제9장 추리통계분석

| 도입 사례 |

8장의 도입사례에서는 우리나라 2015년도의 TOEIC시험 평균점수가 677점이라는 자료를 살펴 본 바 있습니다. 즉, 2015년 한 해 동안 우리나라에서 TOEIC에 응시한 인원은 약 200만 명(중복응시는 없다고 가정 하지요)으로 집계되었고, 각 응시자마다 천차만별인 약 200만 명의 TOEIC점수 합계를 전체 응시자로 나눈 수가 677점이라는 것을 알려주는 자료이지요. 그런데 약 200만 명의 평균점수와 약 5,100만 명에 달하는 대한민국 국민 전체(만일 전체 국민이 TOEIC시험을 본다면)의 평균점수와는 차이가 날까요? 차이가 난다면 얼마나 날까요? 또 단순히 대한민국 전체의 TOEIC평균점수를 알기 위해 약 200만 명이라는 많은 사람들이 응시를 해야 할까요? 적은 수의 응시자들(예: 1,000명)의 시험결과로 대한민국 전체의 TOEIC평균을 예측할 수는 없을까요?

생각해 볼 문제 ————————————

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

② 추리통계분석과 가설검정은 어떠한 관계가 있을까요?

1. 추리통계분석

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

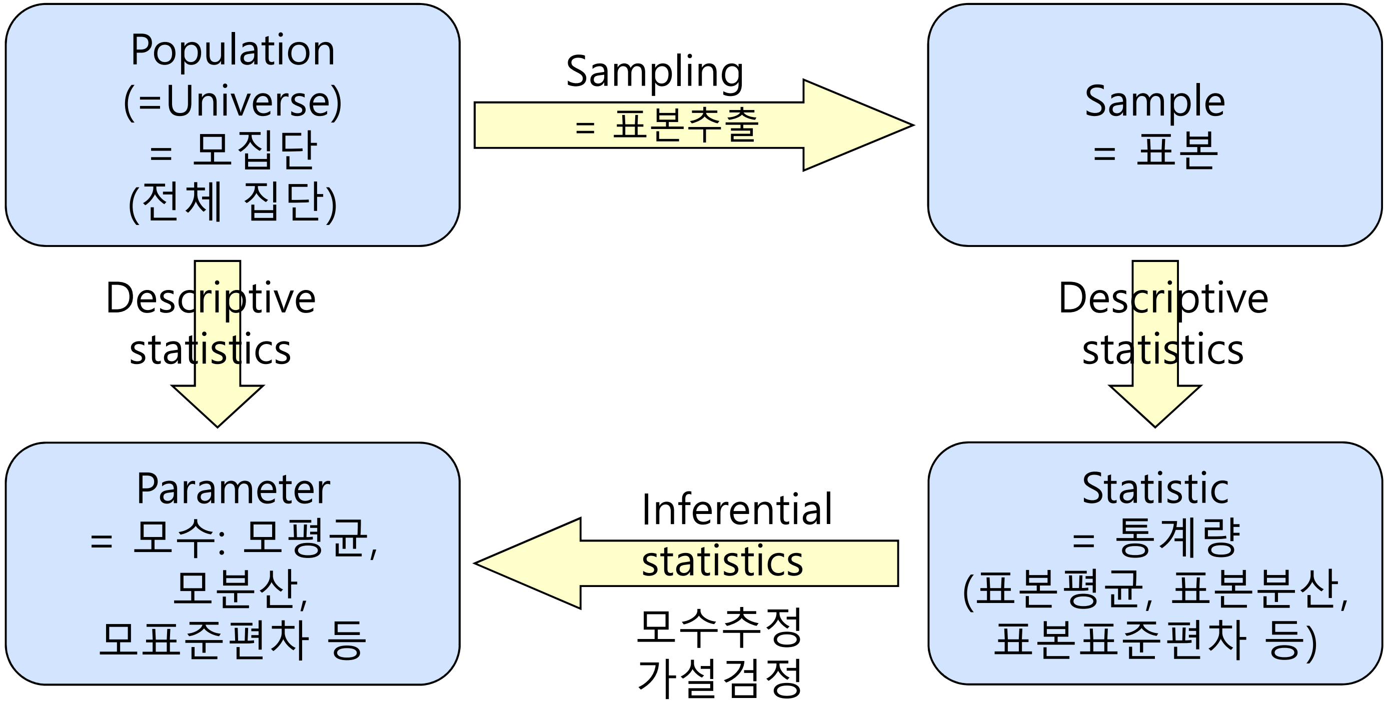
- 기술통계(記述統計: descriptive statistics)
  - 측정된 현상의 특징을 설명(즉, 기술)하고 요약해 주는 정보를 다루는 통계학의 분야
  - 모/표본평균(mean)과 모/표본분산(variance) 등 모수(모집단을 요약·설명해 주는 기술통계도구)와 표본통계량(표본을 요약·설명해 주는 기술통계도구)을 계산해내는 통계학의 분야
  - 기술통계는 전수조사와 표본조사의 경우에 동일하게 적용가능
  - 현실적으로 전수조사는 수행되지 않음(모수는 미지수)

- 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 모집단의 특징을 요약·설명하는 모수(parameter)를 추정하는 역할을 하는 통계학의 분야

- 기술통계는 전수조사와 표본조사의 경우에 동일하게 적용가능하나 현실적으로 전수조사는 수행되지 않음

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

[그림 9-1] 기술통계와 추리통계



2. 통계량과 모수

♦ 전수조사를 한다면, 기술통계분석만으로도 정확하게 현상을 설명하거나 현상간의 관계를 알아낼 수 있을 것임

♦ 실질적으로 전수조사는 거의 이루어지지 않음

☞ 표본조사를 이용해서 모집단에서 작동하는 현상의 특징을 추측해야 함

♦ 바람직한 표본조사

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

♦ **표본/표집오차**

▫ 표본통계량과 모수간의 차이

표본오차 = 표본통계량 – 모수 (식 9-1)

모수 = 표본통계량 – 표본오차 (식 9-2)

♦ 표본조사의 목표

▫ 통계량으로 모수를 추측

☞ 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

→ <u>표본오차는 확률변수</u>

☞ 통계학에서의 (확률)분포이론(distribution theory) 등을 통해 표본오차가 어떠한 크기를 가질지를 추측(infer) 가능

☞ 표본통계량(=표본조사를 통해 계산)과 표본오차(=분포이론을 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

→ <u>표본통계량의 분포</u> 즉 **표본분포**(sampling distribution) 이해 필요

1] 분포의 유형

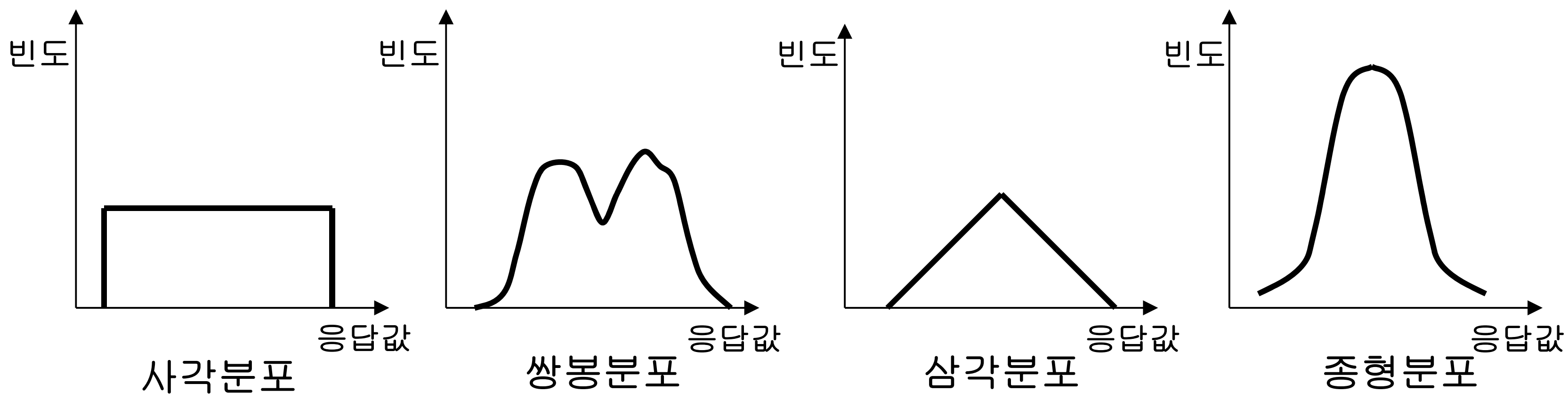
- ♦ 먼저 분포는 몇 가지 기준에 따라 아래와 같이 다양한 분포유형으로 구분할 수 있습니다.

(1) 대칭분포(symmetric distribution)

- ▫ 중간에 해당하는 응답값을 중심으로 중간 위나 중간값의 아래 응답값의 빈도가 동일한 분포

- ♦ 종형분포(bell-shaped distribution)의 예
  - ▫ 정규분포(normal distribution), *t* 분포(*t*-distribution) 등

[그림 9-2] 대칭분포의 예



(2) 비대칭분포(asymmetric distribution)

- 중간에 해당하는 응답값을 중심으로 그 위 혹은 아래에 해당하는 응답값의 빈도가 동일하지 않은 분포

  예) 편중/꼬리분포(skewed distribution)

♦ 우편중분포(<u>positive</u> skewness: skewed distribution to the right)

- 오른쪽 꼬리가 길게 나타나는 형태의 편중분포

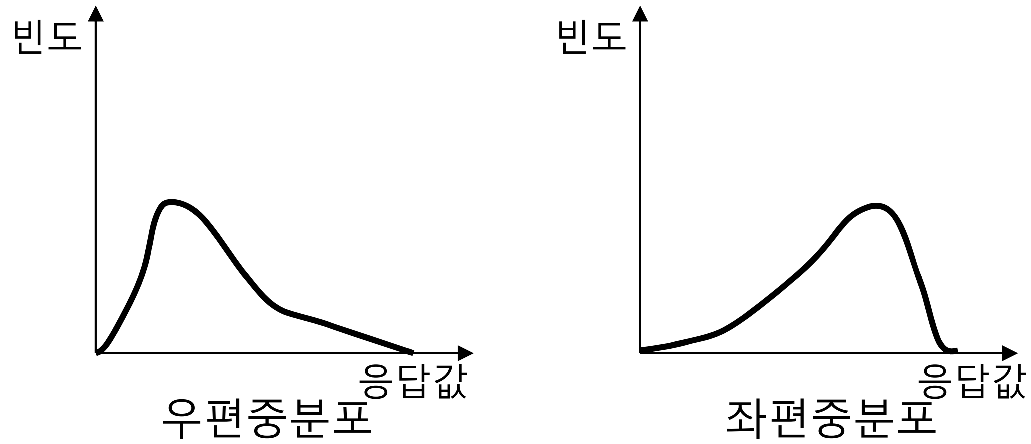
  예: 분포 (chi-square distribution)

♦ 좌편중분포(<u>negative</u> skewness: skewed distribution to the left)

- 왼쪽 꼬리가 길게 나타나는 형태의 편중분포

– 정규분포의 경우 왜도값과 첨도값이 0임

[그림 9-3] 비대칭분포의 예



2) 정규분포

(1) 정규분포

- 대칭분포 중 중간(즉, 평균)에 해당하는 응답값의 빈도가 제일 많고 평균에서 위 혹은 아래로 멀어질수록 응답값의 빈도가 적어지는 형태의 종모양의 분포(bell-shaped distribution) 중, 특정한 응답값(즉, 관찰값)이 나타날 확률이 아래의 <식 9-3>을 따르는 분포

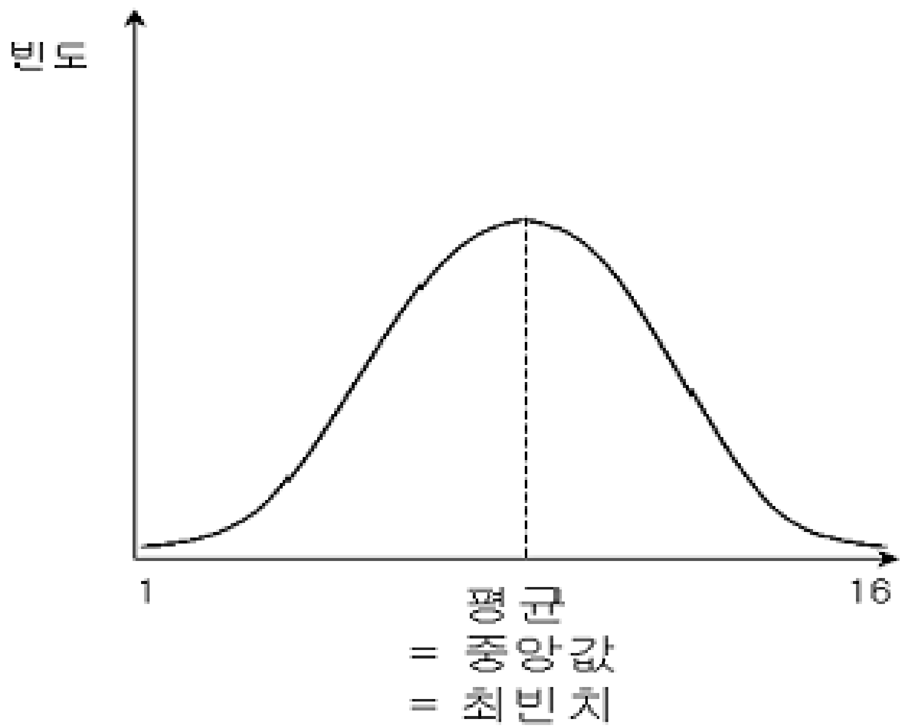
$$Y = \frac{N}{\sigma\sqrt{2\pi}} e^{\frac{-(X_i-\mu)^2}{2\sigma^2}} \qquad (식\ 9\text{-}3)$$

단, Y = 관찰값 X_i에 대응하는 분포곡선의 높이(=빈도)
N = 전체 사례수(=표본의 크기)
μ = $\overline{X}$(=관찰값 X_i 의 평균)
σ = 표준편차
π = 원주율(=3.1416....)
e = 자연 대수(natural logarithm)의 기초(=2.7183....)

- ♦ 프랑스의 De Moivre(1733년): 정규분포곡선의 공식을 수학적으로 유도
- ♦ 벨기에의 Quetelet(9세기 중엽 경): 정규분포를 여러 사회현상의 설명에 선구적으로 적용

▫ 정규분포 곡선의 가로축은 응답값의 범위를 보여주고 세로축은 각 응답값이 발생한 빈도(f) 혹은 비율(f/N) 을 보여주고 있음

[그림 9-4] 정규분포의 예



* 정규분포가 중요한 이유

- 연속된(continuous) 수로 측정 가능한 많은 사회현상은 측정값의 빈도가 정규분포를 따르는 것으로 알려져 있음

- 많은 연속적이지 않은 분포(discrete probability distribution)의 근사값을 구하는 데도 사용될 수 있음

♦ 정규분포의 중요성

▫ 정규분포가 고전적인 통계적 추론(classical statistical inference)의 기초를 제공

☞ 정규분포란 이상적인 분포(ideal distribution)

☞ 현실에서의 정규분포란 완벽한 정규분포에 끝없이 접근하는 '유사'정규분포

☞ 표본의 평균과 표본의 분산(표준편차)에 따라 정규분포의 구체적인 모습은 각각 다를 수 있음-> 정규분포를 이용한 가설검정 때마다 별도의 판단기준(임계치) 계산 불편

☞ 통계학자들: 원 응답자료(raw data or raw score)를 변환 표준화된 자료(표준점수=Z score)로 만들고 표준화된 자료의 정규분포곡선을 고안 (평균이 0, 표준편차가 1이 됨)

☞ 서로 다른 두 종류의 자료(예: kg으로 측정된 3학년 1반의 몸무게와 lb로 측정된 3학년 2반의 몸무게)를 직접 비교 가능

♦ 표준정규분포(unit_normal distribution)

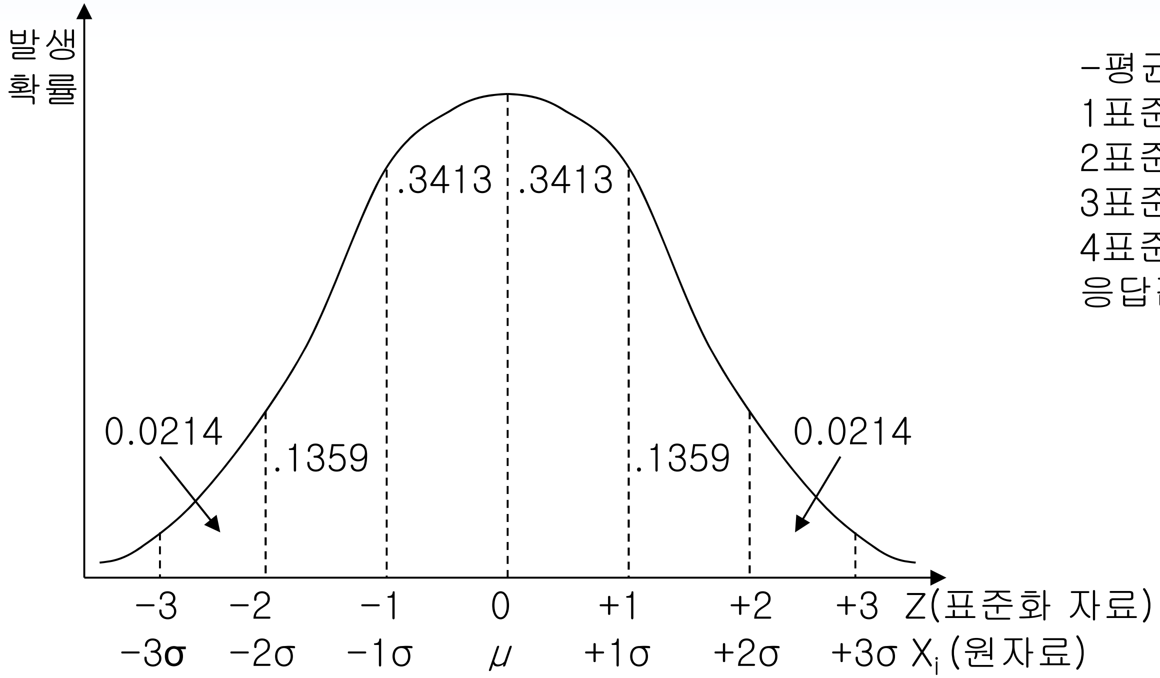
$$Z = (X_i - \bar{X})/S \qquad \text{(식 9-4)}$$

