

평균 하나로 전체 학생을 설명할 수 있을까?

1쪽 / 여섯 명의 기록으로 알 수 있는 것

질문: 어느 과목 수강생 120명의 편도 통학 시간을 알고 싶다. 월요일 오전 정문에서 자원한 학생 6명에게 통학 시간을 물었다. 응답은 10, 15, 15, 20, 25, 35분이다. 모든 값은 같은 날의 편도 통학 시간이며 빠진 응답은 없다.

모집단은 알고 싶은 대상인 수강생 120명이다. 이번에 관찰한 표본은 응답한 6명이다. 정문에 있었고 자원한 사람만 조사했으므로, 전체 수강생에서 무작위로 뽑은 표본이라고 할 수 없다.

산술평균은 관찰값의 합을 관찰값 수로 나눈 값이다. 합계는 120분이므로 표본평균은 $120 \div 6 = 20$ 분이다. 평균 20분은 여섯 명의 기록을 요약한 값이지, 120명 전체를 조사한 결과가 아니다.

중앙값은 값을 작은 순서로 정렬했을 때 가운데를 나타낸다. 값이 6개이므로 세 번째 15분과 네 번째 20분의 평균을 구한다. 중앙값은 $(15 + 20) \div 2 = 17.5$ 분이다. 가운데 두 값 중 하나만 고르지 않는다.

이 자료로 쓸 수 있는 문장: '월요일 오전 정문에서 자원한 6명의 평균 편도 통학 시간은 20분이고 중앙값은 17.5분이었다.' 이 자료만으로 쓸 수 없는 문장: '수강생 120명의 평균 통학 시간은 20분이다.'

평균 하나로 전체 학생을 설명할 수 있을까?

2쪽 / 조건 하나를 바꾸고 다시 판단하기

변경 A: 마지막 값 35분을 95분으로 바꾼다. 새 학생을 추가하는 것이 아니라 같은 여섯 값 중 하나를 교체한다. 자료는 10, 15, 15, 20, 25, 95분이다. 합계는 180분, 평균은 30분이다. 가운데 두 값은 여전히 15와 20이므로 중앙값은 17.5분이다.

이 예제에서 큰 값 하나의 변화는 평균을 10분 올렸지만 중앙값은 바꾸지 않았다. 이를 ‘중앙값은 어떤 자료에서도 절대 변하지 않는다’는 규칙으로 외우면 안 된다. 정렬했을 때 가운데 값이 달라지면 중앙값도 달라진다.

변경 B: 처음 자료에서 35분 응답이 누락됐다고 가정한다. 누락은 통학 시간이 0분이라는 뜻이 아니다. 확인된 다섯 값의 합은 85분이고 평균은 17분이다. 보고할 때는 ‘응답이 확인된 5명의 평균’이라고 쓰고 누락 1건을 남긴다. 85를 6으로 나누지 않는다.

결론의 범위: 평균과 중앙값을 정확히 계산해도 자원 응답 표본이라는 한계는 남는다. 전체 수강생을 설명하려면 조사 대상과 추출 방법을 다시 검토해야 한다. 표본 수를 늘리는 것만으로 모든 선택 편향이 사라진다고 단정하지 않는다.

확인 문제: ① 처음 자료의 평균과 중앙값을 계산하고 분모/가운데 위치를 설명하라. ② 변경 A에서 두 대표값이 어떻게 달라지는가? ③ ‘전체 수강생의 평균이 20분’이라는 결론의 문제를 설명하라. ④ 변경 B에서 누락을 0분으로 처리하면 왜 틀리는가?