Contents

Chapter 8 정규분포

01 연속확률분포

- 연속확률변수
  : 구간의 모든 값을 가질 수 있음
  주어진 구간에서 확률이 어떻게 분포하는지 함수를 이용해 표현

- 확률밀도함수 (probability density function, pdf)

  1) 모든 x값에 대해 $f(x) \geq 0$
  2) $P(a \leq X \leq b) = \int_a^b f_X(x)dx$
  3) $P(-\infty \leq X \leq \infty) = 1$

  를 만족하는 $f(x)$를 X의 확률밀도함수라고 한다.

- 예제 1, 예제 2

02 정규분포의 일반적인 성질 및 확률계산

- 정규분포 (normal distribution)

① 연속확률변수 X의 확률밀도함수가 $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ $-\infty < x < \infty$ 인 확률분포

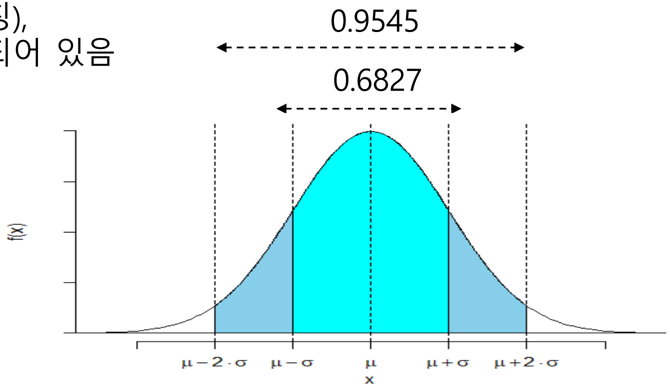
② 평균 μ와 분산 σ^2에 의해 분포가 확정됨

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

③ 특징

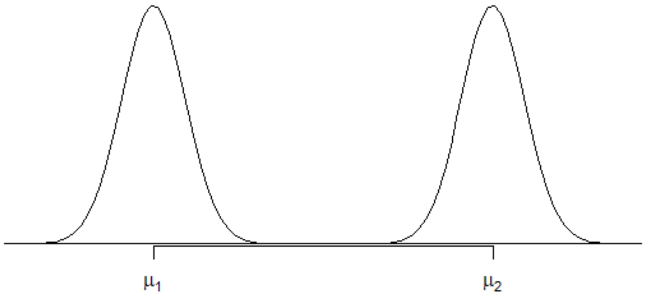
1) 평균 = 최빈값 = 중앙값 (평균을 중심으로 좌우대칭),

2) μ를 중심으로 $\pm 3\sigma$안에 확률이 거의 (0.9973) 집중되어 있음 (평균으로 부터 멀어지면 함수값이 급격히 작아짐)