단, X_i = 표본에서의 각 응답값

$\bar{X}$ = 표본평균

S = 표본표준편차

[그림 9-5] 표준정규분포



-평균을 중심으로
1표준편차 내에 68.26%,
2표준편차 내에 95.44%,
3표준편차 내에 99.74%,
4표준편차 내에 거의 100%의 응답값이 위치함

(2) 표본분포

♦ 빈도분포(frequency distribution)
- ▫ 조사대상 변수가 가지는 응답범주(response category)별 응답자수의 분포
- ▫ 조사대상 현상을 이해하는 기초자료
- ▫ 하나의 표본(one sample)을 대상으로 해서 도출된 결과

♦ 하나의 모집단에서 표본을 추출할 수 있는 경우의 수

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

$_{25}C_5$개= 53,130가지의 표본을 구성

☞ 충분히 큰 모집단(예: 우리나라 휴대전화 가입자 3,000만명)에서 적절한 크기의 표본(예: 1,000명)을 추출할 때에는, 실제로 표본을 추출할 수 있는 경우의 수 (=$_{3,000만}C_{1,000}$)는 거의 무한(infinite)함

☞ <u>어떠한 표본을 추출하느냐에 따라서 각기 다른 표본평균과 표본분산이 도출</u>

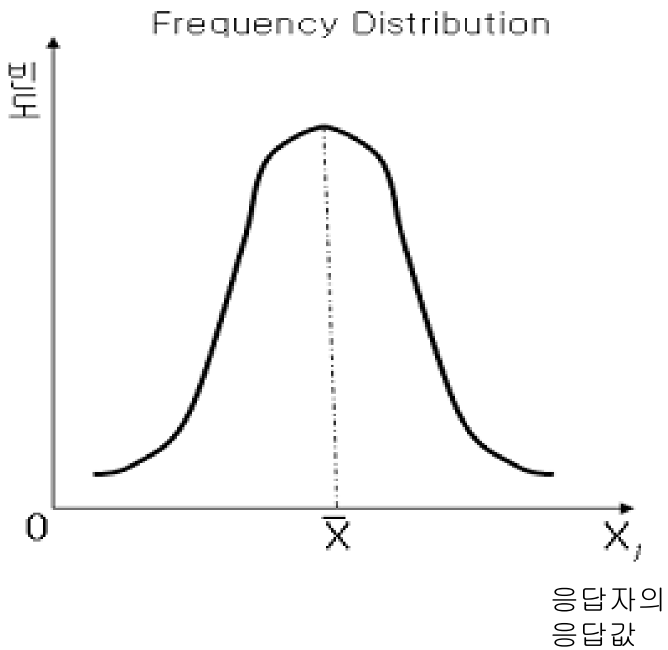
☞ 여러 번 표본을 추출하여 표본조사를 하는 경우에는 추출한 표본의 수만큼 표본평균 자료가 계산되고, <u>표본평균들의 분포</u>(sampling distribution of sample means), 즉 <u>표본분포</u>를 구할 수 있게 됨

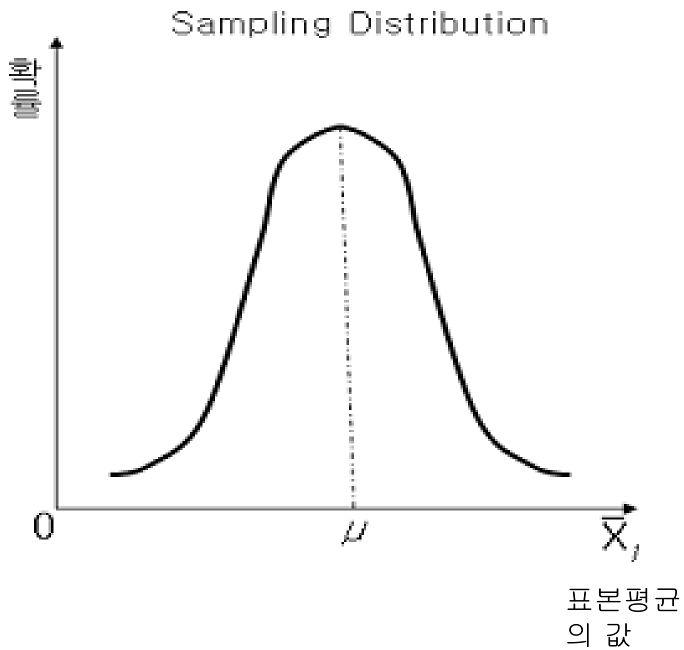
☞ 모집단으로부터 30회 이상의 표본을 추출하는 경우에, 30개 이상의 표본평균들(sample means)로 이루어진 표본분포는 정규분포(normal distribution)의 특성을 따른다고 하는 점이 밝혀짐

♦ 빈도분포와 표본평균의 표본분포

☞ 두 분포가 모두 정규분포이나 가로축과 세로축의 내용은 다름

[그림 9-6] 빈도분포와 표본분포





- 표본분포(sampling distribution)
  - 표본통계량(sample statistic) 값들의 분포
    - ☞ 표본평균(sample means)의 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 (표본의 크기가 큰(N>30) 경우 이 표본이 추출된 모집단의 분포의 유형과는 무관하게 이러한 표본의 반복추출로 구성되는 표본평균의 표본분포는 정상분포를 따름)
    - ☞ 모집단의 분포유형과는 무관하게 표본평균의 분포의 특징을 알게 된다면, 표본오차(sampling error)를 알 수 있게 되는 것

제1절 추리통계분석이란 무엇인가?

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

- **제1정리** : 모집단의 분포가 *정규분포이면* 표본평균(=$\overline{X}$)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

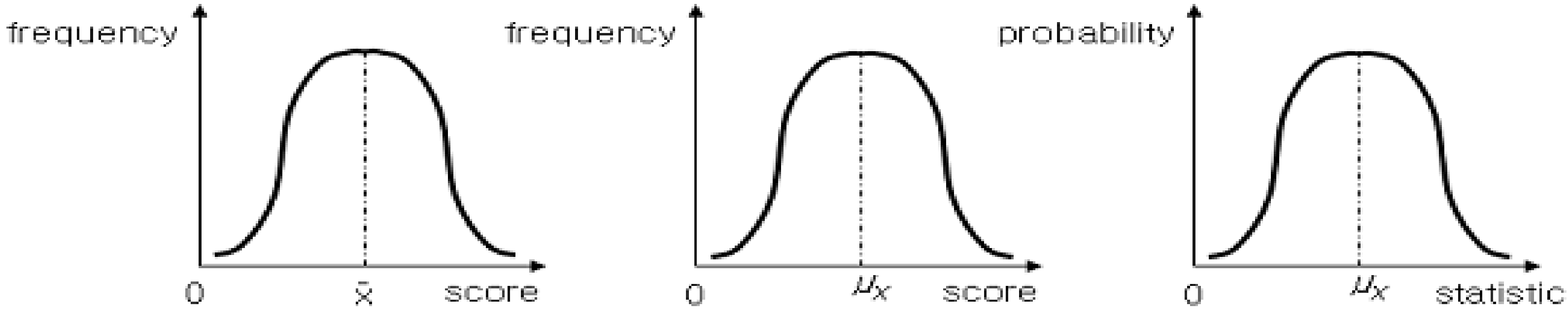
➪ 유용성

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 9-7] 표본의 빈도분포, 모집단의 빈도분포 및 표본분포

<표본의 빈도분포>	<모집단의 빈도분포>	<표본분포>
(예: 조사론수강생 50명)	(예: H대학생 6,000명)	(예: ${}_{6{,}000}C_{50}$)



① 평균 :	$\bar{X} = \dfrac{\sum_{i=1}^{n} X_i}{N}$	μ_X	μ_X
② 분산 :	$S^2 = \dfrac{\sum (X_i - \bar{X})^2}{N-1}$	σ^2	
③ 표준편차 :	$S = \sqrt{S^2}$	σ	
④ 표본크기 :	$N = 50$	$N = 6{,}000$	
⑤ 표본오차 :			$\dfrac{\sigma}{\sqrt{N}}$
⑥ Standardized variate :	$Z_i = \dfrac{(X_i - \bar{X})}{S_X}$	$Z_i = \dfrac{X_i - \mu_X}{\sigma_X}$	$Z_i = \dfrac{(\bar{X}_i - \mu)}{\sigma_{\bar{X}}}$

(3) 표본오차

- 표본평균 표본분포의 평균
  - 모집단의 평균과 동일<(식 9-5) 참조>

- 표본평균 표본분포의 표준편차(=표준오차=오차한계)
  - 모집단의 표준편차를 표본의 크기(사례수)로 나눈 값<(식 9-6) 참조>
  - 표본추출이 완벽하지 못하기 때문에 특정 표본평균이 모평균으로부터 떨어진 정도

표본평균 표본분포의 평균$(=[\bar{X}_1 + \bar{X}_2 + \cdots + \bar{X}_n]/N) = \mu$ (식 9-5)

표본평균 표본본포의 표준편차 (=표준오차)$= \sigma / \sqrt{N}\ (= s/\sqrt{N}\ )$ (식 9-6)

- 모집단의 표준편차를 알아야 표준오차, 즉 표본오차가 계산됨
  - ☞ 통계학자들은, <u>하나의 표본에서 계산된 표본표준편차((식 9-6)의 괄호 안의 식))가 표본오차의 근사치라고 함을 밝혀냄</u>
  - ☞ 결국, 표본평균, 표본표준편차 및 정규분포를 따르는 표본분포의 특징을 알게 되면 표본오차(=표준오차)를 알게 됨
  - ☞ 표본평균도 알고, 표본오차도 알게 되므로, 표본평균이 모수와 얼마나 유사한지도 알게 됨

3) *t* 분포

♦ 표본분포가 정규분포를 따르지 않을 때에도 추리통계를 수행 가능

예) 표본의 사례수(즉, 표본의 크기)가 매우 작은 경우

♦ *t* 분포(*t*-distribution)

▫ 표본의 크기가 30 이하인 경우에 사용 가능

▫ 영국의 Gosset이 1908년에 학생(student)이라는 익명으로 소표본(N < 30)에서의 표본분포는 정규분포를 따르지 않는다는 것을 발견하여 알려지게 됨

☞ *t* 분포는 '학생의 *t* 분포(student's *t*-distribution)'라고 불리기도 함

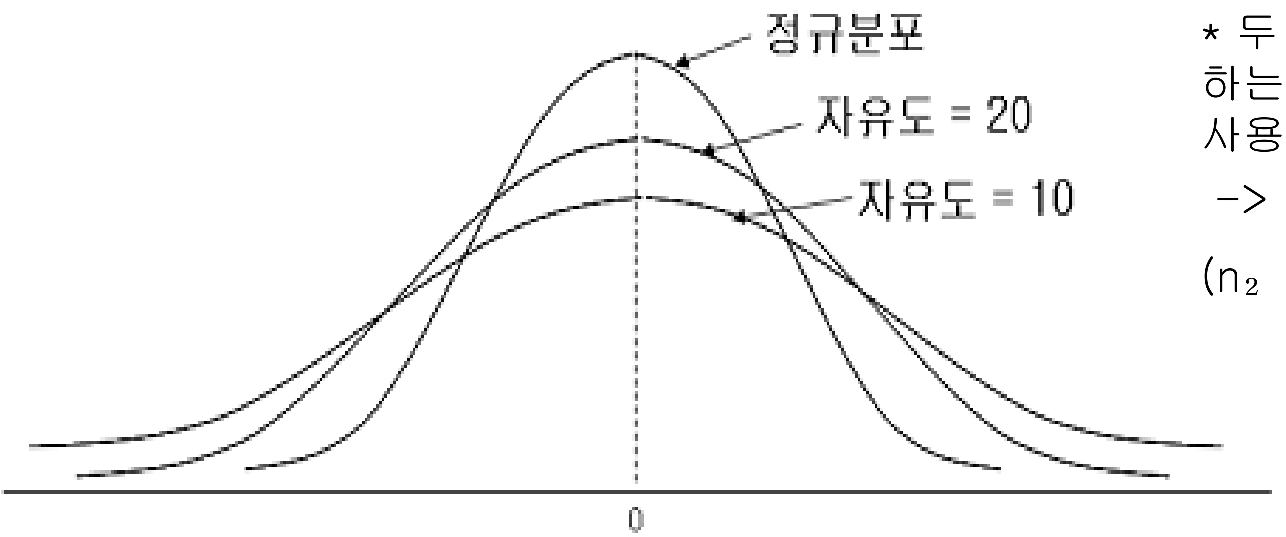
♦ *t* 분포란 표준정규분포처럼 단일분포를 보이는 것이 아니고 표본의 크기에 따라 표본분포(sampling distribution)가 변하는 특징을 보유

☞ 즉, 자유도(degrees of freedom: df)에 따라 표본분포곡선의 모습이 변화([그림 9-8] 참조).

- 두 집단간 평균 차이에 대한 가설검증의 경우에 주로 사용

- 자유도 (degrees of freedom: df)
  - *t* 분포뿐만 아니라 다른 유형의 통계검정에도 자주 사용되는 중요한 통계개념
  - 개념적으로 자유도란 표본분포(sampling distribution)를 구성하기 위해 자유롭게 반복해서 추출할 수 있는 표본(repeated random sample)의 수
  - 구체적으로 자유도는 표본크기에서 표본에 부여되는 제약조건의 수를 차감해서 계산(pp.279-280)

[그림 9−8] 표본의 크기와 *t* 분포



* 두 개의 표본간에 존재하는 평균의 차이검정에 사용

−> 자유도 : $(n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2$

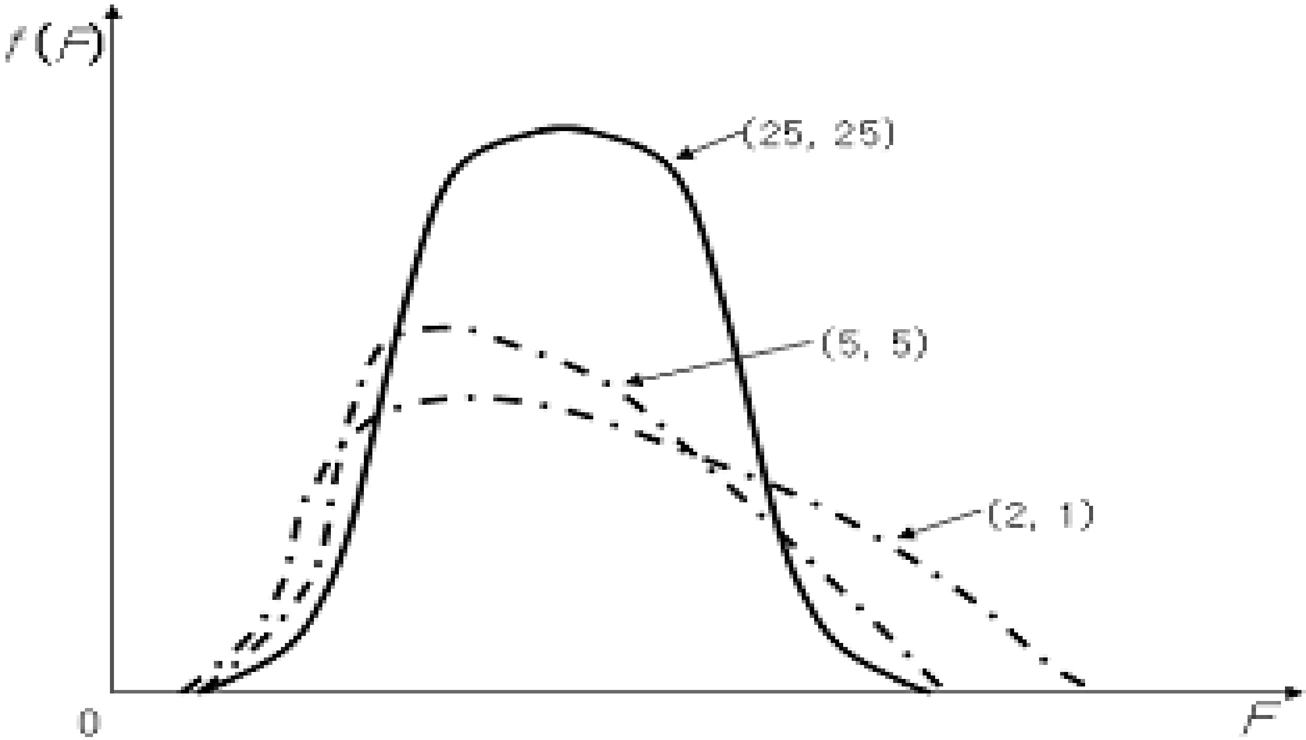
4) *F* 분포

- *F* 분포(*F* -distribution)
  - ▫ *F* 값을 나타내는 분자와 분모 각각의 자유도에 의하여 규정되는 분포
  - ▫ *F* 분포의 표본분포(sampling distribution)는 자유도의 값에 따라 다양한 수의 분포 가능
  - ▫ <u>비교집단이 2개보다 큰 경우</u>(즉, 3개 이상 집단간의 비교)에도 집단간의 차이를 설명할 수 있는 표본분포
  - ▫ 분산분석(analysis of variance)에서 대표적으로 사용되고 있는 분포

