기초통계

목 차

확률분포 – 복습

❖ 이산확률분포

- 초기하분포(hypergeometric distribution)
- 이항분포(binomial distribution)
- 포아송분포(Poisson distribution)

❖ 연속확률분포

- 일양 또는 균등분포(uniform distribution)
- 지수분포(exponential distribution)
- 정규분포(normal distribution)
- 표준정규분포(standard normal distribution)

연속확률분포

제 7장에서는 이산확률분포 중 가장 기본적이고 실제로 많이 적용되고 있는 이항분포와 그에 관계된 분포들에 대한 공부를 하였다.

이 장에서는 셀 수 있는 것이 아니라, 시간 몸무게 또는 0 과 1사이에 실수처럼 셀수 없는 연속적인 값을 가지는 연속확률분포들 중에서 대부분의 통계학 이론의 기본이 되는 정규분포에 관하여 공부해 보기로 한다.

❖ 연속확률변수와 관련된 확률은 그 확률변수의 확률밀도함수 $f(x)$에 의해 결정 (공식 유도 page 214참조)

1. 모든 x의 값에 대하여 **f(x)>=0**
2. 확률변수 X가 두 수 a와 b 사이에 놓일 확률은 $f(x)$의 아래 a와 b 사이에 있는 면적과 같다.

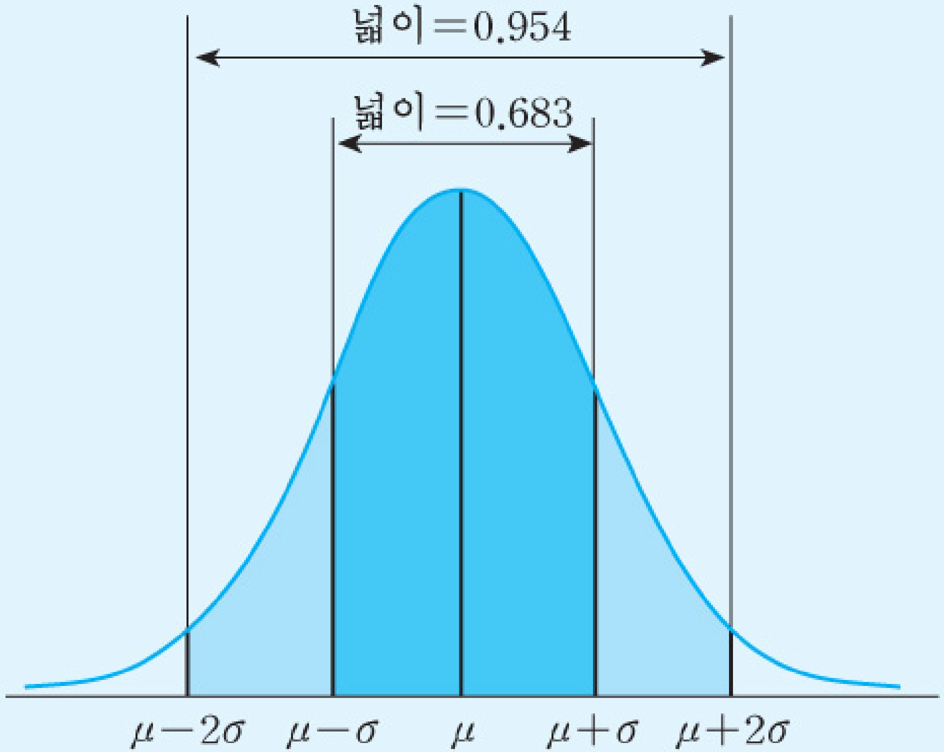
$$P(a \le X \le b) = \int_a^b f(x)dx$$

3. $f(x)$ 아래에 있는 전체 면적은 1이다.

