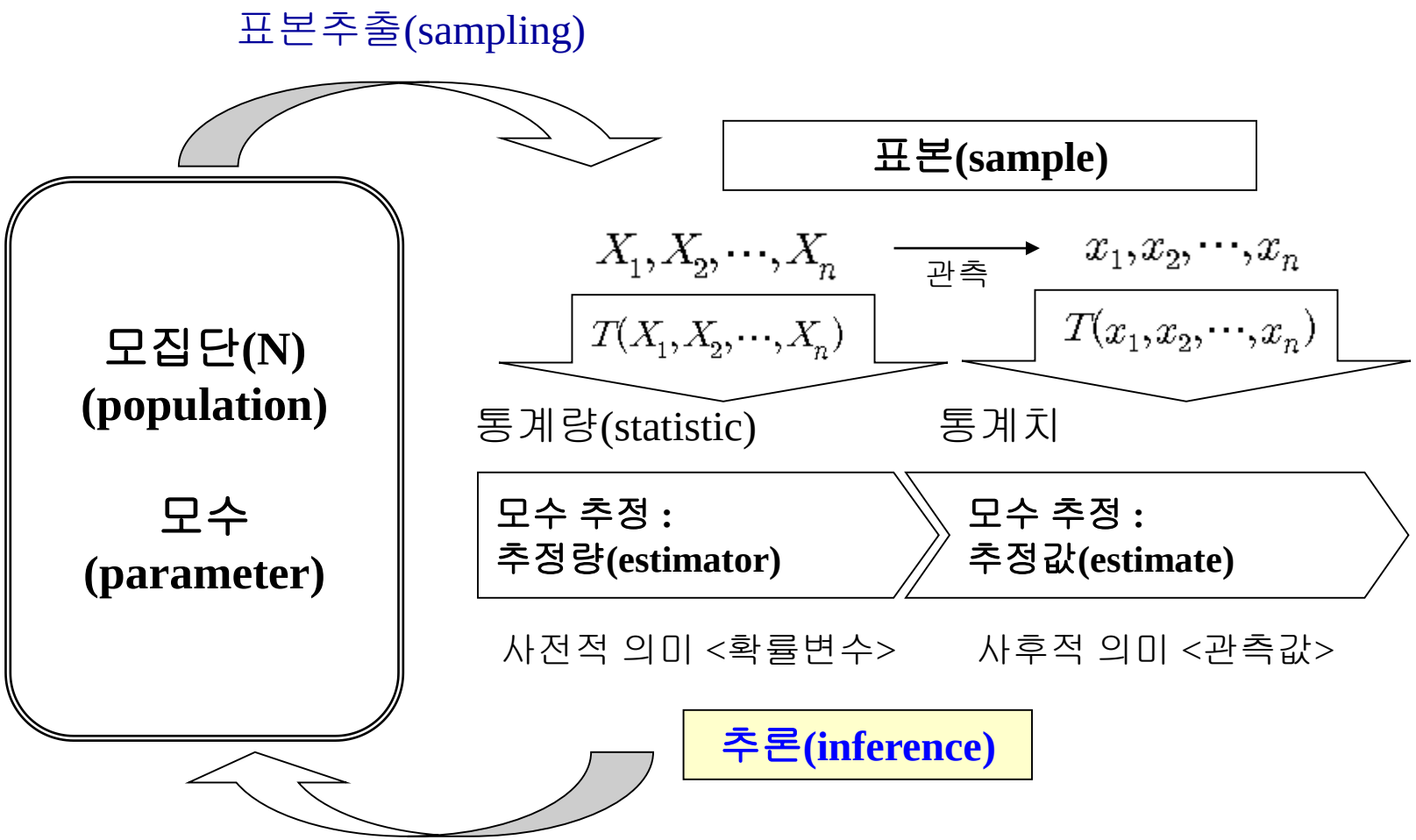
7장 표본추출방법과 표본분포

1. 모수와 통계량
2. 표본추출
3. 표본평균의 표본분포와 중심극한정리
4. 추가분포
5. 표본비율의 표본분포

1 모수와 통계량


표본추출(sampling)
표본(sample)
모집단(N)
(population)
모수
(parameter)
$X_1, X_2, \cdots, X_n$
관측
$x_1, x_2, \cdots, x_n$
$T(X_1, X_2, \cdots, X_n)$
$T(x_1, x_2, \cdots, x_n)$
통계량(statistic)
통계치
모수 추정 :
추정량(estimator)
모수 추정 :
추정값(estimate)
사전적 의미 <확률변수>
사후적 의미 <관측값>
추론(inference)

7장 표본추출방법과 표본분포

표본오차 :

표본을 조사하여 그 결과를 모집단의 결과라고 일반화함으로써 발생하는 오차.

(1) 모집단을 대표하지 못하는 비전형적인 표본을 뽑았기 때문에 생긴 오차. (잘못된 표본추출법에 의한 오차)

(2) 표본의 크기 때문에 생긴 오차.

비 표본오차 :

표본의 추출방법과 관계없이, 즉 전수조사를 하더라도 발생하는 오차를 말하며

이는 사람의 부주의, 무지, 측정도구 및 기타 자료의 잘못으로 생긴 오차.

