[도입사례]

우리나라 2005년도의 TOEIC시험 평균점수가 593점이고, 2005년 한 해 동안 우리나라에서 TOEIC에 응시한 인원은 1,856,307명(중복응시는 없다고 가정 하지요)이라고 집계되었다고 합니다. 즉, 응시자마다 천차만별인 1,856,307명의 개별 TOEIC점수를 모두 합해서 전체 응시자로 나눈 결과가 593점이라는 것을 알려주는 자료이지요. 그런데 1,856,307명의 평균점수와 4,900여 만 명에 달하는 대한민국 국민 전체(만일 전체 국민이 TOEIC시험을 본다면)의 평균점수와는 차이가 날까요? 차이가 난다면 얼마나 날까요? 또 대한민국 전체의 TOEIC평균점수를 알기 위해 1,856,307명이라는 많은 사람들이 응시를 해야 할까요? 적은 수의 응시자들(예: 1,000명)의 시험결과로 대한민국 전체의 TOEIC평균을 예측할 수는 없을까요?

생각해 볼 문제

① 표본을 관찰한 자료에서 유도된 표본통계량이 전체 모집단을 대표하기 위한 조건은 무엇일까요?

② 추리통계분석과 추정 그리고 가설검정은 어떠한 관계가 있을까요?

• 출처 : 신민철 저 “경영경제 통계학의 기초, 창민출판사, 2010년.

제1절 확률과 통계학

1. 추리통계분석

✦ 통계학(statistics)

- 계량적 자료(quantitative data)를 분석하는 이론 및 방법을 다루고 있는 학문분야

➪ 기술통계와 추리통계로 구분

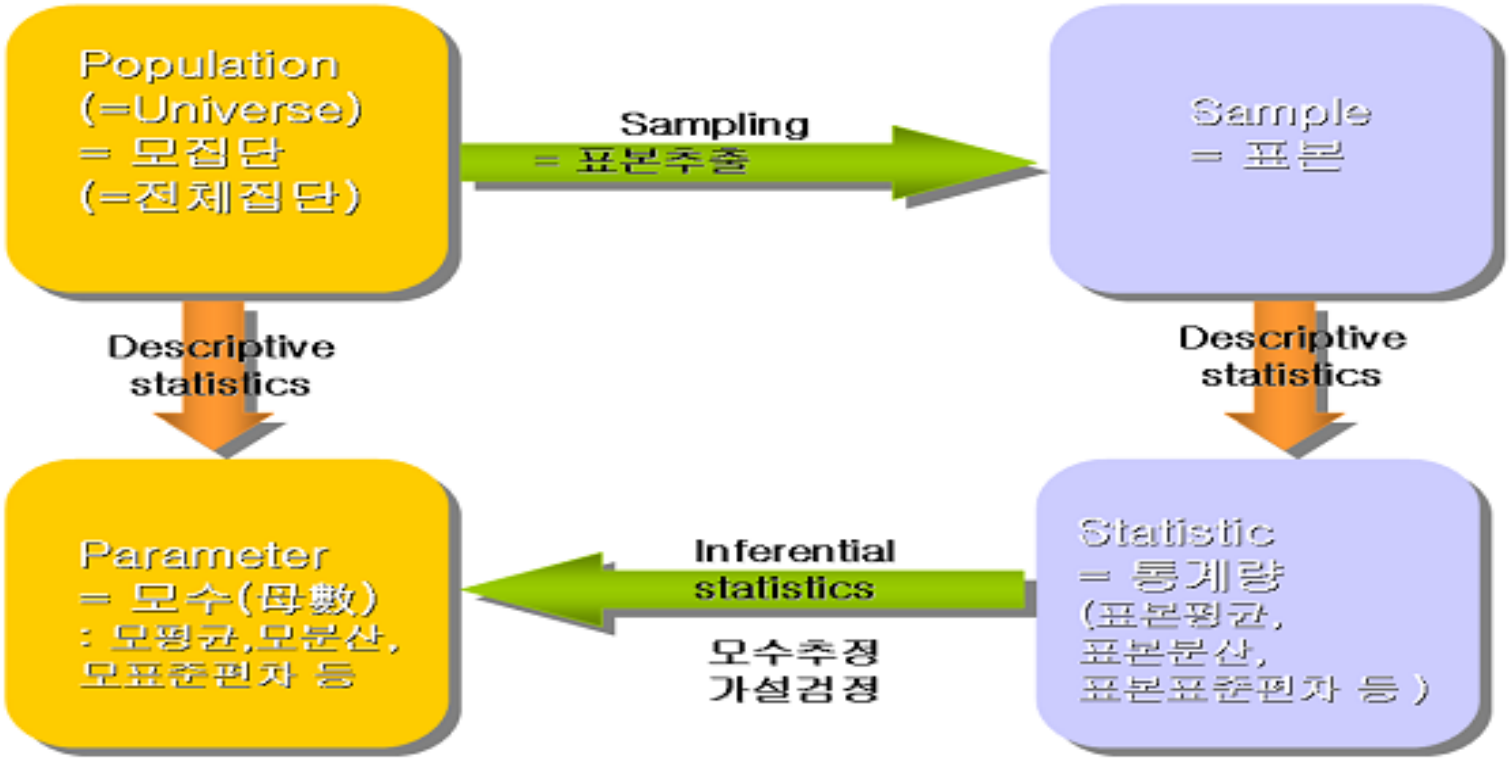
✦ **기술**통계(記述統計: descriptive statistics)

- 측정된 현상의 특징을 설명(즉, 기술)하고 요약해 주는 정보를 다루는 통계학의 분야
- 모수와 표본통계량을 계산해내는 통계학의 분야
  - **모수**(母數: population parameter)
    = 모집단을 요약·설명해 주는 기술통계도구
    예) 모평균(population mean)과 모분산(population variance) 등
  - **표본통계량**(sample statistic)
    = 표본을 요약·설명해 주는 기술통계도구
    예) 표본평균(sample mean)과 표본분산(sample variance) 등

- 기술통계는 전수조사와 표본조사의 경우에 동일하게 적용가능하나 현실적으로 전수조사는 수행되지 않음

➪ 현실적으로 모수(black box 속에 가려져 있는 존재)는 미지수

[그림 8-1] 기술통계와 추리통계



✦ **추리통계**(推理統計 혹은 推測統計: inferential statistics)

- 알려진 부분적인 정보를 바탕으로 미지의 전체 정보를 추측하는 기능을 하는 통계학의 분야
- 표본통계량을 가지고 모수(parameter)를 추정하는 역할을 하는 통계학의 분야

➪ 모수추정과 가설검정으로 구분

- **모수추정**(parameter estimation)
  = 표본통계량으로 모수를 추측하는 추리통계분야

- **가설검정**(hypothesis testing)
  = 모수에 대해 특정한 가설을 설정한 후 표본통계량을 기초로 이 가설의
  진위를 판단하는 추리통계분야

2. (표본)통계량과 모수

1) 모 수

- 전수조사를 한다면, 기술통계분석만으로도 정확하게 현상을 설명하거나 현상간의 관계를 알아낼 수 있을 것임
- 그러나 실질적으로 전수조사는 거의 이루어지지 않음
  - **유한**모집단의 경우
    - → 이론적으로는 전수조사 가능
    - → 유한모집단의 크기가 큰 경우에는 현실적으로 전수조사 불가능

※ 유한모집단(finite population)
= 모집단의 요소(즉, 구성원) 수가 유한한 모집단

▫ **무한**모집단의 경우

→ 이론적으로도 실제 전수조사는 불가능

※ 무한모집단(infinite population)

= 모집단 요소의 수가 무한한 모집단

예) 복원추출과 같이 이론상 무한대의 시행이 가능한 사건으로 구성된 모집단

유한모집단 평균(= μ_X) = $\sum X_i / N$ (식 8-1[=식 6-1A])

유한모집단 분산(= σ^2) = $\sum (X_i - \mu)^2 / N$ (식 8-2[=식 6-2A])

무한모집단 평균(= μ_X) = $E(X) = \sum X_i P(X_i)$ (식 8-3[=식 7-14])

무한모집단 분산(= σ^2)

$= Var(X) = \sum [X_i - E(X)^2] P(X_i)$

$= E(X_i^2) - [E(X)]^2$ (식 8-4[=식 7-17])

⇨ 확률변수의 평균/분산과 무한모집단의 평균/분산은 동일한 개념

2) 표본통계량

◆ 유한모집단의 경우도 실질적으로 전수조사는 이루어지지 않음(∵비용 등)

⇨ 표본조사를 이용해서 모집단에서 작동하는 현상의 특징을 추측해야 함

표본평균(= $\bar{X}$) = $\sum X_i/n$ (식 8-5[=식 6-1B])

표본분산(= S^2) = $\sum(X_i - \bar{X})^2/(n-1)$ (식 8-6[=식 6-2B])

3) 표본오차

- 바람직한 표본조사
  - 표본통계량(sample statistic)과 모수(population parameter)간의 차이를 최소화하는 표본조사

- **표본/표집오차**(sampling error)
  - 표본통계량과 모수간의 차이

표본오차 = 표본통계량 - 모수 (식 8-7)

- 표본조사의 목표
  - 표본통계량으로 모수를 추측

모수 = 표본통계량 - 표본오차 (식 8-8)

⇨ 표본오차를 모르는 한 모수 추정 불가

→ 표본오차의 값은 표본통계량의 값에 의존

→ 표본통계량의 값은 모집단에서 추출되는 표본(의 특성)에 의존

→ 특정 표본이 추출될 가능성은 확률로 정의

→ 표본오차는 확률변수

⇨ 확률분포이론(probability distribution theory) 등을 통해 표본오차가 어떠한 크기를 가질지를 추측(infer) 가능

⇨ 표본통계량(=표본조사를 통해 계산)과 표본오차(=확률분포이론을 통해 추측)를 알게 되면 모수를 알(=추정할) 수 있게 됨

→ 표본통계량의 분포는 해당 확률분포의 유형과 특성(즉, 대표값과 산포도)을 파악하면 알 수 있게 됨

⇨ 표본오차를 추측하기 위해 확률분포이론에 대한 이해 필요

→ 표본통계량의 분포 즉 **표본분포**(sampling distribution) 이해 필요

(1) 표본분포

✦ 빈도분포(frequency distribution)

- 조사대상 변수가 가지는 값[=응답범주(response category)]별 빈도(=응답자수)의 분포
- 조사대상 현상을 이해하는 기초자료
- 하나의 표본(one sample)을 대상으로 해서 도출된 결과를 그림으로 표현한 것

➪ 표본통계량인 표본평균과 표본분산(혹은 표본 표준편차)도 각각 1개씩 계산

➪ 표본통계량이 모수와 얼마나 유사한지(즉, 표본오차가 얼마나 되는지) 추정 필요

✦ 하나의 모집단에서 표본을 추출할 수 있는 경우의 수

예) 25명으로 이루어진 모집단에서 5명으로 이루어진 표본을 추출하는 경우의 수:

