통계학
– 제 5 장 정규분포

0011 0010 1010 1101 0001 0100 1011

연속확률 변수와 확률분포

- 연속확률 변수
  - 연속확률 변수
    - 확률변수가 취할 수 있는 가능한 값들이 구간 내의 연속적인 값
    - 확률변수가 취하는 특정 구간에 대한 확률만 계산 가능
  - 확률밀도함수 (probability density function, p.d.f.)
    - 연속확률변수 X 가 취할 수 있는 구간에 대해 확률을 대응시켜주는 함수
    - 확률밀도함수의 성질
      - 모든 x 값에 대해 $f(x) \geq 0$
      - $P(a \leq X \leq b) = \int_a^b f(x)\, dx$
      - $P(-\infty \leq X \leq \infty) = \int_{-\infty}^{\infty} f(x)\, dx = 1$

- 연속확률변수의 기대값과 분산
  - 연속확률변수의 기대값

$$E\left[X\right] = \int_{-\infty}^{\infty} x f(x)\ dx = \mu$$

  - 연속확률변수의 분산

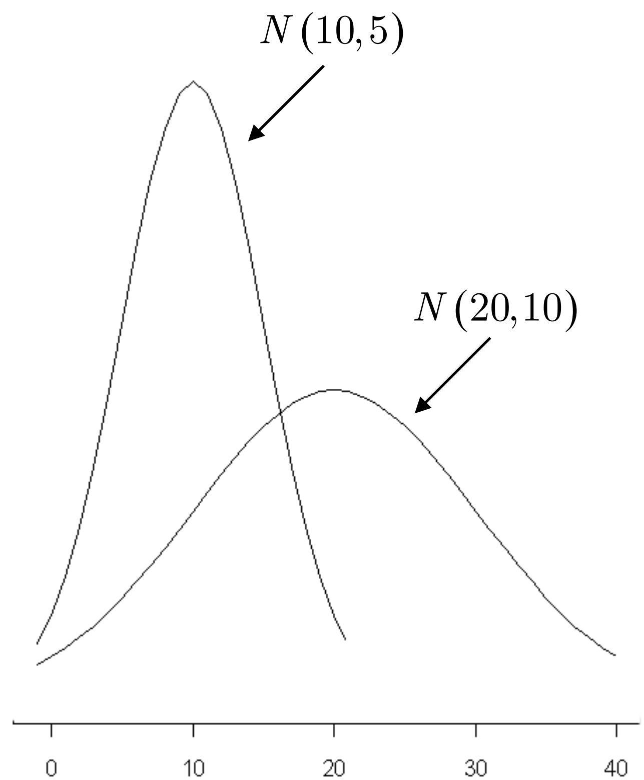
$$\begin{aligned} Var\left[X\right] &= E\left[(X-\mu)^2\right] \\ &= \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\ dx = \sigma^2 \end{aligned}$$

