Chapter
02

모집단과 표본 Population and Sample

목 차

01 모집단과 표본추출

:: Keywords 모집단(모수) | 표본(통계량) | 표본추출

모집단과 표본

■ 모집단

모집단(population) : 통계분석 방법을 적용할 관심 대상의 전체 집합

예 모든 대한민국 여성
202X년에 수입된 모든 쇠고기
A 쇼핑몰 회원 전체
B 통신 회사 전체 가입자

→ 물리적인 한계로 인해 모집단 전체를 전수조사하기는 쉽지 않다.

■ 표본

표본(sample) : 과학적인 절차를 적용하여 모집단을 대표할 수 있는 일부를 추출하여 직접적인 조사 대상이 된 모집단의 일부

모집단과 표본

■ 모수

모수(parameter) : 모집단을 분석하여 얻어지는 결과 수치

ex. 모평균, 모분산, 모표준편차, 모비율

■ 통계량

통계량(statistic) : 표본을 분석하여 얻어지는 결과 수치

ex. 표본평균, 표본분산, 표본표준편차, 표본비율

모집단		표본	
모수	모평균(μ) 모분산(σ^2) 모표준편차(σ) 모비율(p)	통계량	표본평균($\bar{x}$) 표본분산(s^2) 표본표준편차(s) 표본비율($\hat{p}$)

모집단과 표본

참고 **평균을 표현하는 기호**

평균을 표현할 때 μ로 표현하면 모수에 해당하는 모평균을, $\overline{x}$로 표현하면 추출한 표본의 평균을 의미한다. 그런데 경우에 따라서는 평균(mean)을 m으로 표현하기도 한다. 모평균이나 표본평균을 구분하지 않고 표현하는 경우에 m을 사용한다.

모집단과 표본

예제 2-1 모집단과 표본의 관계 설명하기

요즘 TV 방송을 비롯해 유튜브, SNS 등에서 먹방이나 음식점 소개가 대세다. 음식을 먹는 자신만의 방법을 소개하기도 하고 음식을 평가하거나 요리 방법을 소개해서 알리기도 한다. 그런데 제작진은 매번 새로운 먹방 식당을 어떻게 찾는 것일까? 시청률을 끌어올리는 먹방 식당 찾기 과정을 모집단과 표본의 관계로 설명하라.

풀이

방송 주제로 특정 음식이 정해지면, 해당 음식을 판매하는 국내의 모든 식당이 모집단이 된다. 선택된 음식을 판매하는 식당의 수는 통계청의 통계 자료나 편람 혹은 국세청 자료를 이용하면 확인할 수 있다. 그러나 그 수가 셀 수 없을 정도로 많고, 자료의 내용도 실제와 다를 가능성이 높다. 왜냐하면 업종 신고 시 판매 음식을 정확하게 표시하지 않는 경우가 많기 때문이다. 보통 일주일 단위로 프로그램이 제작된다는 것을 감안하면 국내의 모든 음식점을 확인하는 것은 불가능하다.

따라서 제작자는 표본조사 방법을 선택한다. 이때 표본은 시청자들의 제보로 얻는데, 이런 경우 어떤 편의(편견, biasness)가 발생할 수 있기 때문에 과학적 접근이라는 측면에서 보면 문제가 있다. 제작진 입장에서는 '특정 음식'이나 '식당'에 지나치게 초점을 맞출 경우 다큐멘터리가 되어 시청율이 떨어질 우려도 있다. 이 때문에 연예인들을 동원해서 '정해진 주제'를 언급하기는 하지만 흥미 위주로 흘러가는 것이 보통이다.[2]

참고 그리스 문자

통계에서는 그리스 문자가 상당히 많이 쓰인다. 읽는 방법을 익히고, 대소문자를 구분할 수 있으면 상당히 편리하다.

[표 2-1] 그리스 문자

대문자	소문자	발음	대문자	소문자	발음	대문자	소문자	발음
A	α	alpha	I	ι	iota	P	ρ	rho
B	β	beta	K	κ	kappa	Σ	σ	sigma
Γ	γ	gamma	Λ	λ	lambda	T	τ	tau
Δ	δ	delta	M	μ	mu	Y	υ	upeilon
E	ε	epsilon	N	ν	nu	Φ	φ	phi
Z	ζ	zeta	Ξ	ξ	xi	X	χ	chi
H	η	eta	O	o	omikron	Ψ	ψ	psi
Θ	θ	theta	Π	π	pi	Ω	ω	omega