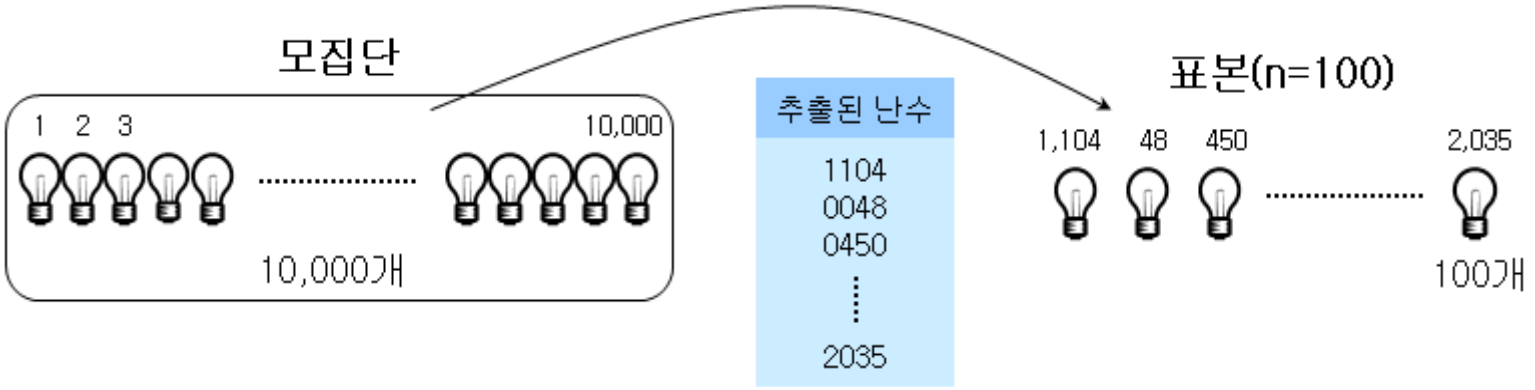
➔ 비 표본오차를 줄이기 위해서는 조사원의 교육, 정확한 측정도구의 설계, 자료의 정밀한 검토 등을 통하여 가능하다.

➔ **표본오차를 줄이기 위해서는 편의(bias)가 없는 표본을 뽑고, 적절한 표본추출 방법을 사용해야 한다.**

7장 표본추출방법과 표본분포

단순임의추출법 (SRS : Simple Random Sampling)

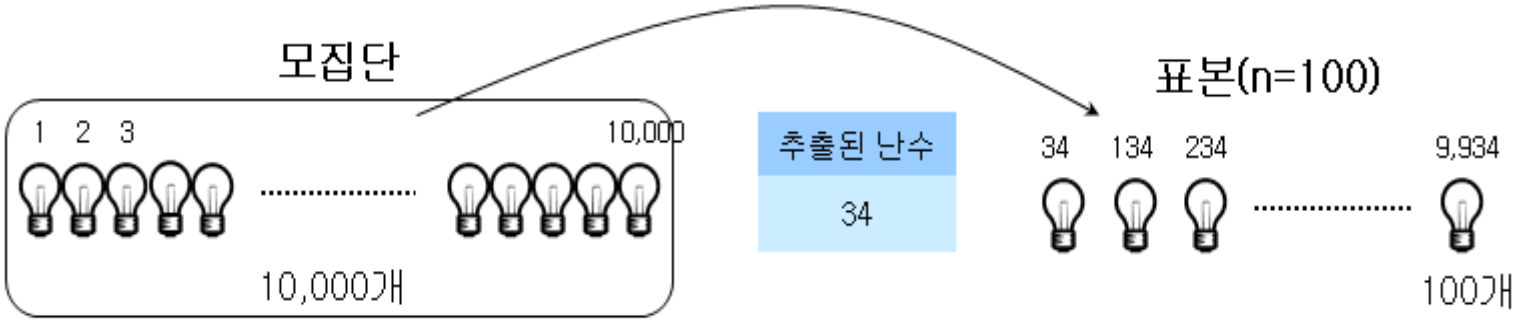
개 념	모집단을 구성하는 각 요소가 표본으로 선택될 확률을 동등하게 부여하여 추출하는 방법
절 차	• 모집단의 각 구성요소에 일련번호 부여 • 랜덤하게 정해진 표본의 수 만큼 개체를 선택 (추첨, 난수표, 컴퓨터 등 사용 ➡ 랜덤성(randomness) 유지)



7장 표본추출방법과 표본분포

계통추출법 (Systematic Sampling)