연속형 확률분포

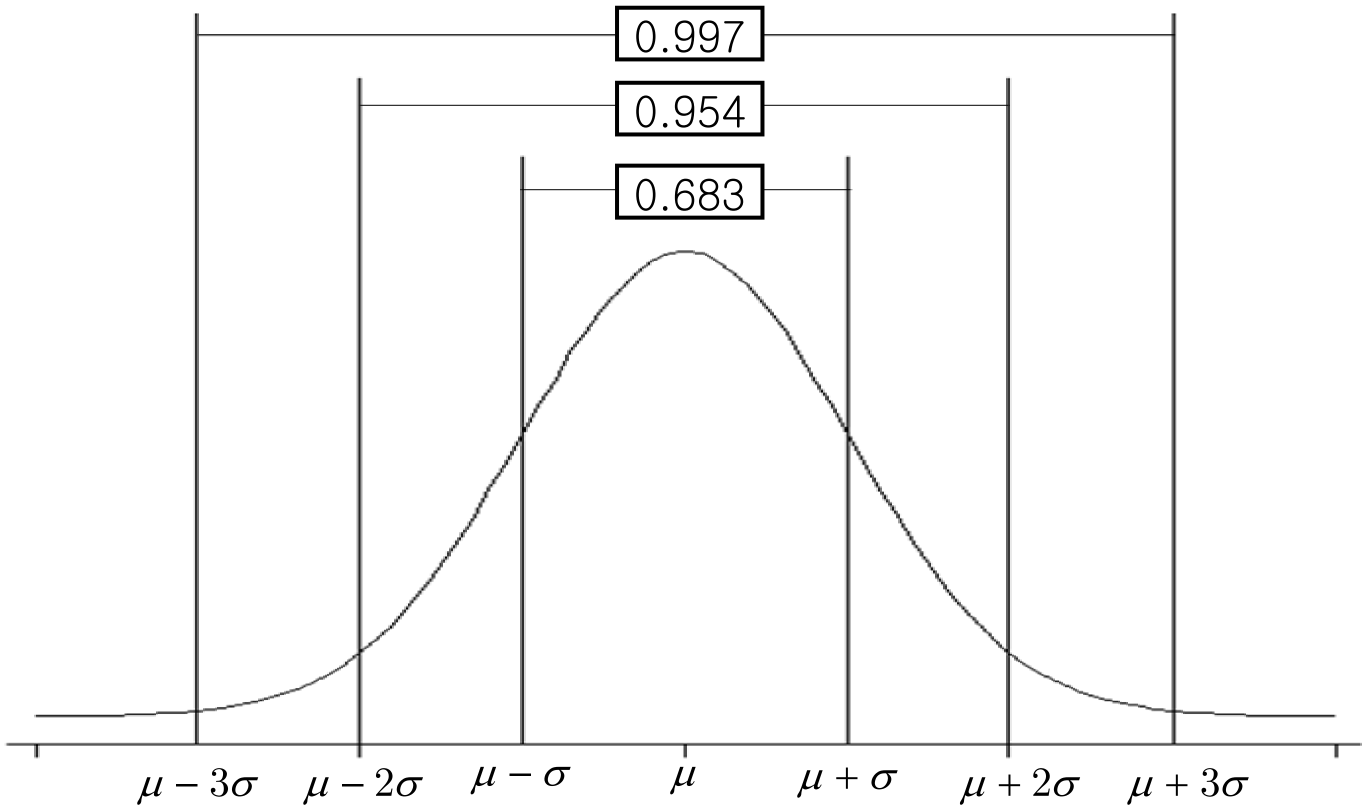
- 정규분포 (Normal Distribution)
  - 정규분포의 정의
    - 확률변수 X가 다음과 같은 확률밀도함수 $f(x)$를 가질 때,
      확률변수 X는 평균 μ, 분산 σ^2인 정규분포를 따른다고 함
    - $$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad -\infty < x < \infty$$
    - $$X \sim N\left(\mu, \sigma^2\right)$$
  - 정규분포가 중요시 되는 이유
    - 자연현상이나 사회현상에서 측정값들의 분포는 정규분포와 유사
    - 정규분포를 따르지 않는 분포도 약간의 변환을 통해
      정규분포와 유사한 형태를 만들 수 있음
    - 표본의 크기가 큰 경우 표본평균은 근사적으로 정규분포를 따름

정규분포의 특징

- 확률밀도함수의 형태는 평균을 중심으로 좌우대칭인 종모양 형태
- 정규분포를 따르는 변수에 상수를 곱하고 더하더라도 정규분포를 따름
- 분산이 크면 종모양이 펑퍼짐하고, 분산이 작으면 종모양이 뾰족함



- 평균으로부터 표준편차의 3배 이상 떨어진 값은 거의 갖지 않음



- 표준정규분포(Standard Normal Distribution)
  - 표준화(standardize) 변환
    - 평균이 μ 이고, 분산이 σ^2 인 확률변수 X 를
      평균이 0 이고, 분산이 1 인 확률변수 Z로 변환

$$Z = \frac{X - \mu}{\sigma}$$

    - 표준화된 확률변수 Z 는 평균이 0 이고, 분산이 1 인 정규분포를 따름

$$X \sim N\left(\mu, \sigma^2\right) \Leftrightarrow Z = \frac{X - \mu}{\sigma} \sim N\left(0,1\right)$$

    - 표준정규분포를 이용하여
      확률변수가 특정범위에 있을 확률을 쉽게 계산 가능
    - 교재 p.327 [부록 <표2> 정규분포표] 참조

- 예 – [모기업의 적성검사]
  - 모기업의 적성검사 점수는
    평균이 506이고 표준편차가 81인 정규분포를 따름
    - 확률변수 X : 적성검사 점수

$$X \sim N\left(506,\ 81^2\right)$$

    - 표준화

$$\frac{X-506}{81} = Z \sim N(0,1)$$

  - 점수가 574점 이하인 응시자들의 비율은?
    - $P(X \le 574) = P\left(\frac{X-506}{81} \le \frac{574-506}{81}\right)$

