사회통계
중심경향 측도

김현우, PhD[1]

[1] 충북대학교 사회학과 조교수

March 13, 2024

진행 순서

1. 요약통계량
2. 자료의 요약(I): 중심경향

요약통계량

요약통계량

자료는 정의상 복수이므로 많은 숫자를 요약해야 할 필요가 있다.

- 상황을 이해하기 위해 실제 자료를 직접 살펴보자.
- eCampus에서 airpollution.zip을 다운받아 압축을 풀자.
- 이 자료는 에어코리아(https://www.airkorea.or.kr)에서 다운받은 2023년 9월까지 우리나라 전역의 대기오염 상태에 관한 자료이다.
- 가장 기본적으로 자료를 요약할때는 파일의 크기를 이야기할 수 있다.
- 2023년 2월.xlsx은 얼마나 큰가? 2023년 6월.xlsx은 얼마나 큰가?
- 2023년 전체에 걸쳐 파일들은 도합 얼마 정도인가?

요약통계량

- 이 자료는 귀엽게도 200 MB에 불과하다. 개별 파일은 30 MB 정도다.
- 오늘날 본격적인 빅데이터 분석에 쓰이는 데이터의 파일 사이즈는 최소한 수 기가바이트(giga byte)에서 수 페타바이트(peta byte)에까지 달한다.
- 가령 미국인구 전체를 포괄하는 마케팅 데이터인 Infutor같은 경우에는 매달 수백 기가짜리 파일이 몇 개씩 생산된다.
- 매우 큰 파일은 그에 적합한 보관, 처리 및 분석 기법을 요구하므로 엑셀로는 다룰 수 없다.
- 우리 수업에서는 (빅데이터는 다루지 않고) 엑셀로 다룰 수 있는 작은 파일들만 분석한다.

요약통계량

실제 자료를 남에게 설명해보자.

- 2023년 3월.xlsx를 엑셀에 불러오자.
- 지난 주에 학습하였듯 자료는 사각형 꼴이다. 좀 더 수학적으로 표현하자면 행렬이다.
- 행은 rows (=극장 따위의 좌석 줄)라고 쓰고 열은 columns (=기둥)이라고 쓴다.
- 변수(variables)는 모두 몇 개인가? 관측치(observations)는 모두 몇 개인가?
- 자료를 남에게 설명할 때 “변수가 몇 개, 관측치가 몇 개” 이런 식으로 표현할 수 있다!

요약통계량

- 기술적으로는 조잡한 표현이지만 현실에서는 종종 이렇게 말한다.

  철수: "영희야, 너희 부서에서 다루는 데이터는 어떻니?"
  영희: "2022년 3월에 약 x메가 짜리 데이터를 생성하고 보관하고 있어."
  철수: "작구나."
  영희: "변수는 y개지만 관측치는 z개가 넘지."
  철수: "어, 다시 들으니 무지 크네."

- 하지만 이런 식으로 자료의 크기를 두고 (본격적인) 요약통계량(summary statistics)이라고는 하지 않는다.

요약통계량

데이터 사이즈는 기하급수적으로 팽창하고 있다!

- 하지만 옛날에는 메가 단위를 엄청 큰 것으로 생각했지만, 지금은 내 손바닥만한 휴대폰도 260 GB를 저장할 수 있다.
- 시대는 급속도로 변하고 있다. 앞으로 더 빨리 변할 것이다.

5MB IBM Hard Drive circa mid-1950s

자료의 요약(I): 중심경향

자료의 요약(I): 중심경향

요약통계량은 자료의 중심이 어디에 있는가를 설명하는 것에서 출발한다.

- "자료를 대표하는(representative) 값은 자료의 가운데 어딘가에 위치해 있다."
- 뭔가 그럴 듯 하지 않나? "자료의 가운데 있는게 대표적이다" 라는 발상!
- 주로 세 가지 통계량이 중심경향(central tendency)을 파악하기 위해 사용된다.

  (1) 평균(mean)
  (2) 중앙값(median)
  (3) 최빈값(mode)

자료의 요약(I): 중심경향

평균은 모든 관측치를 모두 더하고 이를 관측치의 갯수로 나누어준 값이다.

- 이 정의는 사실 좀 더 일반화될 수 있는 여지를 남기고 있다.
- 평균의 다른 이름은 기대값(expected value)이다.
- 모평균(population mean)은 표본평균(sample mean)과 기호가 살짝 다르다.