    (종종 분산분석을 **F검정**이라고 부르기도 함)

- *F* 값(*F* ratio)
  - ▫ 집단간 분산의 추정값을 집단내 분산의 추정값으로 나눈 것

[그림 9-9] 자유도에 따른 *F* 분포곡선



5) X^2 분포

♦ X^2 분포

– 1900년에 Pearson에 의해 개발

▫ X^2 값이 따르는 표본분포(sampling distribution)

▫ t 분포와 마찬가지로 자유도에 의하여 구체적 분포가 결정됨

▫ 일반적으로 크기 N 인 하나의 표본에서 X^2 자유도는 $N-1$로 결정됨

▫ X^2의 표본분포(sampling distribution)는 자유도의 값에 따라 다양한 수의 분포가 가능

♦ X^2 값

▫ 실제로 관찰된 빈도가 기대한 빈도와 얼마나 가까운가를 검정하는 도구

➪ 명목척도로 측정된 두 변수간의 상관관계 검정 시 χ^2 검정 이용

➪ χ^2 검정은 χ^2 분포를 사용하여 가설의 진위를 판단

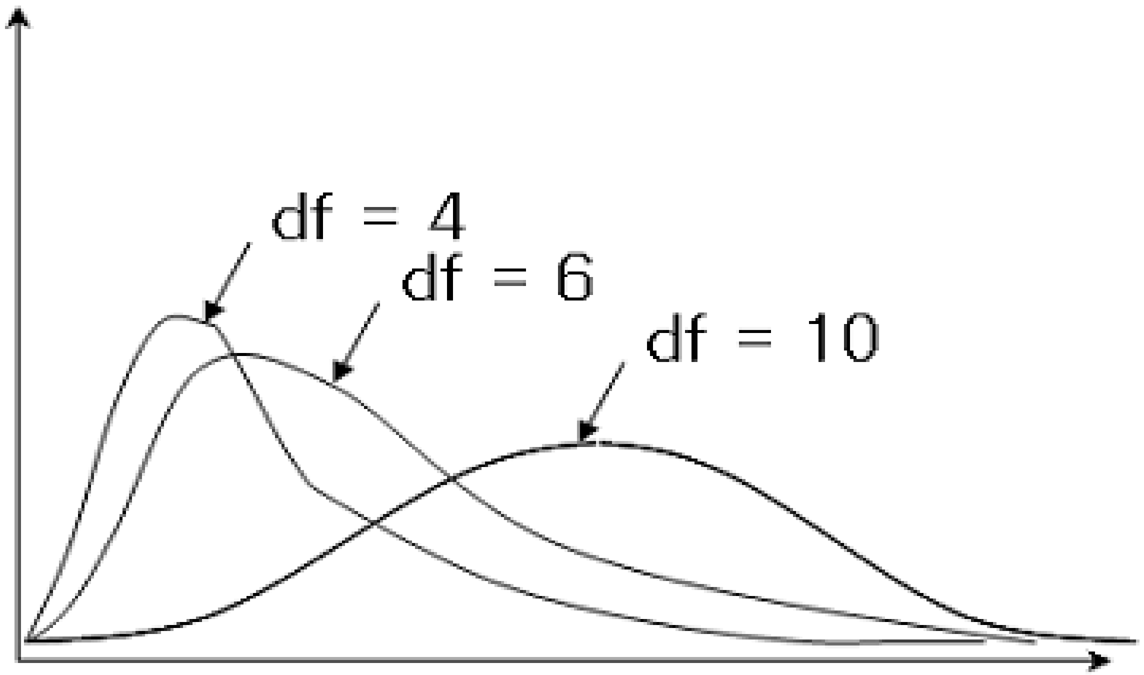
X^2 값 계산 (식 9-7)

$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

단, $E_{ij} = ij$ 첫번째 칸(cell)의 기대빈도

$O_{ij} = ij$ 번째 칸의 실제 빈도

[그림 9-10: X^2 분포 곡선]



♦ 추리통계분석의 목적

▫ 표본통계량(sample statistic)으로 모수(population parameter)를 추정하는 것

♦ 통계적 추론의 구체적 예는 매우 다양

☞ 표본통계량의 유형에 따라 구체적인 표본분포는 상이

* 추리통계분석의 구분 : 추정(parameter estimation)과 가설검정(hypothesis testing)

☞ 표본분포와 무관하게 통계적 추론의 정확성을 판단해 줄 공통의 도구가 필요

☞ 1종 오류(Type I error)와 2종 오류(Type II error)

<표 9-1> 1종 오류와 2종 오류

		전수조사시 알 수 있는 정보	
		H_0 참	H_0 거짓
표본조사시 연구자의 결론	H_0 수용	$1-\alpha$ (정확한 결론)	β (2종 오류)
	H_0 기각	α (1종 오류)	$1-\beta$ (power: 정확한 결론)

♦ 1종 오류와 2종 오류

▫ 표본통계량으로 연구자가 모집단에 대한 추론을 할 때 잘못된 결론이 야기할 수 있는 두 가지의 위험(risk)

♦ 1종 오류(Type I error 혹은 alpha error)

▫ H_0를 잘못 기각할 경우에 발생하는 오류

☞ 1종 오류가 발생할 확률(probability of type I error)을 통계적 유의도 혹은 유의수준(level of significance=100% – 신뢰수준[confidence level])이라고 하고 α 로 표시함

☞ α 란 표본을 기초로 한 통계적 추론의 결과가 우연에 의한 확률일 가능성을 지칭

♦ 한편 표본에서 실제로 관찰된 유의수준(observed significant level)은 **p값(p-value)**이라고 지칭

☞ α 란 연구자가 사전적으로 허용하는 유의수준

☞ **p 값**이란 표본에서 사후적으로 관찰된 유의수준

- 2종 오류 (Type II error 혹은 beta error)
  - H_0를 잘못 수용하는 경우에 발생하는 오류
  - 이러한 오류의 가능성을 **β** 라고 함
  - ☞ 정확한 결론(=1–β)=통계적 power(power of a statistical test)

- 표본조사는 현실적으로 피할 수 없는 선택이고, 표본조사시 표본의 크기를 무한정 늘릴 수도 없을 것임
  - ☞ 연구자가 사전에 받아들일 수 있는 1종 오류의 수준(예: α =.01 혹은 .05)을 정하고, 표본조사에 의거한 통계적 추론의 결과(p값)가 미리 정한 1종 오류의 수준을 충족하는지를 비교해서 통계적 추론의 상대적 정확성의 정도를 판단하게 됨
    - 관례적으로 1종 오류(=α)를 0.05 혹은 0.01 수준으로 설정하는데, 관찰된 1종 오류(=***p***)가 0.01 이하 수준인 통계적 추정(가설검정) 결과는 상대적으로 정확한 결론으로 받아들이고 있음

* 일정 수준의 α 에서 표본의 크기를 키워서 β를 줄이려는 것(즉, 일정한 수준의 1- α 에서 1- β를 크게 하는 것)이 연구자들이 가설검정시 일반적으로 선호하는 접근방법임

1. 추리통계와 가설검정

1) 표본분포

♦ 표본분포

▫ 표본통계량의 분포(distribution of sample statistics)

☞ 표본평균(sample means: $\overline{X}_1$, $\overline{X}_2$ 등)의 표본분포, 표본분산(혹은 표본표준편차)의 표본분포, 표본비율(proportions)의 표본분포, 혹은 표본평균 차이(예: $\overline{X}_1 - \overline{X}_2$)의 표본분포 모두 표본분포의 유형

☞ 모수추정(parameter estimation)과 가설검정(hypothesis testing)이 근본적으로 동일

2) 가설검정 방법

♦ 가설

▫ 두 개 이상의 현상간의 관계에(대한) 논리적인 예측

☞ 구체적인 가설의 내용은 매우 다양

☞ 가설검정의 구체적인 유형도 매우 다양

☞ 변수(현상)들 간의 관계 + 하나의 현상에 대한 가설도 검정할 필요 발생 가능

-> 복합가설(부등호로 표현되는 복합적인 범위에 대한 검정을 수행)과 단순가설(특정한 값을 검정) 모두 가설검정의 대상이 됨

제3절 추리통계분석은 어떻게 실행하는가?

