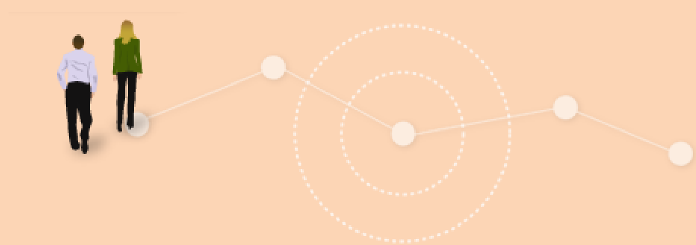
2 장

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	A 쇼핑몰 회원 전체
2016년에 수입된 모든 쇠고기	B 통신회사 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

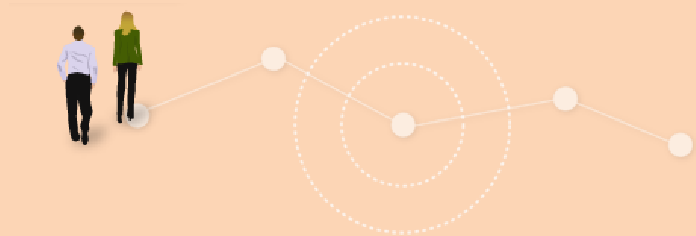
모수
- 모평균(μ)
- 모분산(σ^2)
- 모표준편차(σ)
- 모비율(p)

표본

통계량
- 표본평균($\bar{x}$)
- 표본분산(s^2)
- 표본표준편차(s)
- 표본비율($\hat{p}$)

02 표본의 분포

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

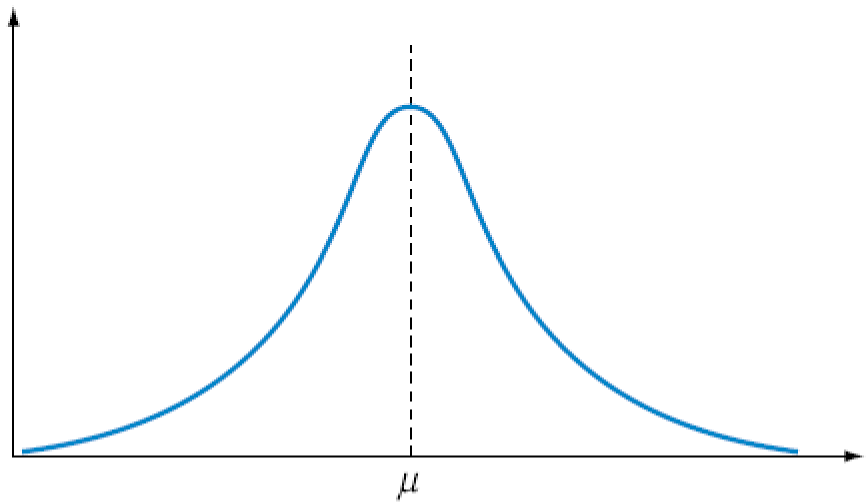


표준화

■ 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.

$$z = \frac{X - \mu}{\sigma / \sqrt{n}}$$

$$(X : \text{측정치}, \;\; \mu : \text{평균}, \;\; \sigma/\sqrt{n} \; : \text{표준오차})$$

표본평균의 확률분포

■ *t*분포

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.

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

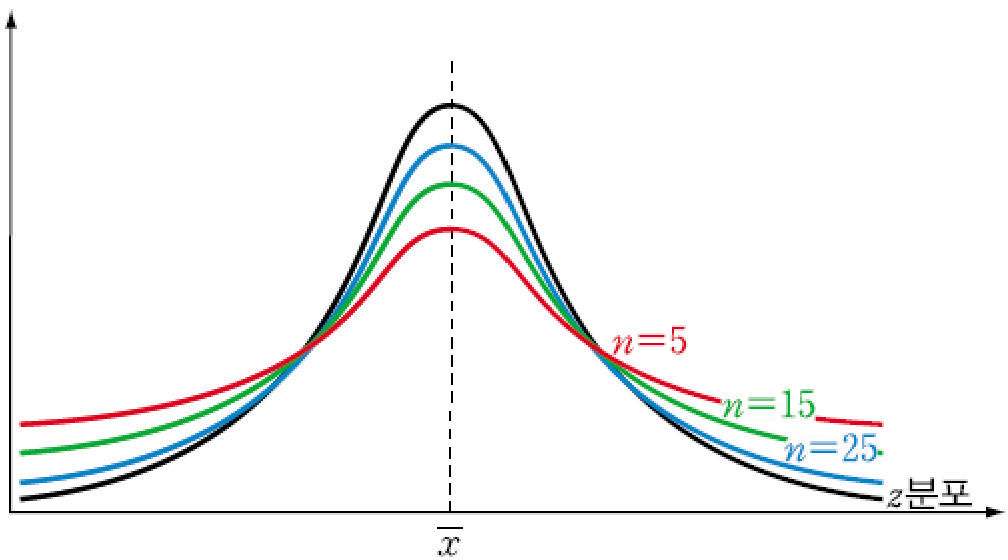
$(X$: 측정치, μ : 평균, $s/\sqrt{n}$: 표준오차, $n-1$: 자유도$)$