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

정규분포의 일반적인 성질 및 확률 계산

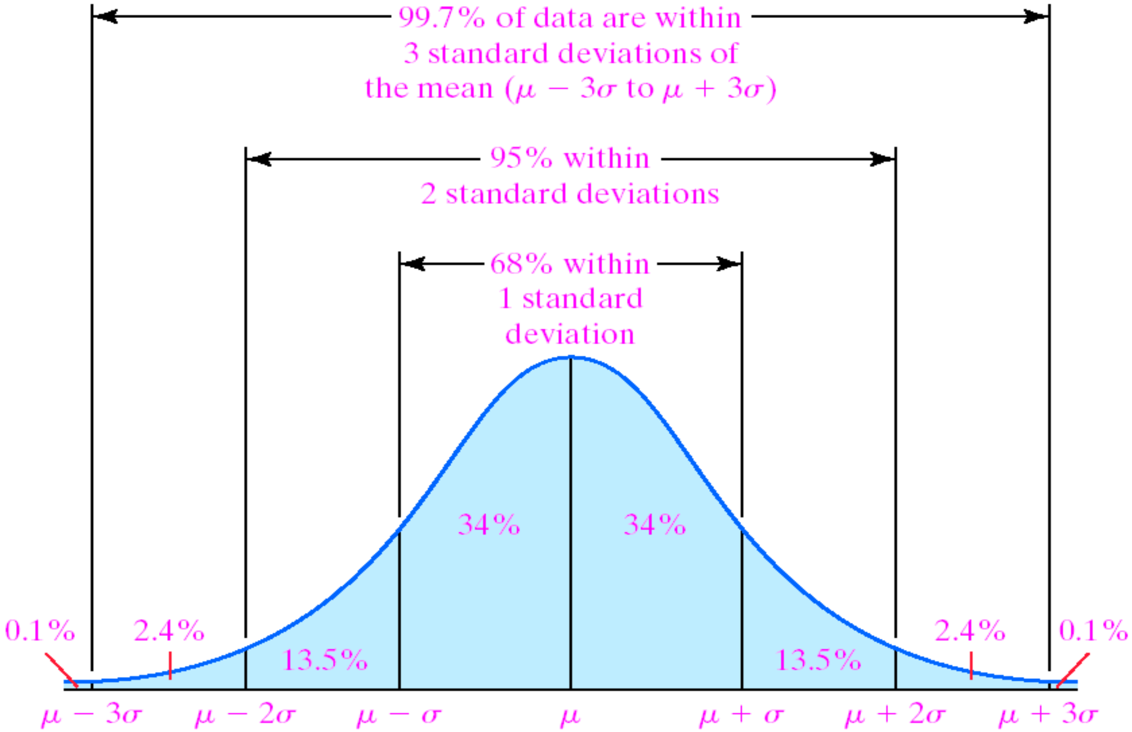
정규분포의 대략적은 특성



| 그림 2 | 정규분포의 확률밀도함수

정규분포의 특징

- 평균(μ)와 분산 (σ^2)에 의해 분포가 결정된다. $\sim N(\mu,\sigma^2)$
- 평균 μ 를 중심으로 대칭을 이루며 종 모양을 한다.
- $\mu \pm 3\sigma$ 안에 거의 모든 확률이 집중되어 있다.



정규분포의 비교

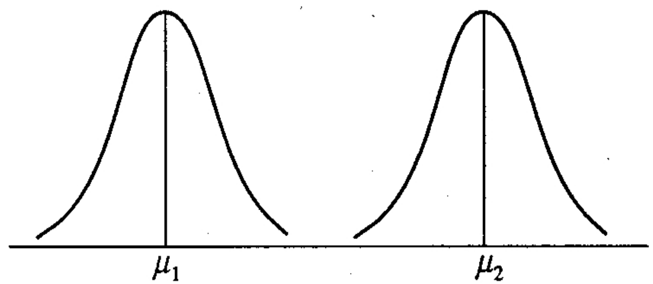


그림 4 : 분산은 같고 평균이 다른 정규분포의 확률밀도함수 비교

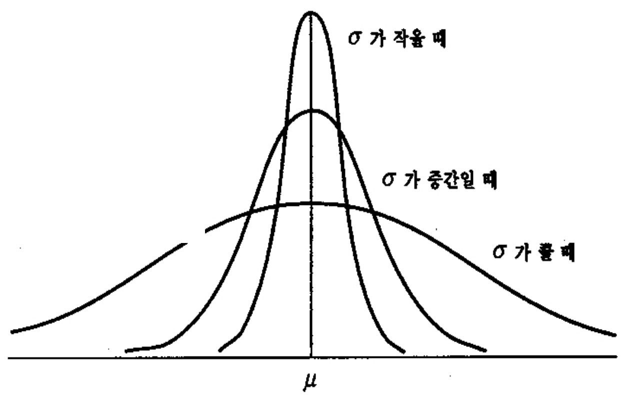


그림 5 : 평균이 같고 분산이 다른 정규분포의 확률밀도함수 비교

- 평균이 다르고, 분산이 같은 경우

  ➔ 수평이동

- 평균이 같고, 분산이 다른 경우

  ➔ 분포의 폭이 변한다.

