사회통계연습
표본과 표집분포

김현우, PhD[1]

[1] 충북대학교 사회학과 조교수

October 8, 2021

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

진행 순서

1. 지난 주 리뷰
2. 표본의 추출
3. 표본평균의 표집분포
4. 표본비율의 표집분포

지난 주 리뷰

지난 주 리뷰

퀴즈 #5 코멘트

- **문제 2.** P국가의 국회의원 선거통계 자료집에 따르면, 선거구별 투표율의 평균은 65%였고 표준편차는 6%였다. 전문가에 따르면 투표율은 정규분포한다. 만일 무작위로 뽑은 한 지역의 투표율이 70% 이상 80% 미만일 확률은 얼마인가?
- 드물게 $P(70 \leq X < 80) = P(X < 80) - P(70 \leq X)$으로 계산하는 경우가 있는데 그게 아니라 $P(70 \leq X < 80) = P(X < 80) - P(X \leq 70)$.
- **문제 3.** K시 지역통계에 따르면 해당 지역의 소규모 상가의 평균 임대료는 제곱미터당 12,000원이고 표준편차는 1,500원이며, 임대료는 정규분포를 따른다. 다파라 부동산 중개인은 임대료의 상위 30%와 상위 10% 사이의 물건을 보고 싶어하는 어느 고객을 맞이하였다. 이 고객이 지불하게 될 임대료는 얼마와 얼마 사이인가?
- 가장 흔하게 틀리는 유형은 NORM.INV(0.7, 0, 1) 대신에 NORM.INV(0.3, 0, 1)을 입력한 경우였다. 이건 하위 30%! 애초에 이렇게 입력하면 (상위 30%를 물어봤는데도) 원점수가 평균보다 낮게 나온다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

지난 주 리뷰

퀴즈 총평

- 숙제를 일단 하면 전반적으로 잘했다. 하나 정도 틀리는 경우가 가끔씩 있는데 둘씩 짝지어 똑같은 방식으로 틀리는 걸로 봐서 잘못 상의를 하는 것 같다...
- 혼동할 것을 염려하여 일부러 이야기하지 않은 방식의 그래프인 누적밀도함수(cumulative density function)를 그려서 제출한 경우도 있었다(당연히 okay).
- 숙제 답안은 엑셀 파일 하나로만 주세요. 숙제 안에도 이름은 써주세요. 숙제 파일 이름에 주차와 학번은 써주세요(e.g., 3주차 20201234.xlsx).

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

표본의 추출

표본의 추출

일엽지추(一葉知秋)

- “뜰 안에 잎이 하나 떨어지는 것을 보아 온 천하에 가을이 왔음을 미루어 안다.” ≪회남자(淮南子)≫ 〈설산훈편(說山訓篇)〉
- 무언가를 알기 위해 (설령 전체를 모두 살펴보지 않아도) 부분을 통해 미루어 짐작할 수 있다.
- 지식의 획득에도 어느 정도 경제적 논리가 작동한다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

표본의 추출

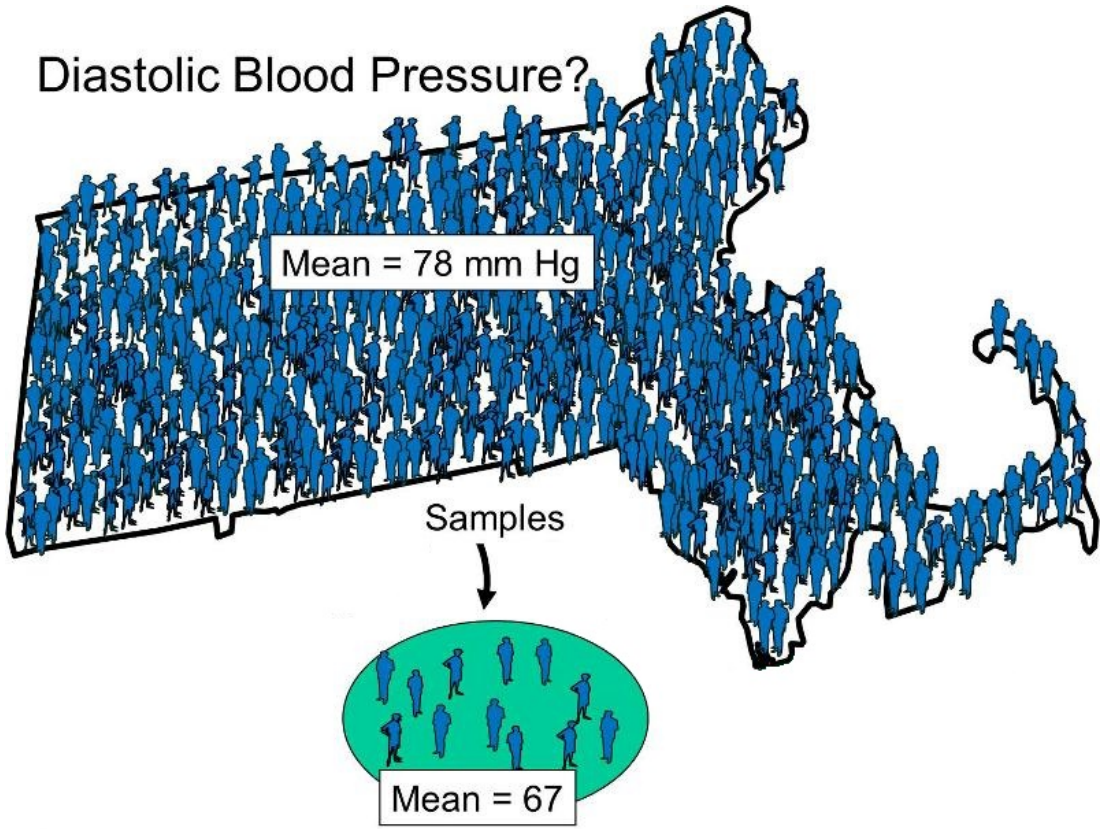
모집단(population)과 표본(sample) 그리고 모수(parameter)와 통계량(statistic)을 짝지어 이해하자.

