



이화여자대학교  
EWha WOMANS UNIVERSITY

# 확률 및 통계학

■ 모집단, 표본, 그리고 정규분포

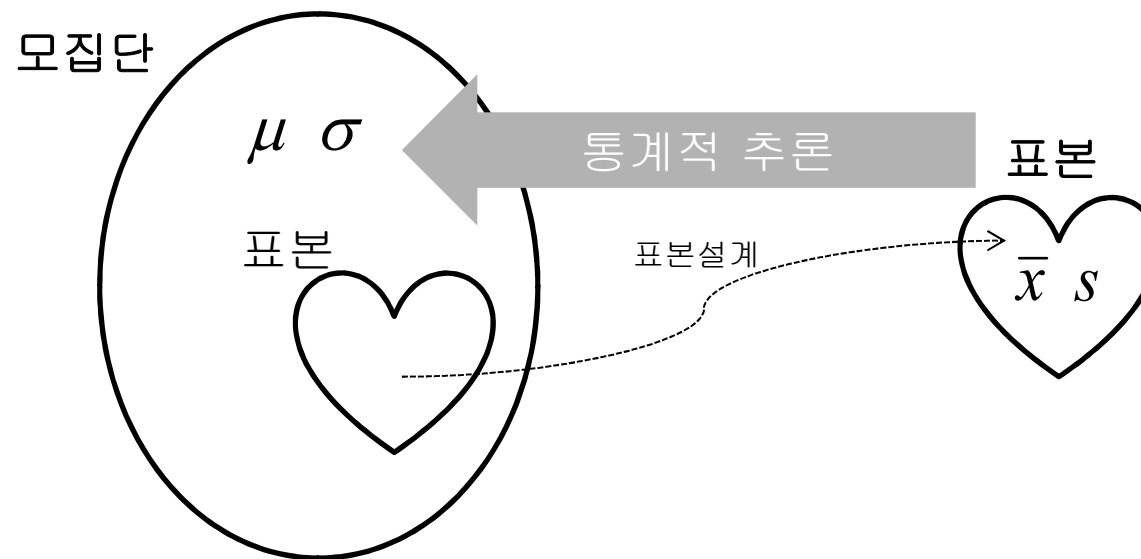
송수민

soominsong@ewha.ac.kr

# 모집단과 표본

- 표본을 사용하는 이유

1. 전수조사 불가능
2. 경제성
3. 조사의 파괴성
4. 시간적 제약



# 표본 추출법

---

- **단순 랜덤 추출법(simple random sampling)**
  - 확률추출법 중 가장 기본적인 방법으로 표본이 선택될 확률이 모두 동일하도록 추출하는 방법
- **층화추출법(stratified sampling)**
  - 모집단을 동질적인 층으로 나눈 다음 각 층에서 단순임의표본을 독립적으로 추출하는 방법
  - 층내부를 동질화, 층간 이질화해야 정확도가 높아짐
  - 예 ) 이화고등학교 학생들의 키 - 각 학년별로 표본추출
- **집락추출법 (cluster sampling)**
  - 모집단을 몇 개의 집락으로 나누어 집락을 표본추출하고 추출된 집락은 전수조사 하거나 다시 표본조사 함
  - 집락 내부는 이질화, 집락간에는 동질화하는 것이 효율적
  - 예) 고등학교 학생들의 월평균 사교육비 - 고등학교 명단에서 고등학교를 표본추출한 후 추출한 고등학교 학생들을 전부 또는 일부 조사
- **계통추출법 (systematic sampling)**
  - 표본을 일정한 간격으로 규칙적으로 추출하는 방법으로 표본추출이 쉬움
  - 주기성을 가진 데이터에 사용시 주의

# 확률분포

## ✓ 변수의 속성에 따른 분류

- 이산확률변수 : 정수, 셀 수 있는 값을 가지는 변수      예) 불량 제품의 수
- 연속확률변수 : 연속된 값을 가지는 변수              예) 매출, 나이