표본추출 방법

가장 정확한 조사 방법 → 모집단을 대상으로 조사하는 것

But, 대부분 표본으로 조사를 진행

- **확률적 표본추출 방법(probability sampling method)**

  표본추출의 방법은 동일한 확률 하에서 표본을 구성

- **비확률적 표본추출 방법(non-probability sampling method)**

  확률과는 상관없이 조사자가 자신의 의지로 표본을 뽑거나 조사 대상이 자발적으로 표본을 구성

확률적 표본추출

■ 단순 무작위 표본추출

모집단에서 일정한 규칙에 따라 표본을 기계적으로 추출하는 방법

ex. 컴퓨터로 추출하거나 난수표를 활용

■ 체계적 표본추출

모집단에 번호를 부여하고 일정한 n개의 간격으로 표본을 추출하는 방법

예 선거 당일 출구에서 투표를 하고 나오는 유권자의 숫자를 세어 1, 11, 21, 31, 41, … 번째(혹은 1, 101, 201, 301, …번째)의 유권자를 대상으로 조사를 하는 경우

확률적 표본추출

■ 비례 층화 표본추출

모집단을 여러 개의 이질적 집단으로 구분한 후, 각 집단의 구성 개수에 비례하도록 추출하는 방법

> 예 총원 10,000명인 A 대학교에서 1, 2, 3, 4학년의 비율은 4 : 3 : 2 : 1이다. 1,000명을 추출하고자 할 때, 각 학년을 구성하는 비율대로 학년별로 각각 400, 300, 200, 100명씩 추출하여 1,000명의 표본을 구성하는 방법

■ 다단계 층화 표본추출

비례 층화 표본추출에서 상-하위 표본 단위를 미리 설정하고 그에 맞추어 다시 추출하는 방법

> 예 총원이 10,000명인 A대학교에서 1,000명을 추출할 때, 먼저 단과 대학별로 구분지은 후 다시 학과별 구성에 맞춰 표본을 추출하는 방법

확률적 표본추출

■ 군집 표본추출

모집단의 구성이 내부 이질적이면서 외부 동질적으로 구성되어 있다면,

모집단 전체를 조사하지 않고 몇 개의 군집을 표본으로 선택해서 조사하는 방법

> 예 서울 시민을 대상으로 자전거의 구매의사를 조사할 때, 25개 구를 모두 조사하지 않고 표본으로 몇 개의 구를 선택하여 조사하는 방법

비확률적 표본추출

■ 편의 표본추출

조사자의 편의에 따라 시간이나 장소에 구애 받지 않고 임의적으로 표본을 추출하는 방법

→ 조사하기 쉽고 비용이 적게 드는 장점에 비해,
모집단에 대한 대표성을 나타내기 힘들며, 실수나 오류가 가장 많이 발생

■ 판단 표본추출

조사자가 적합하다고 판단한 구성원들을 표본으로 선택하는 방법

→ 편의 표본추출과 다른 점은 표본으로 선택할지의 여부를 조사자가 판단

비확률적 표본추출

■ 할당 표본추출

모집단의 속성을 대표할 만한 연령, 학력, 직업 등의 구분을 결정하고, 각각에 대한 표본의 개수를 미리 결정한다. 이후 조사자가 결정한 표본의 개수에 따라 임의적으로 표본을 추출하는 방법

■ 자발적 표본추출

조사자의 의지와는 별개로 응답자가 원하여 조사에 응하는 경우를 표본으로 선택하는 방법

→ 관여도(관심도)가 높은 사람들이 주로 조사에 응하게 될 것이므로 결과의 왜곡이 발생할 가능성이 크다.

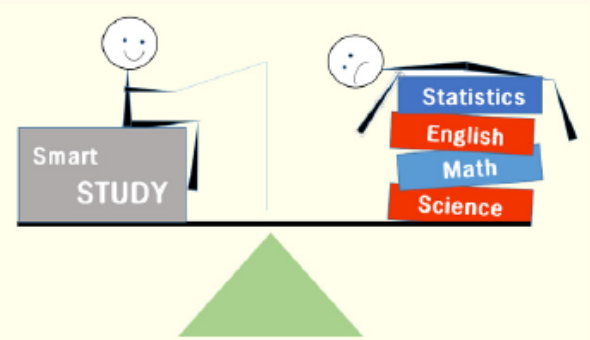
02 표본의 분포

Keywords | 정규분포 | 표준화 | z분포 | t분포 | χ^2 분포 | F분포 | $\hat{p}$ 분포 |

표본을 추출한 후, 표본의 특성을 파악하기 위해 표본분포의 확인이 필요표본평균, 표본분산, 표본비율을 통해 이후에 배우는 통계분석을 무리 없이 진행할 수 있다.

누가 공부를 더 잘한 것일까?

[사례 1] A는 영어를 80점, B는 수학을 60점 맞았다. 이 둘을 비교했을 때, A와 B 중 누가 공부를 더 잘한 것일까? 영어의 과목 평균이 90점, 수학의 과목 평균이 50점이라면, 실제 점수가 더 낮더라도 수학을 60점 맞은 B가 더 잘했다고 할 수 있을 것이다.



[그림 2-2] 누가 더 공부를 잘할까?