표본평균의 확률분포

■ Z분포와 t분포의 관계

$z = \dfrac{X - \mu}{\sigma / \sqrt{n}}$ 와 $t_{n-1} = \dfrac{X - \mu}{s / \sqrt{n}}$ 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

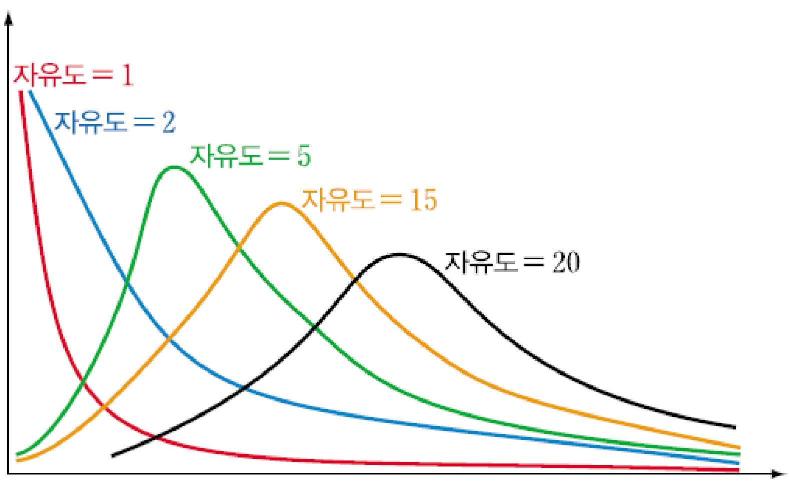
∴ 항상 0보다 큰 값

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

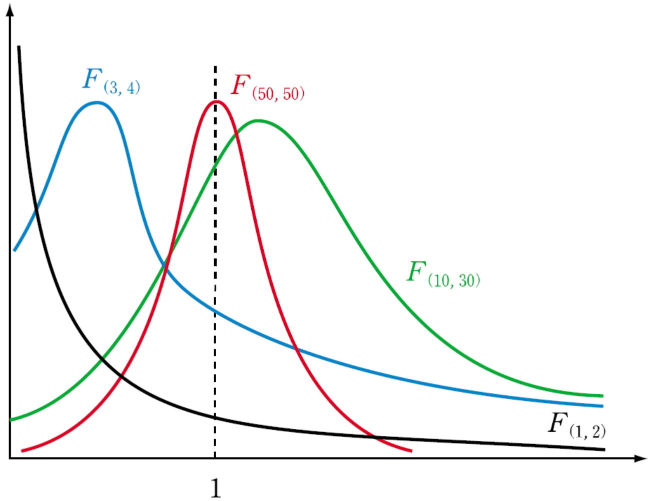
■ F분포

F분포는 두 개의 분산에 관한 추론 → $F(v_1, v_2)$

$\therefore v_1, v_2$는 각각의 X^2에 대한 분산

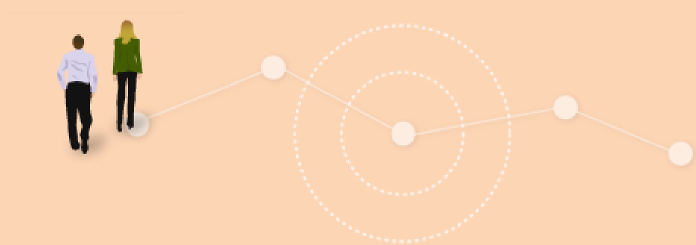
분산이 같은 모집단에서 $Xn, \mathcal{Y}m$만큼 표본을 구하고, 각각의 분산이

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



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)