- 주희는 인구가 72 만명인 네오청주에서 의료기기 스타트업을 운영한다. 사업 기획상 네오청주 시민의 "(이완)혈압의 평균"을 알고 싶다.
- 이때, 네오청주의 모든 시민은 모집단(population)이 된다.
- 그러나 일개 스타트업 사장인 주희 입장에서 모든 네오청주 전체 시민(=모집단)을 조사하기엔 시간과 비용을 감당할 수 없다.
- 대신 주희는 네오청주 전체 시민 중 일부만을 랜덤하게 골라 표본(sample)을 싼값으로 추출할 수 있다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

표본의 추출

- 네오청주의 모든 시민의 "혈압의 평균"은 모수(parameter)라고 부른다. 모수인 "혈압의 평균"은 그 값을 알 수 없지만 상수(constant)이다. 왜냐하면 답은 이미 정해져 있기 때문이다.
- 표본에서 얻은 네오청주 시민의 "혈압의 평균"은 통계량(statistic)이라고 부른다. 통계량은 그 값이 모집단으로부터 랜덤하게 추출된 표본(sample)에 따라 그때그때 달라지는 확률변수(random variable)이다. 상수가 아니라는 말이다.



충북대학교
CHUNGBUK NATIONAL UNIVERSITY

표본의 추출

표집(sampling)이란 표본의 추출을 의미한다.

- 주희는 투망을 쳐서 네오청주에 살던 시민 30명을 붙잡아 표본(sample)을 추출하였다.
- 그 30명의 혈압을 조사하여 "표본(sample)의 평균(mean)"을 계산해 보았더니 67 mmHg라는 평균값을 얻었다.
- 이 67 mmHg라는 추정값(estimate)은 주희가 붙잡은 네오청주 시민의 표본(sample)으로부터 얻은 값에 지나지 않으며 표본을 달리 뽑으면 그때그때 달라질 수 밖에 없는 값이다.

표집(sampling)은 그 자체로 상당히 까다로운 테크닉이다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

표본의 추출

다음 학기 사회조사방법론 수업에서 표집(sampling)을 좀 더 배우게 된다.

- 표집에는 복원 방식에 따라 크게 두 가지 방식이 있다: 복원추출(sampling with replacement)과 비복원추출(sampling without replacement).
- 표본의 추출 과정에서 오차(error)가 생기면 표본은 모집단을 대표하지 못하게 된다.
- 표집과정에서 오차는 크게 표집오차(sampling error)와 비표집오차(non-sampling error)로 구분된다.
- 표집오차(sampling error)는 "표본의 선정 과정에서 생기며" 모집단을 대표하는 전형적인 구성요소를 표집하지 못하여 생기는 오차이다. 이는 다시 우연(by chance)에 의한 오차와 어긋남(bias)에 의한 오차로 구분된다. 우연에 의한 오차는 표본 크기(sample size)를 늘리면 자연스럽게 해소된다. 어긋남에 의한 오차는 모집단의 특정 성격이 체계적으로 과대 또는 과소대표되어 발생한다. 이는 표본 크기를 늘린다고 해소되지 않는다.
- 비표집오차(non-sampling error)는 "표본의 선정 과정 이외에서 생기며" 특히 조사자 교육, 응답자의 피로 등 실제 조사의 상황과 관련하여 생기는 문제인 측정오차(measurement error)가 중요하다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

표본의 추출

통계 분석에서는 특정한 수준 이상의 데이터의 질을 전제로 한다.