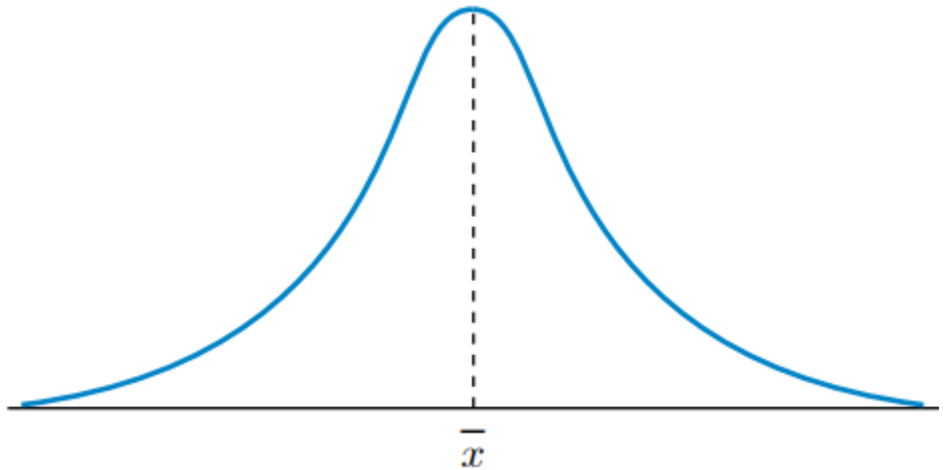
[사례 2] 과목 점수가 60점 미만이면 과락이 적용되어 진급할 수 없다는 기준이 있다고 하자. 만약 A반과 B반의 반평균이 똑같이 70점이라고 할 때, 어느 반이 공부를 더 잘한 것일까? A반은 모두 65점 ~ 75점 사이의 점수를 받았고, B반은 0점 ~ 100점까지 고르게 분포되어 있다면, A반은 낙제생이 없고, B반은 낙제생이 있다. 진급을 중요한 기준이라고 한다면 A반이 공부를 더 효율적으로 잘했다고 할 수 있을 것이다.

표준화

■ 정규분포

표본분포 중 가장 단순하면서 많이 나타나는 형태의 분포

→ 어떤 사건이 일어난 빈도(frequency)를 계산하여 그래프로 나타내면 중심(평균)을 기준으로 좌우가 대칭되는 분포

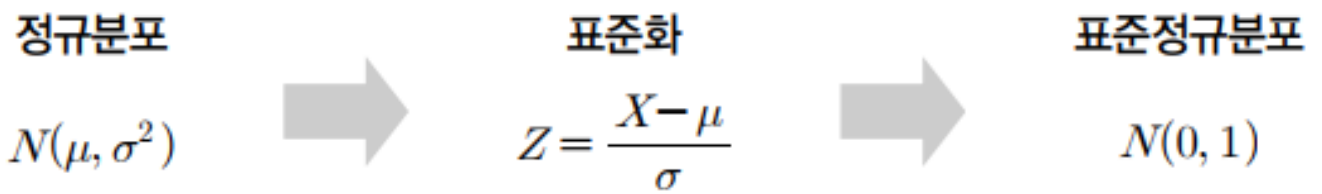


표준화

■ 표준화

단순한 현상은 정규분포만을 이용해도 결과를 알아내는 데 문제가 없지만 대부분의 연구에서는 복잡한 관계에 대한 분석 결과가 필요하므로, 여러 특성에 대한 분석 결과들을 서로 비교할 수 있도록 만드는 과정

→ 표준화란 기준점을 동일하게 맞춰 조사자가 자료들을 쉽게 비교할 수 있도록 만드는 과정으로, 표준정규분포는 평균은 0, 표준편차는 1로 만든다.



표본평균의 확률분포

■ Z분포

표본의 개수가 충분할 때 표준화 과정을 거친 정규분포를 표준정규분포(standard normal distribution), 혹은 z분포라고 한다.

→ 표준정규분포는 '평균=0, 분산=1'인 정규분포를 따른다.

$$Z = \frac{X-\mu}{\sigma} = \frac{X-\mu}{\sigma/\sqrt{n}}$$ [9]

9 표본평균들의 분산인 경우로 확장해보면 분산을 n으로 나누어야 하므로 분모는 $\frac{\sigma}{\sqrt{n}}$가 된다. 따라서 $\frac{X-\mu}{\sigma}$와 $\frac{X-\mu}{\sigma/\sqrt{n}}$는 같다.

표본평균의 확률분포

Note 정규분포의 표준화 과정

정규분포 $N(\mu, n\sigma^2)$에서 알고자 하는 값은 $\mu+n\sigma$이므로 n을 알면 그 값을 알 수 있다.[10] 그러므로 정규분포 $N(\mu, n\sigma^2)$를 따르는 확률변수 X에 대해 $X=\mu+n\sigma$라 하고 n으로 정리를 하면

$$X=\mu+n\sigma$$

$$X-\mu=n\sigma$$

$$\therefore\ n=\frac{X-\mu}{\sigma}$$

이때, 확률변수 X가 평균인 $X=\mu$라면 $n=\frac{X-\mu}{\sigma}$의 값은 0이다. 즉, 확률변수 n의 표준편차를 k라 하면 $n=0+kn$이고 $k=1$이므로, 확률변수 n의 표준편차는 1이다. 그러므로 확률변수 n은 평균$=0$, 표준편차$=1$이고, 표준화공식은 $Z=\frac{X-\mu}{\sigma}$이다.