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1}{N}\left(x_1 + x_2 + \cdots + x_N\right)$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\left(x_1 + x_2 + \cdots + x_n\right)$$

자료의 요약(I): 중심경향

- 앞서 계산한 방식을 산술평균(arithmetic mean)이라고 부르며 가장 폭넓게 쓰인다.
- 하지만 가중평균(weighted mean), 기하평균(geometric mean), 조화평균(harmonic mean)과 같은 색다른 평균 개념들도 있다.
- 가중평균에서 w_i는 i번째 관측치에 대한 가중치(weight)이다.

$$\bar{x} = w_1 x_1 + w_2 x_2 + \cdots + w_N x_N$$

- 산술평균은 다음의 조건이 성립하는 가중평균의 특수 사례에 불과하다(Why?).

$$w_1 = w_2 = \cdots = w_N = \frac{1}{N}$$

- 다른 평균 개념은 교과서를 통해 공부하자.

자료의 요약(I): 중심경향

중앙값은 자료를 크기 순서로 정렬할 때 중앙에 위치한 값이다.

- 일단 자료를 먼저 크기 순서대로 정렬(sort)한다.
- (관측치의 수에 따라 중앙의 위치가 달라지므로) 계산 방식은 관측치의 수가 홀수인 경우와 짝수인 경우가 다르다.

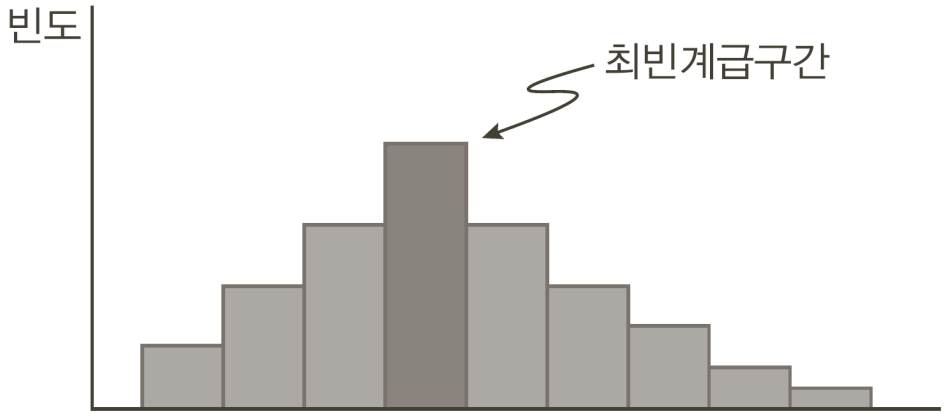
$$M_e = \frac{n+1}{2} \text{ 번째 위치한 값} \qquad \text{(홀수인 경우)}$$

$$M_e = \frac{(n/2) \text{ 번째 위치한 값} + [(n/2)+1] \text{ 번째 위치한 값}}{2} \qquad \text{(짝수인 경우)}$$

자료의 요약(I): 중심경향

최빈값은 자료에서 최고로 빈도가 높은 값이다.

- 최빈값(mode)은 자료에서 최고의 빈도수를 가진 관측치를 알려주기 때문에 아무래도 질적 변수를 요약할 때 특히 유용하다.
- 규모가 큰 양적 자료의 경우 최빈값보다는 구간으로 범주화된 최빈 계급구간(modal class interval)이 더 유용할 수도 있다.



자료의 요약(I): 중심경향

예제 1. 다음과 같은 자료가 주어졌을 때 평균, 중앙값, 최빈값을 각각 구하시오.

ID	AGE
1	31
2	12
3	25
4	25
5	20

자료의 요약(I): 중심경향

- 엑셀에서 함수나 계산은 언제나 =로 시작한다. 함수를 입력할 때는 괄호를 잊지 말 것!
- 평균은 엑셀에서 다음과 같이 입력하여 계산할 수 있다:
  $= (31 + 12 + 25 + 25 + 20)/5$.
- 엑셀 함수로 AVERAGE$(\cdot)$ 를 사용할 수 있다.
- 중앙값을 구하기 위해 자료를 정렬하면 12, 20, 25, 25, 31이다. 여기서 딱 한 가운데 값은 25이고 이것이 중위값이다.
- 엑셀 함수로 MEDIAN$(\cdot)$ 을 사용할 수 있다.
- 12가 한 번, 20이 한 번, 25가 두 번, 31이 한 번 나왔다. 그러므로 최빈값은 25이다.
- 엑셀 함수로 MODE$(\cdot)$ 를 사용할 수 있다.

자료의 요약(I): 중심경향

예제 2. 2023년 3월.xlsx를 다시 여시오. 2023년 3월 13일 14시 충북 청주시의 도시대기를 기준으로 다음의 값들을 모두 구하시오.

(1) 미세먼지(PM10)의 평균
(2) 일산화탄소(CO)의 최빈값
(3) 이산화질소(NO2)의 중앙값

자료의 요약(I): 중심경향

평균, 중앙값, 최빈값 가운데 어느 것을 사용해야 할까?