♦ **모수검정**(parametric test)

▫ 연구대상 현상을 등간척도 혹은 비율척도의 수준으로 측정한 경우와, 표본이 정규분포를 따르는 모집단에서 추출된 경우에 적용하는 통계적 검정방법

* 정규분포나 t분포 등을 이용

♦ **비모수검정**(non-parametric test)

▫ 등간척도 이상의 척도에 대한 가정이나 혹은 정규분포에 대한 가정을 필요로 하지 않는, 따라서 분포를 확정하기 힘든(distribution-free) 경우에 사용하는 통계적 검정방법

☞ 엄밀하게는 <u>모수통계적 접근을 이용하는 비모수검정방법</u>이라고 할 수 있음

* χ^2 분포 등을 이용

* 추리통계분석은 가설검정과 실질적으로는 동일하다고 이해 가능

2. 가설검정의 단계

☞ 가설검정의 단계에는 연역적 추론과 귀납적 추론이 상호보완적으로 사용됨

<그림 9-11> 가설검정의 절차

(통계)가설 H_1과 H_0의 설정

↓

통계분석방법 및 검정통계량 선택

↓

통계적 유의 수준(a) 결정

↓

통계분석(검정통계량 계산)

↓

계산된 검정통계량의 p값을 유의 수준과 비교

↓

H_0의 기각 혹은 수용 결정

1) 가설설정

▫ 첫 단계는 연구문제의 정의에 따른 가설의 설정(formulation of hypothesis)

(1) 귀무가설(H_0) 형태의 통계가설을 설정

♦ 실질가설 vs. 통계가설

♦ 연구가설 vs. 귀무가설

* 연구가설(H1)이 참임을 직접 검정하기보다는 귀무가설(Ho)이 거짓임을 검정하는 것이 논리적으로 더 타당함
* 가설의 통계적 검정은 표본통계량에 기초한 모수에 대한 논리적 예측을 다룸

(2) 단측검정/양측검정 결정

♦ 단측/일방향 검정(one-tailed test)

▫ 연구가설(따라서 귀무가설)의 내용이 현상들간의 관계의 방향(+ 방향인지 혹은 – 방향인지)까지 포함하는 경우 수행하는 검정방법

♦ 양측/양방향 검정(two-tailed test)

▫ 연구가설(따라서 귀무가설)의 내용이 예측방향을 포함하지 않는 경우 수행하는 검정방법

☞ 일반적으로 단측검정이 양측검정보다 더 강력한(엄격한) 결론을 보여 줌

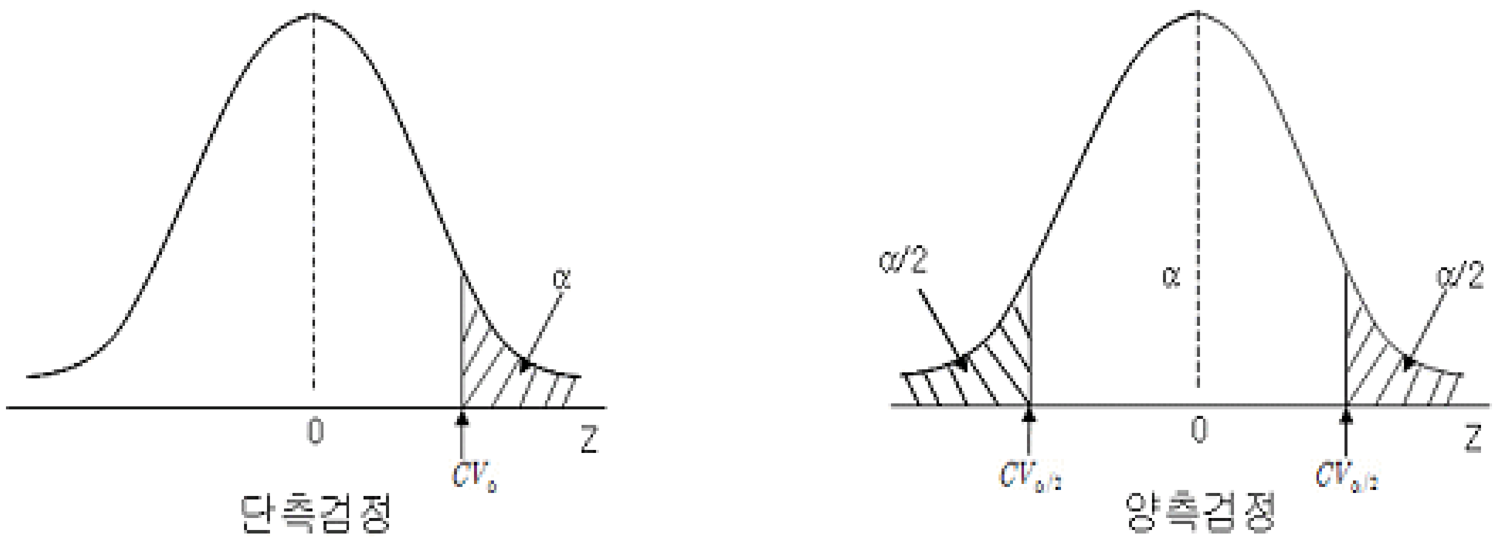
예) 한편 [그림 9-12]는 표본분포가 정규분포를 따를 때, 단측검정과 양측검정의 차이점을 정규분포곡선을 이용하여 설명

- ☞ [그림 9-12]의 단측검정의 가상의 예에서는 <표 9-2>의 (a)의 경우를 도시
- ☞ 영가설이 기각되려면, 표본에서의 '구독자소득평균($\bar{X}_1$) – 非구독자소득평균($\bar{X}_2$)'의 값이 $\mu_1 - \mu_2$ =0(모집단에서의 소득평균의 차이)에서 최소 +1.65표준오차(p =.05수준에 부합하는 표준오차의 크기)만큼 떨어져 있어야 함을 보여주고 있음
- ☞ 양측검정의 경우(<표 9-2>의 (c))에는 +1.96표준오차보다(p =.025수준에 부합하는 표준오차의 크기, 즉 임계치의 크기) 크거나 혹은 -1.96표준오차보다 작아야 함을 보여주고 있음

<표 9-2> 단측검정과 양측검정

	연구가설(H_1)	**귀무가설(H_0)**
단측검정 (a)	구독자소득 > 非구독자소득	구독자소득 ≤ 非구독자소득
단측검정 (b)	구독자소득 ≤ 非구독자소득	구독자소득 > 非구독자소득
양측검정 (c)	구독자소득 ≠ 非구독자소득	구독자소득 = 非구독자소득

<그림 9-12> p=.05 수준에서의 단측검정과 양측검정



2) 통계분석방법 및 검정통계량 선택

- 둘째 단계는 이 가설을 검정하기 위해 적절한 통계분석방법을 선택
  - ☞ 특정 가설은 다양한 통계방법으로 검정할 수 있는 것이 일반적
  - ☞ 연구자는, 근원척도의 유형, 표본통계량(sample statistic)의 표본분포(sampling distribution), 검정통계량(test statistic) 등을 고려하여 가설검정에 적절한 통계분석기법을 선택

♦ 검정통계량(test statistic)

- 표본통계량(sample statistic)이 H_0에서 모수(parameter)에 대해 예측하는 수준에 얼마나 근접했는지를 판단하게 해주는 도구
- 추리통계분석기법에 따라 다양한 유형이 개발되어 있음
  - ☞ SPSS Statistics등을 이용하여 통계분석기법을 실행하는 경우에는 연구자가 선택한 통계기법에 맞는 검정통계량을 SPSS에서 계산해 주기 때문에 연구자는 검정결과를 적절히 해석하기만 하면 됨

가설검정의 목적은 귀무가설의 타당성 여부 판단에 있는데, 이 결정의 기준이 되는 표본통계량-> **검정통계량**(test statistic)
예) 모평균에 대한 가설검정-> 검정통계량은 모평균에 대한 통계량인 표본평균으로부터 구할 수 있음

3) 통계적 유의수준 결정

▫ 가설검정의 세 번째 단계에서는 통계적 유의수준(level of significance: α)을 결정

☞ 가설검정의 마지막 단계에서 H_0을 기각하거나 혹은 수용하는 결정을 내리게 되는데, 이러한 결정이 잘못 내려질 가능성이 상존합니다. 이때 결론을 잘못 내릴 가능성을 어느 정도 수준까지 허용할 것인가를 연구자가 주관적으로 결정을 해야 함

☞ 관례적으로 1종 오류가 0.01 이하 수준인 통계적 추정결과는 상대적으로 정확한 결론 ; 통계적 유의도(α)를 0.01 혹은 0.05로 정하는 것이 관행