- 우리가 사회통계연습을 공부하는 단계에서는 전문적인 조사의 질 관리를 통해 주어진 데이터 속 비표집오차가 아주 작다고 전제한다.
- 데이터는 임의(random)로 추출되어 모집단의 대표성이 있다(representative)고 전제한다.
- 물론 이것은 현실과는 다르다. 현실에서는 random sampling이 대단히 어렵다. 여러가지 구체적인 대안과 실행 절차 및 장단점에 대해서는 다음 학기에 사회조사방법론을 통해 학습한다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

표본의 추출

어떤 의미에서 사회통계학에서 표집의 논리는 점점 낡은 것이 되어가고 있다.

- 오늘날 빅데이터는 모집단(population) 자체를 대상으로 하는 경우가 많다. 그런 경우 추리(inference)의 문제가 예전과는 크게 달라진다.
- 게다가 오늘날에는 새로운 형태의 계량적 사회과학이 출현하고 있고 있는 중이다. 이들에 따르면 우리가 배우고 있는 추정의 논리는 빈도주의자 추리(frequentist inference) 방식에 불과하다.
- 새로운 계량적 사회과학에서는 베이즈 통계, 인과추정, 예측, 시각화(visualization), 코딩(R 또는 Python), 기계학습(machine learning) 등을 더 강조하고 있다. 사회과학적 실험 방법의 재부상도 주목할 만 하다.
- 어쩌면 몇 년 안으로 지금과 같은 커리큘럼은 낡은 것이 될지도 모른다.
- "뭐 그래도 몇 년 안에 다들 졸업 하겠죠? 일단 얼른 취업해서 필요하면 그때 재교육 받죠!"

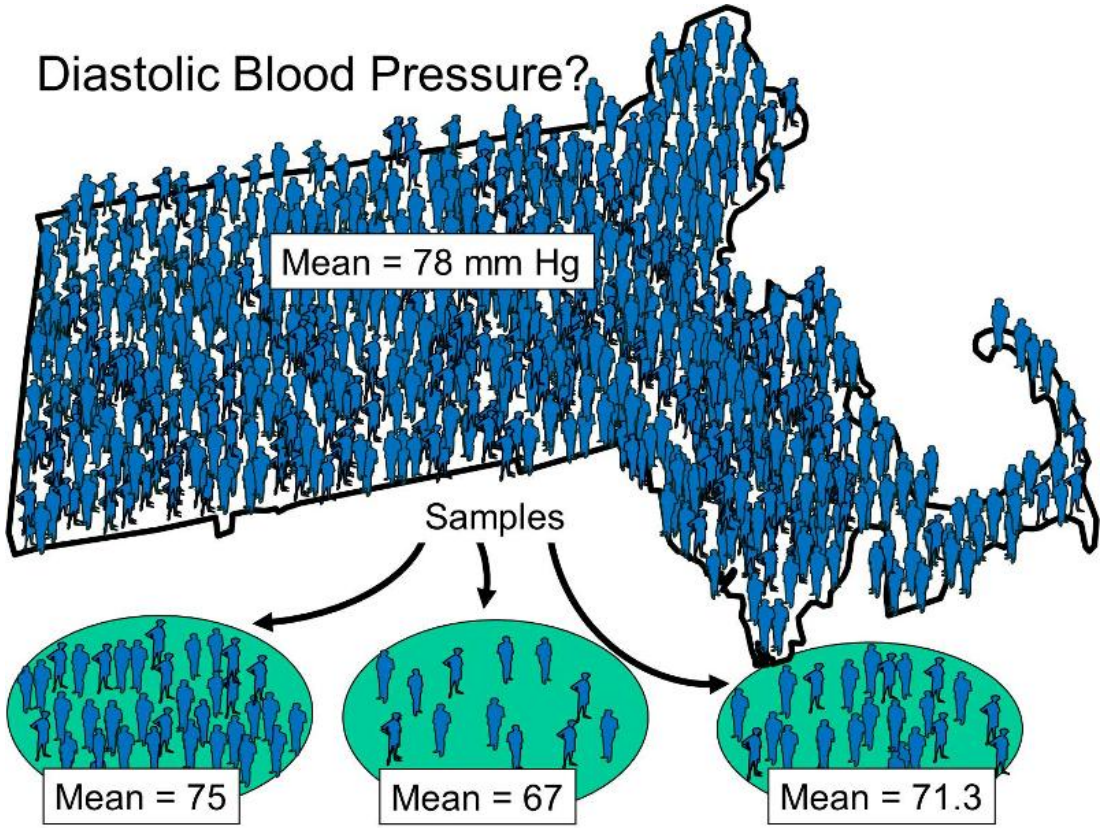
충북대학교
CHUNGBUK NATIONAL UNIVERSITY

표본평균의 표집분포

표본평균의 표집분포

"너의 표본은..."