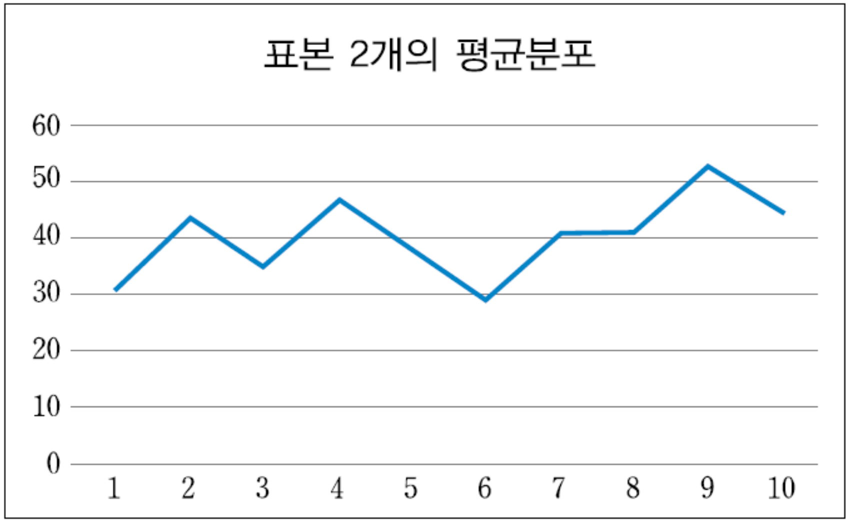
→ 총 10가지 3개 추출한 경우의 수는 ${}_5C_3 = \frac{5!}{3! \cdot 2!} = 10$

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

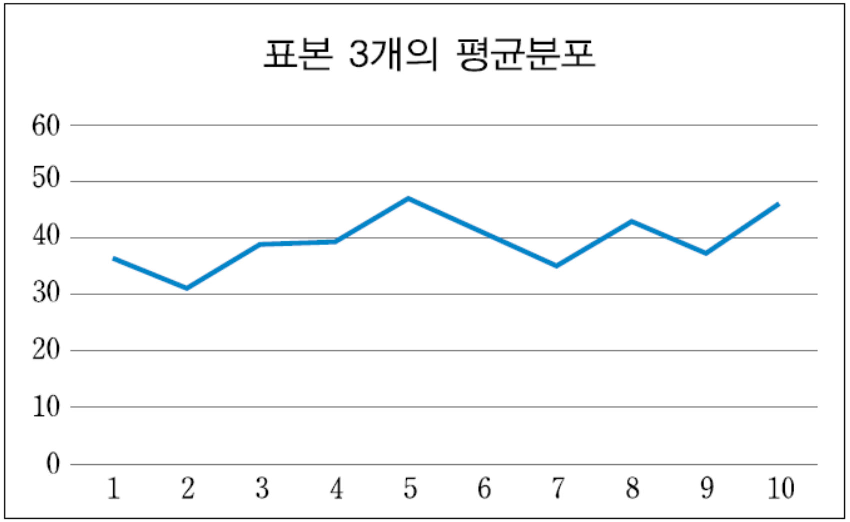
표본분포

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

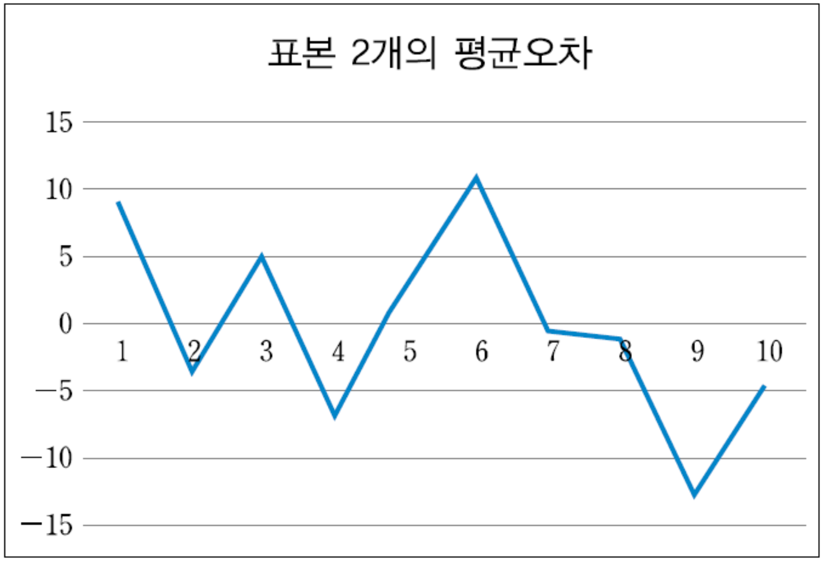
표본평균의 오차

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

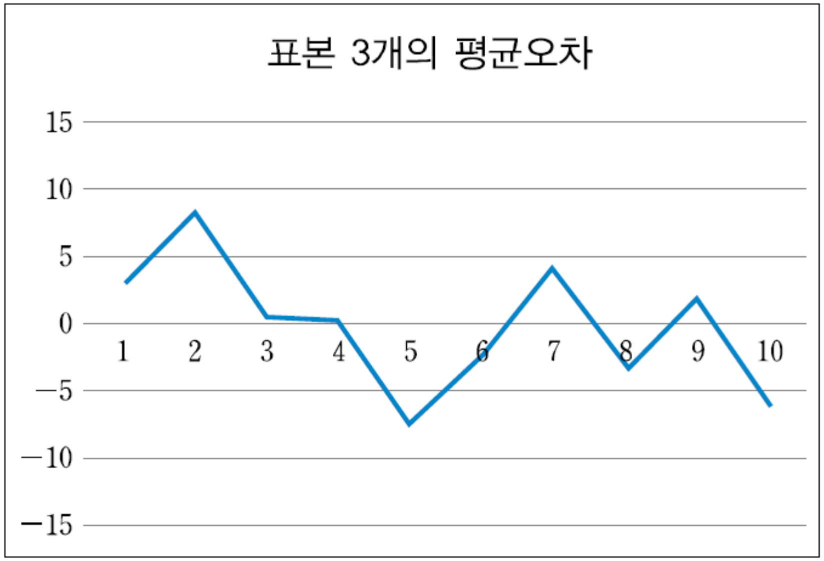
구분		① 모평균	② 표본			③ 표본평균	④ 오차 (①−③)	표준편차
표본 2개	경우의 수 1	40	37	25		31	9	7.187952884
	경우의 수 2		37	49		43	-3	
	경우의 수 3		37	33		35	5	
	경우의 수 4		37	56		46.5	−6.5	
	경우의 수 5		25	49		37	3	
	경우의 수 6		25	33		29	11	
	경우의 수 7		25	56		40.5	−0.5	
	경우의 수 8		49	33		41	−1	
	경우의 수 9		49	56		52.5	−12.5	
	경우의 수 10		33	56		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



(a) 표본 2개의 평균오차

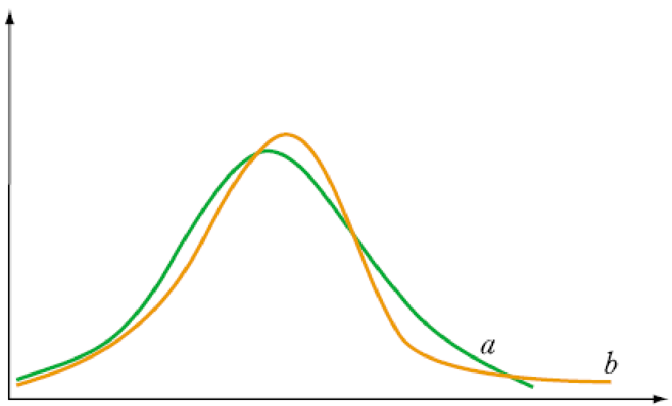


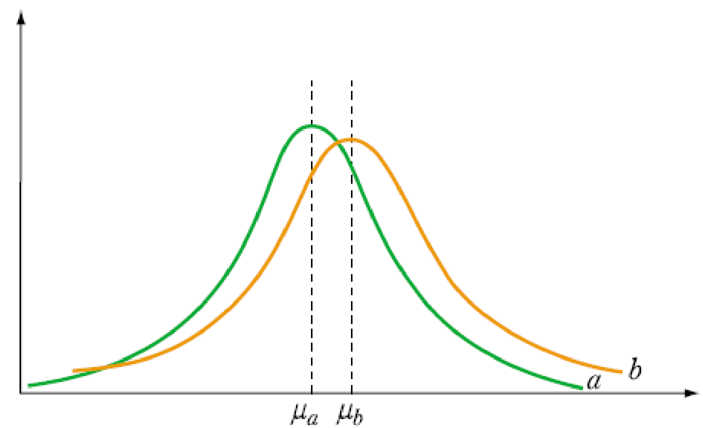
(b) 표본 3개의 평균오차

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.