$_{25}C_5$개= 53,130가지의 표본 추출가능

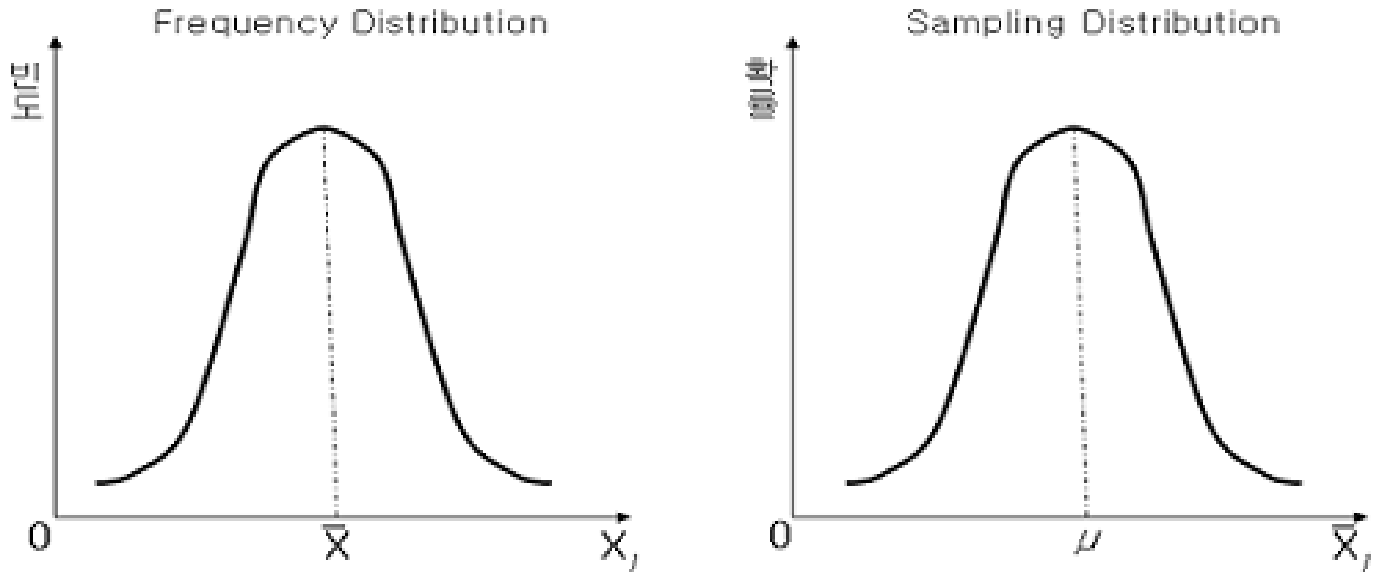
➪ 어떠한 표본을 추출하느냐에 따라서 각기 다른 표본평균과 표본분산이 도출됨

➪ 표본통계량들(=표본평균/분산들)의 분포(sampling distribution of sample means/variances), 즉 표본분포(=표집분포)를 구할 필요

✦ 빈도분포와 표본분포

➪ 두 분포가 모두 정규분포이나 가로축과 세로축의 내용은 다름

[그림 8-2] 빈도분포와 표본분포



✦ **표본분포**(sampling distribution)

◆ 표본통계량(sample statistic) 값들의 분포

➪ 표본평균의 표본분포, 표본분산(표본표준편차)의 표본분포, 표본비율(proportions)의 표본분포, 혹은 표본평균 차이(예: $\bar{X}_1 - \bar{X}_2$)의 표본분포 모두 표본분포에 해당

✦ **표본분포의 특징** -> **중심극한 정리에서 유도된 내용**

- 표본분포는 정규분포를 따름
- 표본평균의 평균(mean of sample means)은 모집단의 평균(population mean)과 동일
- 표본평균의 표준편차는 모평균의 표준편차를 표본크기의 제곱근으로 나눈 값($\sigma/\sqrt{n}$)
- 표본의 크기가 증가할수록 표본평균의 표본분포(sampling distribution of sample means)의 변화폭(variability), 즉 표준편차(=표준오차)의 크기가 작아짐

⇨ 표본의 크기가 커질수록 그 표본평균은 모집단평균과 가까워지고, 표본오차 (sampling error)는 작아짐

⇨ 모집단의 분포유형과는 무관하게 표본평균의 분포의 특징을 알게 된다면, 표본오차(sampling error)를 알 수 있게 되는 것

✦ **중심극한정리 (Central Limit Theorem; CLT)** : 표본평균의 극한분포에 대한 정리

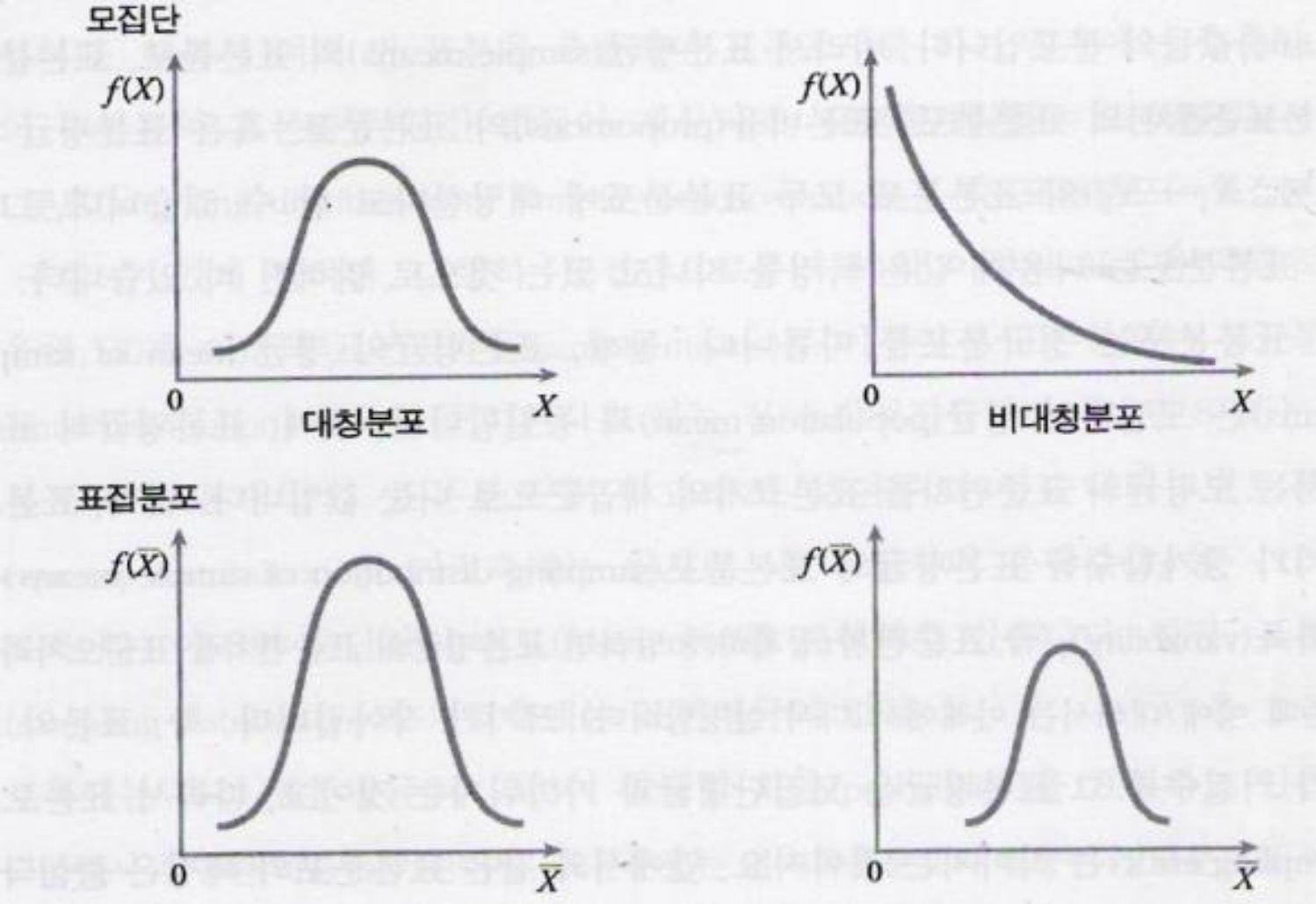
- **제1정리** : 모집단의 분포가 *정규분포이면* 표본평균($=\bar{X}$)은 표본 크기에 상관없이 정규분포를 이룬다.
- **제2정리** : 모집단의 분포가 *정규분포가 아니더라도* 표본의 크기가 점차 커질수록 표본평균의 분포는 근사적으로 정규분포를 이룬다
  => 일반적으로 이를 중심극한정리라 부름
- **제3정리** : 이항/포아송/카이제곱 분포도 n(표본 수)이 클 때 정규근사한다.
  cf. t분포의 정규근사는 **대수의 법칙**(표본의 크기가 커질수록 모평균 u에 근사한 표본평균을 얻을 확률이 커진다)에 의한 것임 (**p.231 참고**)

⇨ **유용성**

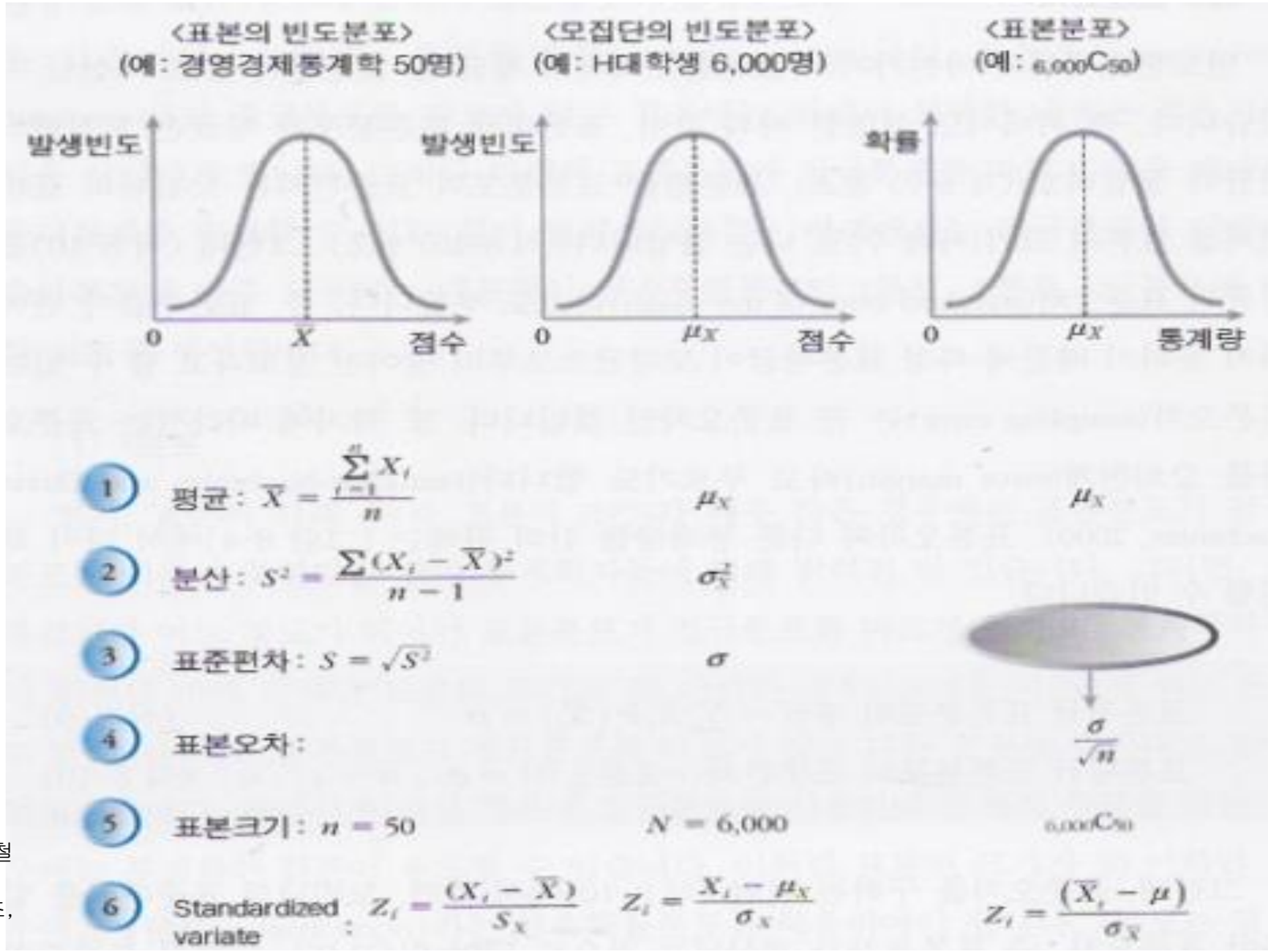
1) 그 결과가 모집단의 확률분포와 무관하다는 점, 모집단이 어떠한 분포이든 관계없이 그 모집단에서 추출된 확률표본의 평균의 분포는 표본의 크기가 증가함에 따라 항상 정규분포에 가까워진다.

2) $\sum X_i$의 분포 역시 n이 증가함에 따라 정규분포에 수렴한다.

[그림 8-3] 중심극한 정리(대수의 법칙)의 예



[그림 8-4] 표본의 빈도분포, 모집단의 빈도분포 및 표본분포



• 출처 : 신민철 저 "경영경제통계학의 기초, 창민출판사, 2010년.