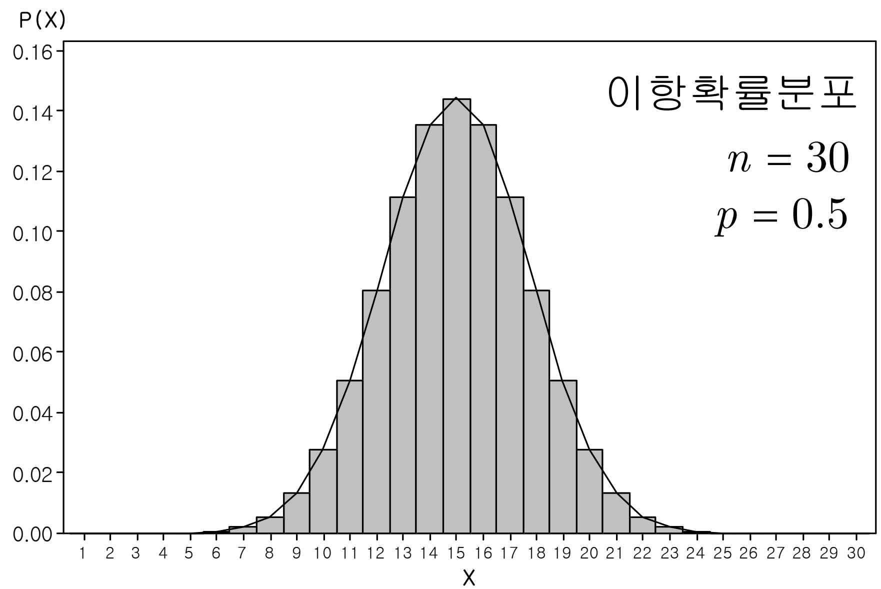
$$= P(Z \le 0.84)$$

$$= 0.7995$$

- 예 – [모기업의 적성검사]
  - 점수의 30번째 백분위수는?
    - $P(X < x) = 0.30$
    - $P(X < x) = P(Z < z) = 0.30$
    - 표준정규분포표 상에서
      $P(Z < -0.52) = 0.3015\,, \quad P(Z < -0.53) = 0.2981$
    - $0.30 \simeq P(Z < -0.524) = P\left(\dfrac{X - 506}{81} < -0.524\right) = P(X < 463.56)$

이항분포의 정규분포근사

- 이항분포의 정규 근사
  - 이항분포에 대한 근사로서 정규분포의 사용
    - 시행회수를 n 이라 하고, 성공확률을 p 라 할 때 일반적으로 np 와 $n(1-p)$ 가 모두 15 이상 인 경우 이항분포의 확률 히스토그램은 정규분포의 확률밀도함수와 같이 대칭인 종모양에 가까워짐



- 이항분포의 정규근사 확률 계산 방법
  - 이항분포에 대한 평균(μ)과 분산(σ^2)을 계산한다
    - $\mu = E[X] = np$
    - $\sigma^2 = Var[X] = np(1-p)$
  - 이산확률변수인 X를 연속확률변수로 변환(연속성수정)
    - 이항분포에서의 확률과 정규분포에서의 확률이 비슷하도록 하기 위해
      이산확률변수 X의 값에 0.5를 더하거나 0.5를 빼는 것
  - 정규분포를 사용하여 필요한 확률 계산

- 예 – [알코올 성분의 음료를 마시는 사람의 수]
  - 어느 도시에서 지난 5년 간의 조사 결과
  - 성인 전체의 30%가 정기적으로 알코올 성분의 음료를 마심
  - 1000명의 성인표본을 추출
    - 1000명 중에서 알코올 성분의 음료를 마시는 사람(S)과 마시지 않는 사람(F)으로 분류 가능

$$\Rightarrow P(S) = 0.3$$

    - 확률변수 X : 알코올 성분의 음료를 마시는 사람의 수

$$\Rightarrow X \sim \mathrm{Bin}(1000,\ 0.3)$$

$$E[X] = np = 1000 \times 0.3 = 300$$

$$Var[X] = np(1-p) = 1000 \times 0.3 \times 0.7 = 210$$

- 예 – [알코올 성분의 음료를 마시는 사람의 수]
  - 알코올 성분의 음료를 마시는 사람의 수가 280명 보다 작을 확률은?
  - 이항분포를 이용하여 계산하는 경우
    - $P(X<280) = P(X=0)+P(X=1)+ \cdots + P(X=279)$

$$= \binom{1000}{0}(0.3)^0(0.7)^{1000} + \cdots + \binom{1000}{279}(0.3)^{279}(0.7)^{721}$$

  - 정규 근사 확률을 이용하여 계산하는 경우
    - $X \sim N(300,\ 210)$
    - 연속성 수정 후 확률 계산

$$P(X<280) = P(X<279.5)$$

$$= P\left(\frac{X-300}{\sqrt{210}} < \frac{279.5-300}{\sqrt{210}}\right)$$

$$\doteq P(Z<-1.41) = 0.0793$$