- 세 값들이 모두 같거나 비슷하면 뭘 써도 상관없다(Why?).
- 일단 평균의 장단점은 비교적 잘 알려져 있다.
- 장점은 평균이 수학적으로 우아하다(elegant)는 사실이다. 다른 요약통계량과는 달리 평균은 간단한 공식을 통해 의미있는 값을 도출해 보인다.
- 반면 평균은 극단값(outliers)에 민감하게 변한다는 단점이 있다.
- 예컨대 자료가 {1, 1, 2, 2, 3, 3, 4, 4, 9999} 와 같이 주어졌다면 평균은 얼마인가? 만일 극단치를 제거하면 평균은 얼마인가? 그 전후 차이는 얼마인가?

자료의 요약(I): 중심경향

평균이 극단값에 의해 쉽게 왜곡된다는 사실은 반드시 기억해야 한다!

- 1985년 UNC-Chapel Hill에서 지리학 학부 전공 초봉 평균은 무려 약 $100,000 였다.
- Michael Jordan이 그 학교 졸업생이였다. 나머지는 걍 지리학과 나와서 흙파먹고 살았다.

통계를 왜곡하는 중인 Michael Jordan

자료의 요약(I): 중심경향

- 2015년 3월에는 우리나라 국회의원이 평균재산이 28억 5천만 원으로 나타나 2012년에 비해 일인당 평균 약 67억원이 줄어들었던 것으로 보도되었다. 갑자기 국회의원들이 청빈을 실천하게 된 것일까?
- 재산이 2조 200억원을 상회하는 정몽준 의원이 2014년 국회의원직을 사임했다.

2조 200억을 흐뭇하게 바라보는 정몽준

자료의 요약(I): 중심경향

중앙값과 최빈값도 완전한 것은 아니다.

- 중앙값과 최빈값은 평균에 비해 극단값에 확실히 덜 민감하다.
- 하지만 중앙값과 최빈값은 평균만큼 수학적으로 우아하지 않다.
- 게다가 최빈값은 아예 값이 없거나 두 개 이상의 값이 나올 수 있다(Why?)!