표준 정규분포를 이용한 확률계산

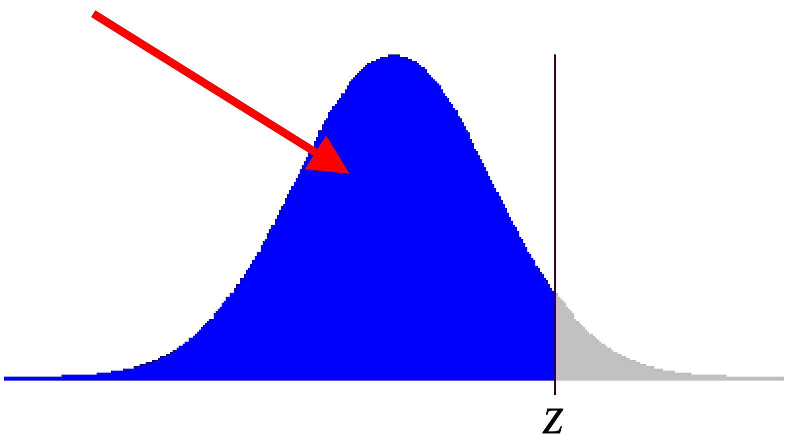
- 평균이 0 이고 분산이 1 인 정규분포를 **표준 정규분포**라 한다.
- 확률변수 Z가 N(0,1)이라 할 때 Z는 0을 중심으로 대칭인 분포를 갖는다.
- 그러므로 z에서 P(Z<=-z)=P[Z>=z]이고 Z는 연속확률분포이므로 P[Z>=z]=P[Z>z]=1-P[Z<=z]가 되어 아래와 같이 표현할 수 있다.

$$P[Z<=-z]=1-P[Z<=z]$$

표준 정규분포를 이용한 확률계산

- 평균이 0 이고 분산이 1 인 정규분포를 **표준 정규분포**라 한다.
- $Z \sim N(0,1)$ 부록 표 참조

$P[Z \le z]$: Cumulative distribution (누적확률)



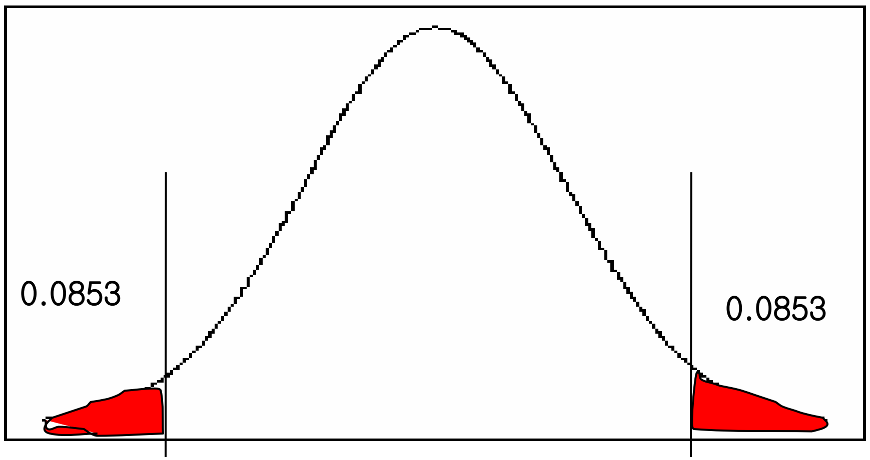
$ex)$ $P[Z \le 1.98] = 0.9761$

$P[Z \le -0.03] = 0.4880$

예 제

표준 정규분포를 이용한 확률계산

문제1) 확률 P[Z<=1.37]와 P[Z>=1.37]을 계산하여라.



표준정규분포 표로부터 P[Z>=1.37]=0.9147이 된다.

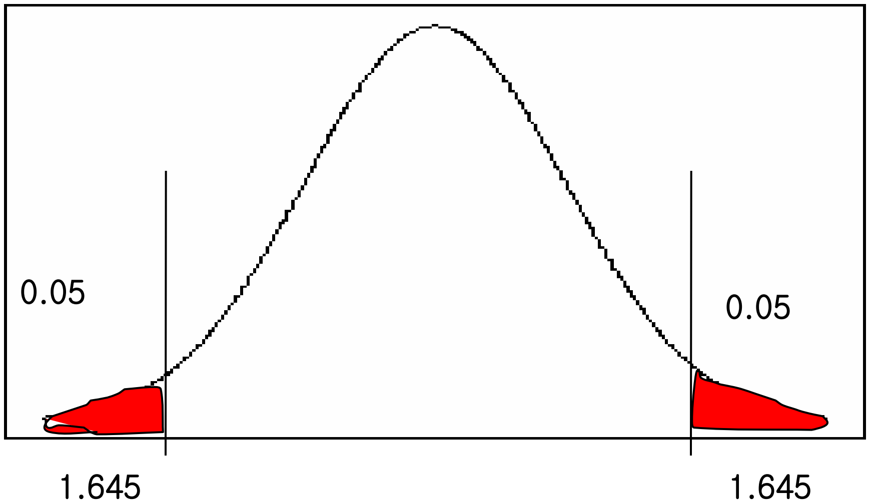
표준 정규분포를 이용한 확률계산

문제2) p=0.05를 만족하는 Z값을 찾는다면?