표본평균의 확률분포

■ t분포

표본이 충분하지 못한 경우, 즉 표본의 개수가 30개를 넘지 못하는 경우에는 t분포를 사용

→ 모집단은 정규분포를 이룬다는 가정이 필요하며,

t분포도 '평균=0, 분산>1'인 정규분포를 따른다.

$$t_{(n-1)} = \frac{X-\mu}{s} = \frac{X-\mu}{s/\sqrt{n}}$$

Note **t분포의 특징**

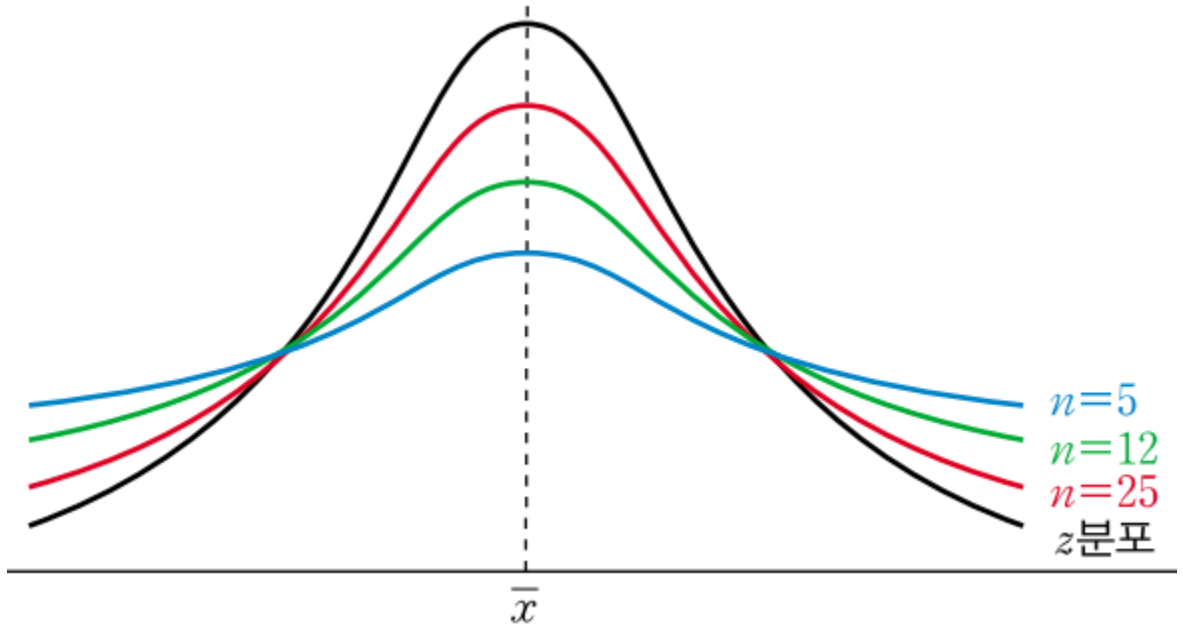
- 표본의 개수가 작은 경우에 사용한다.
- 모분산(σ^2)을 모르는 경우 표본표준편차(s)를 사용한다.
- t분포는 정규분포를 따른다.
- 분산(σ^2)이 클수록 그래프의 봉우리가 완만하고, 분산이 작을수록 봉우리 모양이 뾰족하다.
- 표준정규분포는 '평균$=0$, 분산$=1$'이지만, t분포는 '평균$=0$, 분산>1'이다.

표본평균의 확률분포

■ Z분포와 t분포의 관계

$z = \frac{X-\mu}{\sigma/\sqrt{n}}$ 와 $t_{n-1} = \frac{X-\mu}{s/\sqrt{n}}$ 는 n과 $n-1$을 제외하고 식이 동일

