여러 개 있을 경우 중요성이 많이 들어짐.}$$

6. 요약 정리

P-118

평균	계산방법	사용환경
평균값	$\frac{\sum X}{n}$	데이터가 비교적 좌우 대칭일 때
중앙값	<ul style="list-style-type: none"> ① 모든 값을 오름차순으로 나열 ② 홀수개인 경우 중앙에 위치한 값 ③ 짝수개인 경우 중앙에 위치한 두 값의 평균값 	<ul style="list-style-type: none"> 데이터가 이상치로 인해 편향된 경우
	<ul style="list-style-type: none"> ① 도수가 가장 높은 값 ② 두 종류 이상의 데이터가 포함되었을 경우 각 그룹에 대한 최빈값 선택 	<ul style="list-style-type: none"> ① 혼주적 데이터에서도 사용 가능 ② 데이터가 두 개 이상의 데이터 그룹을 포함하는 경우

7. 연봉 문제

P.119

어떤 회사의 사장이 직원들의 연봉을 올려주고자 했.

그러면서 두 가지 방식을 제안함.

방식①: 모두의 연봉을 동일하게 그백만원씩 올리기

방식②: 각자의 연봉을 10%씩 올리기

한때 직원들의 평균값, 중앙값, 최빈값은 다음과 같다.

평균값 = 오천만원

중앙값 = 이천만원

최빈값 = 천만원

문제 a : 방식①을 따를 경우 평균값, 중앙값, 최빈값을 계산하라

답) 평균값, 중앙값, 최빈값 모두 일정하게 그천만원씩 오른다.

즉,

평균값 = 오천이백만원

중앙값 = 이천이백만원

최빈값 = 천이백만원

문제 b : 방식②을 따를 경우 평균값, 중앙값, 최빈값을 계산하라

답) 평균값, 중앙값, 최빈값 모두 일정하게 그천만원씩 오른다.

즉,

평균값 = 오천오백만원

중앙값 = 이천이백만원

최빈값 = 천백만원

문제 c : 연봉이 평균값에 가까울 경우와 최빈값에 가까운 경우에 따라

연봉인상 방식에 대한 선호도가 다를 수 있다.

어떻게 달라지나?

답) - 중앙값 이상인 경우 : 10% 인상 선호

- 중앙값 이하인 경우 : 이천만원 인상 선호