☞ 한편, SPSS Statistics를 사용하는 가설검정에 있어서는, "영가설(H_0)이 참이라는 전제하에 표본에서 계산된 검정통계량값이 관찰될(표본분포에서 발생할) 확률" 이라고 할 수 있는 유의수준(p값)이 바로 제시됨

☞ 연구자는 이 수준이 자기가 미리 설정한 유의도(α)보다 작으면 H_0를 기각하고 H_1을 지지하는 결론을 내리게 됨

4) 통계분석(검정통계량 계산)

▫ 네 번째 단계에서는 둘째 단계에서 정한 통계분석기법을 실행하고 가설검정에 필요한 검정통계량(test statistic)을 계산

☞ 특정 통계기법의 분석명령을 SPSS가 실행하게 되면, 관련된 검정통계량이 동시에 계산되고 특정한 검정통계량에 상응하는 **관찰된 유의수준**(단측/양측 검정), 즉 ***p*값**도 동시에 제시됨

$$* \text{ 검정통계량} = \frac{\text{표본통계량} - \text{귀무가설에서 설정된 모수값}}{\text{표본통계량의 표준오차}}$$

5) 계산된 검정통계량과 임계치(검정통계량의 p값과 유의수준) 비교

- 다섯 번째 단계에서는, 산출된 검정통계량에 적합한 표본분포표(정규분포표, t 분포표, 카이제곱 분포표 등)를 참조하여 이미 계산된 검정통계량값을 임계치(critical value)와 비교

♦ **임계치**

- 가설기각과 비기각(수용)을 구분하는 검정통계량의 값
  - ☞ 연구자가 결정하는 유의수준(α =.01 혹은 α =0.05)에 의해 결정
  - ☞ SPSS를 실행하면 검정통계량값, 임계치, 검정통계량값에 부응하는 p값이 바로 제시됨
  - ☞ 계산된 검정통계량의 값이 그 검정통계량의 임계치보다 크면 H_0를 기각
  - ☞ **즉, 검정통계량의 수준에 부응하는 p값이 연구자가 미리 생각해둔 유의수준(예: α =.05)보다 더 작으면(예: p =0.01) 영가설(H_0)을 기각**

6) 가설검정 결론

- 다섯 번째 단계에서도 설명한 바와 같이 가설검정의 마지막 단계에서는 연구자가 미리 결정한 통계적 유의수준(예: α =.05 혹은 α < .05) 혹은 이 수준에 부응하는 임계치를 기준으로 하고, 통계분석을 통해 계산된 검정통계량이나 혹은 검정통계량의 *p*값(예: p =.015)을 확인하여 H_0을 기각("검정통계량의 *p*값 ≤ 유의수준" 혹은 "검정통계량 값 ≥ 임계치")하거나 혹은 수용("검정통계량의 *p*값>유의수준" 혹은 "검정통계량 값 < 임계치")결정을 하게 됨

☞ 통계가설을 기각/수용하게 된 후에는, 실질가설의 입장(즉, 언어적인 표현)으로도 가설검정결과를 설명하는 것이 바람직함

3. 추리통계분석(가설검정)기법의 유형

(1) 상호관계분석

- 현상(변수)들간의 관계를 검정

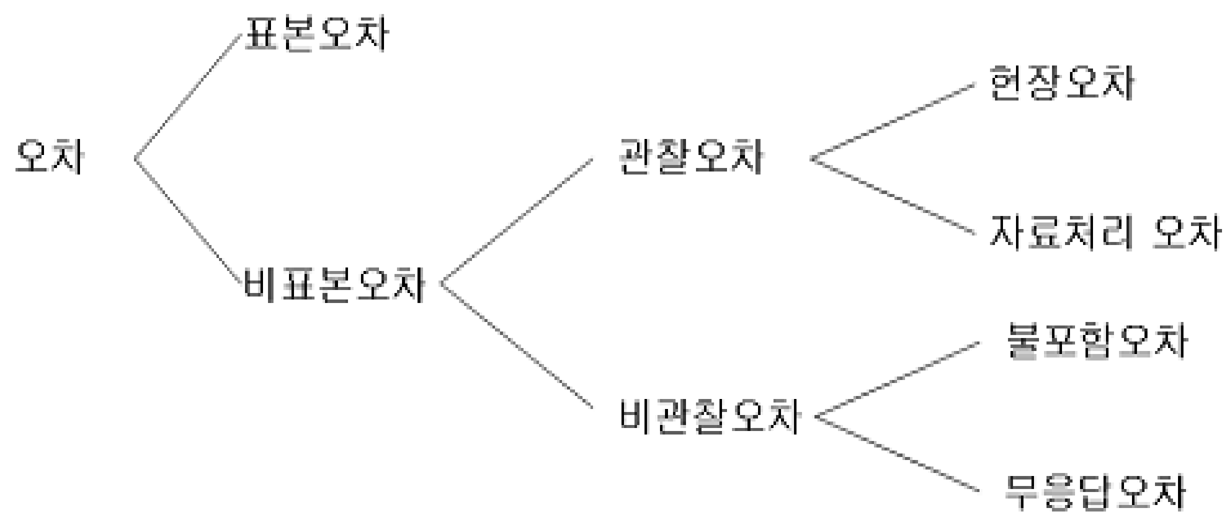
(2) 인과관계분석

- 현상들간의 원인-결과 관계를 검정
  - ☞ 통계적 증거, 이론적 지지(theoretical support), 적절한 연구설계(research design) 등 수반 필요
  - ☞ 본서에서 사용하는 인과관계분석이라는 용어의 의미는, "인과관계를 지지할 수 있는 통계적 증거를 제공하는 것"으로만 한정

1) 추리통계분석의 기본가정

- 모수추정(parameter estimation)의 기반이 되는 표본(sample)은 당연히 모집단으로부터 추출되어야 함
- 확률표본추출(probability sampling)에 따른 표본추출 필요
- 추리통계에서는 표본오차 혹은 표본오류(sampling error)만을 다룸
  - ☞ 표본오류의 가능성이 상대적으로 아주 작을 경우에는 오히려 비표본오차가 상대적으로 더 연구의 타당성을 낮출 가능성
- 100%의 응답률

[그림 9-13] 표본조사의 오차



2) 비표본오차

(1) 관찰오차

♦ 현장오차

▫ 사회현상의 관찰과정에서의 면접원과 응답자의 상호작용에서 발생하는 오차

예) 응답자가 자신이 느끼는 대로가 아니고 면접원이 원한다고 생각하는 대로 응답을 하게 되는 경우에는 조사대상 현상을 제대로 반영하지 못하는 현장오차가 발생

♦ 자료처리오차

▫ 조사결과를 잘못 기록하거나 기록된 자료를 잘못 처리할 때 발생하는 오차

(2) 비관찰오차

♦ 불포함오차

▫ 표본프레임이 불완전해서 모집단을 제대로 반영하지 못하는 경우 발생

예) 전화번호부를 표본프레임으로 사용했을 경우 전화번호부에 누락된 응답자들은 표본에 추출될 수 없게 됨

♦ 무응답오차

▫ 응답자가 응답을 거부했을 때 발생하는 오차

[보충사항]

1. 추 정

✦ 50명 학생들의 통계학 평균점수가 100점 만점 기준 50점(= $\bar{X}$)일 때, 6,000명 모집단의 평균점수(= μ_X) 추정

⇨ 모집단 평균점수(= μ_X) 추측의 예

(a) 모수는 50점일 것이다. ➪ 이상적 추론?

(b) 모수는 25점에서 75점 사이일 것이다. ➪ 상대적으로 신뢰성 있는 추론?

(c) 모수는 0점에서 100점 사이일 것이다. ➪ 100% 확실한 추론?

⇨ 어떻게 추정의 정확성을 증가(or 오류가능성을 감소)시킬 것인가?