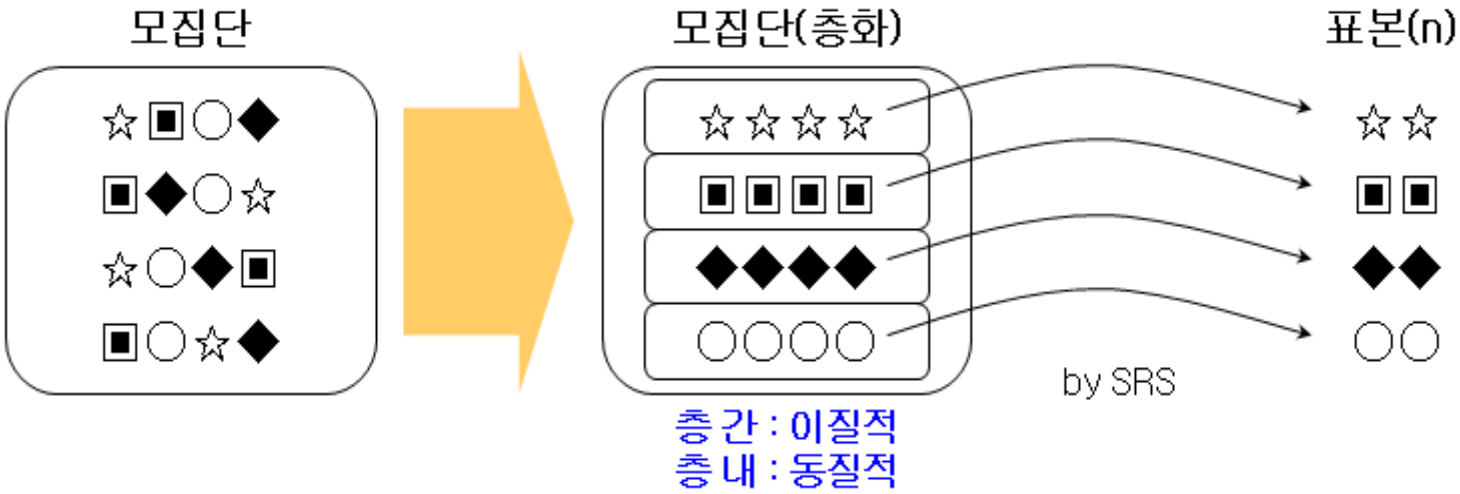
개 념	모집단의 구성 요소들이 자연적 순서 또는 일정한 질서에 따라 배열된 목록에서 일정간격을 두고 추출하는 방법
절 차	• 모집단의 각 구성요소에 일련번호 부여 • 1~k까지의 개체들 중에서 난수를 이용하여 한 개체를 선택 (k = 모집단 수 / 표본 수) • 그 개체에 부여된 일련번호에 k 씩 더해가며 해당 개체를 선택



7장 표본추출방법과 표본분포

층화 추출법 (Stratified Sampling)

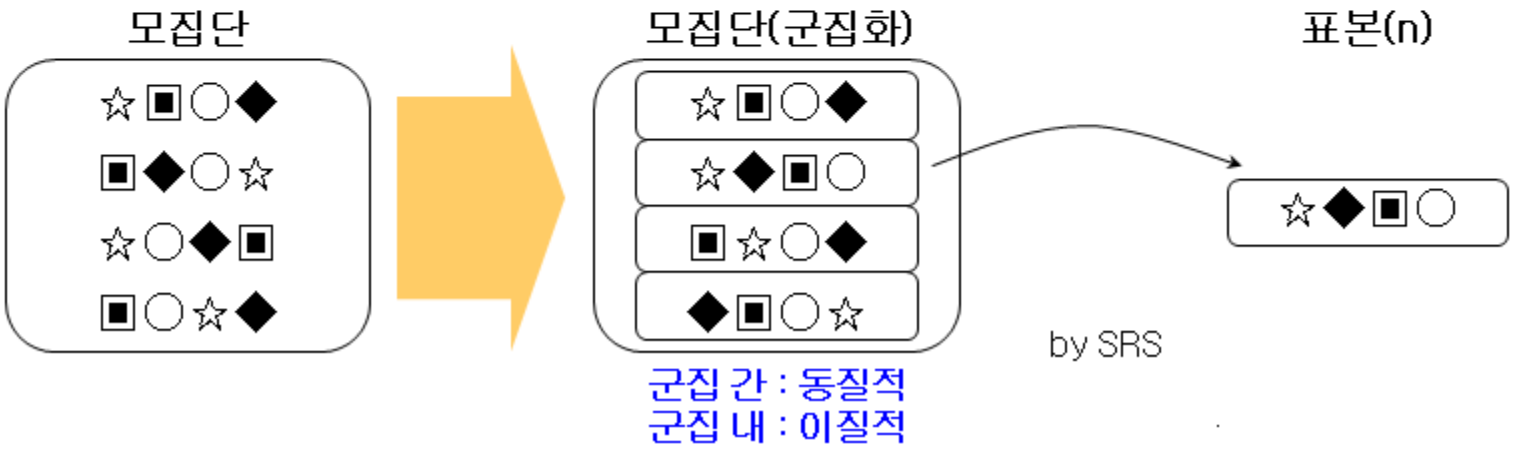
개 념	모집단을 특성에 따라 동질적인 몇 개의 층(strata)으로 구분하고, 각 층으로부터 단순임의추출법에 의해 표본을 추출하는 방법
절 차	• 모집단을 특성에 따라 몇 개의 층으로 구분 • 각 층으로부터 SRS 방법에 의해 일정 수 만큼의 개체를 선택



7장 표본추출방법과 표본분포

군집추출법 (Clustering Sampling)

개 념	모집단을 이질적인 몇 개의 군집(cluster)으로 구분한 다음, 구분된 군집을 추출단위로 하여 랜덤하게 몇 개의 군집을 추출하고 이를 전수조사 또는 단순임의추출법에 의해 표본을 추출하는 방법
절 차	• 모집단을 몇 개의 군집으로 구분하여 각 군집에 번호를 부여 • SRS 방법과 같이 군집들 중에서 랜덤하게 하나를 선택 • 선택된 군집에 있는 모든 개체를 선택 • 원하는 표본 수를 얻을 때까지 계속



7장 표본추출방법과 표본분포

비 확률표본추출법 :

- 표본추출과정에서 표본설계자의 주관성이나 판단이 내포되어 있는 표본추출방법.

