사회통계연습
Measures of Central Tendency

김현우, PhD[1]

[1] 충북대학교 사회학과 조교수

September 10, 2021

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

진행 순서

1. 지난 주 리뷰
2. 요약통계량이란 무엇인가?
3. 데이터의 요약(I): 중심성향(Central Tendency)
4. 중심성향 통계의 한계점

지난 주 리뷰

지난 주 리뷰

퀴즈 1번 문제 코멘트

- 온도(e.g., 체온)은 등간척도임. 섭씨에서는 증류수의 얼음이 언다는 것을 기준으로 삼았으나 측정하는 본질, 즉 "따뜻함"의 0에 대응하지 않음.
- 연령(만나이)이나 소요시간 따위는 비율척도임. 나이 또는 소요시간이 0은 일상적으로는 의미가 와닿지 않지만 수학적으로는 문제가 없음. "당신은 나보다 나이가 두 배 많다"도 make sense. 그런데 의외로 연령은 서열척도라는 대답이 많았음. 유교문화?
- 당신의 학년(1/2/3/4)은 서열척도임. 명목이라는 응답이 좀 많음. 서열(ordinal)은 우열 뿐 아니라 순서(order)라는 의미도 있음. 1학년 다음에야 2학년이 될 수 있음.
- 신입/경력직 여부는 명목척도임. 근데 서열이라고 한 학생들도 좀 있었음. 왜인지 생각해보면 웃픔.
- 선호하는 색깔(빨/파/노)는 명목척도임. 서열이라고 말한 학생이 제법 많음.
- 당신의 사회통계 학점(A/B/C/D)는 서열척도임. 등간이 제법 많음.
- 기업 주가지수는 등간척도임. 지수는 개념상 단지 다른 여러 숫자들을 조합해서 만든 또다른 숫자임. 비율척도라고 하기엔 당연히 0을 가질 수 없음. 일부러 틀리라고 낸 문제인데 맞춘 친구들은 뭐지?

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

지난 주 리뷰

퀴즈 2번 문제 코멘트

- 등간 예제로 "아파트 계단 층 수"를 쓴 친구들이 몇몇 있었음. 어째서 아파트 계단 층 수가 자꾸 나오지 하고 구글링해보니 웬 블로그에 이걸 등간으로 써놓음. 아님. 설령 "층고"를 기준으로 하더라도 1층은 2층 등보다 1.5배 정도 높음. 4층이 없는 경우도 있음. 애초에 1, 2, 3, 4층 할때는 순서를 따지려는 의도(1층을 지나 2층으로 가고 하는 식)이지 정확히 높이를 근거로 층 단위를 측정하는 것이 아님.
- 등간 예제로 리커트 척도를 예제로 든 경우가 있었음. 리커트 척도는 본문 중에 설명이 되었듯 서열임.
- 등간 예제로 수능등급제를 예제로 든 경우가 있었음. 일단 아이디어는 매우 참신함! 이런 아이디어를 떠올린 건 좋은 일. 그런데 안타깝지만 등급에 따라 각 누적백분위가 다름. 일부러 그렇게 만들어 수능등급은 전체 응시자가 정규분포를 만들도록 유도함. 그러므로 서열임.
- 관념(e.g, 시간이나 장소 따위)을 척도의 예제로 든 경우가 소수 있었음. 이것은 측정의 척도(scales of measurement)에 관한 문제임. 즉 100m 주파시간이라던가 유튜브 일별 이용시간을 측정하는 것이지, 시간(Time)이라는 관념이 무슨 척도인가를 고민하는 것은 논점에서 빗나간 것.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

지난 주 리뷰

총평

- 전반적으로 매우 잘했음! 충북대 사회학과 학생들에 대한 기대가 높아짐!
- 한두 문제 맞고 틀리는 것에 집착하지 말 것. 발표 한 번 하면 바로 커버됨.
- 경고: 문제 지우지 말 것. 이번만 넘어감.
- 이제부터 파일 이름에 이름을 쓰지 말 것. 파일 이름에는 몇 주차인지와 학번만 쓸 것. 답안지 맨 끝에만 이름과 학번을 쓸 것.
- 숙제 메일은 hxk271@cbnu.ac.kr로만 보낼 것. 취합 용이.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

지난 주 리뷰

"엑셀이냐! JASP냐!"