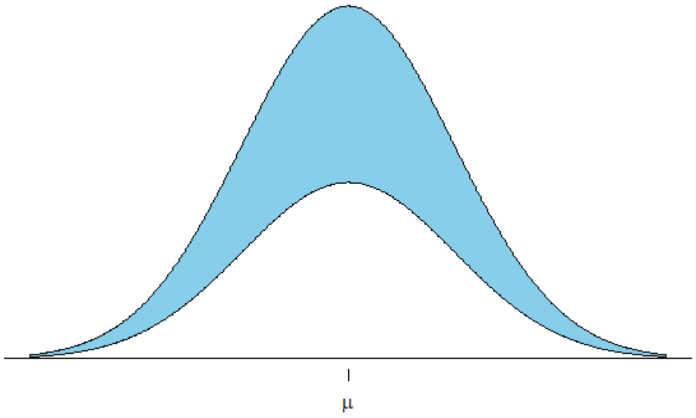


Chapter 8 정규분포

분산은 같고
평균이 다른 두 정규분포



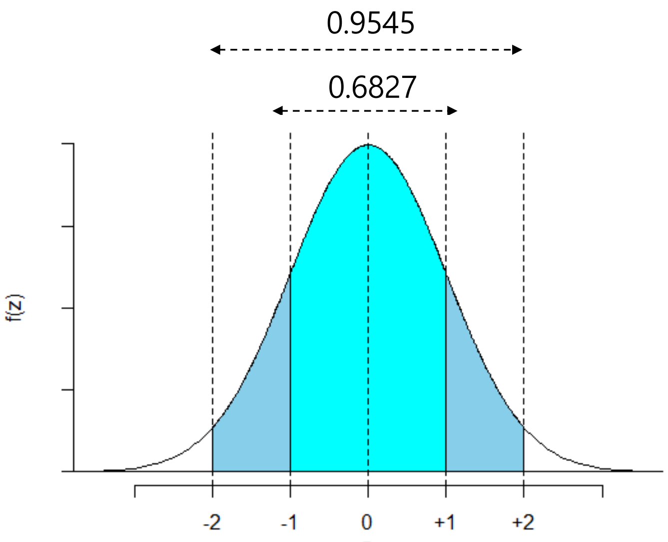
평균은 같고
분산이 다른 두 정규분포



정규분포

- 표준정규분포(standard normal distribution)
  : 평균이 0이고 분산이 1인 정규분포

$$Z \sim N(0,\ 1)$$



Chapter 8 정규분포 if $X \sim N(\mu, \sigma^2)$:

- 정규분포의 확률계산 · $E\left(\frac{X-\mu}{\sigma}\right) = E\left(\frac{1}{\sigma}\cdot X - \frac{\mu}{\sigma}\right) = \frac{E(X)-\mu}{\sigma} = \frac{\mu-\mu}{\sigma} = 0$

$$P(Z \le -z) = P(Z \ge z) = P(Z > z) = 1 - P(Z \le z)$$

- Z가 구간 $[a, b]$에 있을 확률

$$P(a \le Z \le b) = P(Z \le b) - P(Z \le a)$$

- 예제 3, 예제 4, 예제 5

$$\cdot\ Var\left(\frac{X-\mu}{\sigma}\right) = \frac{1}{\sigma^2} \cdot \underbrace{Var(X)}_{\sigma^2} = 1$$

- 표준정규확률변수

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

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

So, $\boxed{\frac{X-\mu}{\sigma}} \sim N(0,1)$

["표준화"]

- 예제 6, 예제 7

n이 크고 np와 n(1-p)가 10보다 크면, 연속성 수정 사용을 의심할 것

$np \geq 10$

$n(1-p) \geq 10$

보통 $p \approx 0.5$

〈N이 충분히 크다!!〉

03 이항분포의 정규분포 근사

"⊕ n이 크고"

- $X \sim Bin(n,p)$ 이고 np나 $n(1-p)$가 모두 클 경우에 (보통 10이상) X는 근사적으로 $N(np, np(1-p))$를 따른다.

$X \overset{*}{\sim} N(np, np(1-p))$

이 표시('근사적임')을 붙이는게 이항분포 핵심.

↑ N이 엄청 크면, 연속성 수정을 해서 구한 확률과 하지 않고 구한 확률은 거의 같다.

- 표준화 : $Z = \frac{X-np}{\sqrt{np(1-p)}} \sim N(0,1)$

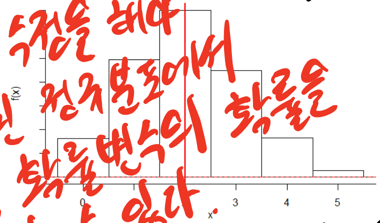
이항분포　　정규분포

- 연속성 수정 (continuity correiction) : $P[a \le X \le b] = P(a-\frac{1}{2} \le X \le b+\frac{1}{2})$

$P(X=x) \doteq P(x-\frac{1}{2} \le x \le x+\frac{1}{2})$　∴ $X \sim N(np, np(1-p))$

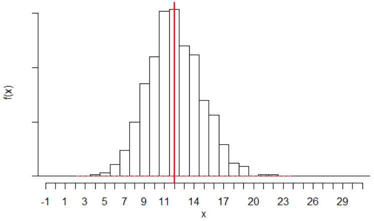
- 예제 8, 예제 9

연속성 수정을 해야 전환된 정규분포에서 해당 확률변수의 확률을 구할 수 있다.