$n \to \infty$ 면 두 분포는 동일한 분포



표본분산의 확률분포

■ χ^2분포

χ^2분포는 정규분포로부터 도출되고, z분포의 제곱에 대한 분포

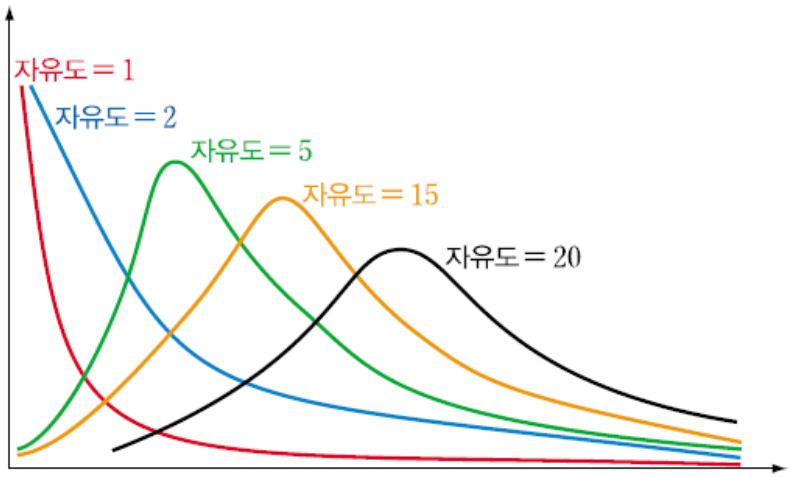
∴ 항상 0보다 큰 값

확률변수 $x_1,\ x_2,\ x_3,\ \cdots,\ x_n$ 이 표준정규분포이면서 독립이라면,

→ $x_1^2,\ x_2^2,\ x_3^2,\ \cdots,\ x_n^2$ 은 새로운 확률변수를 구성하게 되고,

이 분포를 자유도가 n인 χ^2분포라 한다.

χ^2분포는 확률변수 $\sum_{i=1}^{n} x_i^2$



표본분산의 확률분포

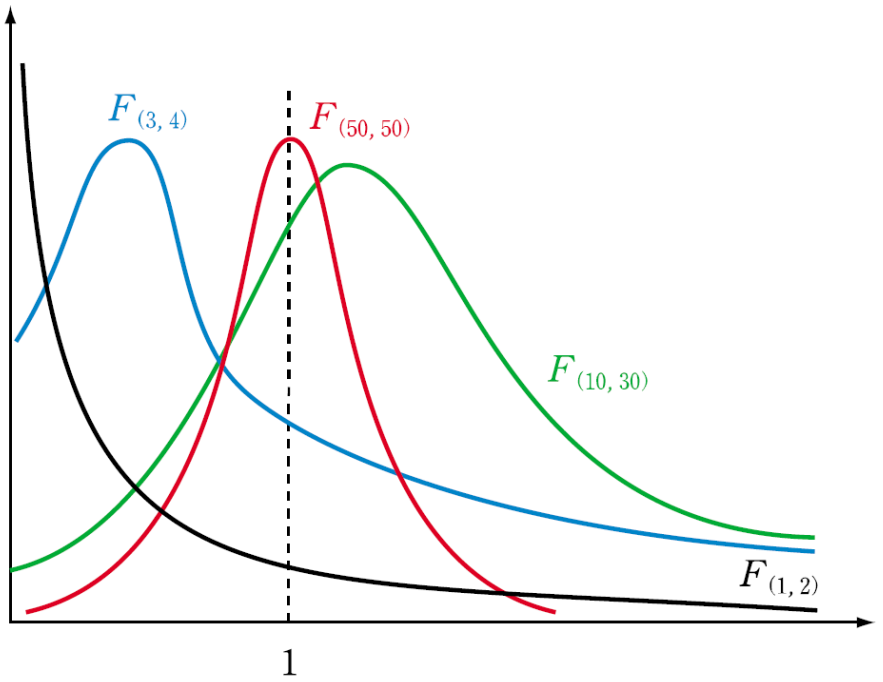
■ F분포

F분포는 두 개의 분산에 관한 추론 → F(v_1, v_2)

∴ v_1, v_2는 각각의 X^2에 대한 분산

분산이 같은 모집단에서 Xn, Ym만큼 표본을 구하고, 각각의 분산이

$v_1 = \frac{S1}{(n-1)}$, $v_2 = \frac{S2}{(m-1)}$ 일 때, $F = \frac{v_1}{v_2}$ 는 각 비율을 나타냄



표본비율의 확률분포

■ $\hat{p}$분포

표본비율 $\hat{p}$은 모집단의 특성 중 모비율을 추정하기 위해 사용됨

주로, 성공 vs. 실패, 남성 vs. 여성, 구매 vs. 비구매 등과 같이 어느 한 사건이 발생하는 베르누이 시행의 이항분포를 활용하여 표본비율의 분포를 구한다.

Note 표본비율의 특성[14]

모집단에서 표본 n개를 추출했을 때, 표본비율 $\hat{p}$는 표본의 개수(N) 중 몇 개(n)의 특정 집단(A)을 의미하는 $n(A)$로 나타낼 수 있으므로 $\hat{p}=\frac{n(A)}{N}$이다. 그러므로 $\hat{p}$의 기대값 $E(\hat{p})=p$이고, 분산 $Var(\hat{p})=\frac{p(1-p)}{n}$, 표준편차 $\sigma(\hat{p})=\sqrt{\frac{p(1-p)}{n}}$이다. $n\cdot\hat{p}\geq 5$이면서 $\hat{p}\leq 0.05$일 때, 표본비율의 분포는 정규분포이며 좌우대칭이다.