(1) 편의추출법 : 연구자가 임의로 모집단에서 표본을 추출하는 방법.

(2) 판단추출법 : 연구자가 가지고 있는 모집단에 대한 지식을 근거로 모집단을 대표할 만한 표본을 추출하는 방법.

(3) 할당추출법: 모집단의 특성의 비율을 근거로 같은 비율을 갖는 표본을 추출하는 방법.

7장 표본추출방법과 표본분포

[Example] 대학교 내에서의 표본추출 : **12,000**명인 모 대학생의 의견조사

➔ 모집단 **12,000**명중 표본을 **240**명 뽑는 경우

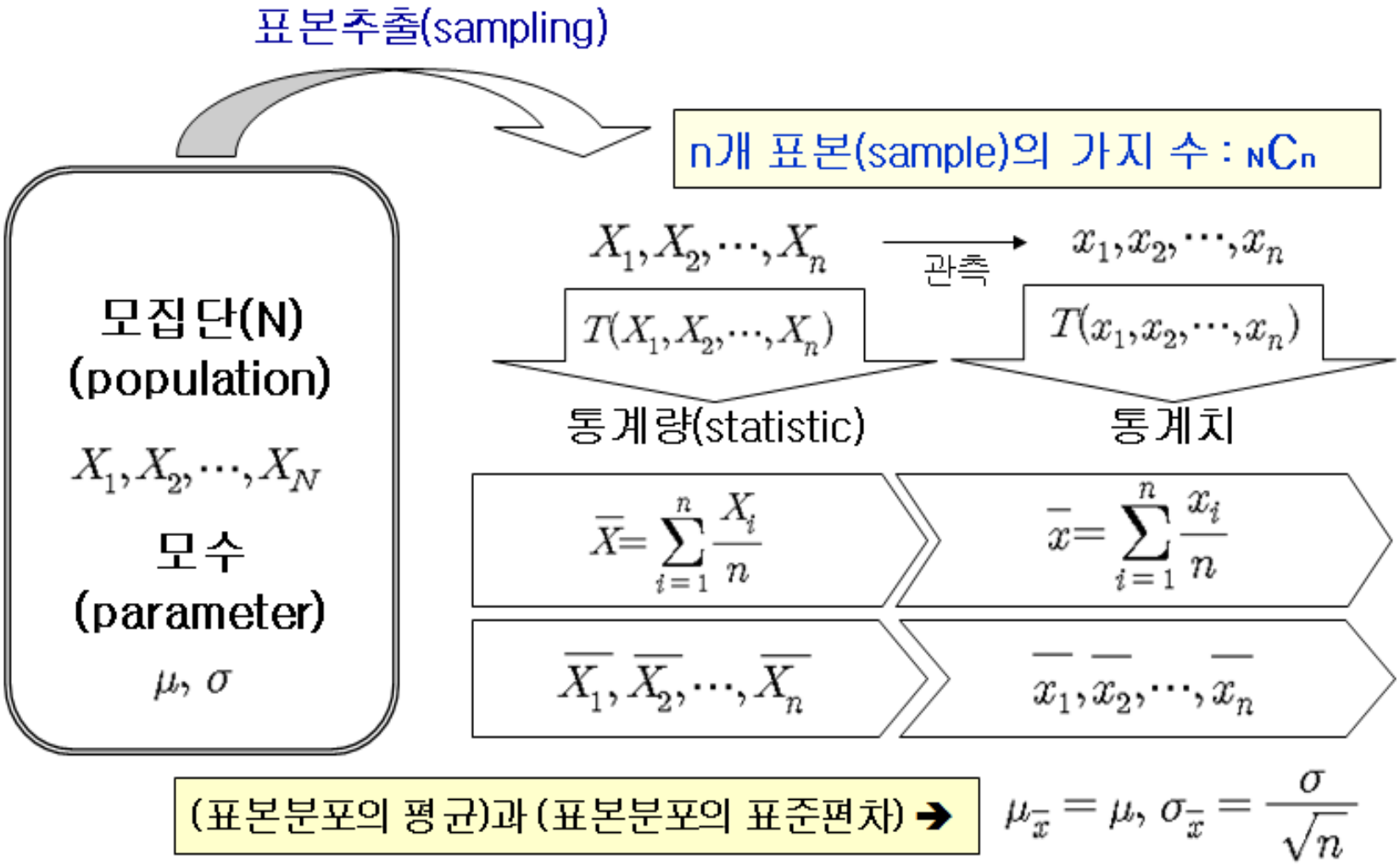
- 층화 추출법 : 단과대학별로 구분하여 집계할 필요가 있는 경우, 전체 표본 240명을 크기에 비례하여 각 단과대학에 배정하기로 하자.

