사회통계
표집분포

김현우, PhD[1]

[1] 충북대학교 사회학과 조교수

진행 순서

1. 표본의 추출
2. 표본평균의 표집분포
3. 표본비율의 표집분포

표본의 추출

표본의 추출

일엽지추(一葉知秋)라는 고사성어가 있다.

- “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.

표본의 추출

모집단과 표본 그리고 모수와 통계량을 짝지어 이해하자.

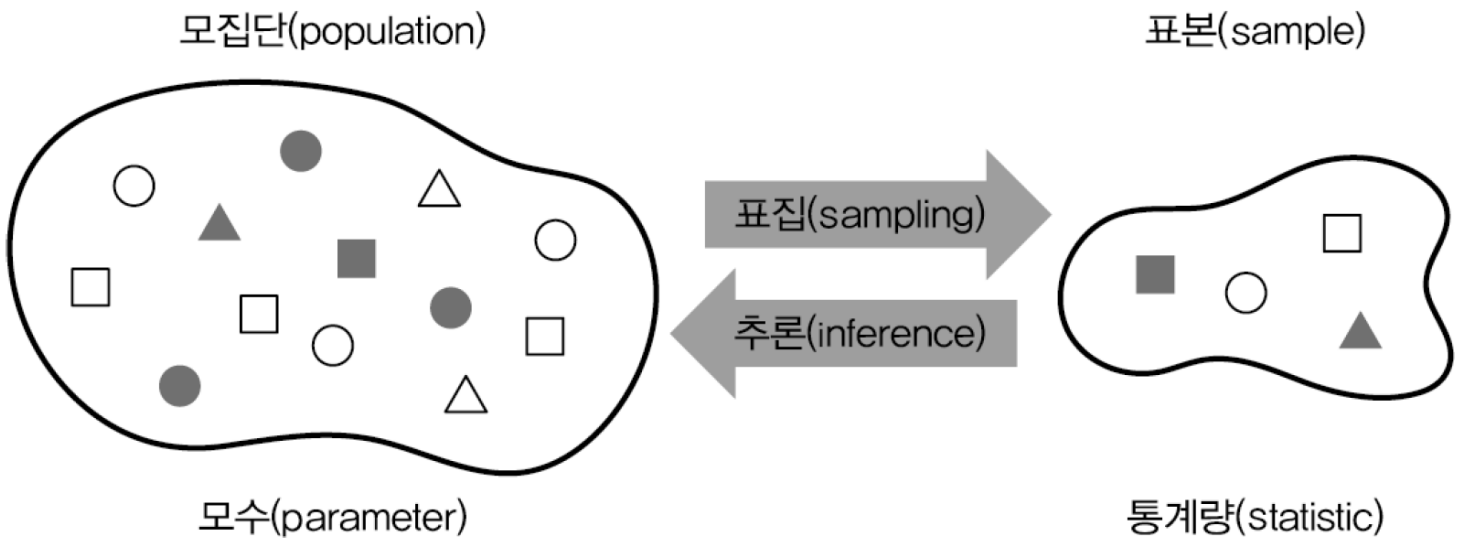
- 고래는 청주에서 의료기기 스타트업을 운영하고 있다. 고래는 지금 시장조사를 위해 모든 청주시민의 “(이완)혈압의 평균”을 알고 싶어 한다.
- 이때, 청주의 모든 시민은 모집단(population)이 된다.
- 스타트업 CEO인 고래 입장에서 청주 전체 시민(=모집단)을 조사하기엔 시간과 비용을 감당할 수 없다.
- 대신 고래 청주 전체 시민 중 일부만을 임의(random)로 골라 표본(sample)을 싼값으로 추출할 수 밖에 없다.

표본의 추출

- 이때 모든 청주시민의 “혈압의 평균”은 모수(parameter)가 된다.
- 모수인 “혈압의 평균”은 그 값을 알 수 없지만 상수(constant)이다. 왜냐하면 답은 이미 정해져 있기 때문이다(Why?).
- 고래의 표본에서 얻은 청주시민의 “혈압의 평균”은 통계량(statistic)이 된다.
- 통계량은 (상수가 아니라) 그 값이 모집단으로부터 임의로 추출된 표본(sample)에 따라 그때그때 달라지는 확률변수(random variable)이다.

