[ProDS] 통계 이론 및 데이터 시각화

확률: 확률의 개념과 특징

Key words

#확률개념
#확률정의 #확률규칙 #여사건 #곱사건 #합사건
#조건부확률 #독립사건

확률의 기본 개념

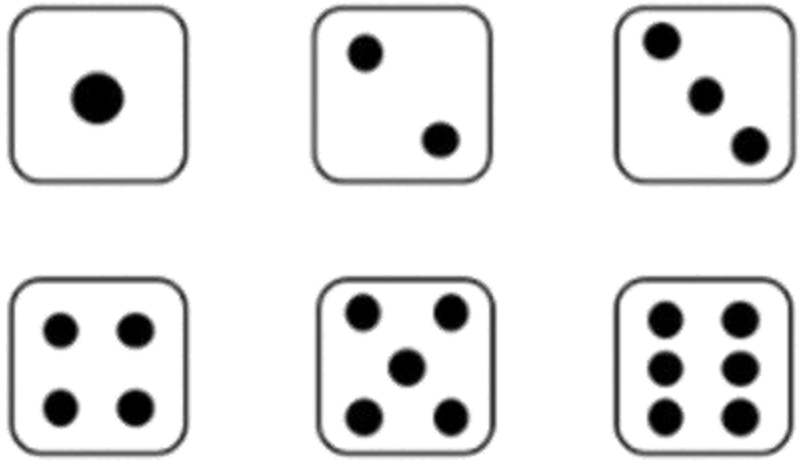
확률모형 (probability model)

- 시행을 반복할 때마다 나오는 결과가 우연에 의존하여 매번 달라지는 현상 또는 실험(확률실험, random experiment)에 대한 수리적 모형

확률의 기본 개념

표본공간 (sample space)

- 확률 실험에서의 모든 관찰 가능한 결과의 집합, S로 표기



확률의 기본 개념

사건(event)

- 표본공간의 임의의 부분집합, A, B 등으로 표기

확률의 정의 및 성질

고전적 접근 (P. Laplace)

- n개의 실험결과로 구성된 표본공간에서 각 실험결과가 일어날 가능성이 같은 경우,
- m $(m \leq n)$ 개의 실험 결과로 구성된 사건 A 의 확률을 아래와 같이 정의함.

$$P(A) = \frac{M}{N}$$

확률의 정의 및 성질

상대적 비율에 의한 접근 (Richard Von Mises)

- n 번의 반복된 실험 중 어떤 사건 A가 발생한 횟수를 m이라고 할 때, 사건 A의 상대빈도는 $\frac{m}{n}$ 으로 구해짐.
- 이 실험의 반복 횟수 n을 무한히 증가했을 때, 사건 A 의 상대빈도가 수렴하는 값을 사건 A 의 확률로 정의하고자 함.

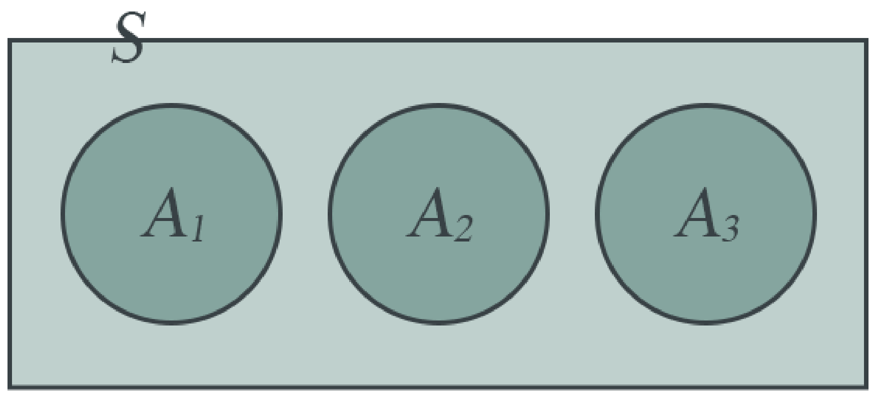
확률의 정의 및 성질

확률의 공리적 정의 (A. N. Kolmogorov)

1) 임의의 사건 A에 대하여 $P(A) \geq 0$

2) $P(S)$=1

3) 표본공간 S 에 정의된 서로 상호배반인 사건 $A_1, A_2, \ldots$ 에 대하여 $P(A_1 \cup A_2 \cup \cdots) = P(A_1) + P(A_2) + \cdots$ 가 성립



확률의 정의 및 성질

공리적 접근방식

- 표본공간을 정의역으로 하며, 위 세가지 공리를 만족하는 함수를 확률로 정의.