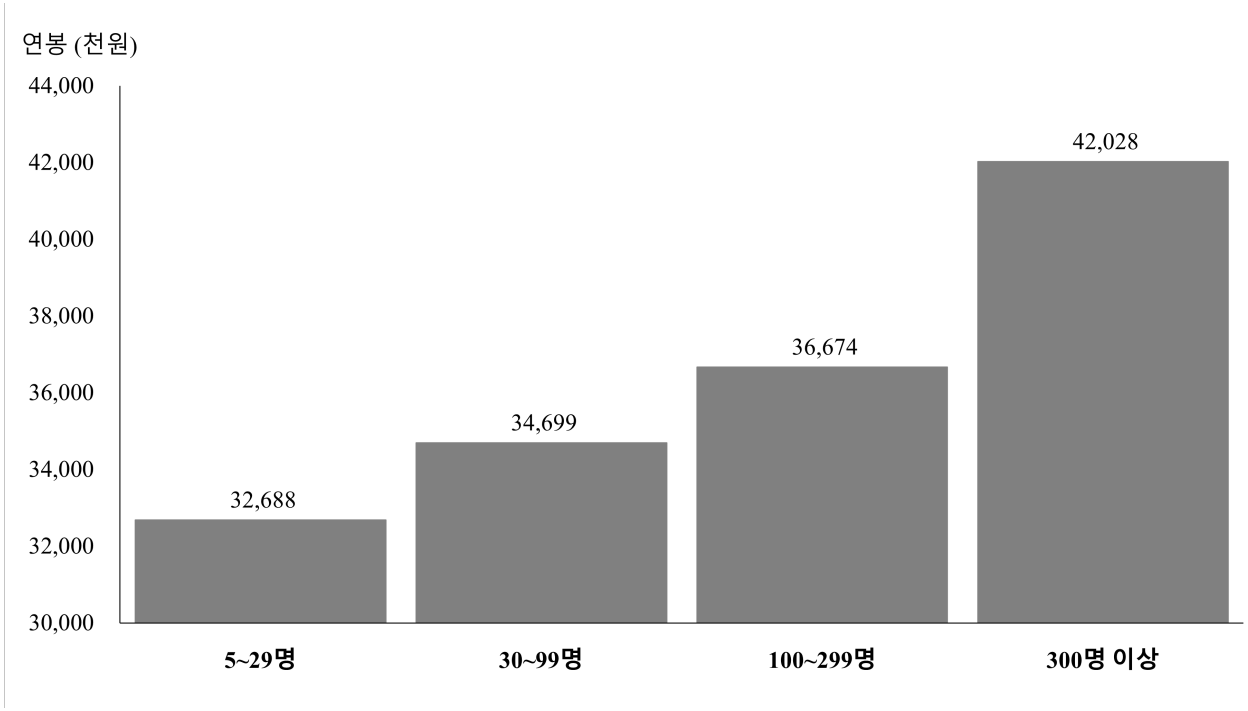
자료의 요약(I): 중심경향

예제 3. coke.csv를 엑셀에 불러오시오. 이 데이터는 10명을 조사하여 이번 한 달동안 얼마나 많은 콜라를 섭취했는가를 파악한 것이다. 평균, 중앙값, 최빈값을 각각 구하시오. 만일 세 요약통계량 중 하나를 선택한다면 그 이유를 설명하시오.

자료의 요약(I): 중심경향

평균은 대중적인만큼 조심히 사용해야 하고 비판적으로 접근해야 한다.

- 만약 2020년 우리나라 대졸 초봉 평균값이 3,536만원이라고 요약한다면, 이 통계는 왜 문제가 되는가?



사업체 규모별 정규직 대졸초임 평균(2023) 고용노동부 임금직무정보시스템

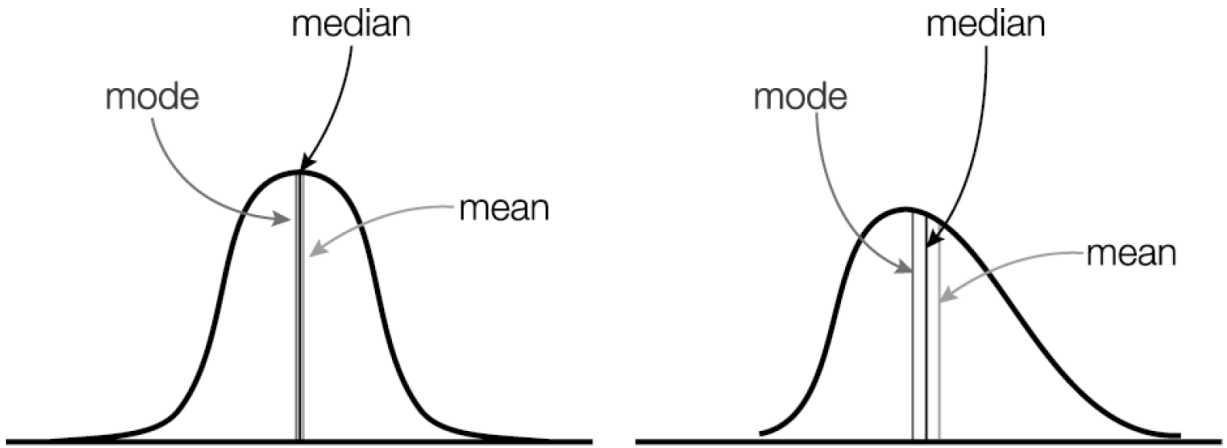
자료의 요약(I): 중심경향

- 우리는 일상에서 너무나 쉽게 "평균소득"이라는 표현을 사용하지만, 평균소득이라는 개념은 현실을 잘 전달하지 못한다.
- 평균소득이 현실을 잘 전달하려면 (1) 중위소득(median income)과 비교하는 맥락에서 쓰이거나, (2) 조직/직무 따위가 동질적이라는 맥락이 먼저 보장되어야 한다 (Why?).
- 그렇기 때문에 제대로 된 통계는 중위소득을 보고하기 마련이다(나중에 "중위소득"을 검색해 보자).
- 늘 요약통계량을 종합적으로 살펴야 하고, 평균값이 중위값이나 최빈치로부터 크게 이탈해 있다면 즉각 통계적 왜곡을 경계해야 한다.

자료의 요약(I): 중심경향

오직 제한된 조건 아래에서만 세 요약통계량이 같다.

- 정규분포(normal distribution)일 때 세 중심경향 요약통계량은 모두 일치한다.
- 정규분포가 흐트러져 오른쪽으로 꼬리가 긴 양(+)의 왜도를 가지면 평균(mean)이 우측으로 이동하고 중앙값(median)도 살짝 따라온다(Why?).



자료의 요약(I): 중심경향

자료유형에 따라 평균이나 중앙값은 사용할 수 없는 경우가 있다.

- 등간척도나 비율척도의 경우에만 평균이 의미를 갖는다(Why?).
- 명목척도인 경우에는 평균 뿐 아니라 중앙값조차 의미가 없다(Why?).
- 연속변수(continuous variable)의 경우에는 (최빈값이 아니라) 최빈계급구간만 의미를 갖는다(Why?).
- 4가지 척도의 구체적인 예를 생각하면서 어떤 요약통계량이 의미있는지 따져보자.