표본의 추출

- 모수는 상수이다! 다만 그 정확한 값을 모를 뿐이다.
- 통계량은 알 수 있다! 다만 (그때그때 표본에 따라 달라지는) 확률변수일 뿐이다.



표본의 추출

표집이란 표본의 추출을 의미한다.

- 고래는 청주시민 30명을 임의표집(random sampling)하기로 했다.
- 그 30명의 혈압을 조사하여 "표본(sample)의 평균(mean)"을 계산해 보았더니 67 mmHg라는 평균값을 얻었다.
- 이때 67 mmHg라는 추정값(estimate)은 순길이 붙잡은 청주시민의 "한 표본(a sample)"으로부터 얻은 값에 지나지 않는다.
- 다시 말해, 표본을 새로 뽑으면 그때그때 달라질 수 밖에 없는 값이다(Why?).
- 사실 표집은 그 자체로 상당히 까다로운 테크닉이다. 우리 학과에서는 서베이 방법론(Survey Methodology)이 개설되지 않는다.

표본평균의 표집분포

표본평균의 표집분포

"너의 표본은..."

- 고래와 같은 동네에서 의료기기 가게를 운영하는 경쟁자인 새우는 청주에서 30명의 표본을 새로 뽑아 각각 75mmHg라는 값을 얻었다.
- 상식적으로 생각해서 인구 85만의 청주에서 30명짜리 표본을 2번 뽑았을 때, 그 표본의 평균들이 정확히 똑같을리가 없다.
- 그러면 대체 어떤 추정을 믿어야 하는걸까?
- 애시당초 "표본의 통계량"을 통해서는 결코 "모집단의 모수"를 알 수 있기는 한 것일까?
- 다행히 이 문제에 대한 이론적인 답이 있고, 그에 따르면 대답은 "Yes!"이다.

표본평균의 표집분포

이제 어떻게 표본을 통해 모집단을 추정할 수 있는지 상상해보자.

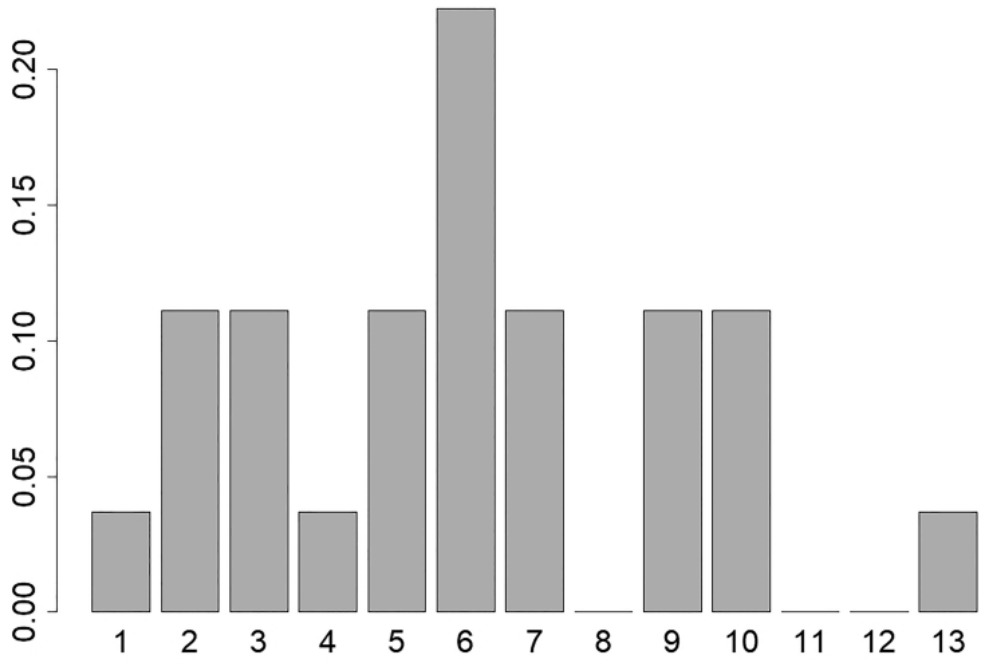
- 나에게 $\{1, 4, 13\}$이라는 모집단이 주어져 있고, 이 모집단에서 $n = 3$인 표본을 뽑는다고 상상해보자.
- 이때 (있을 수 있는) 모든 표본들의 조합은 아래와 같다(직접 만들어보자).