- 그런데 사실 신영은 같은 동네에서 이미 3년째 의료기기 가게를 운영해 왔다. 신영은 경쟁자인 주희의 조사에 영 믿음이 가지 않았고, 네오청주의 또다른 30명을 붙잡아 새로 샘플을 뽑아 혈압을 쟀더니 이번엔 75mmHg라는 값을 얻었다.
- 민기도 배가 아파 의료기기 사업을 시작했다. 그는 또 새로운 30명의 샘플을 뽑아 혈압을 조사해 이번엔 71.3mmHg라는 결론을 얻었다.



충북대학교
CHUNGBUK NATIONAL UNIVERSITY

표본평균의 표집분포

상식적으로 생각해서 72 만명이 거주하는 네오청주에서 표본을 세 번 뽑아봤는데 그 추정값(estimates)이 정확히 똑같을리가 없다.

- 모수(parameter)는 상수이다. 다만 모를 뿐이다. 추정값(statistic)은 알 수 있지만, 그때그때 표본에 따라 달라지는 확률변수이다.
- 그러면 표본을 통해서는 결코 모수를 알 수 없는 것일까? 다행히 이 문제에 대한 "이론적인 답"이 있고, 그에 따르면 "알수 있다" 이다. 이 논리를 이해하려면 이제부터 상상력이 중요하다!
- 72 만명이 거주하는 네오청주에서 "무한히" 30명 짜리 표본을 뽑아보자.
- 여기서 중요한 건, 똑같은 크기(e.g., 30명)의 표본을 추출하고, 또 추출하고, 또 추출하고... 이 짓을 무한히 반복하는 것이다.
- 그러면서 첫번째 표본에서 평균(sample mean)을 구하고, 두번째 표본에서 평균(sample mean)을 구하고, 또 세번째 표본에서 평균(sample mean)을 구하고... 이 짓도 무한히 반복하는 것이다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

표본평균의 표집분포

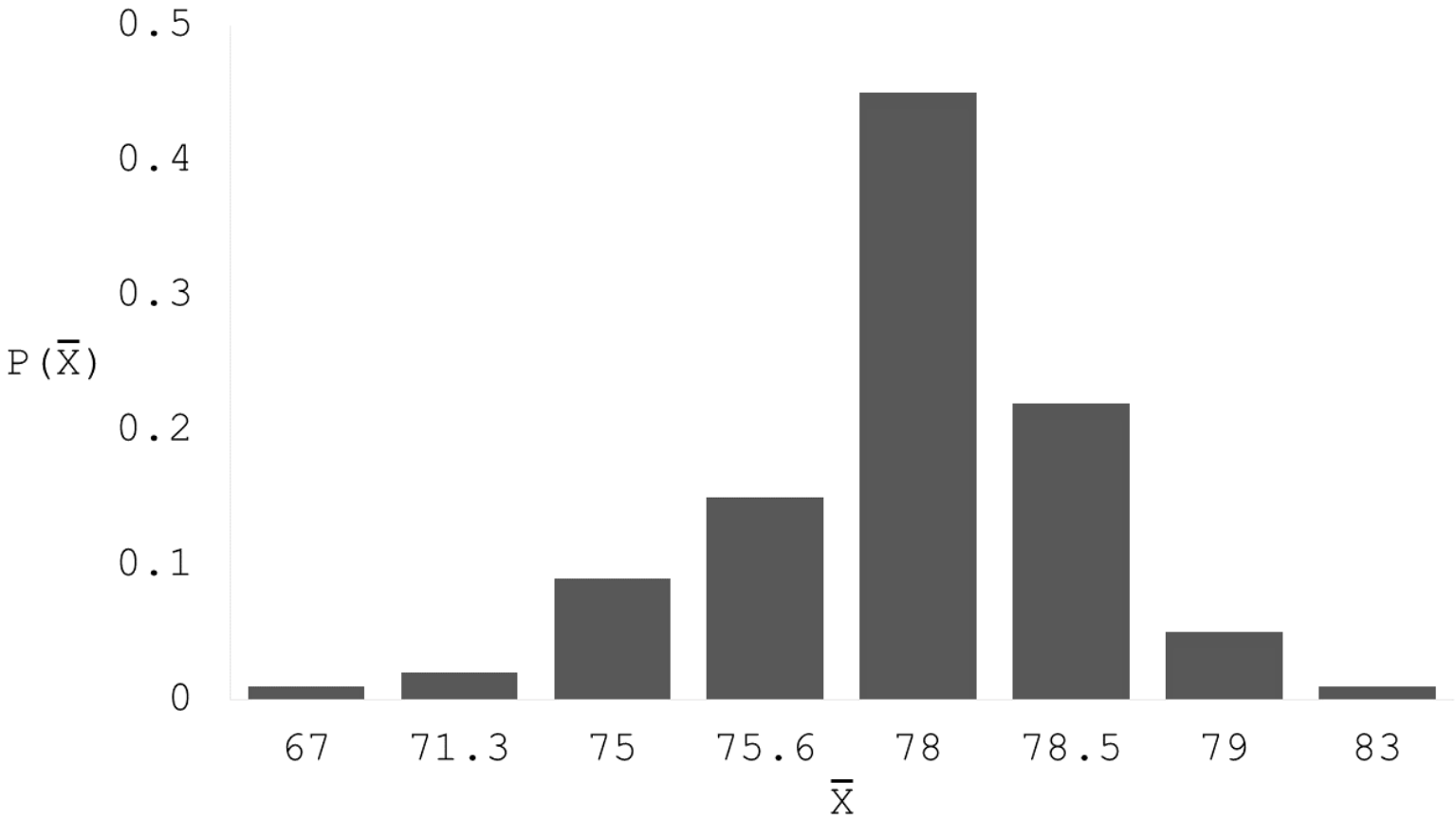
- (무한히 뽑던 도중) 딱 100번째로 표본을 뽑았을때 그동안의 표본평균들(sample means)을 구해서 확률분포(probability distribution)를 한 번 그려본 예제이다.