P[Z<=-1.64]=0.0505, P[Z<=-1.65]=0.0495

(1.64+1.65)/2 = 1.645

(내분점(중앙값) 이용하여 원하는 값을 구할 수 있음)



표준 정규분포를 이용한 확률계산

문제2) p=0.05를 만족하는 Z값을 찾는다면?

보통은 내분점을 사용하지 않고 가까운 점을 사용함.

정규분포에서의 확률 변수

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

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

$$P\,[X \le x] = P\,[\,\frac{X-\mu}{\sigma} \le \frac{x-\mu}{\sigma}\,] = P\,[\,Z \le \frac{x-\mu}{\sigma}\,]$$

평균이 0이고 분산이 1인 정규분포 N(1,0)을 따름

위 공식을 이용하면 일반 정규분포를 갖는 X에 대한 확률은 Z에 대한 확률로 표현 할 수 있음.

정규분포에서의 확률 변수

문제)확률변수 X가 N(60,16)을 따른다고 했을 때 P[55<=X<=63]을 계산하여라.

풀이)

X를 표준화시키면 $Z = \dfrac{X-60}{4}$ 가 된다. 그러므로

$$P[55 \le X \le 63] = P(\frac{55-60}{4} \le \frac{X-60}{4} \le \frac{63-60}{4})$$
$$= P[-1.25 \le Z \le 0.75]$$
$$= P[Z \le 0.75] - P[Z \le -1.25]$$

표준 정규분포표로 부터 P[Z<=0.75]=0.7734, P[Z<=-1.25]=0.1056이므로
확률은 0.7734-0.1056 = 0.6678

정규분포에서의 확률 변수

문제)어느 대학교의 일반수학 중간고사 성적은 분포가 평균 63이고 분산이 100인 정규분포를 따른다고 하자. 이때 다음의 물음에 답하라.

(1) 50점 이하의 학생은 몇 퍼센트 되겠는가?

풀이)

X를 표준화시키면 $Z = \dfrac{X-63}{4}$ 가 된다. 그러므로

$$P[X \leq 50] = P(\frac{X-63}{10} \leq \frac{50-63}{10})$$

$$= P[Z \leq -1.3] = 0.0968 = 9.68\%$$

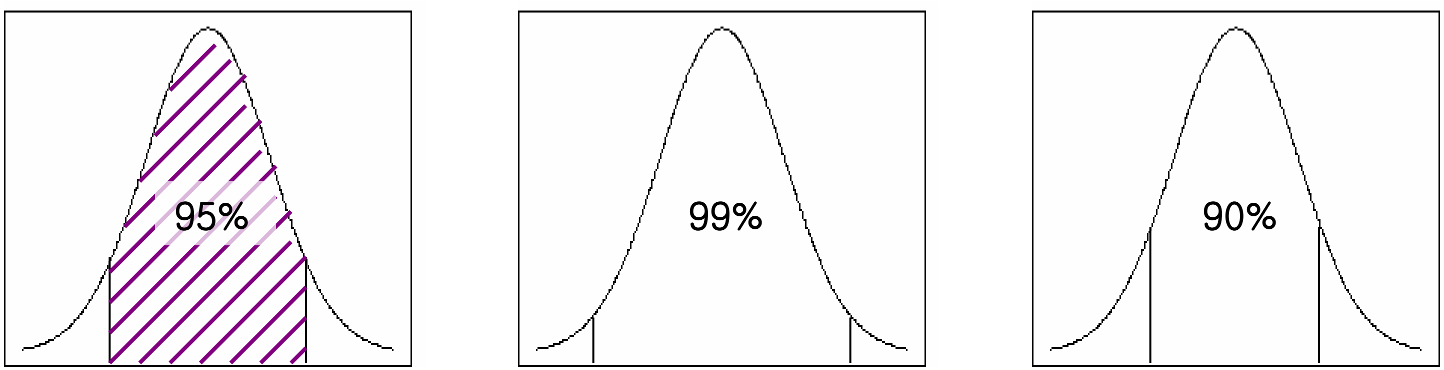
정규분포에서의 확률 변수

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

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

$$P[X \le x] = P\left[\frac{X-\mu}{\sigma} \le \frac{x-\mu}{\sigma}\right] = P\left[Z \le \frac{x-\mu}{\sigma}\right]$$

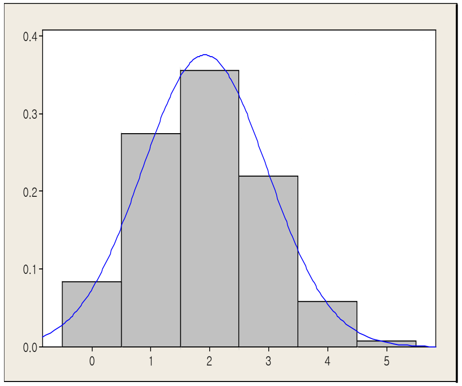
❖ 기억해둬야 할 수치



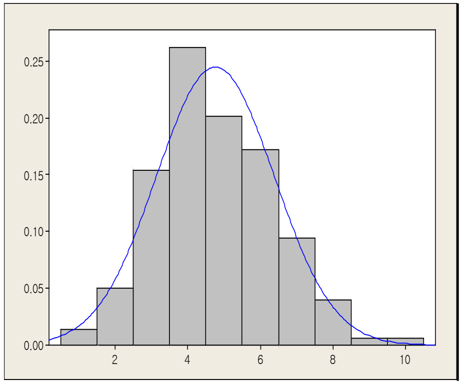
−1.96 1.96 −2.575 2.575 −1.645 1.645

이항분포의 정규분포 근사

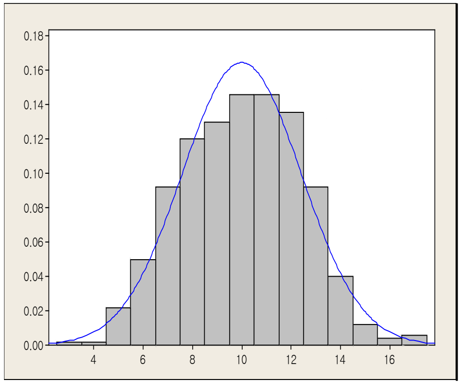
- 이항분포는 성공할 확률이 p인 베르누이 실행을 n번 반복할 때 성공의 횟수가 갖는 분포이다.
- 이항분포의 n이 아주 크고 p가 0이나 1에 가깝지 않을 때, 즉 np와 n(1-p)가 모두 클 때 그 분포가 정규분포에 가까워짐을 설명한다.



$n = 5 \quad p = 0.4$



$n = 12 \quad p = 0.4$



$n = 25 \quad p = 0.4$

이항분포의 정규분포 근사

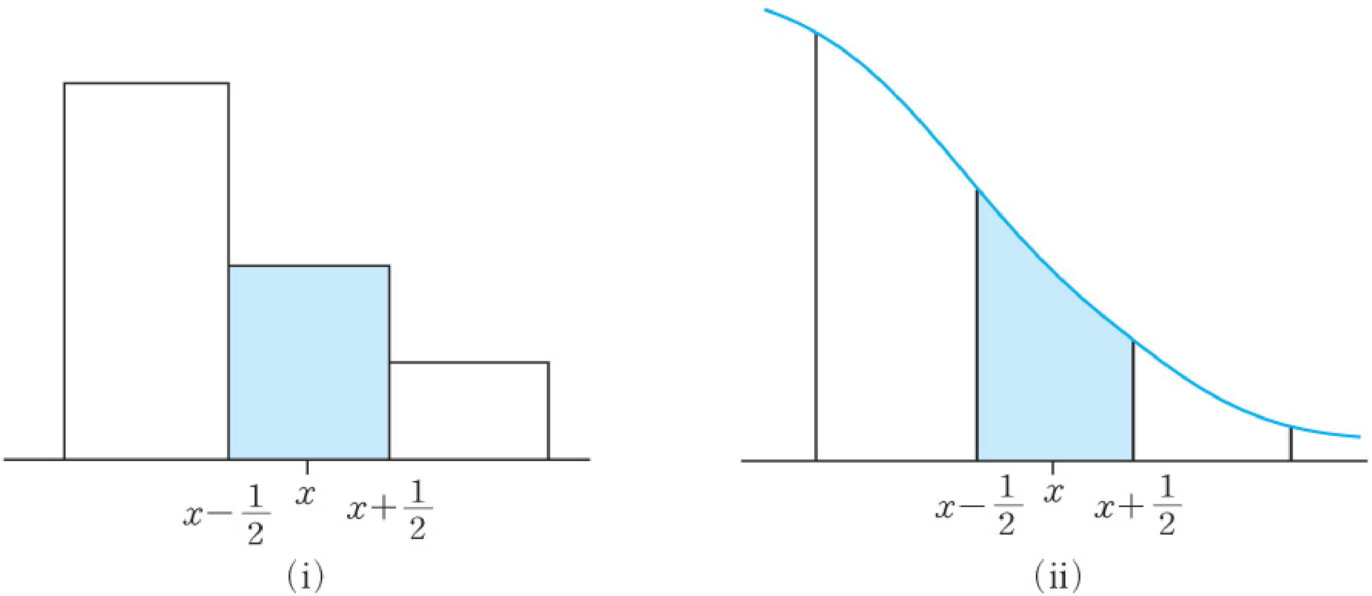
❖ 이항분포의 정규분포 근사

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

$$Z = \left(\frac{X - np}{\sqrt{npq}}\right)$$

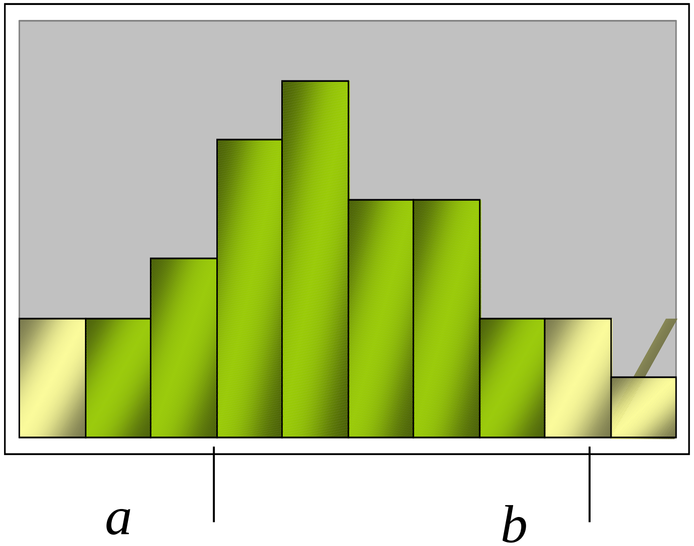
이항분포의 정규분포 근사

❖ 연속성 수정(continuity correction)

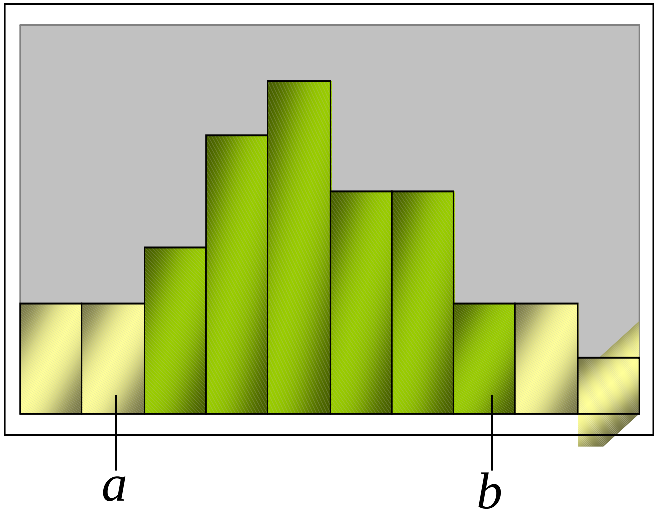


| 그림 10 | 이항분포 확률의 정규분포 확률로의 근사

가감하는 과정에서 이산분포에서 연속분포로 변경됨



$$a \le X \le b \quad \Rightarrow \; a-\frac{1}{2} \;\le\; X \;\le\; b+\frac{1}{2}$$



$$a < X \le b \quad \Rightarrow \; a+\frac{1}{2} \;\le\; X \;\le\; b+\frac{1}{2}$$

이항분포의 정규분포 근사

문제) 확률변수 X는 Bin(150, 0.6)의 분포를 갖는다고 하자. 이 때 다음의 확률을 근사적으로 구하라.

(1) 확률변수 X가 82이상 101이하일 확률은?

풀이) 이 경우 np, n(1-p)가 모두 10이상이므로 X는 근사적으로 정규분포를 따르게 된며, 평균은 μ = np = 150 x 0.6 = 90이고, 분산은 σ^2 = np(1-p) = 150 x 0.6 x 0.4 = 36이 된다.

그러므로 구하고자 하는 근사확률은 정규분포를 이용하여 다음과 같이 계산할 수 있다.

(1) P[82<=X<=101]=P[81.5<=X<=101.5]

$$= P\left(\frac{81.5-90}{6} \le \frac{X-90}{6} \le \frac{101.5-90}{6} \right)$$

$$\approx P[-1.42 \le Z \le 1.92]$$

$$= 0.9726 - 0.0778$$

$$= 0.8948$$

정규분포가정의 조사

이 절에서는 정규분포 가정의 조사에 대해서 알아본다.

방법)

1) 그래프를 그려본다.
2) 실제 자료의 범위가 자료의 비율 68.27%, 95.45%, 99.73%와 비교

이번 절에서는 정규분포의 가정을 조사하는 효과적인 그림으로서 정규점수그림, 또는 정규확률그림에 대해여 공부해본다.

정규분포 가정의 조사

❖ 정규분포 가정의 조사

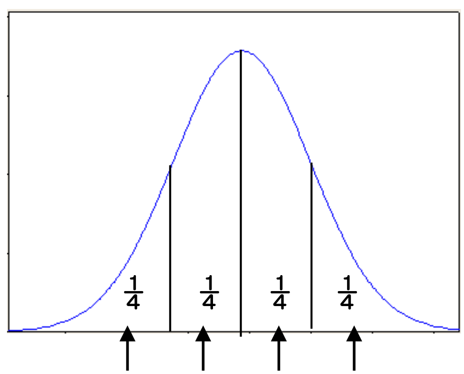
- 히스토그램 : 정규분포의 히스토그램과 유사한지 비교
- $(\bar{x} \pm s), (\bar{x} \pm 2s), (\bar{x} \pm 3s)$ 구간에 포함되는 비율을 정규분포의 비율과 비교

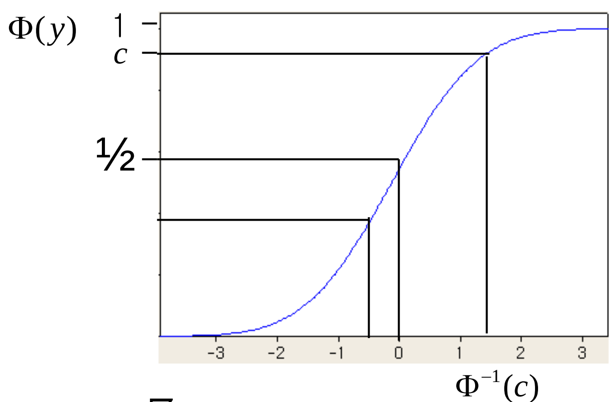
❖ 정규확률그림 (Normal Probability Plot)

- eg : 4개의 관측치 – 가장 이상적인 표준정규분포의 관측치 (정규점수)

  – 4개의 구간에서 확률을 반으로 나누는 값.

(1)

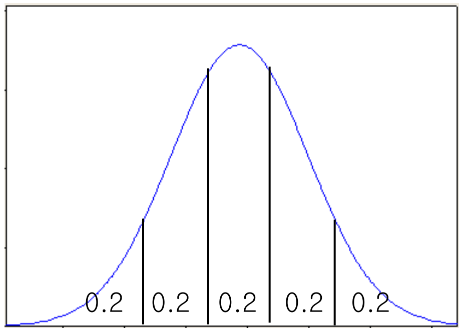




$$\Phi^{-1}\left(\frac{1}{8}\right),\ \Phi^{-1}\left(\frac{3}{8}\right),\ \Phi^{-1}\left(\frac{5}{8}\right),\ \Phi^{-1}\left(\frac{7}{8}\right)$$

$$n : \Phi^{-1}\left(\frac{1}{2n}\right),\ \Phi^{-1}\left(\frac{3}{2n}\right),\ \cdots\ ,\ \Phi^{-1}\left(\frac{2n-1}{2n}\right)$$

(2)



$$\Phi^{-1}(\frac{1}{5}),\ \Phi^{-1}(\frac{2}{5}),\ \cdots,\ \Phi^{-1}(\frac{4}{5})$$

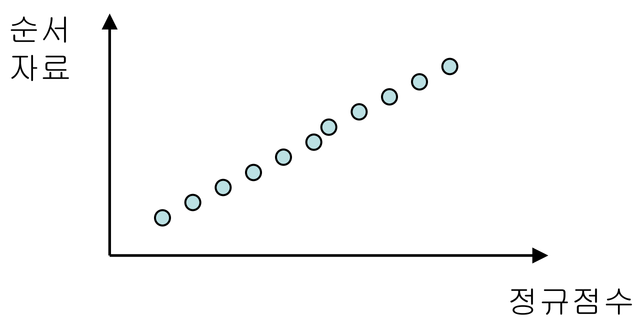
$$n : \Phi^{-1}(\frac{1}{n+1}),\ \cdots\ ,\ \Phi^{-1}(\frac{n}{n+1})$$

✓ 책에는 방법(2) 채택.

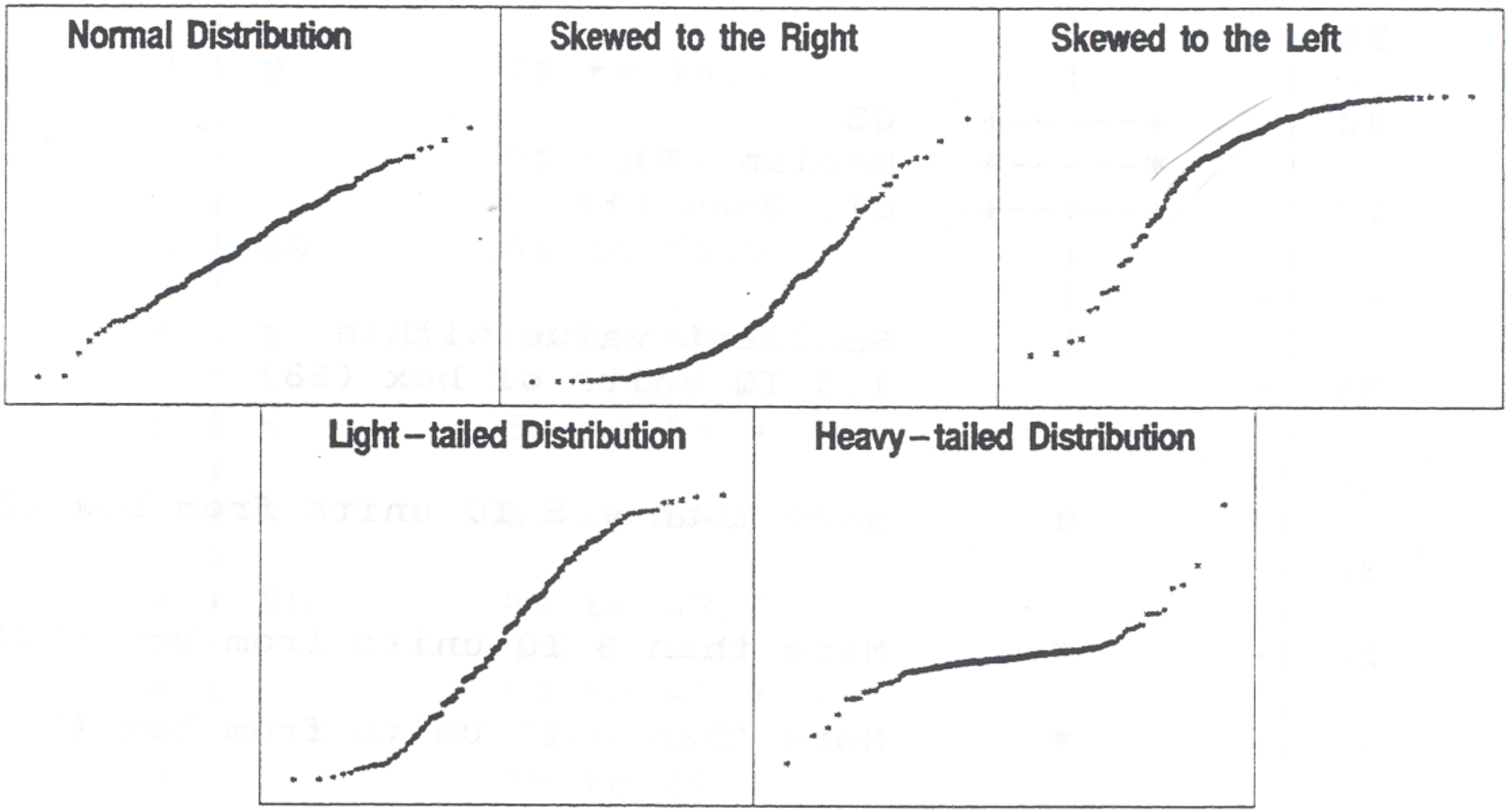
그리는 방법 : (1)자료를 크기 순으로 나열하고,

(2) 각 자료에 해당하는 점수를 계산한다.

(3) i번째 순서의 자료와 i번째 순서의 정규점수를 하나의 쌍으로 2차원 공간상에 나타낸다.



곡선이 나오는 여러 가지 예들



응용문제

❖ S 기업의 신형 복사기에 대한 손익 분기점은 3,500원이고 경험적으로 알려진 평균과 표준편차는 4,000원과 500원이다. 신형 복사기를 판매하는 것이 좋은 것인가? 의사 결정을 하시오.

break-even point	평균	표준편차	표준화	함수
3500	4000	500	0.84	1-normdist(A2,B2,C2,ture)

Thank You!