표본	$\bar{x}$	표본	$\bar{x}$	표본	$\bar{x}$
{1, 1, 1}	1	{4, 1, 1}	2	{13, 1, 1}	5
{1, 1, 4}	2	{4, 1, 4}	3	{13, 1, 4}	6
{1, 1, 13}	5	{4, 1, 13}	6	{13, 1, 13}	9
{1, 4, 1}	2	{4, 4, 1}	3	{13, 4, 1}	6
{1, 4, 4}	3	{4, 4, 4}	4	{13, 4, 4}	7
{1, 4, 13}	6	{4, 4, 13}	7	{13, 4, 13}	10
{1, 13, 1}	5	{4, 13, 1}	6	{13, 13, 1}	9
{1, 13, 4}	6	{4, 13, 4}	7	{13, 13, 4}	10
{1, 13, 13}	9	{4, 13, 13}	10	{13, 13, 13}	13

표본평균의 표집분포

- 각 표본들의 평균을 표본평균(sample mean)이라고 부른다. 이것을 확률변수 $\bar{X}$로 생각하여 확률분포를 표현해보자.
- 이런 확률분포를 표본평균의 표집분포(sampling distribution of the sample mean)라고 부르고, 특별히 표집분포(sampling distribution)라고 줄여 말한다.

$\bar{x}$	$P(\bar{X}=\bar{x})$
1	1/27
2	3/27
3	3/27
4	1/27
5	3/27
6	6/27
7	3/27
9	3/27
10	3/27
13	1/27



표본평균의 표집분포

예제 1. 모집단 {1, 4, 13}이 주어졌을 때, 모든 표본평균을 확률분포로 나타낼 수 있다. 이 확률분포의 평균과 분산을 구하시오.

표본평균의 표집분포

- 표본평균의 확률분포, 즉 표집분포의 평균은 6이고 분산은 8.667이다.

$\bar{x}$	$P(\bar{X}=\bar{x})$	$\bar{x}\cdot P(\bar{X}=\bar{x})$	$(\bar{x}-\mu)^2\cdot P(\bar{X}=\bar{x})$
1	0.037	0.037	0.926
2	0.111	0.222	1.778
3	0.111	0.333	1.000
4	0.037	0.148	0.148
5	0.111	0.556	0.111
6	0.222	1.333	0.000
7	0.111	0.778	0.111
9	0.111	1.000	1.000
10	0.111	1.111	1.778
13	0.037	0.481	1.815
	합계	6	8.667

표본평균의 표집분포

여기서 멈추지 말고 좀 더 생각해보자.

- (일반적인 상황과는 달리) 이 예제에서 우리는 모집단을 알고 있다.
- 당연히 모평균(population mean)과 모분산(population variance)을 계산할 수 있다!

$$\mu = 6, \quad \sigma^2 = 26$$

- 이제 다음의 두 가지 흥미로운 사실을 확인할 수 있다(직접 확인해보자).

  (1) 표집분포의 평균은 모평균과 같다.

$$\mu_{\bar{X}} = \mu$$

  (2) 모분산을 표본 크기로 나눈 값은 표집분포의 분산과 같다.

$$\sigma^2_{\bar{X}} = \frac{\sigma^2}{n}$$

표본평균의 표집분포

이것은 우연이 아니다.

- 85만 명이 거주하는 청주에서 30명 짜리 표본을 무한히 뽑아 각각 평균을 내보자.
- 이게 무슨 말일까? 세 가지 중요한 요소가 있다.

  (1) 똑같은 크기(e.g., 30명)의 표본을 추출하고, 또 추출하고, ..., 또 추출한다.

  (2) 첫번째 표본에서 평균(a sample mean)을 구하고, 두번째 표본에서 평균을 구하고, ..., n번째 표본에서 평균을 구한다.

  (3) 위 과정을 "무한히" 반복한다.

- 이 무한히 많은 표본평균들을 확률변수 $\bar{X}$로 판단하여, (1) 표집분포의 평균을 구하면 정확히 μ와 일치하고, (1) 표집분포의 분산을 구하면 정확히 σ^2/n과 일치한다.