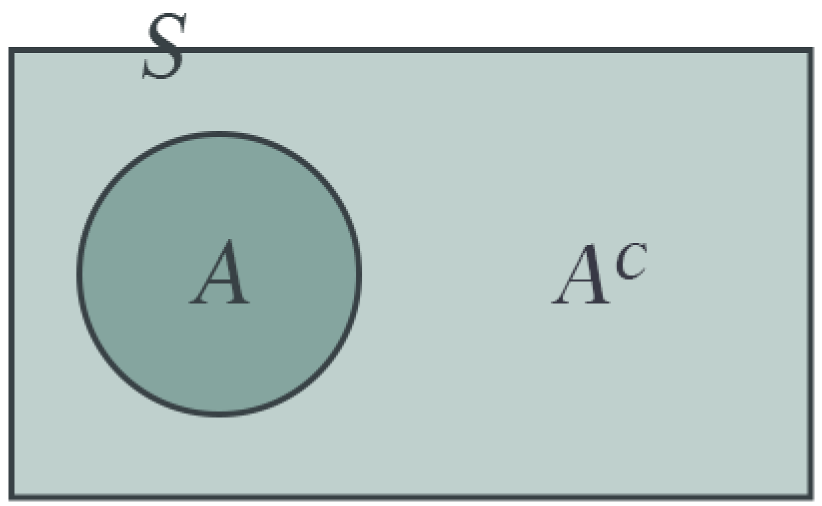
확률의 정의 및 성질

확률의 규칙

- 여사건의 확률

-$P[A^C]$: 사건 A 를 제외한 나머지 사건의 확률

-$P[A^C] = 1 - P[A]$

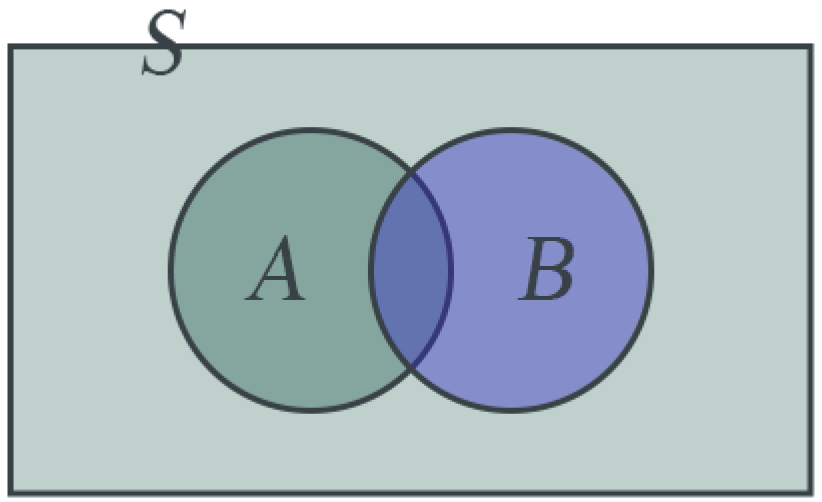


확률의 정의 및 성질

확률의 규칙

- 곱사건의 확률

  - $P[A \cap B]$: 사건 A 와 사건 B가 동시에 발생할 확률

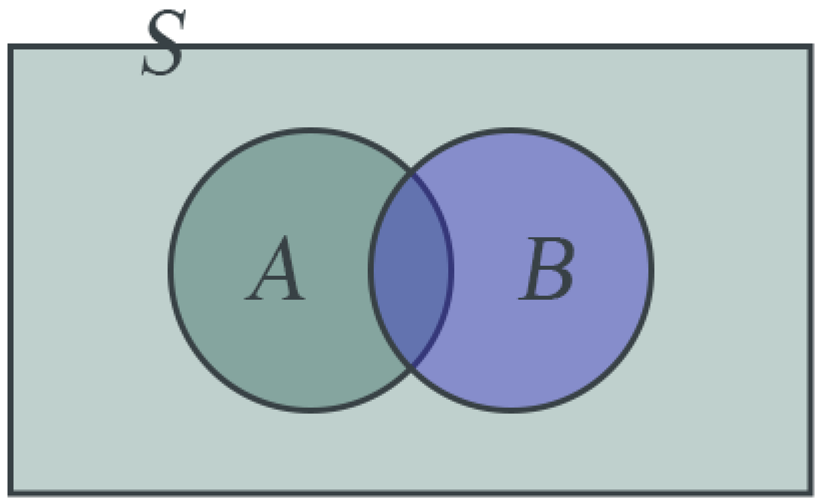


확률의 정의 및 성질

확률의 규칙

- 합사건의 확률

-$P[A \cup B]$: 사건 A 또는 사건 B가 발생할 확률
-$P[A \cup B] = P[A] + P[B] - P[A \cap B]$

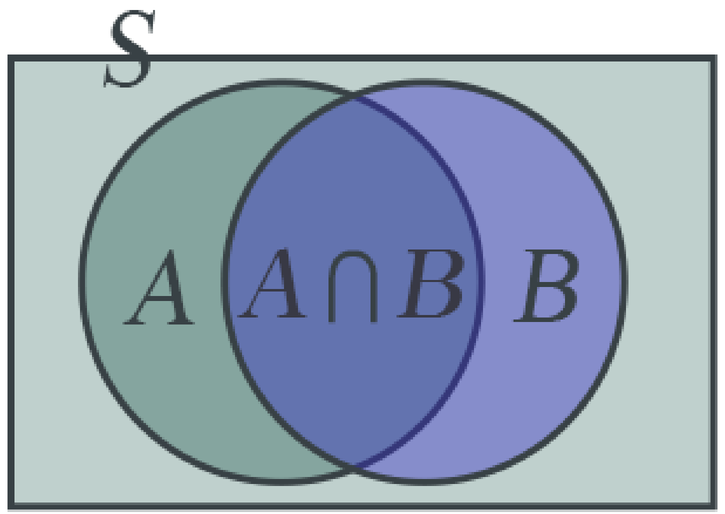


조건부 확률과 독립

조건부 확률의 정의

- 사건 A 와 B 가 표본공간 S 상에 정의되어 있으며 $P(B) > 0$ 라고 가정. 이 때 B가 일어났다는 가정 하의 사건 A 가 일어날 조건부확률은 다음과 같이 정의됨.

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$



조건부 확률과 독립

독립 사건의 정의

- 두 사건 A와 B가 다음 중 하나를 만족시키면 서로 독립이라고 함.
(단, $P(A) > 0, P(B) > 0$)

1) $P(A|B) = P(A)$

2) $P(A \cap B) = P(A) \cdot P(B)$

3) $P(B|A) = P(B)$

확률: 베이즈 정리

Key words

#분할 #전확률공식
#베이즈정리

표본공간의 분할과 전확률 공식

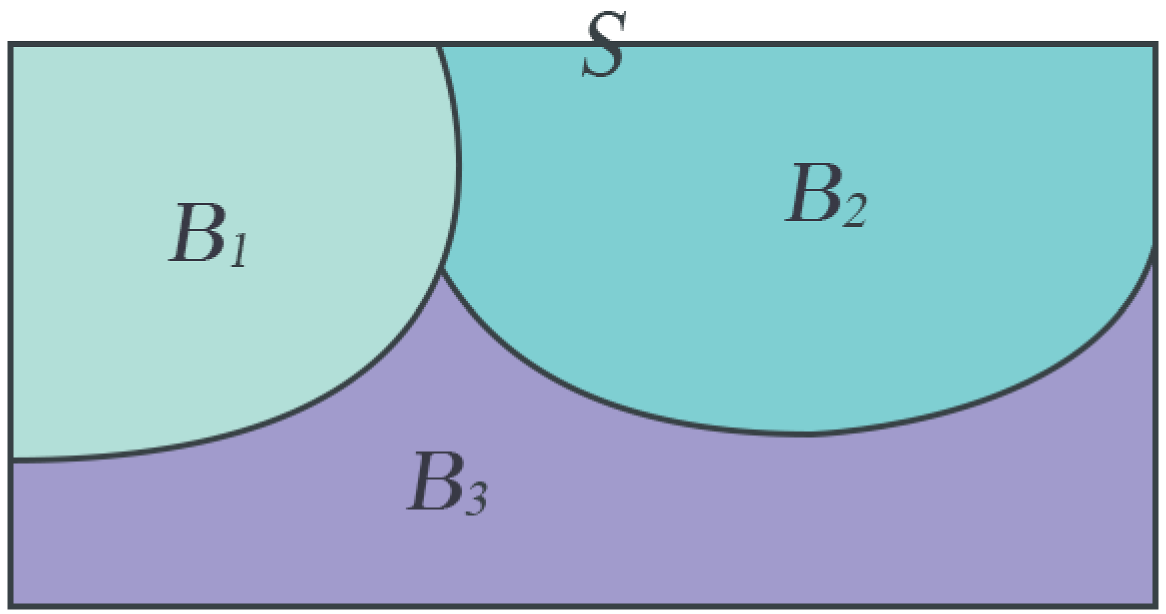
표본공간의 분할

- $B_1, \ldots, B_k$가 다음 조건을 만족하면 표본 공간 S의 분할이라고 함.

  - 서로 다른 i, j에 대해 $B_i \cap B_j = \emptyset$: 상호배반

  - $B_1 \cup B_2 \cup \cdots \cup B_k = S$

- $k = 3$인 경우.



표본공간의 분할과 전확률 공식

전확률공식

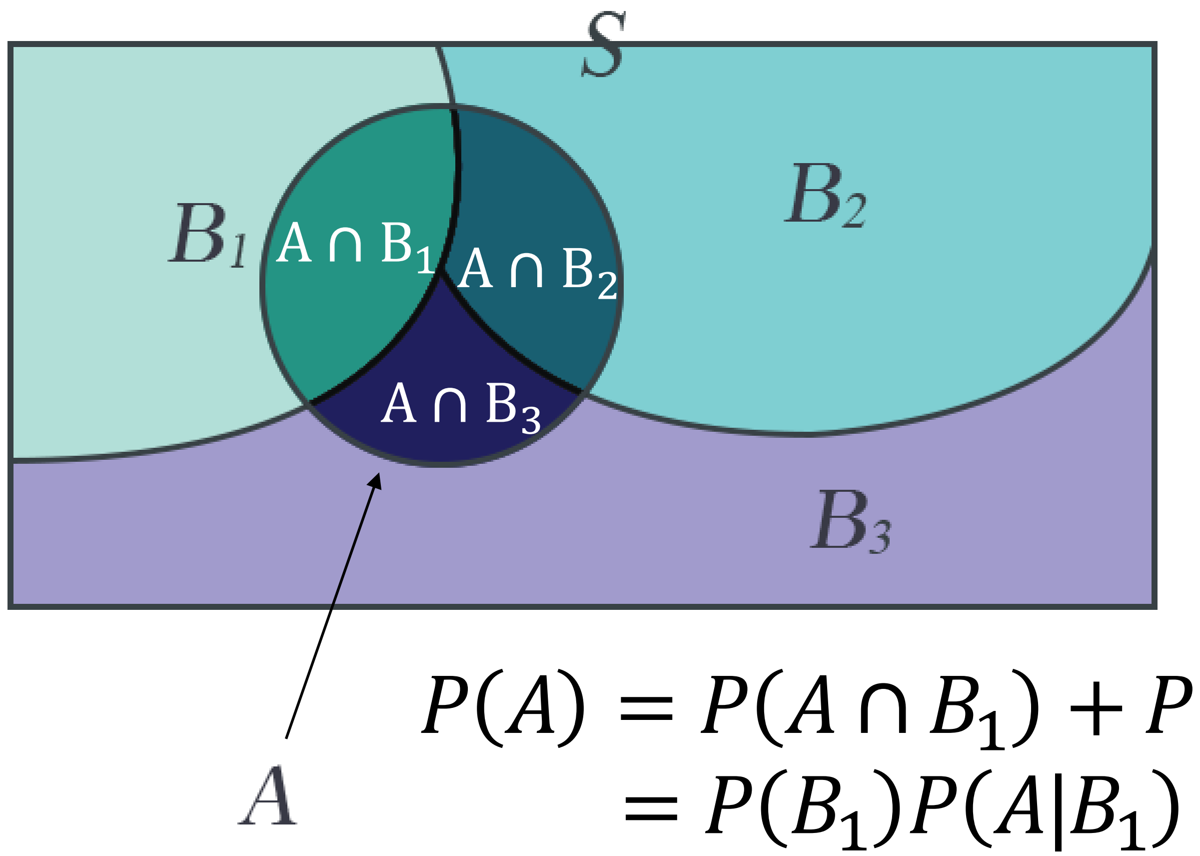
- 사건 $B_1, B_2, \ldots, B_k$는 상호배반이며, $B_1 \cup \cdots \cup B_k = S$라고 함.
- 이 때 S에서 정의되는 임의의 사건 A에 대하여 다음이 성립.

$$
\begin{aligned}
P(A) &= P(A \cap B_1) + \cdots + P(A \cap B_1) \\
&= P(B_1)P(A|B_1) + \cdots + P(B_k)P(A|B_k)
\end{aligned}
$$

표본공간의 분할과 전확률 공식

전확률공식

- $k = 3$인 경우.



$$P(A) = P(A \cap B_1) + P(A \cap B_2) + P(A \cap B_3)$$
$$= P(B_1)P(A|B_1) + P(B_2)P(A|B_2) + P(B_3)P(A|B_3)$$

베이즈 정리

베이즈 정리

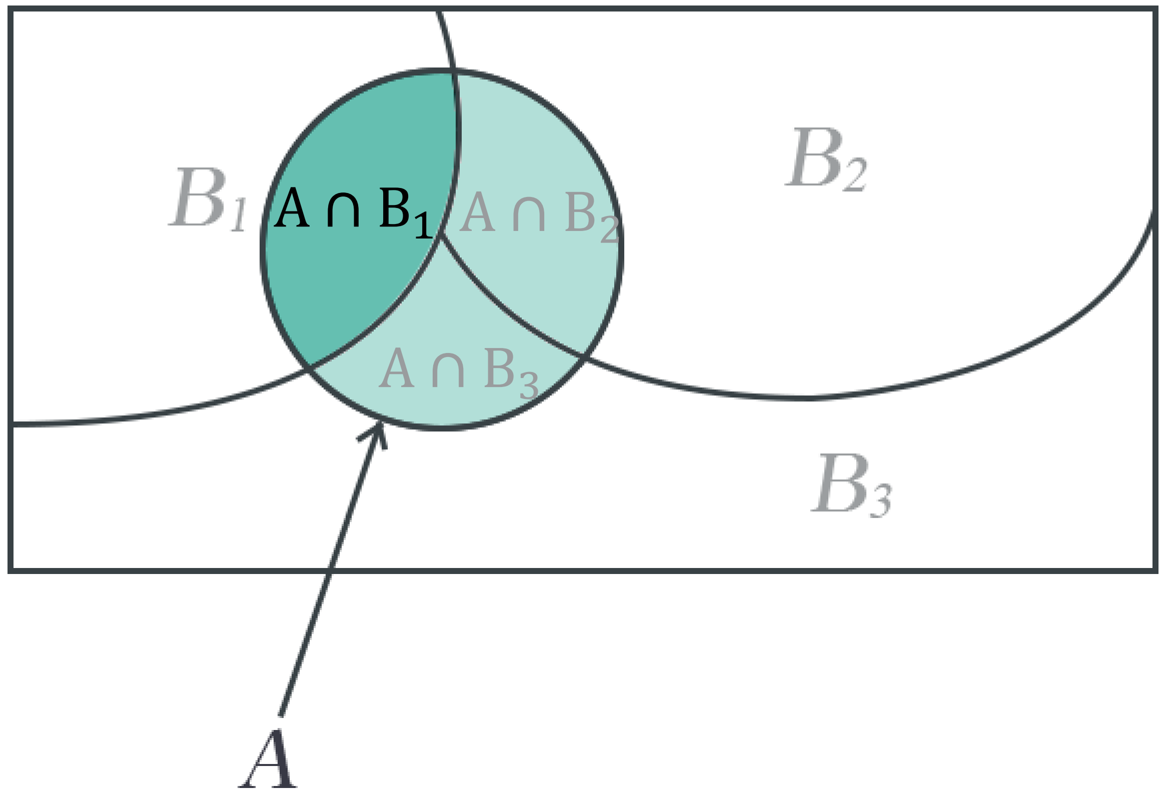
- 사건 $B_1, B_2, \ldots, B_k$는 상호배반이며, $B_1 \cup \cdots \cup B_k = S$라고 함.
- 이 때 사건 A가 일어났다는 조건 하에서 사건 B_i가 일어날 확률은 다음과 같음.

$$P(B_i|A) = \frac{P(A \cap B_i)}{P(A)} = \frac{P(B_i)P(A|B_i)}{P(B_1)P(A|B_1) + \cdots + P(B_k)P(A|B_k)}$$

베이즈 정리

베이즈 정리

- $k = 3$인 경우.

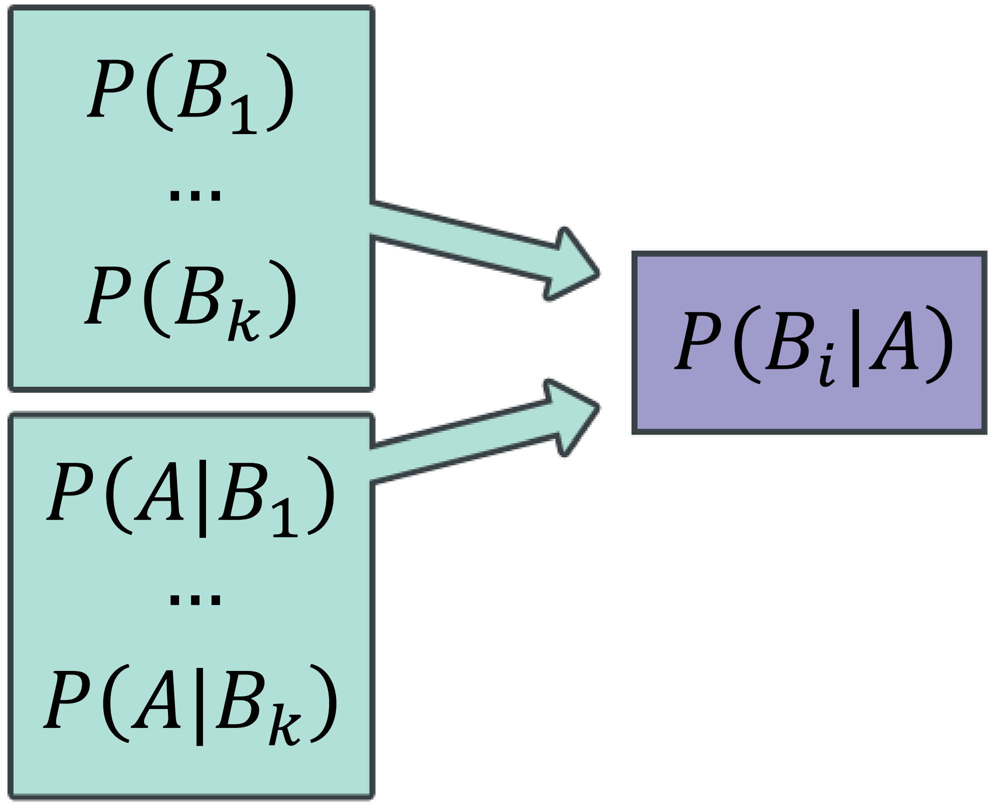


$$P(B_1|A) = \frac{P(A \cap B_1)}{P(A)}$$

베이즈 정리

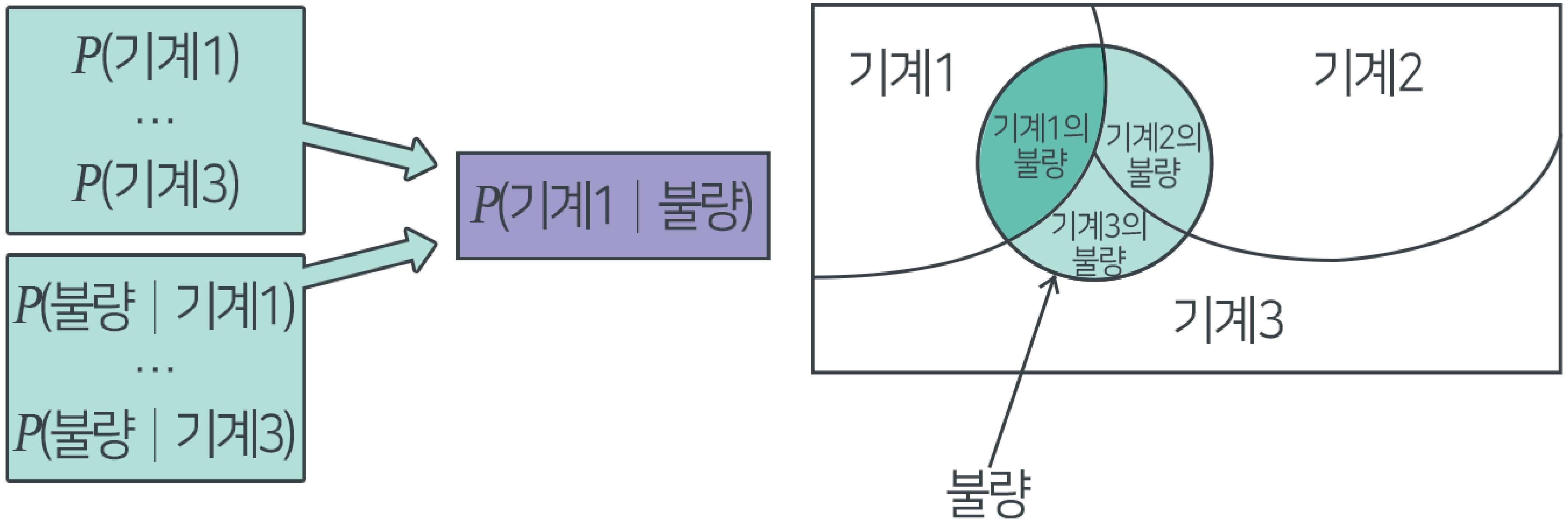
베이즈 정리 활용

- $B_1, B_2, \ldots, B_k$ 으로 분할된 사건의 각 확률을 알고,
  각 B_i를 전제로 했을 때의 사건 A가 발생할 조건부 확률을 알 때,
  사건 A를 전제로 한 각 B_i의 조건부 확률을 구하기 위한 정리.



베이즈 정리

베이즈 정리 활용



확률과 확률분포:
확률변수와 확률분포, 분포의 특성치

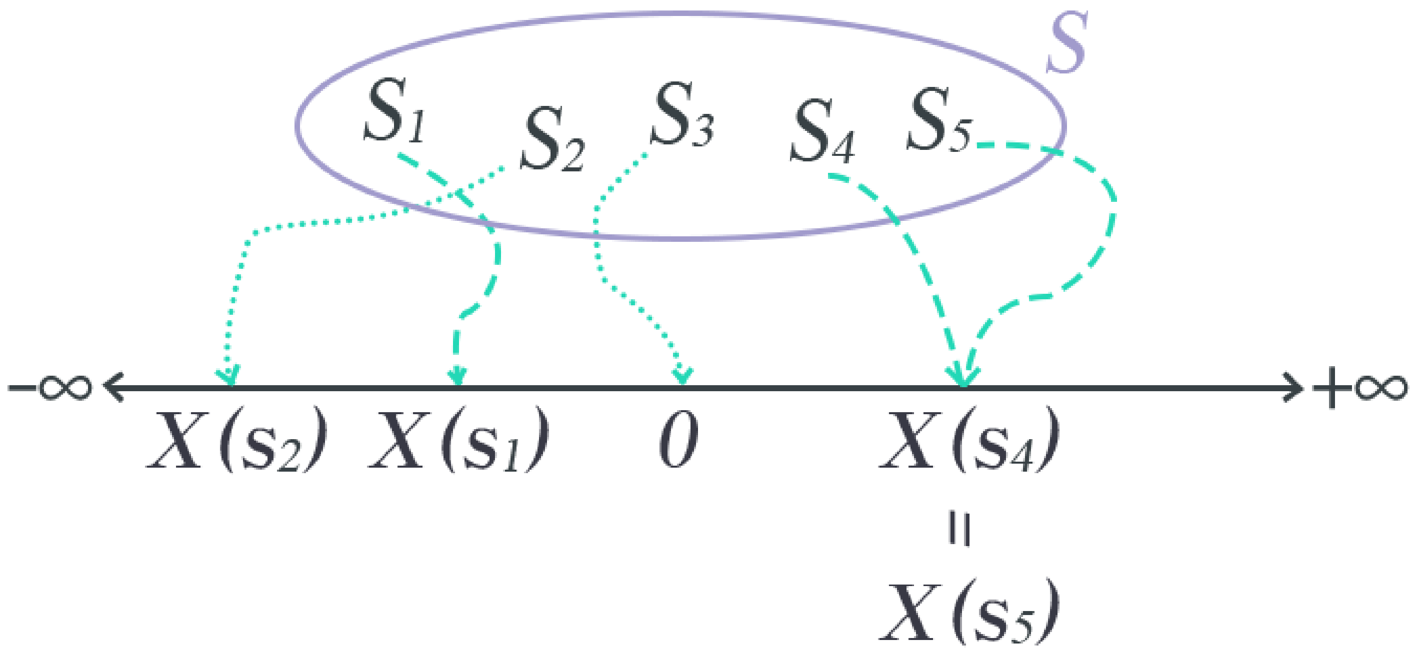
Key words

#확률변수 #확률질량함수
#확률밀도함수 #누적분포함수
#기대값 #분산 #표준편차

확률변수와 확률분포

확률변수

- 확률변수 : 표본공간에서 정의된 실수값 함수.