확률분포란 변수가 어떤 실험이나 관찰의 결과로 나타날 확률을 구하는 규칙

## ● 이산확률분포

- 이산확률변수가 따르는 확률규칙
- 이산확률분포 중 이론적 확률분포에는 이항분포, 포아송분포, 초기하분포 등이 있음
- $x=0, 1, \dots, n$ 이고 성공확률이  $p$ 인 이항분포

$$X \sim b(n, p) \quad f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad 1 > p > 0$$

## ● 연속확률분포

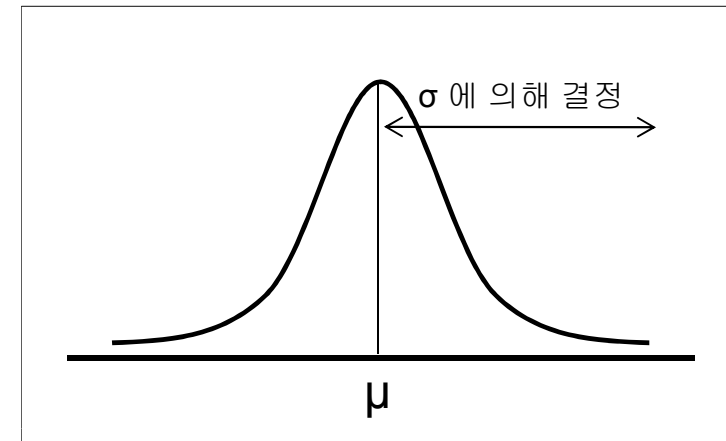
- 연속확률변수가 따르는 확률규칙
- 연속확률분포 중 이론적 확률분포에는 정규분포, T분포 등이 있음

# 정규분포 (1)

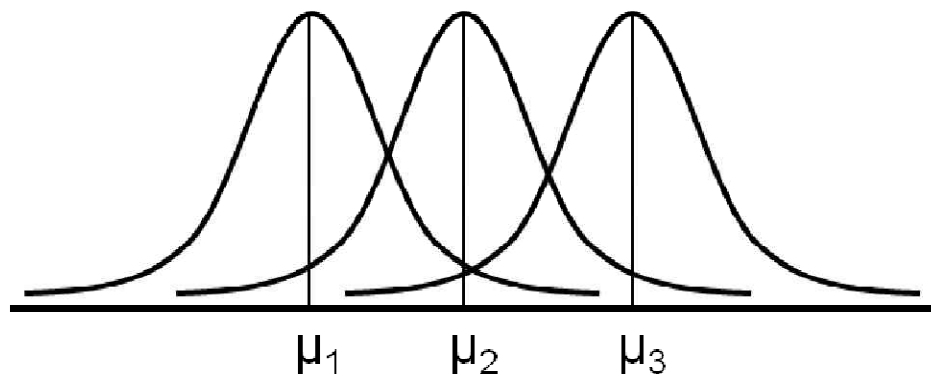
- 평균이  $\mu$ 이고 분산이  $\sigma^2$ 인 정규분포

$$X \sim N(\mu, \sigma^2)$$

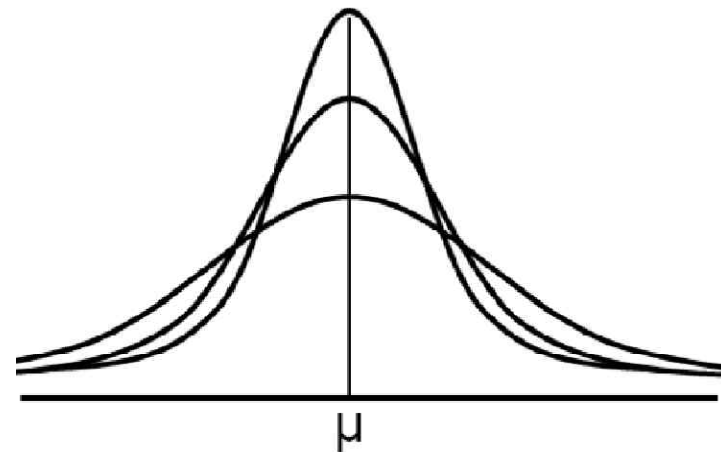
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2}}, \quad -\infty < x < \infty$$