표본평균의 표집분포

표집분포에는 두 가지 중요한 특징이 있다.

- 각각의 표본이 $n \geq 30$라는 조건을 충족한다고 전제하자.
- 우리에게 주어진 무한히 많은 표본평균들을 가지고, 이 표본평균들(sample means)의 평균(mean)을 계산할 수 있다. 말하자면 이것은 "평균들의 평균"이다.
- 그런데 이 표본평균들의 평균은 모집단의 평균(population mean), 즉 모평균에 무한히 근접한다(=일치한다).

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

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 표본 크기(여기서는 $n = 30$)로 나눈 값에 무한히 근접한다(=일치한다)는 것이 증명될 수 있다.

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

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

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

표본평균의 표집분포

그리고 여기 통계학사상 가장 위대한 발견이 있다.

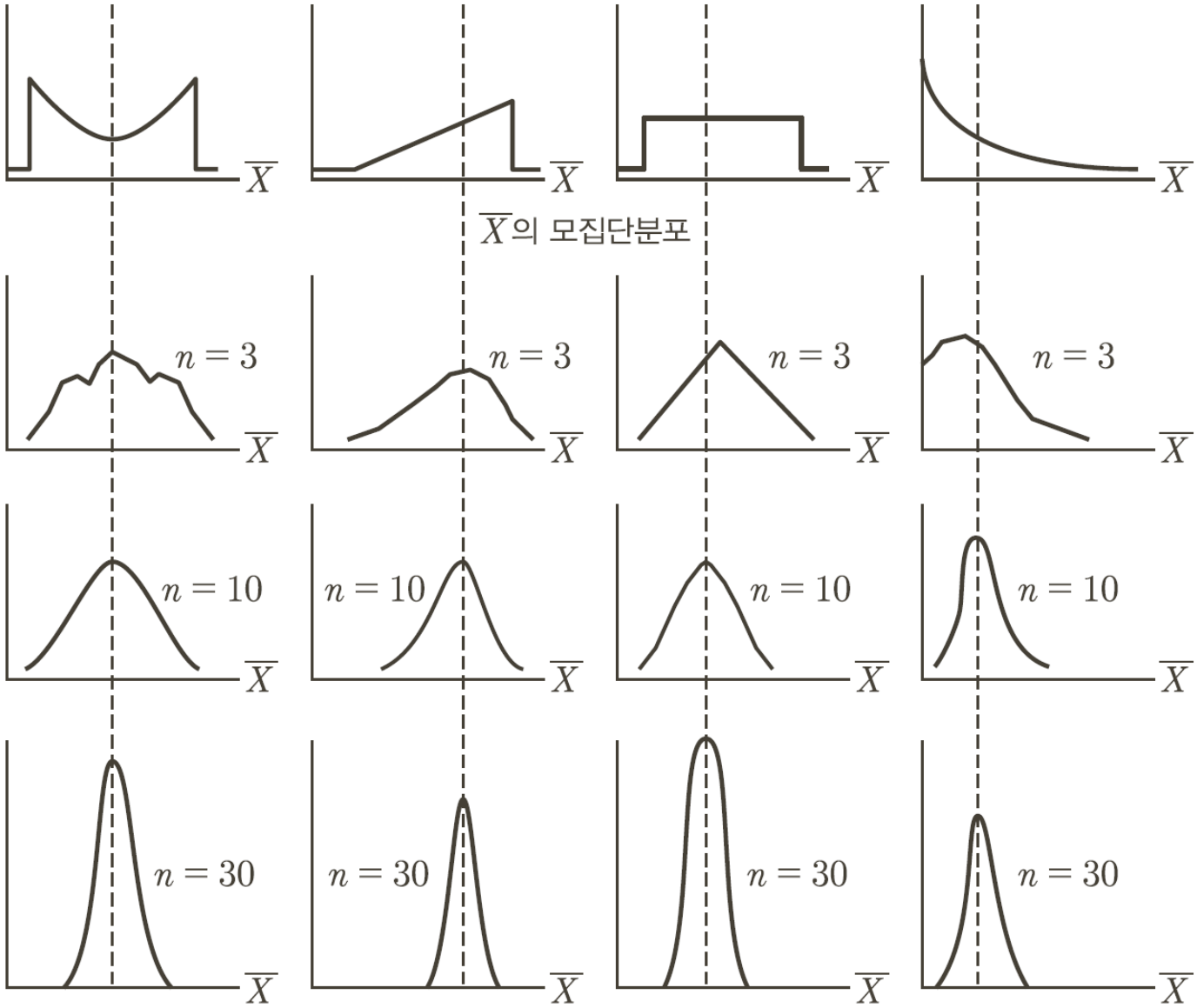
- 중심극한정리(central limit theorem)에 따르면,
- 모집단의 분포가 "어떤 꼴이든 상관없이"
- 표본 크기(sample size)가 충분히 크면($n \geq 30$)
- 표집분포(sampling distribution), 즉 표본평균들의 확률분포는
- 평균이 μ이고 분산이 σ^2/n인 정규분포에
- 근사한다(approximate).