- 타학과 다른 교수님들과도 논의를 했고, 심리학과 교수님 한 분은 JASP를 추천함. 하지만 최종적으로 숙고한 끝에 엑셀로 정했음.
- 일단 잡마켓에서 엑셀이 너무 강세. 이번 기회를 놓치면 다음이 어려울 수도 있음. 대부분 4학년 때는 너무 바쁘고 3학년 때 밖에 없는데 그땐 또 그 때대로 전공 듣는다고 바쁠듯.
- 내년부터는 1학기 사회통계에서 엑셀을, 2학기 사회통계연습에서 JASP (또는 R 아니면 Python)을 가르칠 예정.
- 학생들 수요만 있다면 1년 안으로 전산실에서 SPSS나 JASP 무료 특강이라도 할테니 와서 들어라. 그러면 졸업할 때 "엑셀, SPSS, JASP, 그리고 R (또는 Python) 능통"으로 이력서 추가. 근데 올까?
- Excel/SPSS/JASP/R 동아리 만들어서 컴활(2급)/사조사(2급)/데이터분석 준전문가(ADsP)/빅데이터분석기사 자격증 준비해서 따자. 목소리를 모아오면 나도 학과/사회대에 요청하겠음.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

요약통계량이란 무엇인가?

요약통계량이란 무엇인가?

실제 대기오염 데이터를 가지고 연습하자.

- eCampus에서 airpollution.zip을 다운받아서 압축을 풀 것.
- 이 자료는 에어코리아(https://www.airkorea.or.kr)에서 다운받은 2020년 우리나라 전역의 대기오염 상태에 관한 데이터이다.
- 가장 기본적으로 데이터를 요약할때는 파일의 크기를 이야기할 수 있다.
- 2020년 3월.xlsx은 얼마나 큰가? 2020년 6월.xlsx은 얼마나 큰가? 2020년 9월.xlsx은 얼마나 큰가? 대체로 이 파일의 크기는 일정하다고 말할 수 있나?
- 2020년 전체에 걸쳐 파일들은 도합 얼마 정도인가?

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

요약통계량이란 무엇인가?

오늘날 본격적인 빅데이터 분석에 쓰이는 데이터의 파일 사이즈는 최소한 수 기가바이트(giga byte)에서 수 페타바이트(peta byte)에까지 달한다.

- 우리 예제는 귀엽게도 210메가 전후에 불과하다. 개별 파일은 20메가 정도다.
- 이런 장난감 데이터(toy data) 말고, 가령 미국인구 전체를 포괄하는 마케팅 데이터인 Infutor같은 경우에는 2020년 1월에만 수백 기가짜리 파일이 몇 개씩 생산된다.

매우 큰 파일은 그에 적합한 보관, 처리 및 분석 기법을 요구하므로 엑셀로는 불가능!

- 그러나 이 수업은 더이상 이 주제를 다루지 않고 엑셀로 다룰 수 있는 작은 파일들로 예제를 한정함.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

요약통계량이란 무엇인가?

데이터 사이즈는 기하급수적으로 팽창하고 있다!

- 하지만 옛날에는 메가 단위를 엄청 큰 것으로 생각했다.
- 지금 내 손바닥만한 휴대폰은 260 GB를 저장할 수 있다.
- 시대가 급속도로 변했다. 앞으로 더 빨리 변할 것이다.

5MB IBM Hard Drive circa mid-1950s

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

요약통계량이란 무엇인가?

이제 2020년 6월.xlsx를 열자.

- 기본적으로 data는 사각형 꼴이다. 이건 굉장히 중요한 발견이다. 좀 더 수학적으로 표현하자면 행렬(matrix)이다.
- 행은 가로고, 열은 세로다.
- 행은 rows (=극장 따위의 좌석 줄)라고 쓰고 열은 columns (=기둥)이라고 쓴다.
- 변수(variables)는 모두 몇 개인가?
- 관측치(observations; records)는 모두 몇 개인가?

데이터를 남에게 설명할 때 “변수가 몇 개, 관측치가 몇 개” 이런 식으로 표현할 수 있다!