$\bar{X}_i$	$P(\bar{X}_i)$	$\bar{X}_i \cdot P(\bar{X} = \bar{x}_i)$	$(\bar{X}_i - \mu)^2 \cdot P(\bar{X} = \bar{x}_i)$
67	1/100	0.67	1.07
71.3	2/100	1.43	0.73
75	9/100	6.75	0.49
75.6	15/100	11.34	0.45
78	45/100	35.10	0.20
78.5	22/100	17.27	0.30
79	5/100	3.95	0.14
83	1/100	0.83	0.32
	1	77.34	3.70

충북대학교 CHUNGBUK NATIONAL UNIVERSITY

표본평균의 표집분포

- 물론 100번째까지만의 표본평균들(sample means)을 가지고 확률질량함수 (probability mass function; PMF) 그래프도 그려볼 수 있다.



충북대학교 CHUNGBUK NATIONAL UNIVERSITY

표본평균의 표집분포

무한히 뽑은 나의 작고 소중한 표본평균들(sample means)...

- 무한한 시간을 버틴 우리에겐 이제 무한히 많은 표본평균들(sample means)이 주어질 것이다.
- 이 수많은 표본평균들(sample means)을 가지고 확률밀도함수(probability density function; PDF)를 그릴 수 있다.
- 이런 확률분포를 표본평균의 표집분포(sampling distribution of the sample mean)라고 부를 수 있는데, 특별히 표집분포(sampling distribution)라고 줄여 말한다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

표본평균의 표집분포

표집분포(sampling distribution)에는 두 가지 중요한 특징이 있다.

- 우리에게 주어진 무한히 많은 표본평균들을 가지고, 이 표본평균들(sample means)의 평균(mean)을 계산할 수 있다. 말하자면 이는 "평균들의 평균"이다.
- 그런데 이 표본평균들의 평균은 모집단의 평균(population mean)에 무한히 근접한다(=일치한다)는 것은 증명될 수 있다.

$$\mathrm{E}(\bar{\mathrm{X}}) = \mu$$

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

표본평균의 표집분포

- 우리에게 주어진 무한히 많은 표본평균들을 가지고, 이 표본평균들(sample means)의 분산(variance)도 계산할 수 있다. 말하자면 이는 평균들의 분산이다.
- 이 표본평균들의 분산은 모집단의 분산을 표본수(여기서는 N=100)로 나눈 값에 무한히 근접한다(=일치한다)는 것이 증명될 수 있다.

$$\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$$

- 이 표본평균들의 표준편차(standard deviation)를 특별히 표준오차(standard error; SE)라고 부른다.

$$\mathrm{se}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

- 사실 표준오차(standard error)는 표본평균(sample means)을 이용해서 의사결정을 할 때 예상되는 오류의 크기를 나타낸다(Why?).

충북대학교 CHUNGBUK NATIONAL UNIVERSITY

표본평균의 표집분포

"그리고 여기 통계학사상 가장 위대한 발견이 있다."

- 모집단의 분포가 "어떤 꼴이든 상관없이"
- 표본 크기(sample size)가 충분히 크면(대략 30정도)
- 표집분포(sampling distribution), 즉 표본평균들(sampling means)의 확률분포는
- 평균이 μ이고 표준편차가 $\frac{\sigma}{\sqrt{n}}$ 인 정규분포(normal distribution)에
- 근사한다(approximate).

이것은 중심극한정리(central limit theorem)라고 불리우며 통계학사상 가장 위대한 발견이다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

표본평균의 표집분포

"왜 그토록 위대한가?"

- "모집단의 분포가 어떻게 생겼든 상관없이" 그것의 (무한히 많은) 표본($n > 30$)의 평균들은 반드시 정규분포함을 증명해 보였기 때문이다.
- 다시 말해, 정규분포에 관한 성질을 가지고 "모든" 표본평균들의 분포를 설명할 수 있다. 그게 무엇에 관한 모집단이건간에!
- 그런데 옛날 교과서 중에는 "모집단(population)이 정규분포인가 정규분포가 아닌가"를 구태여 나누어 설명하는 경우도 있다. 그런 것은 중심극한정리(central limit theorem) 덕택에 사실 아무런 의미도 없다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

표본평균의 표집분포

예제 1. 정부의 최근 발표에 따르면 네오청주의 대학생은 평균적으로 28,650천원의 학자금 대출 부채를 가지고 졸업한다. 그 표준편차는 7,000천원인데, 자산소득 양극화로 인해 전혀 정규분포는 아니라고 한다. 도균은 기말과제를 수행하기 위해 네오청주에서 30명의 랜덤한 대학생에게 학자금 대출액을 물어보았다.

- 도균의 표집분포는 어떤 형태를 갖게 될 것인가?
- 이 표집분포의 평균과 표준오차(standard error)를 구하시오.

$$\mathrm{E}(\bar{\mathrm{X}}) = 28650, \mathrm{se}(\bar{\mathrm{X}}) = \frac{7000}{\sqrt{30}}$$

- 도균이 물어본 30명의 평균 부채가 27,000천원보다 클 확률은 얼마인가?

$$\mathrm{P}(\bar{\mathrm{X}} > 27000) = \mathrm{P}\left(\mathrm{Z} > \frac{27000 - 28650}{7000/\sqrt{30}}\right) = 1 - \mathrm{P}(\mathrm{Z} < -1.29) = 0.90$$

예전에는 $\mathrm{Z} = \frac{\mathrm{X}-\mu}{\sigma}$ 였지만 지금은 $\mathrm{Z} = \frac{\mathrm{X}-\mu}{\sigma/\sqrt{\mathrm{n}}}$ 이다(Why?)

충북대학교 CHUNGBUK NATIONAL UNIVERSITY

표본비율의 표집분포

표본비율의 표집분포

아까는 표본평균(sample mean)을 살펴보았고 이번엔 표본비율(sample proportion)을 살펴보자.

- 아까 표본평균을 했는데 왜 표본비율은 또 할까?
- 표본평균은 숫자형(numerical) 척도로 측정된 변수에 대해서만 의미를 갖는다. 예컨대 숫자형 척도인 키, 몸무게의 평균은 의미를 갖지만, 범주형(categorical) 척도인 성별(0=남자; 1=여자), 인종(1=백인; 2=흑인; ...), 종교(0=없음; 1=기독교; 2=불교; ...)의 평균에는 의미가 없다.
- 반면 범주형(categorical) 척도는 비율이 의미를 갖는다. 예컨대 여성의 비율, 백인의 비율, 기독교의 비율 하는 식으로.
- 표본평균은 정규분포(normal distribution)에 직결되어 있으나, 표본비율은 이항분포(binomial distribution)에 직결되어 있다

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

표본비율의 표집분포

다행스럽게도 비율의 표집분포 논리도 결국 똑같다!

- 최근 네오청주의 취업난이 심상치 않다. 민녕은 네오청주의 실업율을 알아보기 위해 20세 이상 구직자 가운데 랜덤하게 50명의 표본을 뽑았고 그 중 30명이 실직 상태임을 확인하였다.
- 민녕의 표본에서 성공 비율 p는 다음과 같이 정의된다. 이때, X는 표본에서 성공 횟수이고 n은 표본 크기이다.

$$p = \frac{X}{n}$$

- 민녕은 이제 일생을 걸고 무한히 많은 50명 짜리 표본을 뽑고 있다. 매번 50명 짜리 표본을 뽑을때마다 위의 p값을 계산했다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

표본비율의 표집분포

- (무한히 뽑던 도중) 민녕이 딱 100번째로 표본을 뽑았을때 그 동안 표본비율들(sample proportions)을 구해서 확률분포(probability distribution)를 한 번 그려본 예제이다.

X (50명 중 관찰된 실직자 수)	p (관찰된 비율)	P(p) (확률)
26	0.52 (=26/50)	0.06
27	0.54 (=27/50)	0.08
29	0.58 (=29/50)	0.11
30	0.6 (=30/50)	0.11
31	0.62 (=31/50)	0.11
32	0.64 (=32/50)	0.10
33	0.66 (=33/50)	0.08

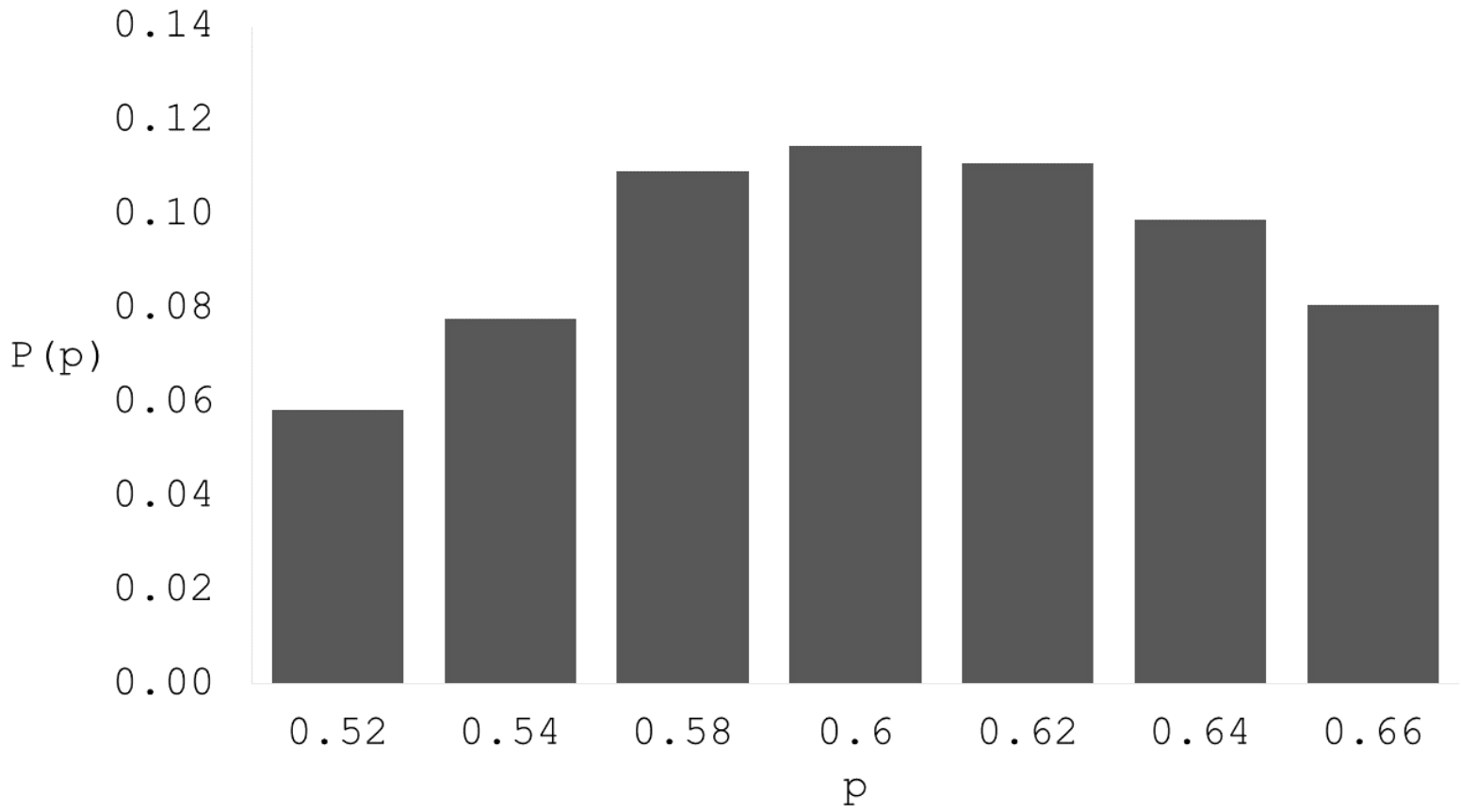
- 수학에 강한 민녕은 P(p)을 구할 때 물론 저번 주에 배운 이항분포 확률질량함수(PMF) 공식을 사용했다.