표본평균의 표집분포

왜 그토록 위대한가?

- "모집단의 분포가 어떻게 생겼든 상관없이" 그것의 (무한히 많은) 표본($n \geq 30$)의 평균들은 반드시 정규분포함을 증명해 보였기 때문이다.
- 다시 말해, 정규분포에 관한 성질을 가지고 "모든" 표본평균들의 분포를 설명할 수 있다. 그게 무엇에 관한 모집단이건간에!
- 교과서 중에는 "모집단이 정규분포인가 정규분포가 아닌가"를 구태여 나누어 설명하는 경우도 있다. 그런 것은 표본 크기가 $n \geq 30$만 넘어도 중심극한정리 덕택에 사실 큰 의미가 없다.

표본비율의 표집분포



표본평균의 표집분포

예제 2. 정부의 최근 발표에 따르면 청주의 대학생은 평균적으로 28,650 천원의 학자금 대출 부채를 가지고 졸업한다. 그 표준편차는 7,000 천원인데, 최근 자산소득 양극화로 인해 정규분포하지 않는다. 고래는 기말과제를 수행하기 위해 청주에서 임의의 대학생 30명의 학자금 대출액을 조사하였다. 고래의 표본에서 계산한 부채의 평균이 27,000 천원보다 클 확률은 얼마인가?

표본평균의 표집분포

- (학자금 대출액의 모집단은 정규분포하지 않을지라도) 고래의 표집분포는 $n \geq 30$ 조건을 충족하므로 정규분포에 근사한다.
- 고래의 표본에서 부채 평균이 27,000 천원보다 클 확률을 먼저 표현해보자.

$$P(\bar{X} > 27000) = P\left(Z > \frac{27000 - \mu_{\bar{X}}}{\sigma_{\bar{X}}}\right)$$

- 지난 주에 $Z = \frac{X - \mu}{\sigma}$ 였지만, 지금은 $Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma/\sqrt{n}}$ 이다(Why?).

표본평균의 표집분포

- 모르는 부분을 채워넣기 위해 이 표집분포의 평균 $\mu_{\bar{X}}$ 과 표준오차 $\sigma_{\bar{X}}$ 를 구해야 한다.

$$\mu_{\bar{X}} = 28,650 \qquad \sigma_{\bar{X}} = \frac{7000}{\sqrt{30}}$$

- 고래의 표본에서 30명의 부채 평균이 27,000 천원보다 클 확률은 얼마인가?

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

- "고래의 표본평균이 27,000 천원보다 클 확률은 90%이다."

표본비율의 표집분포

표본비율의 표집분포

아까는 표본평균을 살펴보았지만 이번엔 표본비율을 살펴보자.

- 아까 표본평균을 했는데 왜 표본비율(sample proportion)은 또 할까?
- 키나 몸무게 같은 양적 변수(quantitative variable)일 때만 평균은 의미를 갖는다.
- 성별(1=여성; 0=남성)이나 종교(0=없음; 1=개신교인; 2=불교인; 3=천주교인; 4=기타)같은 질적 변수(qualitative variable)일 때 평균에는 아무런 의미도 없다.
- 반면 비율(proportion)은 질적 변수일 때만 의미를 갖는다.
- 예컨대 여성의 비율, 기독교인의 비율은 의미를 갖는다. 몸무게의 비율이나 월 소득의 비율이라는 말은 어감부터 뭔가 이상하다.
- 표본평균은 정규분포와 관련되어 있으나, 표본비율은 이항분포(binomial distribution)와 관련되어 있다.