✦ **추정의 구분**

- **점추정**(point estimation)

  = 하나의 값으로 모수를 추측하는 것; 예) (a)

- **구간추정**(interval estimation)

  = 모수가 포함되는 구간을 추측하는 것; 예) (b), (c)

1) 점추정

✦ 추정량과 추정치

- 추정량(estimator)

  = 모수를 추정하기 위해 이용하는 <u>표본통계량</u>, 확률변수임

- 추정치/추정값(estimate)

  = 모수를 추정한 구체적인 값

✦ 바람직한 점추정량

◆ 개념적으로는 추정량 $\hat{\theta}$과 모수 θ의 차이, 즉 $(\hat{\theta}-\theta)$를 최소화하는 추정량

⇨ 구체적으로는 평균제곱오차(mean squared error: MSE)가 작은 추정량

$$MSE(\theta) = E[(\hat{\theta}-\theta)^2] = Var(\hat{\theta}) + [E(\hat{\theta})-\theta]^2$$

- **불편성**(unbiasedness)
  - $\hat{\theta}$의 평균[$=E(\hat{\theta})$]이 θ와 동일한 경우
    → 불편추정량(unbiased estimator)

    예) 표본평균 $\bar{X}$는 모집단 평균($=\mu$)의 불편추정량

- **효율성**(efficiency)
  - $\hat{\theta}$ 의 분산이 최소인 특성
    → 최소분산 불편추정량(minimum variance unbiased estimator)
    (가장 효율적인 불편추정량)

- **일치성**(consistency)
  - 표본의 크기($=n$)가 한없이 증가할 경우 추정량 $\hat{\theta}$이 모수 θ에 한 없이 가까워지는 특성

* 충분성 : 표본자료가 내포하고 있는 모수에 대한 정보와 지식을 포괄적으로 요약해주는 추정량, (표본평균 $\bar{X}$, *중앙값)*

- 점추정의 상대적 신뢰도 정보는 없음

⇨ $E(\overline{X}) = \mu$ 이지만 $\overline{X} = \mu$ 이 얼마나 신뢰할 수 있는지 판단할 수 없음

2) 구간추정

✦ 구간추정(interval estimation)

- 모수가 존재할 구간을 제시
- 이 구간이 얼마나 믿을만한지에 대한 정보도 제시

※ 점추정량

- '평균적'으로는 모수를 대변
- <u>점추정치가 상대적으로 얼마나 믿을 만한 지에 대한 정보가 없음</u>

- **유의수준**(significance level)
  - α
  - 모수가 특정 구간 내에 포함되지 않을 가능성

- **신뢰수준/신뢰도**(confidence level)
  - (1− α)
  - 모수가 특정 구간 내에 포함될 가능성

- 100(1− α)%의 <u>신뢰구간</u>(confidence interval)
  - 모수 θ가 포함될 가능성이 100(1− α)% 인 구간

    예) 관행적으로 95%(즉 α =0.05) 혹은 99%(즉 α =0.01)의 신뢰구간을 사용

- 표본분포를 포함한 모든 정규분포는 표준정규분포([그림 8-8] 참조)로 변환 가능

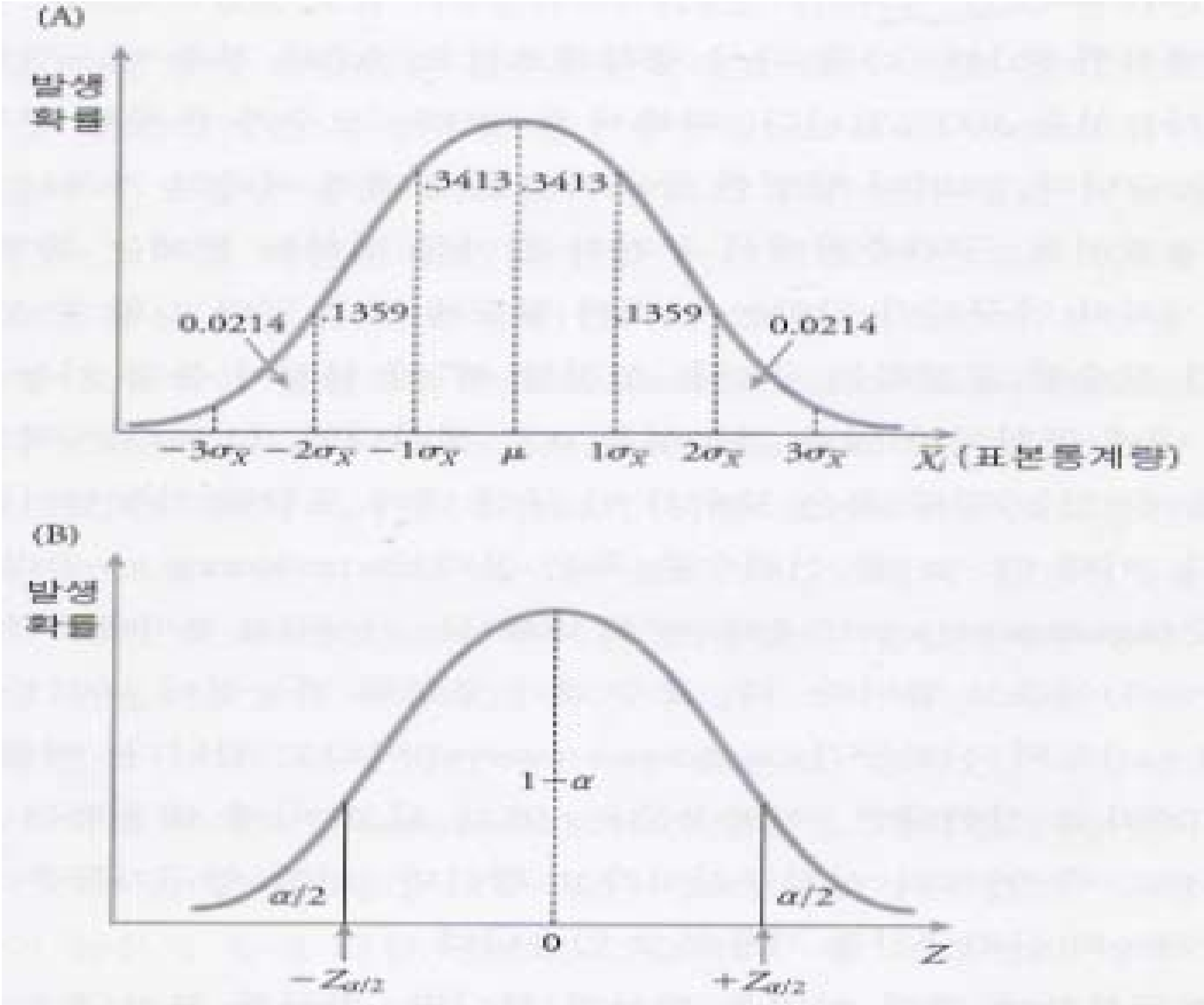
  예1) 모수 중심 좌우 1$\sigma_{\bar{X}}$ 내에는 68.26%,

  좌우 2$\sigma_{\bar{X}}$ 내에는 95.44%,

  좌우 3$\sigma_{\bar{X}}$ 내에는 99.74%,

  좌우 4$\sigma_{\bar{X}}$ 내에는 거의 100%의 표본통계량(즉, 표본평균) 값 위치

[그림] 표본분포와 표준정규분포



예2) 특정한 표본평균 $\bar{X}$가 $[\mu - 2\sigma_{\bar{X}}, \mu + 2\sigma_{\bar{X}}]$ 구간 내에 속할 확률

$= P(\mu - 2\sigma_{\bar{X}} \leq \bar{X} \leq \mu + 2\sigma_{\bar{X}}) = .9544$

➪ 이 확률계산식을 모수를 중심으로 변형하면 (식 8-13)이 됨

$$P(\bar{X} - 2\sigma_{\bar{X}} \leq \mu \leq \bar{X} + 2\sigma_{\bar{X}}) = .9544 \qquad \text{(식 8-13)}$$

→ 구간추정량 $[\bar{X} - 2\sigma_{\bar{X}}, \bar{X} + 2\sigma_{\bar{X}}]$이 모수 μ를 포함할 확률은 .9544

- 모수(=표본통계량의 평균)가 포함될 90%/95%/99% 신뢰구간 선정
  - 모수를 포함할 확률이 90%/95%/99%인 구간추정량의 크기를 정하는 것
  - 구간추정량이 모수를 포함할 확률이 90%/95%/99%에 대응하는 ***Z* 값**을 구하는 것

✦ 모평균 μ에 대한 100(1−α)% 신뢰구간 공식

모평균 μ 에 대한 100(1− α)% 신뢰구간 (식 8-14)

$$P(\overline{X} - Z_{\alpha/2}\, \sigma_{\overline{X}} \le \mu \le \overline{X} + Z_{\alpha/2}\, \sigma_{\overline{X}}) = 1 - \alpha$$

where σ가 알려진 경우와 σ가 미지이나 대표본($n \ge 30$)인 경우

예) 모평균 μ 에 대한 99% 신뢰구간

$P(\overline{X} - Z_{\alpha/2}\, \sigma_{\overline{X}} \le \mu \le \overline{X} + Z_{\alpha/2}\, \sigma_{\overline{X}}) = .99$를 만족하는 Z값을 특정하면 구해짐

➪ 표준정규분포곡선에서 $\pm Z_{\alpha/2}$ 로 둘러싸인 범위의 면적(=1−α)이 .99인 값?

➪ $[\overline{X} \pm 2.57\sigma_{\overline{X}}]$가 99% 신뢰구간에 해당하는 구간추정량

σ가 미지이고 소표본인 경우 μ 에 대한 100(1−α)% 신뢰구간 (식 8-15)

$$P(\overline{X} - t_{\alpha/2}\, S_{\overline{X}} \le \mu \le \overline{X} + t_{\alpha/2}\, S_{\overline{X}}) = 1 - \alpha$$