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

요약통계량이란 무엇인가?

우리는 데이터를 종종 크기로 요약한다.

철수: "영희야, 너희 부서에서 다루는 데이터는 어떻니?"
영희: "2020년 12월에 약 20메가 짜리 데이터를 생성하고 보관하고 있어."
철수: "작구나."
영희: "변수는 12개지만 관측치는 300,000개가 넘지."
철수: "어, 다시 들으니 무지 크네."

기술적으로는 대단히 crude (=거칠고 조잡하다)한 표현이지만 현실에서는 종종 이렇게 말한다.

하지만 이런 식으로 데이터의 크기를 두고 (본격적인) 요약통계량이라고는 하지 않는다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

데이터의 요약(I): 중심성향(Central Tendency)

데이터의 요약(I): 중심성향(Central Tendency)

보다 본격적인 요약통계는 데이터의 중심(central)이 어디에 위치하고 있는가를 설명하는 것에서 출발한다.

- 데이터를 대표하는(representative) 값은 데이터의 가운데 어딘가에 위치해 있다라는 아이디어.
- 뭔가 그럴 듯 하지 않나? "데이터의 가운데 있는게 대표적이다" 라는 발상!

주로 세 가지 통계가 중심성향을 파악하기 위해 사용된다.

- 평균(average; mean)
- 중앙치; 중위수; 중위값; 중위치; 중간수, 메디안(median)
- 최빈수; 최빈치; 최빈값(mode)

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

데이터의 요약(I): 중심성향(Central Tendency)

쉽게 말하자면, 평균(mean)이란 모든 관측치(observations)을 모두 더한 뒤, 이를 observations의 갯수로 나누어준 값이다.

- 이 정의는 사실 좀 더 일반화될 수 있는 여지를 남기고 있지만 조금만 있다가 이야기하기로 하자.
- 처음엔 쉬운 게 중요하기 때문이다.

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

데이터의 요약(I): 중심성향(Central Tendency)

만약 데이터가 다음과 같이 주어졌다고 하자.

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

그러면 평균은 (31+12+25+25+20)/5 이다.

엑셀에서 함수나 계산은 언제나 =로 시작한다. 함수를 입력할 때는 괄호를 잊지 말 것.

엑셀 함수로 average() 를 사용할 수 있다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

데이터의 요약(I): 중심성향(Central Tendency)

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

다음으로 중위값(median)은 관측치를 크기 순서대로 정렬(sort)해 놓았을 때 그 수열의 딱 가운데 오는 값이다.

위의 데이터를 먼저 sort하면 12, 20, 25, 25, 31이다. 여기서 딱 한 가운데 값은 25이고 이것이 중위값이다.

엑셀 함수로 median() 을 사용할 수 있다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

데이터의 요약(I): 중심성향(Central Tendency)

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

마지막으로 최빈값(mode)은 데이터에서 최고로 빈도가 높은 값이다.

12가 한 번, 20이 한 번, 25가 두 번, 31이 한 번 나왔다. 그러므로 최빈값은 25이다.

엑셀 함수로 mode() 를 사용할 수 있다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

데이터의 요약(I): 중심성향(Central Tendency)

아까 eCampus에서 다운받은 airpollution.zip의 파일인 2020년 9월.xlsx을 다시 열자.

[데이터]-[필터]를 사용해 다음의 조건을 특정하자: 〈지역〉은 충북 청주시, 〈망〉은 도시대기, 측정일시는 〈2020.09.10.11〉.

필터링된 자료를 복사하여 다른 새 탭에 붙여넣고 작업하자.

- Q1. Find the mean of CO.
- Q2. Find the mode of PM10.
- Q3. Find the median of NO2.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

중심성향 통계의 한계점

중심성향 통계의 한계점

eCampus에서 coke.csv를 다운받아서 열자. 이 데이터는 10명을 조사하여 이번 한 달동안 얼마나 많은 콜라를 섭취했는가를 파악한 것이다.

Q1. Find the mean.

Q2. Find the median.

Q3. Find the mode.

Q4. 여러분의 보스는 둘 이상의 숫자는 싫어한다. 이 많은 숫자들(=데이터)을 요약하는 단 하나의 숫자가 무엇인지를 물었다. 어떻게 답할 것인가?