모집단		표본	
: 제 1단과대학	2000명 (2/12),	: 제 1단과대학	40명 (2/12),
제 2단과대학	1000명 (1/12),	제 2단과대학	20명 (1/12),
제 3단과대학	3000명 (3/12),	제 3단과대학	60명 (3/12),
제 4단과대학	2000명 (2/12),	제 4단과대학	20명 (2/12),
제 5단과대학	4000명 (4/12),	제 5단과대학	80명 (4/12),
합계	12,000명	합계	240명

➔ 제 **1**단과대학에서 어떻게 **40**명을 추출할까 ? : 만일 **1**단과대학의 **8**개 학과의 정원이 거의 비슷하게 구성되어 있다면 학과가 학생들로 이루어진 군집(집락)이 됨

➔ **층화 3단집락추출 :** 층(단과대학), 1단계(4개 학과), 2단계(2개 학년), 3단계(5명) 추출 **(검토 : 4개학과 * 2개 학년 * 5명씩 = 40명)**

7장 표본추출방법과 표본분포



(표본분포의 평균)과 (표본분포의 표준편차) ➔ $\mu_{\bar{x}} = \mu,\ \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

7장 표본추출방법과 표본분포

표본평균 $\overline{X}$의 평균과 분산

평균 μ와 분산 σ^2을 갖는 무한모집단으로부터 단순확률추출에 의해 크기 n인 표본을 추출한다면, 표본평균 $\overline{X}$의 기댓값과 분산은 각각 다음과 같다.

$$E(\overline{X}) = \mu \ , \ Var(\overline{X}) = \frac{\sigma^2}{n}$$

중심극한정리

평균 μ와 분산 σ^2을 갖는 무한모집단으로부터 단순확률추출에 의해 크기 n의 표본을 추출한다면, 표본평균 $\overline{X}$는 표본크기 n이 커짐에 따라 근사적으로 평균이 μ이고 분산이 σ^2/n인 정규분포를 따른다.

따라서 표준화 확률변수 $Z = \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}}$의 분포는 표본크기 n이 커짐에 따라 근사적으로 표준정규분포 $N(0,1)$을 따른다.

$X_1, X_2, \cdots, X_n$: 평균 μ, 분산 σ^2 인 임의의 모집단으로부터의 확률표본

➡ n이 충분히 클 때, $\overline{X} \overset{\cdot}{\sim} N\left(\mu, \dfrac{\sigma^2}{n}\right) \Leftrightarrow Z = \dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}} \overset{\cdot}{\sim} N(0,1)$

주1) 대체적으로 n≥25(또는 30)이면, 근사 정도가 만족할 만 하다.

7장 표본추출방법과 표본분포

$X_1, X_2, \cdots, X_n$: 모평균 μ, 모분산 σ^2인 모집단으로부터의 확률표본

➡ 표본평균 $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

● **표본평균의 분포에 대한 성질 :** $E(\overline{X}) = \mu$, $Var(\overline{X}) = \frac{\sigma^2}{n}$, $sd(\overline{X}) = \frac{\sigma}{\sqrt{n}}$

표본의 크기가 클 때 $\overline{X}$는 모집단의 평균인 μ 근처에 밀집되어 분포!

[1] 정규분포 $N(\mu, \sigma^2)$인 경우 ➡ $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

[2] 정규분포가 아닌 경우 ➡ ?

예제1 : 충북대학교의 전체 교수의 수는 800명이고 평균연령은 45.8세이고 표준편차가 15세이다. 만약 크기가 100명인 표본을 모두 추출할 경우 이 평균연령의 표본분포의 평균과 표준편차는? 또한 평균연령의 확률분포는 ?

이 대학교의 전체 교수가 모집단, 모집단의 평균이 바로 표본분포의 평균이므로

45.8세가 표본분포의 평균이고, 표준편차는 $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{100}} = 1.5$ 이 된다.

$\overline{X} \sim N(\mu, \frac{\sigma^2}{n}) = N(45.8, 1.5^2)$

7장 표본추출방법과 표본분포

예제2 신안전 타이어 공업주식회사에서는 새로운 형태의 광폭 타이어를 생산하고 있다. 이 타이어의 **평균수명이 50,000km**이고 **표준편차는 14,000km**인 **정규분포**를 하고 있다고 알려져 있다.

(1) 수명이 60,500km이상인 타이어는 전체 생산량의 몇 %가 되는가?

(2) 만일 크기가 49개인 표본을 뽑는다면 이 표본의 평균이 48,000km 이하가 될 확률은 얼마인가?