- 분산이 같고 평균이 다른 정규분포

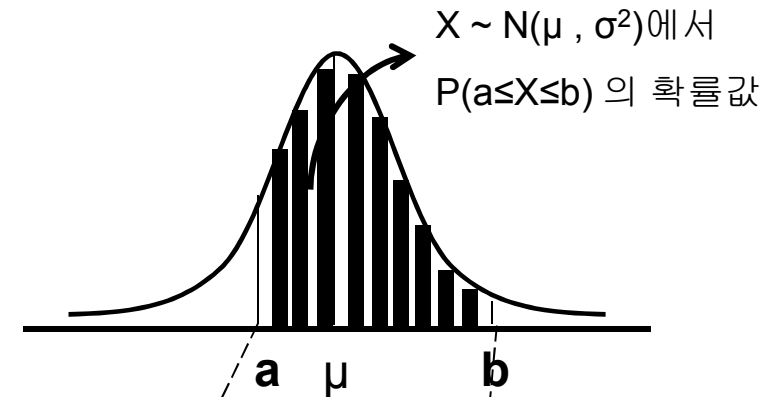


- 평균이 같고 분산이 다른 정규분포



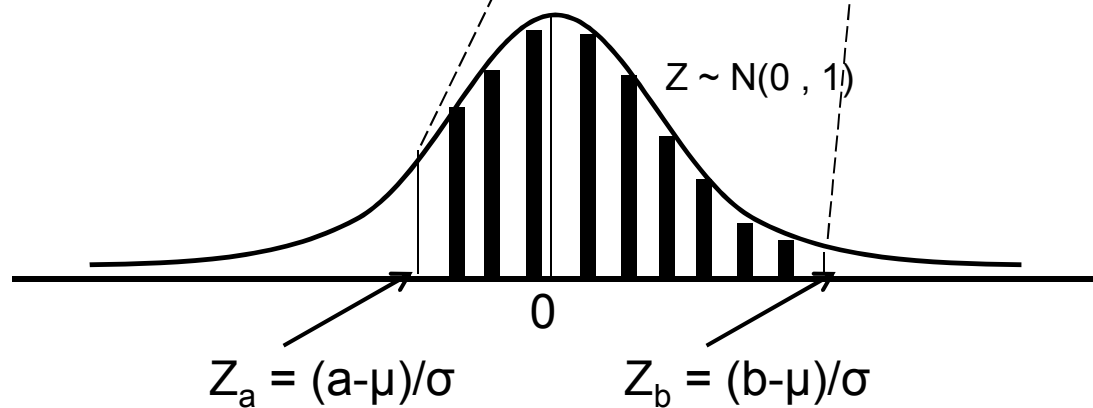
## 정규분포 (2)

- $X \sim N(\mu, \sigma^2)$ 일 때,  
변수  $X$ 가  $a$ 와  $b$  사이에 있을 확률은  
해당구간에서 정규분포 곡선이 차지하는 면적



- 평균이 0이고 분산이 1인 정규분포가 표준정규분포
- 모든 정규분포를 따르는 변수를 표준화하면 표준정규분포로 전환 가능

표준점수 :  $Z = \frac{X - \mu}{\sigma}$



## 정규분포 (3)

- 표준정규분포의 확률은 교재 **189**페이지 『표1. 표준정규분포표』 참고

✓ 표준정규분포의 관계식

- $P[a < Z < b] = P[Z > a] - P[Z > b]$
- $P[Z < a] = 1 - P[Z > a]$
- $P[Z < 0] = 0.5$
- $P[Z < -c] = 1 - P[Z < c]$

표준정규분포표에서  
다음 확률을 구하라.

(1)  $P[0 < Z < 1.25]$

(2)  $P[Z > -1.96]$

(3)  $P[-1.96 < Z < 1.96]$

(4)  $P[-2.33 < Z < 1.25]$

표준정규분포표에서 다음을 만족하는  
양수 **a**값을 구하라.