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

중심성향 통계의 한계점

셋 중 어느 것이 옳은가?

- 평균(mean), 중위값(median), 최빈값(mode)이 모두 같거나 비슷하면 뭘 써도 상관없다.
- 하지만 꼭 기억해야 한다. 평균(mean), 중위값(median), 최빈값(mode)이 모두 다르고 심지어 별로 비슷하지도 않을 수도 있다!

일단 평균에는 명확한 장단점이 있다는 것이 잘 알려져 있다.

- 장점은 평균이 수학적으로 우아하다(elegant)는 사실이다.
- 다른 요약통계량과는 달리 평균은 간단한 공식을 통해 의미있는 값을 도출해 보인다.
- 단점은 극단값(outliers; extremes)에 민감하다는 사실이다.
- 예컨대 자료가 1, 1, 2, 2, 3, 3, 4, 4, 9999 하고 주어졌다면 평균은?

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

중심성향 통계의 한계점

평균이 극단값(outliers; extremes)에 의해 쉽게 왜곡(bias)된다는 사실은 반드시 기억해야 한다!

1985년 UNC-Chapel Hill에서 Geography 전공한 졸업생의 초봉 데이터를 요약하는 평균값은 무려 약 $100,000 였다.

- Michael Jordan이 그 학교 졸업생이였다. 나머지는 걍 지리학과 나와서 흙파서 먹었다.

통계를 왜곡하는 중인 Michael Jordan

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

중심성향 통계의 한계점

2015년 3월에는 국회의원이 평균재산이 28억 5천만 원으로 보도되어, 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

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

중심성향 통계의 한계점

또다른 데이터가 다음과 같이 주어져 있다: [0, 1, 2, 3, 4, 199, 199]

- Q1. 이 많은 숫자들(=데이터)의 평균은 얼마인가?
- Q2. 이 많은 숫자들(=데이터)의 최빈치는 얼마인가?
- Q3. 이 많은 숫자들(=데이터)의 중위값은 얼마인가?
- Q4. 당신의 직관에 따른다면 이 많은 숫자들(=데이터)를 잘 대표하는 숫자는 무엇인가?

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

중심성향 통계의 한계점

coke.csv를 다시 열자. 이 데이터는 10명을 조사하여 이번 한 달동안 얼마나 많은 콜라를 섭취했는가를 파악한 것이다.

Q1. 여러분의 보스는 둘 이상의 숫자는 싫어한다. 이 많은 숫자들(=데이터)을 요약하는 단 하나의 숫자가 무엇인지를 물었다. 어떻게 답할 것인가?

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

중심성향 통계의 한계점

Q1. 만약 2020년 2월 기준으로 우리나라 대졸 초봉은 3,382만원이라고 요약한다면, 이 요약통계량은 왜 문제가 되는가?

충청신문 2021년 3월 4일자에 따르면 잡코리아 샘플 우리나라 대기업 대졸 초봉의 평균은 4,124만원이다. 중소기업 대졸 초봉의 평균은 2,793만원이다.

- https://www.dailycc.net/news/articleView.html?idxno=636669
- Q2. 왜 이렇게 나누어 보고하고 있다고 생각하나?

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

중심성향 통계의 한계점

Take-away 1. 데이터를 요약하는 첫 번째 방법은 mean, median, mode로 요약하는 것이다.

Take-away 2. 평균은 우아하다. 하지만 극단치에 쉽게 끌리고 마는 약점이 있다.

- "평균소득"이라는 개념은 기본적으로 사기다. 평균소득이 make sense 하려면 (1) 중위소득과 비교하는 맥락에서 쓰이거나, (2) 조직/직무 따위가 동질적이라는 맥락이 먼저 보장되어야 한다.
- 하지만 우리는 일상에서 너무나 쉽게 평균소득이라는 말을 한다.
- 그렇기 때문에 제대로 된 통계는 중위소득(median income)을 보고한다. 당장 인터넷 브라우저에 "중위소득"이라고 검색해 보라.
- 평균값이 중위값이나 최빈치로부터 크게 이탈해 있다면 즉각 경계해야 하고 요약통계량을 종합적으로 살펴야 한다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY