3.1장. 확률분포

박명현

서울대학교 학부대학

확률분포

- 확률변수의 대표적인 확률분포들를 소개하고 그의 성질들을 파악해보자.
- 이산형 확률분포
  - 베르누이분포
  - 이항분포
  - 초기하분포
- 연속형 확률분포
  - 균일분포
  - 정규분포

이산형 확률분포

베르누이분포

- 어떤 확률실험이 두 개의 결과만을 가질 수 있다고 가정하자.

  예)입시에서 합격과 불합격,

  스포츠 경기에서 승리와 패배,

  수술 후 환자의 치유 여부 등

- 베르누이시행(Bernoulli trial)

(1) 시행의 결과가 '성공(s)' 또는 '실패(f)' 두 가지뿐이고, 따라서 표본공간은 $S = \{s, f\}$로 원소가 두 개인 집합이다.

(2) 성공의 확률을 $p = P\{s\}$, 실패의 확률을 $q = P\{f\}$로 나타내며, 따라서 $0 \leq p \leq 1$이고 $q = 1 - p$가 된다.

베르누이분포

- 베르누이확률변수(Bernoulli random variable):

  베르누이시행의 표본공간 $S = \{s, f\}$에서 시행의 결과에 따라 0 또는 1의 값을 대응시키는 확률변수

$$X = \begin{cases} 1, & if\ Success \\ 0, & if\ Fail \end{cases}$$

- 성공의 확률이 p인 베르누이분포는 다음과 같다.

$$p(x) = p^x(1-p)^{1-x} = \begin{cases} p, & if\ (X=1) \\ 1-p, & if\ (X=0) \end{cases}$$

- 베르누이분포의 평균과 분산

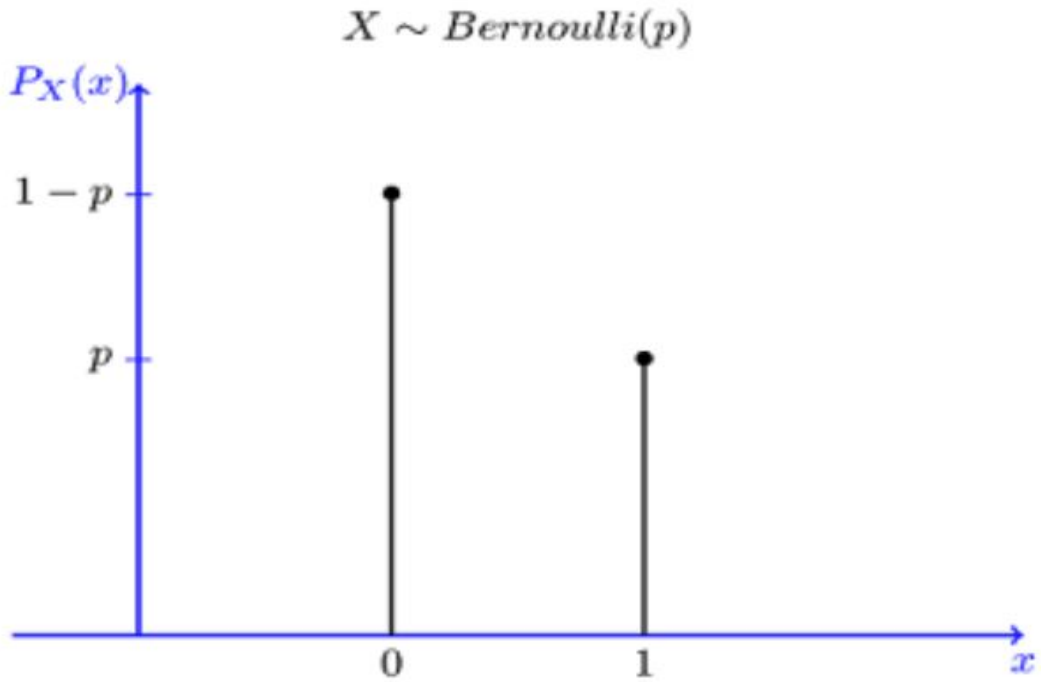
$$E(X) = 1 \cdot p + 0 \cdot (1-p) = p$$

$$Var(X) = E(X^2) - E(X)^2 = 1^2 \cdot p + 0^2 \cdot (1-p) - p^2 = p(1-p)$$

베르누이분포

표 4-1 베르누이 분포

x	0	1
$p(x)$	$1-p$	p



이항분포(binomial distribution)

- 시행의 결과가 "1"(성공) 또는 "0"(실패) 와 같이 이원적으로 분류되는 (무한)모집단에서 랜덤표본 $X_1, X_2, \cdots, X_n$을 추출하였다.
- 추출된 확률변수 $X_1, X_2, \cdots, X_n$은 서로 독립이며 성공의 확률이 p인 베르누이분포 $B(1, p)$를 따른다:

$$X_i = \begin{cases} 1, & (success) \\ 0, & (fail) \end{cases} \qquad for\ i = 1, \cdots, n$$

- 이때 총 n번의 시행 중 성공의 횟수(표본에서 "1"의 개수)는 다음과 같이 X_i들의 합인 X로 표현되고, 이때 확률변수 X는 이항분포를 따른다.

$$X = X_1 + \cdots + X_n \sim B(n, p)$$

이항분포

- 이항분포: 성공률 p인 베르누이시행을 n번 독립적으로 반복시행할 때 성공 횟수 X의 분포.

  - 이항분포의 확률질량함수는 다음과 같다.

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x};\ x = 0, 1, \cdots, n$$

  - x번의 실험에서 성공, $n-x$번의 실험에서 실패할 확률: $p^x(1-p)^{n-x}$
  - x번의 성공과 $n-x$번의 실패를 한 줄로 늘어 놓는 경우의 수: $\binom{n}{x}$

- 이항계수(binomial coefficient): $\binom{n}{x} = \frac{n!}{(n-x)!x!}$
- 확률변수 X는 총 n번의 시행 중 성공한 횟수로 정의한다.
- 확률변수 X가 모수 (n, p)를 따를 때, $X \sim B(n, p)$라 표현한다.
- $n = 1$일 때 $B(n, p)$는 "1"의 확률이 p인 베르누이분포가 된다.

이항분포의 평균과 분산

- $X = \sum_{i=1}^{n} X_i, \quad X_i \sim B(1, p)$의 성질을 이용하여 유도가능.
- 이항분포의 평균: $E(X) = np$

$$
\begin{aligned}
pf) \quad E(X) &= E(X_1) + \cdots + E(X_n) \\
&= p + \cdots + p \\
&= np
\end{aligned}
$$

- 이항분포의 분산: $Var(X) = np(1-p)$

$$
\begin{aligned}
pf) \quad Var(X) &= Var(X_1 + \cdots + X_n) \\
&= \sum_{i=1}^{n} Var(X_i) + 2 \sum_{i<j} Cov(X_i, X_j) \\
&= \sum_{i=1}^{n} Var(X_i) \\
&= np(1-p)
\end{aligned}
$$

이항분포

- $n = 6$일 때, $p = 0.5, 0.3, 0.7$값의 변화에 따른 이항분포의 모양변화

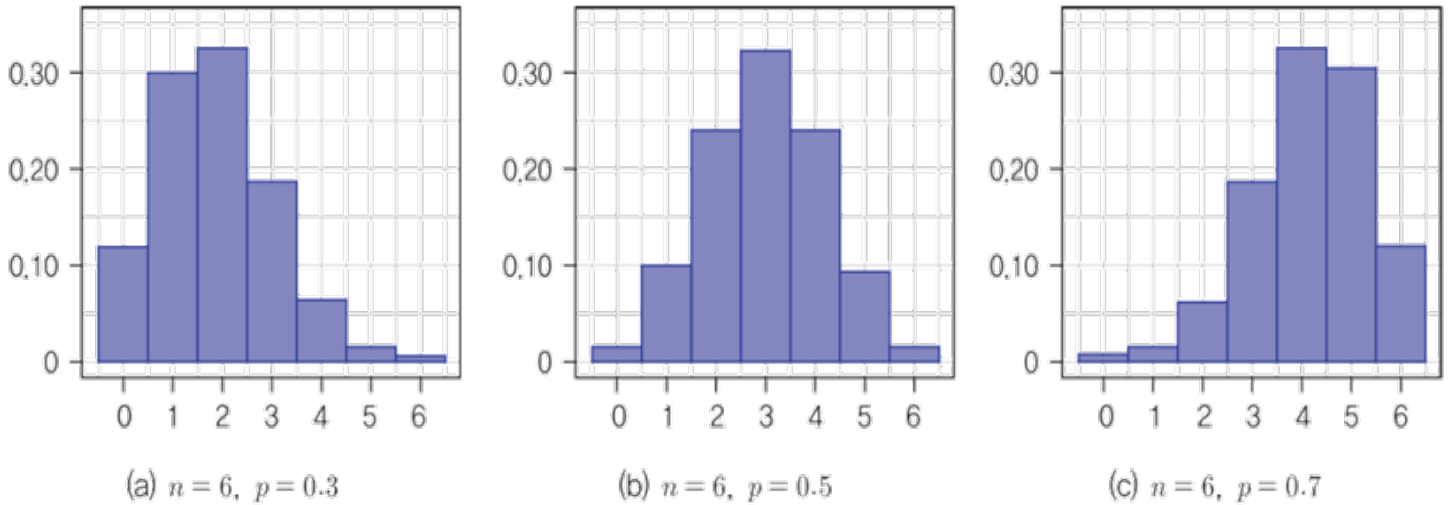


그림 4-4 $n = 6$일 때 이항분포

- $p = 0.5$: 분포의 모양이 대칭,
  $p < 0.5$: skewed to the right, $p > 0.5$: skewed to the left.

이항분포의 확률계산

- 이항분포의 누적확률분포표를 이용하여 계산.
  => $P(X \leq c)$의 값들이 여러 가지 n, p, c값에 대하여 주어져 있음.

Figure: 이항분포의 누적확률분포표

$$P\{X \leqq c\} = \sum_{x=0}^{c} \binom{n}{x} p^x (1-p)^{n-x}$$

		p										
		.05	.10	.20	.30	.40	.50	.60	.70	.80	.90	.95
	c											
$n=1$	0	.950	.900	.800	.700	.600	.500	.400	.300	.200	.100	.050
	1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$n=2$	0	.902	.810	.640	.490	.360	.250	.160	.090	.040	.010	.002
	1	.997	.990	.960	.910	.840	.750	.640	.510	.360	.190	.097
	2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$n=3$	0	.857	.729	.512	.343	.216	.125	.064	.027	.008	.001	.000
	1	.993	.972	.896	.784	.648	.500	.352	.216	.104	.028	.007
	2	1.000	.999	.992	.973	.936	.875	.784	.657	.488	.271	.143
	3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$n=4$	0	.815	.656	.410	.240	.130	.063	.026	.008	.002	.000	.000
	1	.986	.948	.819	.652	.475	.313	.179	.084	.027	.004	.000
	2	1.000	.996	.973	.916	.821	.688	.525	.348	.181	.052	.014
	3	1.000	1.000	.998	.992	.974	.938	.870	.760	.590	.344	.185
	4	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

이항분포의 확률계산

- 다양한 형태의 확률도 $P(X \leq c)$꼴로 바꾸어 계산할 수 있다.
- a, b가 정수일 때,

  $P(X = a) = P(X \leq a) - P(X \leq a - 1)$

  $P(a \leq X \leq b) = P(X \leq b) - P(X \leq a - 1)$

  $P(a < X < b) = P(X \leq b - 1) - P(X \leq a)$

  $P(X \geq a) = 1 - P(X \leq a - 1)$

- binomial distribution calculator
  https://homepage.divms.uiowa.edu/ mbognar/applets/bin.html

이항분포

Example

5지 선다형 문제로 구성된 20문항의 시험에서 무작위로 답을 선택할 때, 다음 확률을 구하라.
(a) 정답을 하나도 맞히지 못할 확률
(b) 정답을 2개 이상 맞힐 확률
(c) 정답을 4개 이상 6개 이하 맞힐 확률

sol) 20문항 중 정답의 수를 X라고 하면, $X \sim B(n = 20, p = 0.2)$이고, 따라서 이항분포표에서 다음의 확률을 얻을 수 있다.

(a) $P(X = 0) = \binom{20}{0} 0.2^0 0.8^{20} = 0.012$

(b) $P(X \geq 2) = 1 - P(X = 0) - P(X = 1) = 0.931$

(c) $P(4 \leq X \leq 6) = P(X \leq 6) - P(X \leq 3) = 0.913 - 0.411 = 0.502$

이항분포

Example

한 택배 회사의 정시 배달 성공률은 95%로 알려져 있다. 배송 20건을 단순랜덤추출하여 조사했을 때, 표본에서 정시 배달률이 90% 이하일 확률을 구하라

sol) 배송 20건 중 정시 배달 실패 건수를 확률변수 X라고 하면, 근사적으로 $X \sim B(n = 20, p = 0.05)$

- 표본에서 정시 배달 성공률이 90% 이하라는 것은 정시 배달 실패율 $\frac{X}{20}$가 10% 이상이라는 뜻이다. 이를 식으로 나타내어 계산하면 다음과 같다.

$$P(\frac{X}{20} \geq 0.1) = P(X \geq 2) = 1 - P(X = 0) - P(X = 1) = 0.264$$

초기하분포(hypergeometric distribution)

예) 어떤 주머니에 D개의 빨간 공과 N-D개의 하얀 공이 들어 있을 때, n개의 공을 랜덤하게 비복원으로 뽑았을 때, X를 빨간 공의 개수라고 하자.

- 특성값이 두 가지로 분류되는 경우
- 특성값 "1"의 개수가 D, "0"의 개수가 $N-D$인 크기 N의 (유한)모집단에서 크기 n인 랜덤표본을 비복원으로 뽑음
- $X = \sum_{i=1}^{n} X_i, \quad X_i \sim B(1, p)$
- n개의 표본에서 "1"의 개수를 통계량 X라고 하자.
- X는 초기하분포를 따르며, 그 확률질량함수는 다음과 같다.

$$p(x) = \frac{\binom{D}{x}\binom{N-D}{n-x}}{\binom{N}{n}}; \; x = 0, 1, \cdots, n \; (n \leq D, n \leq N-D)$$

초기하분포(hypergeometric distribution)

- 모비율 $p = \frac{D}{N}$의 추정은 표본비율 $\hat{p}$을 이용

  $\hat{p}$=(표본에서 "1"의 개수)$/n = \frac{X}{n}$

- 초기하분포의 평균과 분산

  - $E(X) = np$ 단, $p = D/N$
  - $Var(X) = np(1-p) \cdot \frac{N-n}{N-1}$

- 초기하분포의 이항분포 근사

: 초기하분포(비복원추출)에서 추출방법을 복원추출로 하는 경우 다음의 조건을 만족하면, 초기하 분포는 이항분포 $B(n, p)$에 수렴한다.

$$N \longrightarrow \infty,\ D \longrightarrow \infty,\ \frac{D}{N} \longrightarrow p$$

초기하분포

Example

신세대 의식 구조 조사를 실시하고자 찬성자의 수가 6명, 반대자의 수가 4명인 크기가 10인 집단에서 3명의 랜덤표본을 뽑았다. 이 중 찬성자수의 평균과 분산을 구하여라.

sol) 찬성자 수 X는 아래의 초기하분포를 따른다.

$$p(x) = \frac{\binom{6}{x}\binom{4}{3-x}}{\binom{10}{3}}; \ x = 0, 1, 2, 3$$

- 따라서, X의 평균과 분산은 다음과 같이 계산된다.

$$E(X) = 0 \cdot \frac{\binom{6}{0}\binom{4}{3}}{\binom{10}{3}} + 1 \cdot \frac{\binom{6}{1}\binom{4}{2}}{\binom{10}{3}} + 2 \cdot \frac{\binom{6}{2}\binom{4}{1}}{\binom{10}{3}} + 3 \cdot \frac{\binom{6}{3}\binom{4}{0}}{\binom{10}{3}} = \frac{9}{5}$$

- 마찬가지 방법으로 $E(X^2) = \frac{19}{5}$. 따라서

$$Var(X) = E(X^2) - E(X)^2 = \frac{14}{25}$$

초기하분포의 이항분포 근사

Example

어떤 제품을 생산하는 공정의 불량률은 5%로 알려져 있다. 오늘 생산한 10,000개의 제품 중에서 20개를 단순랜덤추출하여 조사할 때 표본의 불량률이 10%이상일 확률을 구하여라.

sol) 20개 중 불량품의 수를 X라고 하면 X는 초기하분포를 따른다.

$$\begin{aligned} P(X/20 \geq 0.1) &= P(X \geq 2) \\ &= 1 - P(X=0) - P(X=1) \\ &= 1 - \frac{\binom{9500}{19}\binom{500}{1}}{\binom{10000}{20}} - \frac{\binom{9500}{20}\binom{500}{0}}{\binom{10000}{20}} = \cdots \end{aligned}$$

- 그런데 X는 근사적으로 이항분포 $B(20, 0.05)$를 따르므로

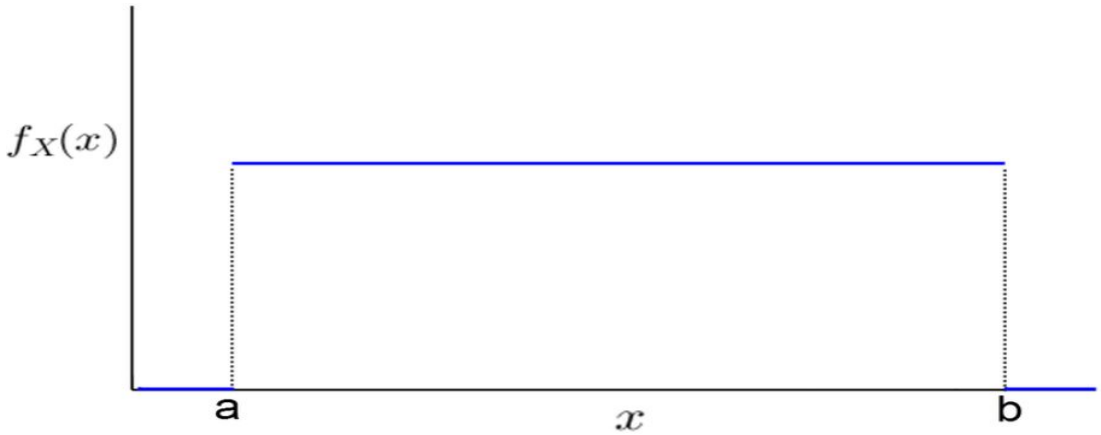
$$P(X \geq 2) = 1 - P(X \leq 1) \approx 1 - 0.736 = 0.264$$

연속형 확률분포

균일분포(uniform distribution)

- 균일분포: 확률변수 X가 어느 구간 (a,b)에서 정의되고, 확률밀도함수의 크기(높이)가 동일한 확률분포

Figure: U(a,b)의 확률밀도함수



예) 약속시간에 늦는 시간이 특별한 경향이 없이 0 ∼ 60분 사이에 랜덤하게 골고루 퍼져 있을 때의 늦는 시간 X

- 컴퓨터로 0과 1사이에서 난수(random number)를 생성하는 경우, 생성된 난수를 X라고 하면 그 값들은 모두 동일한 확률로 나타남

균일분포(uniform distribution)

- 균일분포의 확률밀도함수:

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \le x \le b \\ 0, & o.w \end{cases}$$

- $E(X) = \int_a^b x \cdot f(x)dx = \int_a^b x \cdot \frac{1}{b-a} dx = \frac{a+b}{2}$
- $E(X^2) = \int_a^b x^2 \cdot f(x)dx = \int_a^b x^2 \cdot \frac{1}{b-a} dx = \frac{a^2+ab+b^2}{3}$
- $V(X) = E(X^2) - E(X)^2 = \frac{(b-a)^2}{12}$
- $a \le c \le d \le b$를 만족하는 c와 d에 대하여

$$P(c \le X \le d) = \int_c^d f(x)dx = \frac{1}{b-a} \int_c^d 1dx = \frac{d-c}{b-a}$$

정규분포(normal distribution)

- 정규분포: 연속형 분포 가운데 가장 널리 쓰이는 확률분포
- 자연현상에서 비롯된 수많은 결과들은 그 값이 평균에 집중되어 있고 평균에서 멀어질 수록 도수가 작아지는 현상을 보인다.
- 정규분포곡선은 평균값을 중심으로 하여 좌우대칭의 형태를 띄고, 평균에서 좌우로 멀어질수록 수평축에 무한히 가까워지는 종 모양을 이룬다.

정규분포

- 정규분포를 결정하는 중요한 두 개의 값: 평균 μ, 표준편차 σ
- 확률변수 X가 평균이 μ, 표준편차가 σ인 정규분포를 따를 때, 이를 $X \sim N(\mu, \sigma^2)$으로 나타낸다. 이의 밀도함수는 다음과 같다.

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

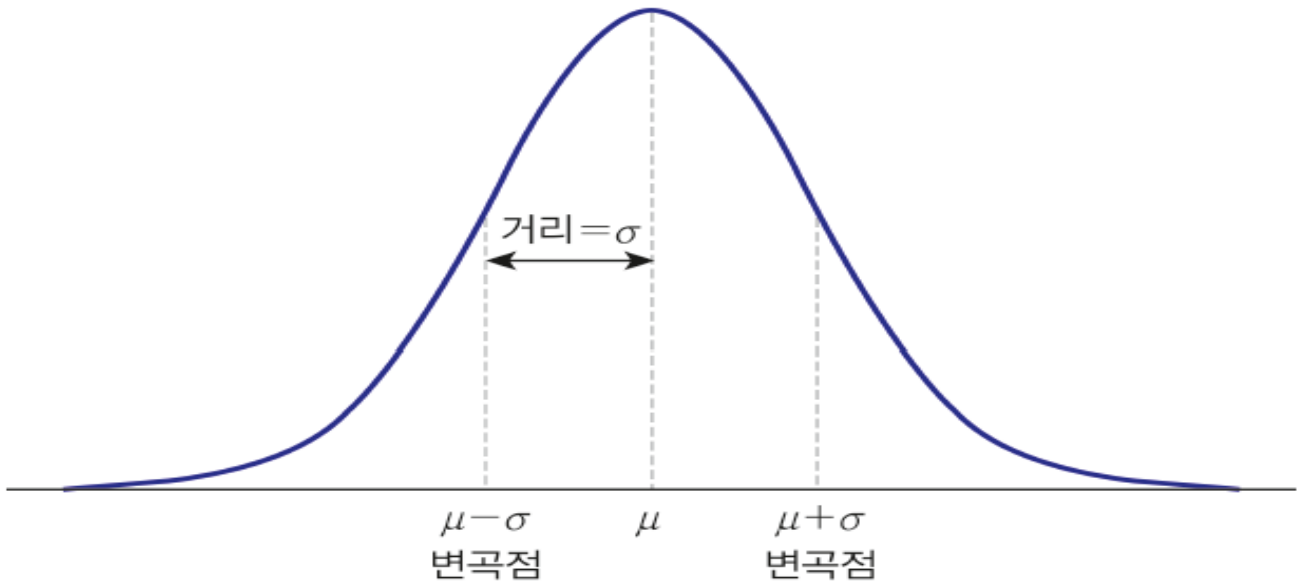


그림 4–1 정규분포의 밀도함수

정규분포의 특성

- 정규분포의 밀도곡선의 특징

  (1) 모평균 μ를 중심으로 좌우 대칭

  (2) 곡선은 대칭점 μ에서 가장 높음

  (3) 대칭점에서 멀어질 수록 곡선의 높이가 점점 낮아짐

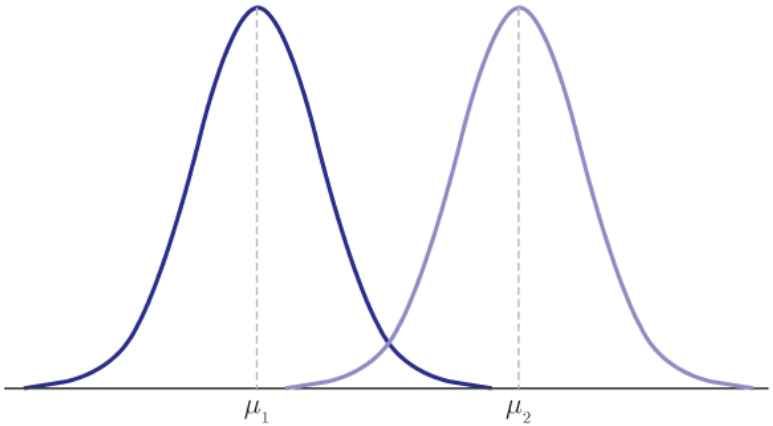
  (4) 곡선의 변곡점은 μ을 기준으로 좌우 각각 σ만큼 떨어져 있음

- $X \sim N(\mu, \sigma^2)$일 때, $P(a \leq X \leq b)$는 정규곡선의 (a,b)구간의 아래 부분의 영역의 넓이에 해당

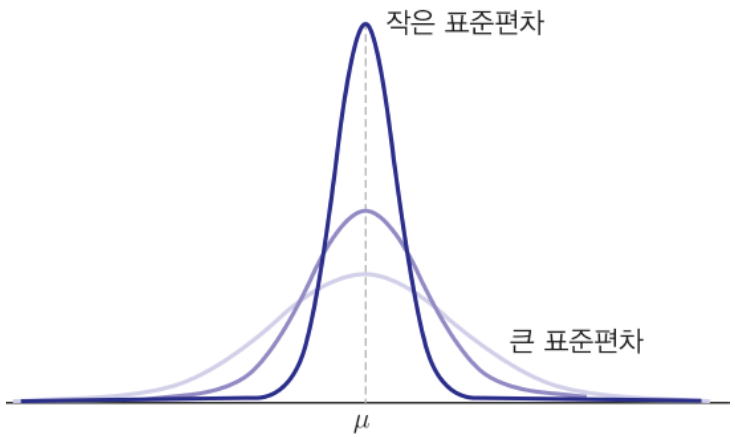
  - 정규곡선 아래 전체의 넓이는 1

정규분포의 특성

- 평균값의 변화는 곡선의 중심위치만 이동시키고, 곡선의 모양에는 영향을 미치지 않음
- σ가 작아질수록 μ 근처의 곡선의 넓이가 커지고, 반대로 σ가 커질수록 곡선은 점차 평평해짐



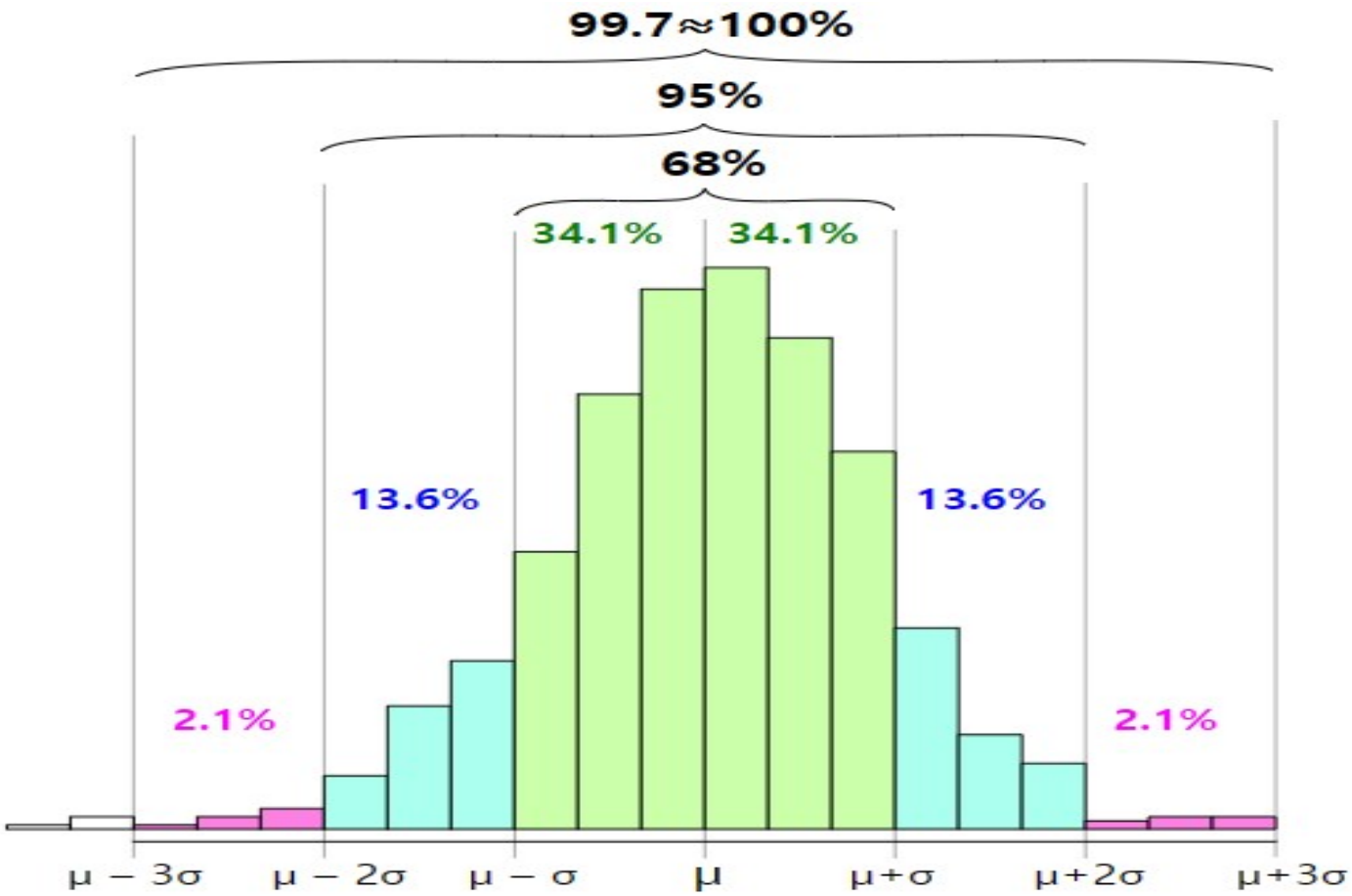
(a) 표준편차는 같고 평균이 다른 경우



(b) 평균은 같고 표준편차가 다른 경우

그림 4-2 평균과 표준편차의 변화에 따른 정규분포의 변화

변동성 척도(normal distribution case)



- $P(\mu - \sigma \leq X \leq \mu + \sigma) \approx 68.27\%$
- $P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 95.45\%$

정규분포의 성질

- 정규분포의 성질 1:
  $X \sim N(\mu, \sigma^2)$일 때 임의의 상수 a,b에 대하여 다음이 성립한다.

$$aX + b \sim N(a\mu + b, a^2\sigma^2)$$

- 정규분포의 성질 2:
  $X_1 \sim N(\mu_1, \sigma_1^2)$, $X_2 \sim N(\mu_2, \sigma_2^2)$이고 X_1과 X_2가 서로 독립이면

$$a_1X_1 + a_2X_2 \sim N(a_1\mu_1 + a_2\mu_2, a_1^2\sigma^2 + a_2^2\sigma_2^2)$$

표준정규분포(standard normal distribution)

- 정규분포 $N(\mu, \sigma^2)$를 따르는 확률변수 X를 표준화하면, 표준화된 확률변수는 평균이 0, 표준편차가 1인 정규분포를 따른다. 즉, $X \sim N(\mu, \sigma^2)$일 때, 다음이 성립

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

- 표준정규분포: 평균이 0, 표준편차가 1인 정규분포, 확률변수 Z로 표기

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

- 표준정규분포의 확률밀도함수:

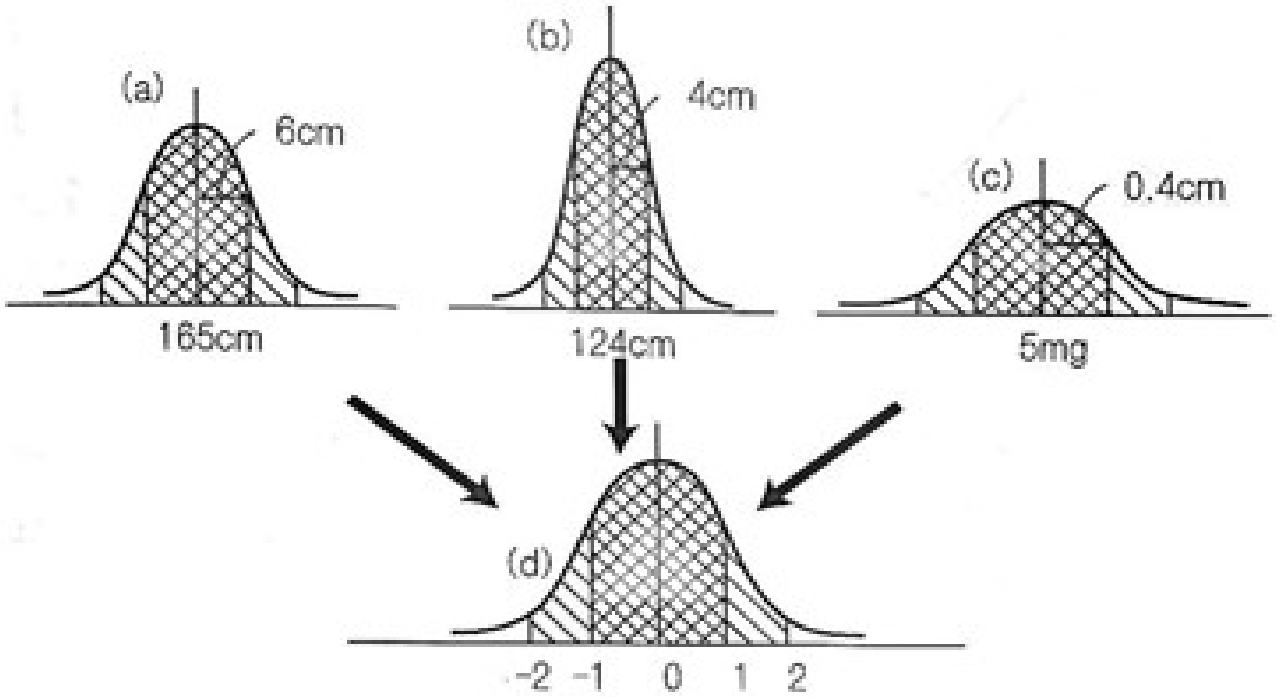
$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}, \ -\infty < z < \infty$$

정규분포의 성질

- 정규분포의 성질 3:
  $X \sim N(\mu, \sigma^2)$이면
  $$Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

- 정규분포의 성질 4:
  $Z \sim N(0, 1)$이면
  $$X = \sigma Z + \mu \sim N(\mu, \sigma^2)$$

정규분포의 표준화



(a) 성인 남자의 신장(cm), (b) 초등학교 3학년의 신장(cm),
(c) 한 알의 위장약에 포함되어있는 판토텐산칼슘의 양(mg)

정규분포의 확률계산

- 정규분포 $N(\mu, \sigma^2)$에 관한 여러가지 확률은 표준정규분포에 관한 확률로 바꾸어 계산할 수 있다.
- 정규분포의 확률 계산: 확률변수 X의 분포가 $N(\mu, \sigma^2)$일 때, $(X-\mu)/\sigma$의 분포는 $N(0,1)$이므로, 다음의 사실이 성립함

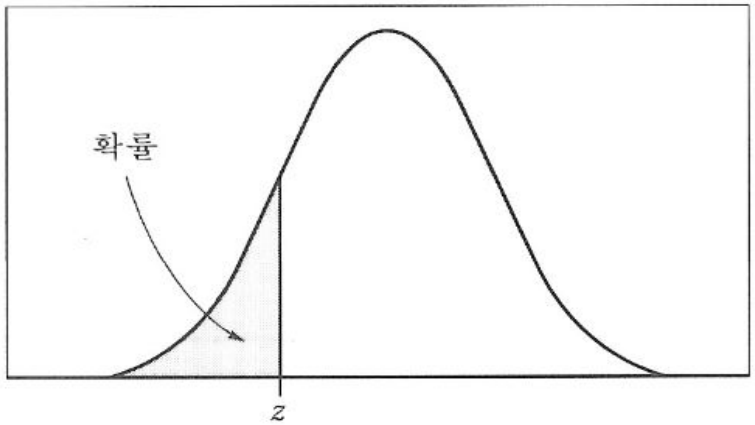
$$P(X \le b) = P\left(\frac{X-\mu}{\sigma} \le \frac{b-\mu}{\sigma}\right) = P\left(Z \le \frac{b-\mu}{\sigma}\right)$$

$$\begin{aligned} P(a \le X \le b) &= P\left(\frac{a-\mu}{\sigma} \le \frac{X-\mu}{\sigma} \le \frac{b-\mu}{\sigma}\right) \\ &= P\left(\frac{a-\mu}{\sigma} \le Z \le \frac{b-\mu}{\sigma}\right) \\ &= P\left(Z \le \frac{b-\mu}{\sigma}\right) - P\left(Z \le \frac{a-\mu}{\sigma}\right) \end{aligned}$$

표준정규분포표

- 표준정규분포의 확률: 표준정규분포표에 여러 가지 z값의 왼쪽 부분의 넓이, 즉 $P(Z \leq z)$의 값이 계산되어 있다.
- 예) $P(Z \leq -3.42) = 0.0003$

Figure: 표준정규분포표



표의 숫자는 z보다 같거나 작은 확률을 나타낸다.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014

표준정규분포의 확률계산

- 표준정규분포표를 이용한 확률의 계산

(1) $P(a \leq Z \leq b) = P(Z \leq b) - P(Z < a)$

(2) $P(Z \leq -z) = P(Z \geq z)$

(3) $P(Z \geq z) = 1 - P(Z < z)$

(4) $P(Z \leq -z) = 1 - P(Z < z)$

(5) $P(Z \leq 0) = P(Z \geq 0) = \frac{1}{2}$

(6) $P(|Z| \leq z) = 2 \times (P(Z \leq z) - 0.5)$

표준정규분포의 확률계산

Example

$P(Z < -1.9$ 또는 $Z > 2.1)$을 구하여라.

sol) $P(Z < -1.9$ 또는 $Z > 2.1) = P(Z < -1.9) + P(Z > 2.1)$

- 정규분포표에서

  $P(Z < -1.9) = 0.0287$

  $P(Z > 2.1) = 1 - P(Z \leq 2.1) = 1 - 0.9821 = 0.0179$
- 따라서 $P(Z < -1.9$ 또는 $Z > 2.1) = 0.0466$

Example

$P(Z > z) = 0.025$가 되는 z의 값을 구하여라.

sol) $P(Z \leq z) = 1 - P(Z > z) = 1 - 0.025 = 0.975$이므로

- 표준정규분포표에서 z의 값은 1.96임을 알 수 있다.

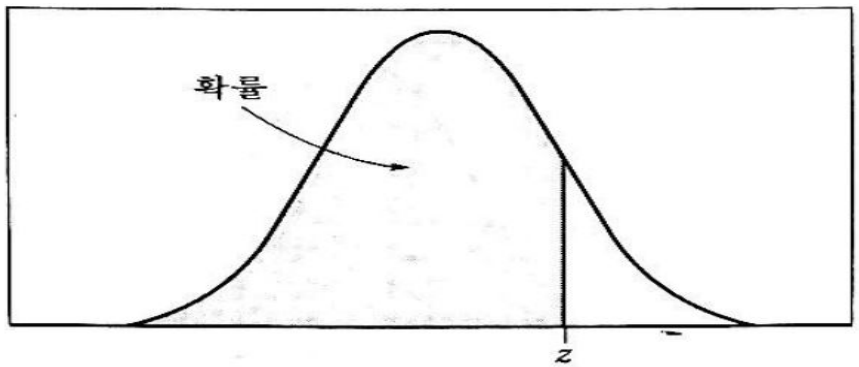
표준정규분포의 확률계산

Example

$P(Z \leq c) = 0.7$이 되는 c의 값을 구하여라.

sol) $P(Z \leq 0.52) = 0.6985 < 0.7 < 0.7019 = P(Z \leq 0.53)$

1. $c = (0.52 + 0.53)/2 = 0.525$
2. 선형보간법을 이용.



표의 숫자는 z보다 같거나 작은 확률을 나타낸다.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549

정규분포의 백분율

- $Z \sim N(0,1)$일 때, 주어진 α값에 대하여 $P(Z > z) = \alpha$를 만족하는 z의 값을 z_α로 표기하고, 이를 표준정규분포의 제 $100(1-\alpha)$백분위수 또는 $(1-\alpha)$분위수(quantile)라고 한다.
- 예) $z_{0.005} = 2.58$, $z_{0.025} = 1.96$, $z_{0.05} = 1.645$

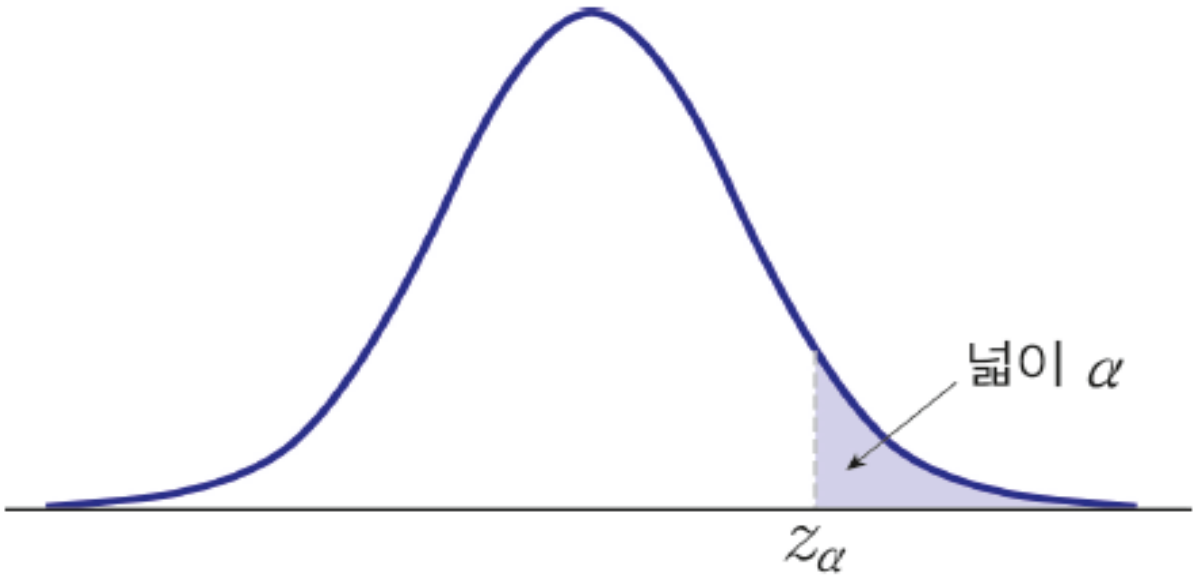


그림 4–3 z_α의 위치

정규분포의 확률계산

Example

통계학 수업을 듣는 학생들의 성적 분포가 근사적으로 $N(60, 10^2)$을 따른다고 한다. 75점 이상을 받은 학생에게 A학점을 부여할 때, A학점을 받을 학생의 비율을 근사적으로 구하라

sol) 통계학 성적을 나타내는 확률변수 $X \sim N(60, 10^2)$라 하면, 75점 이상을 받게 될 학생의 비율은 근사적으로 다음과 같다.

$$\begin{aligned} P(X \geq 75) &= P\left(\frac{X-60}{10} \geq \frac{75-60}{10}\right) \\ &= P(Z \geq 1.5) \\ &= 0.0668 \end{aligned}$$

정규분포의 확률계산

Example

매일 아침 운동으로 달리기와 자전거 타기를 한다고 하자. 달리기로 소모하는 칼로리는 평균 200kcal, 표준편차 15kcal인 정규분포를 근사적으로 따르고, 자전거 타기로 소모하는 칼로리는 평균 80kcal, 표준편차 5kcal인 정규분포를 근사적으로 따른다고 한다. 이때 하루 운동하여 300kcal 이상을 소모할 확률을 구하라.

sol) X_1: 달리기로 소모하는 칼로리

X_2: 자전거 타기로 소모하는 칼로리

$S = X_1 + X_2$: 운동으로 소모하는 총 칼로리

$$X_1 \sim N(200, 15^2), X_2 \sim N(80, 5^2)$$

- 이때, 운동으로 소모하는 총 칼로리 S도 정규분포를 따르고 이때의 평균과 분산은 각각 다음과 같다.

$$E(S) = E(X_1) + E(X_2) = 280$$

$$Var(S) = Var(X_1) + Var(X_2) = 15^2 + 5^2 = 250$$

- 따라서 S는 다음과 같은 정규분포를 따른다.

$$S \sim N(280, 250)$$

- 하루 운동으로 300kcal 이상을 소모할 확률은

$$\begin{aligned} P(S > 300) &= P\left(\frac{S - 280}{\sqrt{250}} > \frac{300 - 280}{\sqrt{250}}\right) \\ &= P(Z > 1.26) = 1 - P(Z \leq 1.26) \\ &= 1 - 0.8962 = 0.1038 \end{aligned}$$

이상치 판단 기준 (Outlier Detection)

1. IQR (Interquartile Range) 기준
   - Q1 = 제1사분위수, Q3 = 제3사분위수
   - IQR = Q3 - Q1
   - 이상치: $x < Q1 - 1.5 \times IQR$ 또는 $x > Q3 + 1.5 \times IQR$

2. Z-score 기준 (정규분포 가정)
   - $z = \frac{x-\mu}{\sigma}$ (μ: 평균, σ: 표준편차)
   - 일반적으로 $|z| > 2$ 또는 $|z| > 3$이면 이상치

3. 시각적 방법
   - Boxplot: 사분위 기반 이상치 확인
   - Scatter plot: 변수 관계에서 튀는 점 확인