(1)  $P[Z > a] = 0.05$

(2)  $P[Z > a] = 0.025$

(3)  $P[-a < Z < a] = 0.95$

## 정규분포 (4)

---

- 정규분포를 따르는 확률변수의 확률을 구하는 문제는  
무조건 표준정규분포로 표준화시켜서 해결!
  - 교재 58-61페이지 예제 3.1, 3.2, 3.3

**$X \sim N(39, 6^2)$ 일 때 다음 확률을 구하라.**

(1)  $P[X > 50]$

(2)  $P[X < 30 \text{ 또는 } X > 55]$



## 정규분포 (5)

---

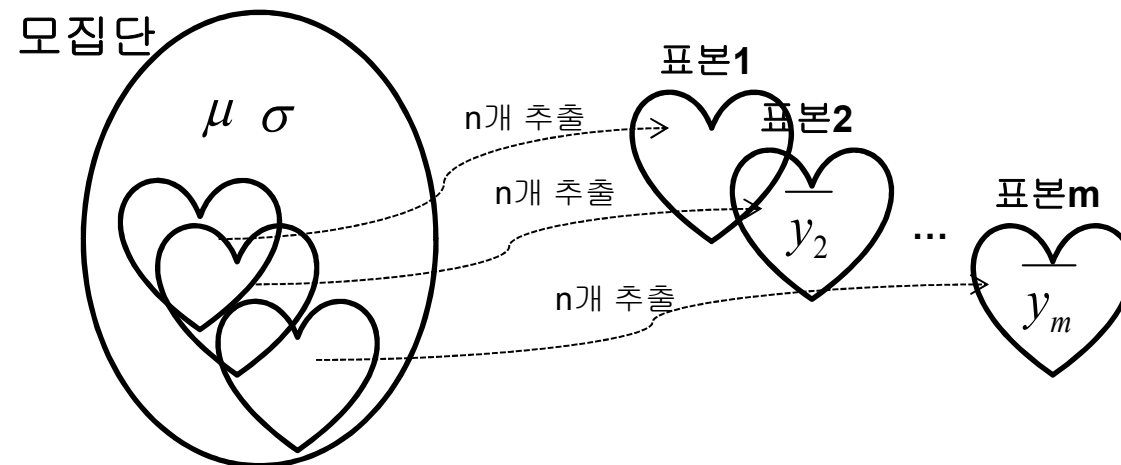
올림픽 대로를 통과하는 차량의 시속은 정규 분포를 따른다고 알려져 있다. 시속 **50km** 이상인 차량이 **86%**, **65km** 이하는 **70%**라고 한다.

(1) 평균과 표준편차를 구하라.

(2) 올림픽 대로는 40km 이하로 주행하면 차량 소통에 지장이 있다고 한다. 몇 %의 운전자가 차량소통에 지장을 주는지 구하여라.

# 중심극한정리 (1)

- 모집단의 분포와 상관없이, 표본평균의 분포는 근사적으로 정규분포를 따른다
  - 평균이  $\mu$ , 분산이  $\sigma^2$  인 모집단에서 크기  $n$ 의 표본을 독립적으로  $m$ 번 추출하였을 때,
    1. 표본평균의 평균은  $\mu$
    2. 표본평균의 표준편차는  $\sigma/\sqrt{n}$  (= 평균의 표준오차)
    3. 모집단의 분포가 정규분포일 때는 표본평균의 분포도 정규분포가 되고, 정규분포가 아니라도 표본의 수가 커지면 정규분포에 근사



## 중심극한정리 (2)

변수  $X$ 의 모집단의 평균은 **200**, 표준편차가 **15**라고 알려져 있다.

(1) 표본 크기가 36일 때, 표본 평균값이 196에서 199 사이에 있을 확률을 구하라.

(2) 표본 크기가 100일 때, 표본 평균값이 197이상일 확률을 구하라.

**00** 항공에서 발표한 자료에 따르면 승객 1인당 짐의 무게의 평균은 **10kg**이고 표준편차는 **2kg** 이라고 한다.

(1) 어느 날 랜덤으로 추출한 100명 승객들의 평균 짐의 무게가 11kg 이상일 확률을 구하라.