표본비율의 표집분포

다행히 비율의 표집분포 논리도 결국 똑같다!

- 먼저 표본비율은 (확률의 고전적 정의와 마찬가지로) 아래처럼 정의된다.

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

- 이때, X는 표본에서 (관심있는 사건의) 성공 횟수이고 n은 표본 크기이다.
- 무한히 많은 표본을 뽑아 그 비율들을 확률변수 p로 하는 확률분포를 상상해보자. 이를 표본비율의 표집분포(sampling distribution of the sample proportion)라고 부르고, 이것도 표집분포라고 줄여 부를 수 있다.

표본비율의 표집분포

위대한 중심극한정리는 여기서도 적용된다.

- (무한히 많이 뽑은) 표본비율의 표집분포의 평균 μ_p는 모집단의 비율(population proportion), 즉 모비율 π에 무한히 근접한다(=일치한다).

$$\mu_p = E(p) = \pi$$

표본비율의 표집분포

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

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

- 이때 은근슬쩍 제5주차에 배운 확률변수의 분산 특성과 제6주차에 배운 이항분포의 분산을 이용하였다.

$$Var(aX) = a^2 Var(X)$$
$$Var(X) = n\pi(1-\pi)$$

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

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

표본비율의 표집분포

다만 표본비율의 표집분포에서 중심극한정리의 성립 조건은 살짝 다르다.

- 표본평균의 표집분포의 경우 각각의 표본이 $n \geq 30$라는 조건을 요구하였다.
- 표본비율의 표집분포의 경우 표본 크기 n 뿐 아니라 표본비율 π도 고려해야 한다. 보다 구체적으로 다음의 두 조건이 성립해야 중심극한정리가 성립한다.

$$np \geq 5 \qquad n(1-p) \geq 5$$

- 다시 말해, 발생 확률이 너무 희박하거나 반대로 너무 흔하면 정규분포를 사용하기 위해 표본 크기가 엄청 커져야 한다(Why?).

표본비율의 표집분포

예제 3. 청주에 본사를 두고 있는 IT기업들의 55%가 작년에 사이버 공격을 당했다고 한다. 그게 정말일까 하고 궁금했던 브래드는 청주 시내 IT기업 30곳을 임의로 방문해 사이버 공격을 당했는지 여부를 조사하였다.
브래드의 표본에서 사이버 공격을 당했다고 응답한 비율이 70%보다 클 확률은 얼마인가?

표본비율의 표집분포

- 브래드가 구하고자 하는 확률을 어떻게 표현할 수 있을지 먼저 생각해보자.

$$P(p > 0.7) = P\left(Z > \frac{0.7 - \pi}{\sqrt{\pi(1-\pi)/n}}\right)$$

- 지난 주에 $Z = \frac{X - \mu}{\sigma}$ 였지만 지금은 $Z = \frac{p - \pi}{\sigma_p} = \frac{p - \pi}{\sqrt{\pi(1-\pi)/n}}$ 이다(Why?).
- 아래의 조건에 따라 사이버 공격을 당했던 비율의 표집분포는 정규분포할 것이다.

$$np = 30 \cdot 0.7 = 21 \qquad n(1-p) = 30 \cdot (1 - 0.7) = 9$$

표본평균의 표집분포

- 모르는 부분을 채워넣기 위해 이 표집분포의 평균 π과 표준오차 σ_p를 구해야 한다.

$$\pi = E(p) = 0.55 \qquad \sigma_p = \sqrt{\frac{.55(1-.55)}{30}}$$

- 브래드의 표본에서 사이버 공격을 당했다고 응답한 비율이 70%보다 클 확률은 얼마인가?

$$P(p > 0.7) = P\left(Z > \frac{0.7 - 0.55}{\sqrt{0.55(1-0.55)/30}}\right) = 1 - P(Z < 1.65) = 0.049$$

- "브래드의 표본에서 사이버 공격을 당했다고 응답할 비율이 0.7 이상으로 나올 확률은 약 4.9% 이다."