모집단 :

$$P(X \geqq 60{,}500) = P(Z \geqq \frac{X-\mu}{\sigma}) = P(Z \geqq \frac{60{,}500-50{,}000}{14{,}000})$$
$$= P(Z \geqq 0.75) = 1-0.7734 = 0.2266(22.66\%)$$

표본집단 :

$$\mu_{\bar{x}} = E(\bar{X}) = 50{,}000km,\ \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{14{,}000}{7} = 2{,}000km$$

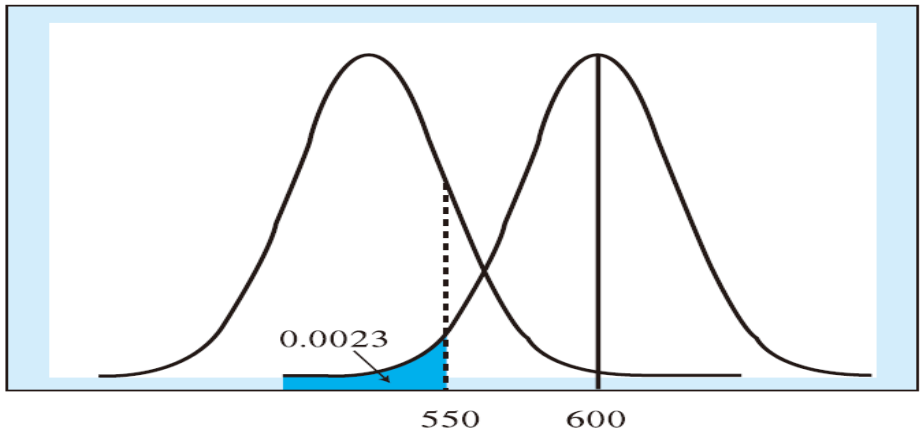
$$P(\bar{x} \leqq 48{,}000) = P(Z \leqq \frac{\bar{x}-\mu_{\bar{x}}}{\sigma_{\bar{x}}}) = P(Z \leqq \frac{48{,}000-50{,}000}{2{,}000})$$
$$= P(Z \leqq -1.0) = 1-0.8413 = 0.1587(15.87\%)$$

7장 표본추출방법과 표본분포

[예제 3] 미국 소재 한 대학의 총장은 대학교 졸업생들의 주당 평균수입이 **$600**이라고 주장한다. 표준편차가 **$100**일 때, 무작위로 추출한 **32**명의 졸업생들의 주당 평균수입이 **$550**이하일 확률은 ?

$$P(\overline{X} \leqq 550) = P(Z \leqq \frac{550-600}{\sqrt{312.5}}) = P(Z \leqq -2.83) = 0.0023$$

[예제 4] 위 예에서 확률표본 **32**명의 졸업생들의 주당 평균수입이 **#550**으로 조사되었다면, 주당 평균수입이 **$600**이라고 주장하는 이 대학의 총장 주장에 대해 어떠한 결론을 내릴 수 있는가 ?



표본평균 $\overline{X}$의 평균이 $550과 $600일 때의 확률분포

7장 표본추출방법과 표본분포-추가분포(T-분포)

$X_1, X_2, \cdots, X_n$이 정규모집단 $N(\mu, \sigma^2)$으로부터의 확률표본일 때, 표본평균 $\overline{X}$에 대하여

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \text{ 즉, } Z = \frac{\overline{X} - \mu_{\overline{x}}}{\sigma/\sqrt{n}}$$

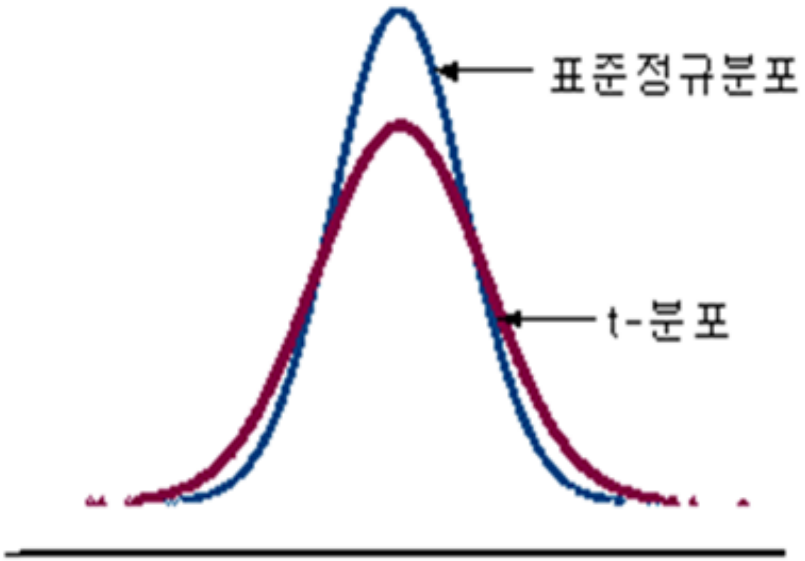
표본의 크기가 작고, 모집단의 σ을 모르는 경우 :

σ => 표본표준편차 $S = \sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2/(n-1)}$, Z => $t = \frac{\overline{X} - \mu_{\overline{x}}}{S/\sqrt{n}}$

$$Z \sim N(0,1) \Rightarrow t \sim t(n-1)$$

7장 표본추출방법과 표본분포-추가분포(T-분포)

아래 그림에서 보는 바와 같이 표준정규분포와 비슷하게 평균이 0이고 분포의 모양이 평균을 중심으로 좌우가 대칭이다. 그러나 t-분포는 표준정규분포보다 평균 주위의 높이가 낮고 양쪽 꼬리 근처가 더 두꺼운 모양을 갖고 있다.



표본의 크기가 (n>30) 커지면 t-분포도 정규분포에 근사 한다.