$$P(p) = \binom{N}{x} \cdot \frac{30}{50}^{x} \left(1 - \frac{30}{50}\right)^{N-x}$$

충북대학교 CHUNGBUK NATIONAL UNIVERSITY

표본비율의 표집분포

- 민녕은 100번째까지만의 표본비율들(sample proportions)을 가지고 확률질량함수(probability mass function; PMF) 그래프도 그려보았다.



충북대학교
CHUNGBUK NATIONAL UNIVERSITY

표본비율의 표집분포

무한히 뽑은 민녕만의 작고 소중한 표본비율들(sample proportions)...

- 무한한 시간을 버틴 민녕에겐 이제 무한히 많은 표본비율들(sample proportions)이 주어질 것이다.
- 이 수많은 표본비율들(sample proportions)을 가지고 확률밀도함수(probability density function; PDF)를 그릴 수 있다.
- 이런 확률분포를 표본비율의 표집분포(sampling distribution of the sample proportion)라고 부르고, 이것도 표집분포(sampling distribution)라고 줄여 말할 수 있다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

표본비율의 표집분포

비율의 표집분포에는 두 가지 중요한 특징이 있다.

- 우리에게 주어진 무한히 많은 표본비율들의 평균은 모집단의 비율(population proportion)에 무한히 근접한다(=일치한다).

$$E(p) = \pi$$

- 우리에게 주어진 무한히 많은 표본비율들의 분산(variance)은 모집단의 분산을 "표본수의 제곱"으로 나눈 값에 무한히 근접한다(=일치한다).

$$Var(p) = \frac{n\pi(1-\pi)}{n^2} = \frac{\pi(1-\pi)}{n}$$

- 이 표본비율들의 표준편차(standard deviation)도 표준오차(standard error)라고 부른다.

$$se(p) = \sqrt{\frac{\pi(1-\pi)}{n}}$$

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

표본비율의 표집분포

위대한 중심극한정리는 여기서도 적용된다.

- 표본비율들의 평균들은 표본 크기(sample size)가 커질수록 평균이 π이고 표준편차가 $\pi(1-\pi)/n$인 정규분포에 무한히 근사한다는 것이 수학적으로 증명되어 있다.
- 그런데 하나 중요한 부분! 중심극한정리를 비율 자료에 대해 적용하려면 표본 크기(N) 뿐 아니라 발생 확률인 p도 고려해야 한다. 즉, $np \geq 5$와 $n(1-p) \geq 5$ 일때 정규분포근사가 정당화된다.
- 다시 말해, 발생 확률이 너무 희박하거나 반대로 너무 흔하면 정규분포를 사용하기 위해 표본 크기가 엄청 커져야 한다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

표본비율의 표집분포

예제 2. 네오청주에 HQ를 두고 있는 IT기업들의 55%가 작년에 사이버 공격을 당했다고 한다. 그게 정말일까 하고 궁금했던 준필은 100개의 네오청주 시내 IT기업들을 랜덤하게 방문해 혹시 사이버 공격을 당했는지 여부를 물어보았다.

- 사이버 공격을 당했던 비율의 표집분포는 어떤 형태를 갖게 될 것인가?
- 표본비율의 기댓값과 표준오차는 얼마인가?

$$\mathrm{E}(\mathrm{p}) = .55, \mathrm{se}(\mathrm{p}) = \sqrt{\frac{.55(1-.55)}{100}}$$

- 준필이 물어본 100개 기업 가운데 사이버 공격을 당했다고 응답한 표본비율이 60% 보다 클 확률은 얼마인가?

$$\mathrm{P}(\mathrm{p} > .6) = \mathrm{P}\left(\mathrm{Z} > \frac{.6-.55}{\sqrt{.55(1-.55)/100}}\right) = 1 - \mathrm{P}(\mathrm{Z} < 1.005) = 0.157$$

예전에는 $\mathrm{Z} = \frac{\mathrm{X}-\mu}{\sigma}$ 였지만 지금은 $\mathrm{Z} = \frac{\mathrm{p}-\pi}{\pi(1-\pi)/\sqrt{\mathrm{n}}}$ 이다(Why?)

충북대학교 CHUNGBUK NATIONAL UNIVERSITY