03 표본분포와 중심극한정리

:: Keywords 표본분포 | 표본평균의 오차 | 중심극한정리

표본분포

표본분포(sample distribution)는 표본에서 도출되는 통계량에 대한 확률분포

→ 표본분포는 모수를 추정하기 위한 표본 통계량의 확률분포 (여러 번 측정)

Ex. 5일간의 통학 시간이 각각 37분, 25분, 49분, 33분, 56분이 소요되었다면,

→ 평균 통학 시간은?

$$5\text{일간의 평균 통학 시간} = \frac{37+25+49+33+56}{5} = 40$$

모집단의 구성이 5개로 되어 있으므로 간단히 표본을 2개 추출하는 경우와 3개 추출하는 경우를 비교해보면...

표본분포

표본을 2개 추출하는 경우의 수를 구하면 다음과 같다.

$$(37,\ 25),\ (37,\ 49),\ (37,\ 33),\ (37,\ 56),\ (25,\ 49)$$
$$(25,\ 33),\ (25,\ 56),\ (49,\ 33),\ (49,\ 56),\ (33,\ 56)$$

→ 총 10가지

2개 추출한 경우의 수는 $_5C_2 = \frac{5!}{2! \cdot 3!} = 10$

표본을 3개 추출하는 경우의 수를 구하면 다음과 같다.[7]

$$(37,\ 25,\ 49),\ (37,\ 25,\ 33),\ (37,\ 25,\ 56),\ (37,\ 49,\ 33),\ (37,\ 49,\ 56)$$
$$(37,\ 33,\ 56),\ (25,\ 49,\ 33),\ (25,\ 49,\ 56),\ (25,\ 33,\ 56),\ (49,\ 33,\ 56)$$

→ 총 10가지

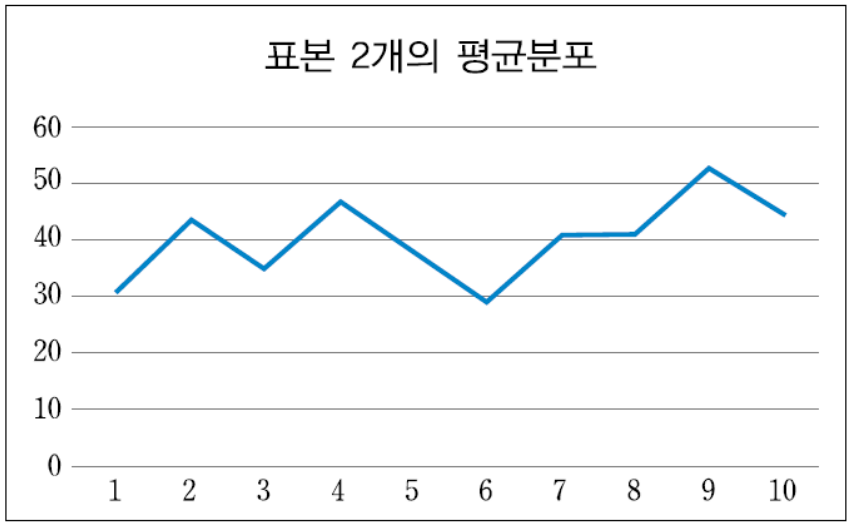
3개 추출한 경우의 수는 $_5C_3 = \frac{5!}{3! \cdot 2!} = 10$

표본을 2개 혹은 3개 추출할 때, 각 경우의 수에 대한 평균을 구해보면...