7장 표본추출방법과 표본분포-추가분포(T-분포)

[예제 5] 어느 유리공장에서 판유리를 생산하는데 최근 생산된 유리를 무작위로 **16장** 뽑아서 조사한 결과 **표준편차는** 0.8mm이었고, 과거에 생산된 유리의 평균두께가 정규분포에 따르며 **평균이** 4mm가 될 때, **표본평균이** 3.58mm 이하가 될 확률은?

[풀이]

$$\mu_{\bar{x}} = \mu = 4,\ S = 0.8,\ \overline{X} = 3.58,\ n = 16$$

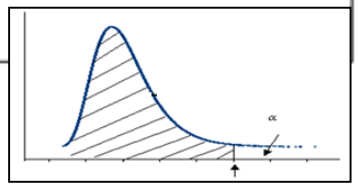
$$t = \frac{\overline{X} - \mu_{\bar{x}}}{S/\sqrt{n}} = \frac{3.58 - 4}{0.8/\sqrt{16}} = -2.1$$

$$\therefore\ P(\overline{X} \leqq 3.58) = P(t \leqq -2.1) = 1 - 0.975 = 0.025$$

t~t(n-1), 즉, 자유도가 15인 t=2.1은 t-분포표에서 t=2.131에 근사 값으로서 확률이 0.975이다.

7장 표본추출방법과 표본분포-추가분포(카이제곱-분포)

분산의 표본분포 : 모집단에서 크기가 n인 가능한 모든 표본을 추출하여 각 표본의 분산들이 이루는 확률분포를 말한다.



모분산 : $\sigma^2 = \sum_{i=1}^{N} \frac{(X_i - \mu)^2}{N}$ 표본분산 : $S^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}$

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

$1-$표본 : $(X_{11}, X_{12}, \cdots, X_{1n})$ ==> $\bar{X_1}$, $S_1^2 = \sum_{i=1}^{n} \frac{(X_{1i} - \bar{X_1})^2}{n-1}$

$2-$표본 : $(X_{21}, X_{22}, \cdots, X_{2n})$ ==> $\bar{X_2}$, $S_2^2 = \sum_{i=1}^{n} \frac{(X_{2i} - \bar{X_2})^2}{n-1}$

⋮ ⋮

$k-$표본 : $(X_{k1}, X_{k2}, \cdots, X_{kn})$ ==> $\bar{X_k}$, $S_k^2 = \sum_{i=1}^{n} \frac{(X_{ki} - \bar{X_k})^2}{n-1}$

(N개의 모집단에서 n개의 표본을 뽑을 가지 수 : ${}_N C_n = k$)

각 표본분산 $S_1^2, S_2^2, \cdots, S_k^2$ 이 이루는 확률분포를 말한다.

헤어드라이어 생산공정의 원래 설계에 따르면 **표준편차는** 30시간인데 이때 **표본** 16개의 표준편차가 36.6시간보다 작을 확률은?

$$P(S \leq 36.6) = P(\chi^2 \leq \frac{(n-1)S^2}{\sigma^2})$$

$$P(\chi^2 \leq \frac{(16-1)36.6^2}{30^2}) = P(\chi^2 \leq 22.3)$$

부록 ➔ 1-0.1= 0.9

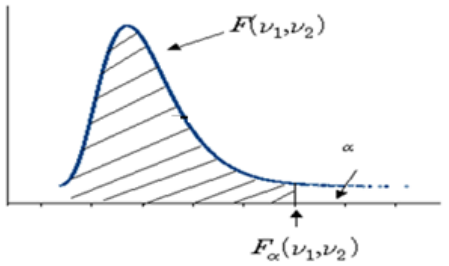
7장 표본추출방법과 표본분포-추가분포(F-분포)

F 분포(F Distribution)의 정의

모집단이 정규분포를 이루고 분산이 각각 σ_1^2, σ_2^2인 두 모집단에서 표본의 크기가 각각 n_1, n_2인 표본을 추출하여 표본분산을 계산한 것이 S_1^2, S_2^2이라 할 때 표본분산과 모분산의 비율로 이루어진 χ^2의 비율이 F-분포를 이룬다.

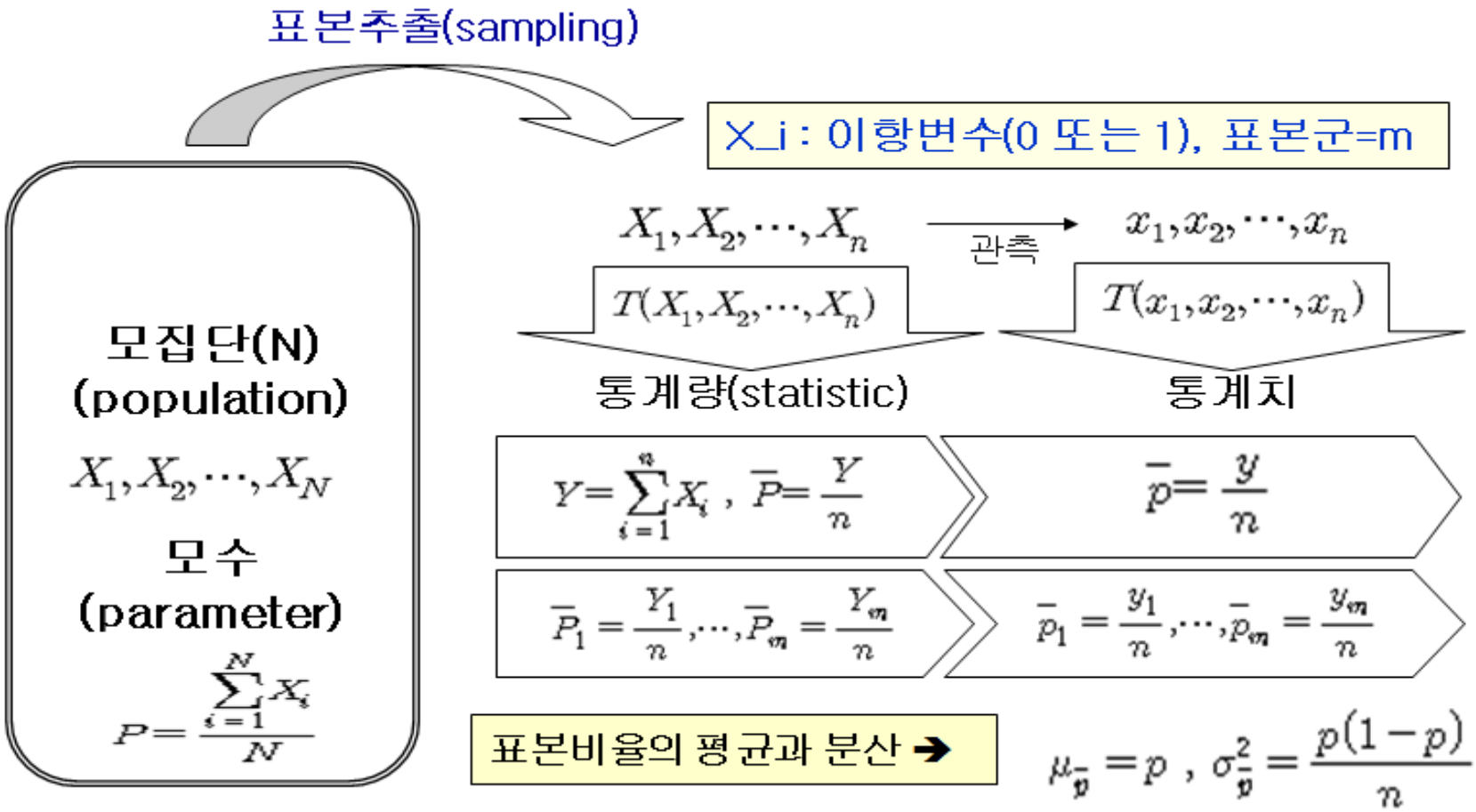
$$\chi_1^2 = \frac{(n_1-1)S_1^2}{\sigma_1^2} \sim \chi^2(n_1-1),\ \chi_2^2 = \frac{(n_2-1)S_2^2}{\sigma_2^2} \sim \chi^2(n_2-1)$$

$$\frac{\chi_1^2/\nu_1}{\chi_2^2/\nu_2} = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(\nu_1, \nu_2)$$



ν_1(분자의 자유도) $= n_1 - 1$, ν_2(분모의 자유도) $= n_2 - 1$

7장 표본추출방법과 표본분포



$X_i \sim ber(p)$, $Y = \sum_{i=1}^{n} X_i$ => $Y \sim B(n, p)$: $E(Y) = np$, $V(Y) = np(1-p)$

=> $E(\frac{Y}{n}) = p$, $V(\frac{Y}{n}) = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$ => $\bar{p} \sim (p, \frac{p(1-p)}{n})$

7장 표본추출방법과 표본분포

표본비율 $\hat{p}=Y/n$의 근사분포

표본비율 $\hat{p}$의 분포는 $np \geq 5$이고 $n(1-p) \geq 5$를 만족하면 표본크기 n이 커짐에 따라 근사적으로 평균이 p이고 분산이 $p(1-p)/n$인 정규분포를 따른다. 따라서 확률변수 $Z=(\hat{p}-p)/\sqrt{p(1-p)/n}$의 분포는 표준정규분포 $N(0,1)$에 근사한다.

[예제 6] 어느 감기약의 치유율은 **90%**라고 알려져 있다. 올해에 유행하고 있는 감기에 대해서도 **90%**의 치유율을 보장할 수 있는가를 알아보기 위해 **100**명의 감기환자에게 감기약을 투여하였다. 이들 중 감기로부터 완전히 회복된 사람들의 비율이 **85%~95%**내에 들어 있을 확률은 **?**

$$np = 100 \times 0.9 = 90 > 5\,,\; n(1-p) = 100 \times 0.1 = 10 > 5$$

$$by\ C.L.T\ :\ \hat{p} \sim N\left(0.9\,,\ \frac{0.9(1-0.9)}{100}\right) \rightarrow P(0.85 < \hat{p} < 0.95) \simeq P(-1.67 < Z < 1.67) = 0.905$$

[예제 7] 우리나라 남자의 약 6%에서 적녹 색각이상이 나타난다고 한다. 우리나라 남자 중에서 임의로 200명을 뽑았을 때 표본에서 적녹 색각이상인 사람이 8%이상 나타날 확률은 얼마인가 ?

$X \sim B(200, 0.06)$ ➡ $\hat{p} \sim N(0.06, 0.00028)$