## 신뢰구간 (1)

---

\*\* 신뢰구간에 대한 신문기사 자료 예시

00일보가 지난 3월 한국갤럽에 의뢰해 서울시민 1006명을 대상으로 실시한 긴급 여론조사(표본오차 95% 신뢰 수준에서  $\pm 3.1\%$ 포인트, 응답률 21.9%) 결과, 안철수 원장은 서울시장 보궐선거에 출마 의사를 밝혔거나 출마가 거론되는 예비후보 10명 중 39.5%의 지지율로 1위를 차지했다. 안철수 원장의 지지율은 2위 한나라당 나경원 최고위원의 지지율 보다 세 배 이상 높았다.

## 신뢰구간 (2)

---

- 평균  $\mu$ 의 **95%** 신뢰구간의 의미
  - 표본을 같은 방법으로 100번 뽑아 신뢰구간을 구했을 때,  
100개의 신뢰구간 중 95개가 모평균  $\mu$ 를 포함
  
- 일반적으로, 평균  $\mu$ 의 **100(1- $\alpha$ )%** 신뢰구간  $CI_{\mu} = \bar{Y} \pm z_{\alpha/2} \sigma / \sqrt{n}$  로 표현
  - (1- $\alpha$ )를 신뢰수준, 신뢰계수라 함
  - 0.90, 0.95, 0.99를 가장 많이 쓰며, 이에 대응되는  $z_{\alpha/2}$  값  
 $z_{0.05} = 1.64, \quad z_{0.025} = 1.96, \quad z_{0.005} = 2.58$
  - 구간의 아래 끝점을 신뢰 하한, 위 끝점을 신뢰 상한이라 함
  - 교재 65페이지 예제 3.4

## 신뢰구간 (3)

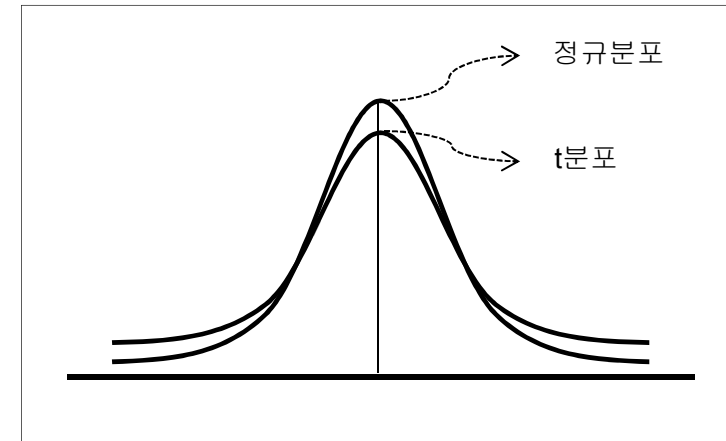
표준편차가 **15**인 모집단에서 뽑은 표본의 평균이 **200**이라고 한다.

- (1) 표본 크기가 36일 때, 모평균의 95% 신뢰구간을 구하여라.
- (2) 표본 크기가 45일 때, 모평균의 95% 신뢰구간을 구하여라.
- (3) 표본 크기가 64일 때, 모평균의 95% 신뢰구간을 구하여라.
- (4) 표본 크기가 64일 때, 모평균의 90% 신뢰구간을 구하여라.
- (5) 표본 크기가 64일 때, 모평균의 99% 신뢰구간을 구하여라.

# t 분포

- 모집단이 정규분포를 따르나,  
모분산의 값을 모르는 경우 표본평균의 분포

$$t = \frac{\bar{X} - \mu}{s_{\bar{X}}} = \frac{\bar{X} - \mu}{s_x / \sqrt{N}} \sim t(n-1)$$



- $\sigma$ 의 추정값으로 표본표준편차  $s$ 를 사용
- 표준정규분포에 비해 중심의 높이는 낮지만 꼬리부분은 두터운 대칭형
- 자유도(degree of freedom) (= 표본크기-1)을 모수로 갖는 분포
- 자유도가 커지면 정규분포와 거의 같아짐