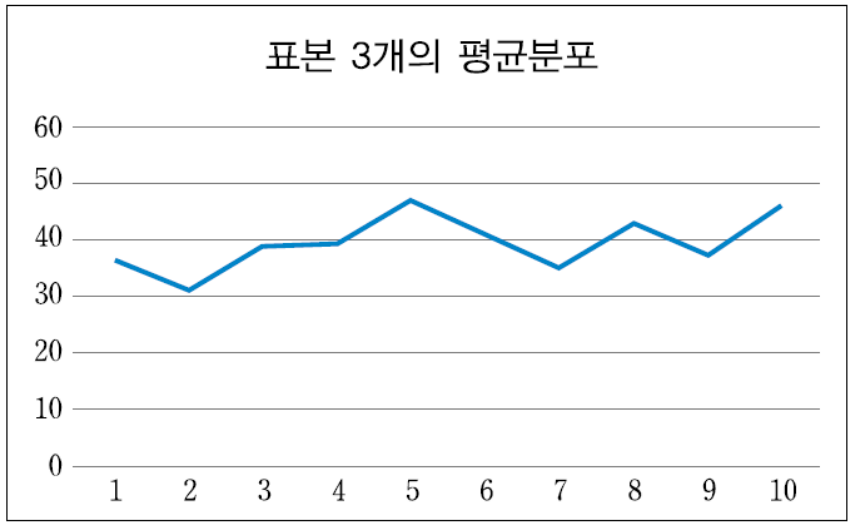
표본분포

구분		표본			표본평균
표본 2개	경우의 수 1	37	25		31
	경우의 수 2	37	49		43
	경우의 수 3	37	33		35
	경우의 수 4	37	56		46.5
	경우의 수 5	25	49		37
	경우의 수 6	25	33		29
	경우의 수 7	25	56		40.5
	경우의 수 8	49	33		41
	경우의 수 9	49	56		52.5
	경우의 수 10	33	56		44.5
표본 3개	경우의 수 1	37	25	49	37
	경우의 수 2	37	25	33	31.7
	경우의 수 3	37	25	56	39.3
	경우의 수 4	37	49	33	39.7
	경우의 수 5	37	49	56	47.3
	경우의 수 6	37	33	56	42
	경우의 수 7	25	49	33	35.7
	경우의 수 8	25	49	56	43.3
	경우의 수 9	25	33	56	38
	경우의 수 10	49	33	56	46

표본분포



(a) 표본 2개의 평균분포



(b) 표본 3개의 평균분포

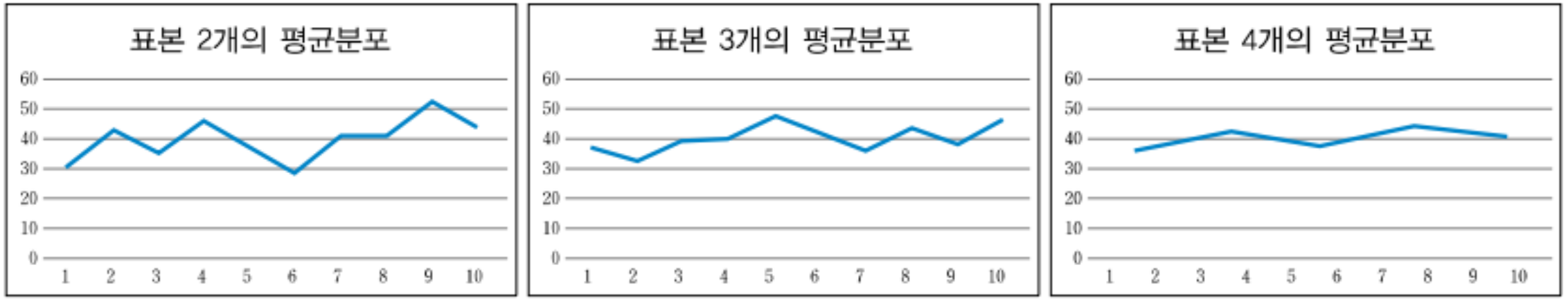
표본평균의 오차

표본평균의 오차 : 표본으로부터 모수를 추정했을 때, 모수와 통계량 간의 차이

구분		① 모평균	② 표본			③ 표본평균	④ 오차 (①-③)	표준편차
표본 2개	경우의 수 1	40	37	25		31	9	7.187952884
	경우의 수 2		37	49		43	-3	
	경우의 수 3		37	33		35	5	
	경우의 수 4		37	56		46.5	−6.5	
	경우의 수 5		25	49		37	3	
	경우의 수 6		25	33		29	11	
	경우의 수 7		25	56		40.5	−0.5	
	경우의 수 8		49	33		41	−1	
	경우의 수 9		49	56		52.5	−12.5	
	경우의 수 10		33	56		44.5	−4.5	
표본 3개	경우의 수 1	40	37	25	49	37	3	4.79196849
	경우의 수 2		37	25	33	31.7	8.3	
	경우의 수 3		37	25	56	39.3	0.7	
	경우의 수 4		37	49	33	39.7	0.3	
	경우의 수 5		37	49	56	47.3	−7.3	
	경우의 수 6		37	33	56	42	−2	
	경우의 수 7		25	49	33	35.7	4.3	
	경우의 수 8		25	49	56	43.3	−3.3	
	경우의 수 9		25	33	56	38	2	
	경우의 수 10		49	33	56	46	−6	