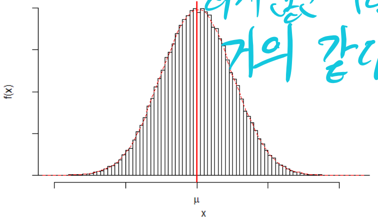


$n = 5, p = 0.4$

$np = 2$



$n = 30, p = 0.4$



$n = \infty, p = 0.4$

$n(1-p)=?$

$\begin{cases} np=12 \\ n(1-p)=18 \end{cases}$

$\therefore X \sim N(12, (2.6833)^2)$

04 정규분포가정의 조사

- 정규분포 가정의 조사
  : 정규분포가정이 맞는지를 확인하는 작업, "정규확률 그림"을 이용

- 정규확률그림(정규점수그림) (=Q-Q plot)
  : 표본이 이상적인 정규분포와 얼마나 유사한지를 보여주는 그림

- 정규확률그림 그리는 순서

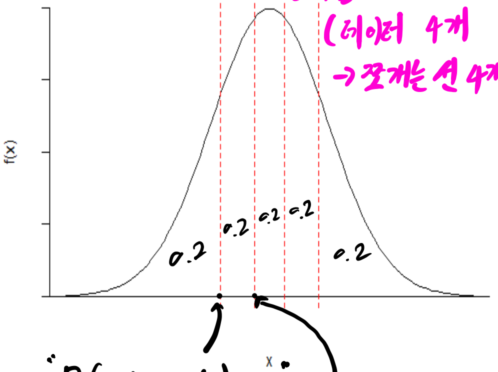
"정규점수"

① 자료를 작은 것부터 크기 순으로 나열한다
② 각 자료에 해당하는 점수를 계산한다
③ i번째 순서의 자료와 i번째 순서의 정규점수를 하나의 쌍으로 2차원 공간상에 나타낸다.

- 정규확률그림을 이용한 정규성 판정
: 정규확률그림이 "직선식"을 나타내면 정규분포의 가정이 타당하고 "곡선형태" 등 직선식을 벗어나면 정규분포의 가정이 의문시된다고 할 수 있다

Chapter 8 정규분포

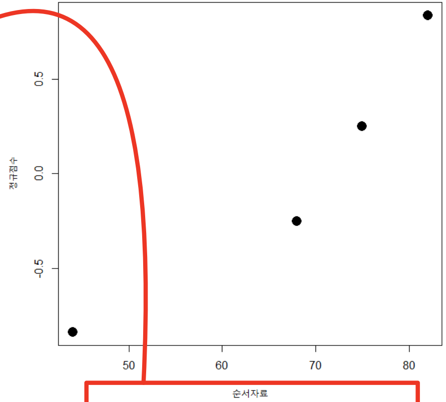
표본내 원소 갯수만큼 등확률로 쪼갠다.
(여기서 4개 → 쪼개는 선 4개)



"$P(z \le -0.84) = 0.2$"
"$P(z \le -0.25) = 0.4$"

이 둘을 매칭하여, 튜플 생성.

정규점수 (m_i)	크기순으로 나열된 관측값 x	이상적인 x 값
m_1 =-0.84	$x_{(1)}$ = 44	$\mu + \sigma m_1$
m_2 =-0.25	$x_{(2)}$ = 68	$\mu + \sigma m_2$
m_3 =0.25	$x_{(3)}$ = 75	$\mu + \sigma m_3$
m_4 =0.84	$x_{(4)}$ = 82	$\mu + \sigma m_4$



"$y = \sigma \cdot m + \mu$"
기울기가 'σ'이고, 절편이 'μ'인 직선

- 자료의 변환
  : 자료가 정규분포를 벗어 났을 때, 자료를 변환하여 정규분포를 따르도록 시도

Ex) x^2, x^3, $x^{\frac{1}{4}}$, $\sqrt{x}$, $\log x$, $\frac{1}{x}$, ...

로그정규분포.