표본평균의 오차

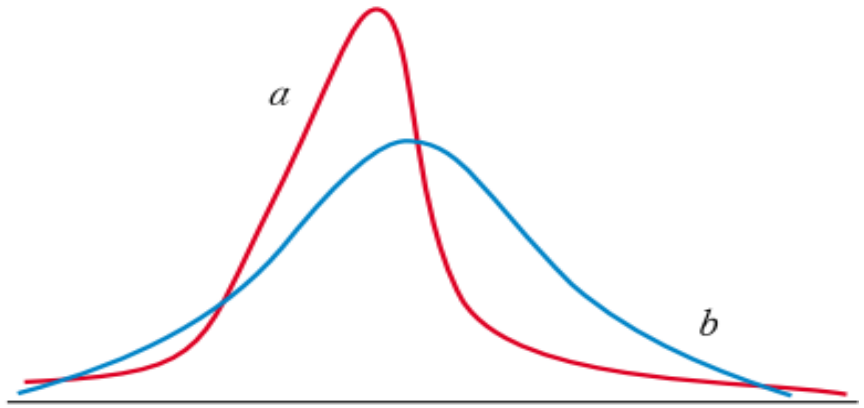
표본의 개수가 늘어날수록 통계량이 모수와 가까워짐

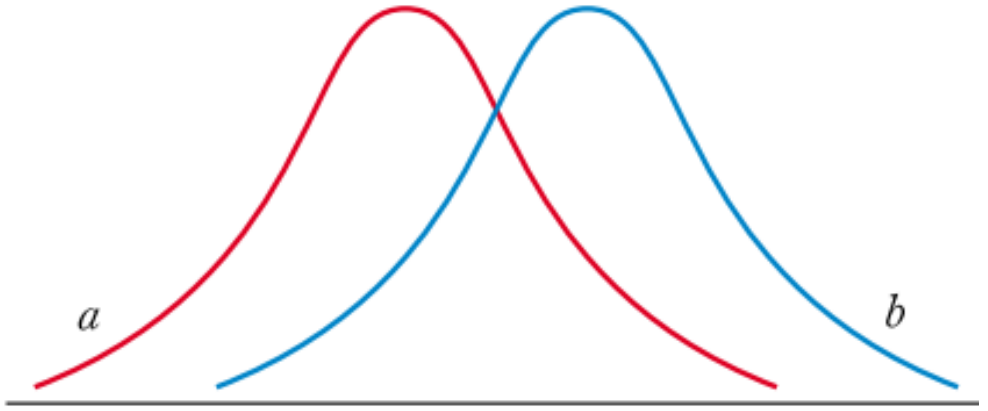


중심극한정리

중심극한정리(Central Limit Theorem : CLT)는 표본의 개수(n)가 충분하다면 모수를 모르는 상황에서도 표본 통계량으로 정규분포를 구성하여 모수를 추정할 수 있다는 것이다.

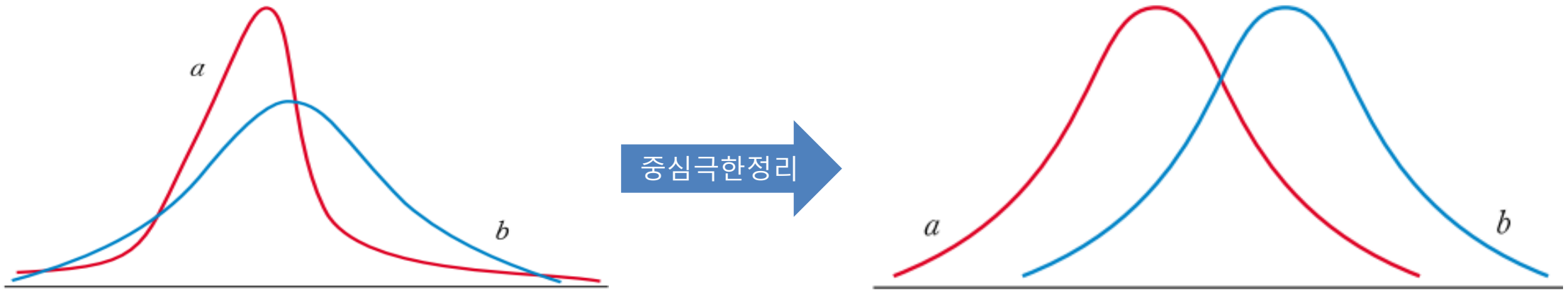
→ 중심극한정리에서는 모집단이 정규분포를 이루지 않아도 표본의 개수가 충분하다면 정규분포를 이루게 된다.





중심극한정리

중심극한정리를 이용하면 정규분포의 모양으로 확인할 수 있어서 평균을 바로 비교할 수 있다. 정규분포로 구성하면 그래프의 가장 높은 상단이 평균이 되므로 평균값을 비교할 수 있다.



Note **적절한 표본의 개수는 어느 정도일까?**

적절한 표본의 개수는 연구 특성이나 연구자의 연구 의도에 따라 다르다. [그림 2-10]과 [그림 2-11]에서는 모집단을 구성하는 측정치의 개수가 5개라는 극단적인 예를 들어서 설명했지만, 보통 30개를 최소 수준으로 이해한다. 그런데 [그림 2-11]에서도 알 수 있듯이 표본의 개수가 커질수록 모수와 조금이라도 더 가까워진다. 따라서 연구자들은 인문/사회과학 분야의 통계조사나 연구에서는 표본의 개수를 가능한 한 크게 해야 할 필요가 있다고 믿는다. 그래야 모집단 정보를 확인할 수 없는 경우가 대부분인 상황에서 오류를 최대한 방지하고 조사 결과에 더욱 신뢰를 줄 수 있다고 생각한다.

Q&A

통계학, 제대로 시작하자!

통계의 쓰임을 이해하고,
실제로 활용할 수 있어야 한다.