확률변수와 확률분포

확률변수

- 확률변수 : 표본공간에서 정의된 실수값 함수.
  - 이산형 확률변수 : 확률변수가 취할 수 있는 값이 셀 수 있는 경우.
  - 연속형 확률변수 : 주어진 구간에서 모든 실수 값을 취할 수 있어 셀 수 없는 경우.

확률변수와 확률분포

확률분포

- 확률질량함수

  - 확률변수 X가 이산형인 경우 X가 취할 수 있는 값 $x_1, x_2, \dots x_n$의 각각에 대하여 확률 $P[X = x_1], P[X = x_2], \cdots, P[X = x_n]$을 대응시켜 주는 관계를 X의 확률질량함수라고 하며 $f(x)$로 표기.

$$f(x_i) = P[X = x_i], \qquad i = 1,2, \dots, n$$

- 확률질량함수의 성질

  1) 모든 $i = 1,2, \dots, n$에 대해 $0 \le f(x_i) \le 1$
  2) $\sum_{i=1}^{n} f(x_i) = 1$

확률변수와 확률분포

확률분포

- 확률밀도함수

  - 확률변수 X가 연속형인 경우 X가 가질 수 있는 구간 $(-\infty, \infty)$위에서의 함수 $f(x)$ 가 다음을 만족할 때, 이를 X의 확률밀도함수라고 함.

$$\int_a^b f(x)\,dx = P[a \le X \le b] \ \ (\text{단}, -\infty < a < b < \infty)$$

- 확률밀도함수의 성질

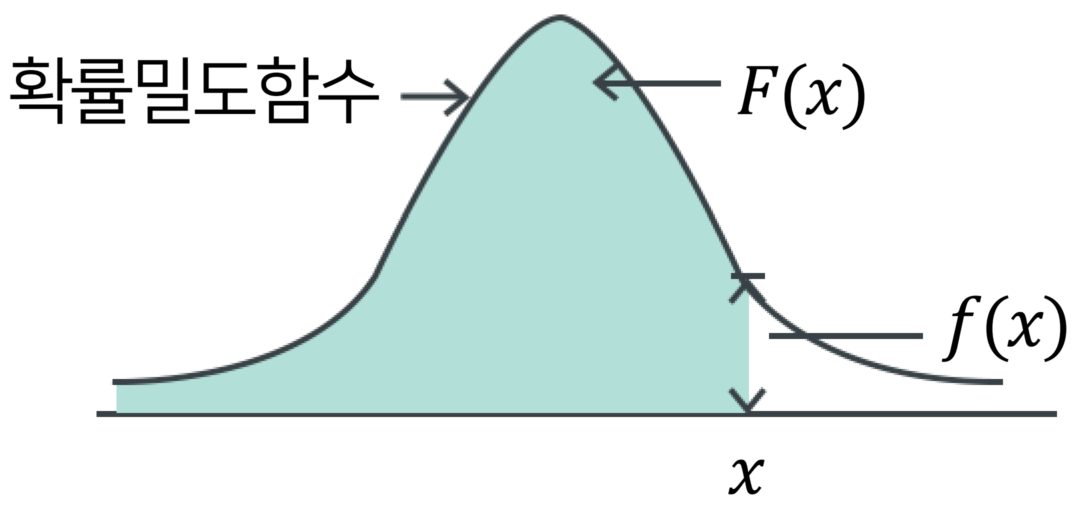
  1) 모든 a, b 에 대해 $0 \le \int_a^b f(x)\,dx \le 1$

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

확률변수와 확률분포

누적분포함수

-X 의 확률밀도함수가 $f(x)$일 때, X 의 누적분포함수 $F(x)$ 는
$X \le x$ 인 모든 X 에 대한 $f(x)$의 적분 값이 됨.



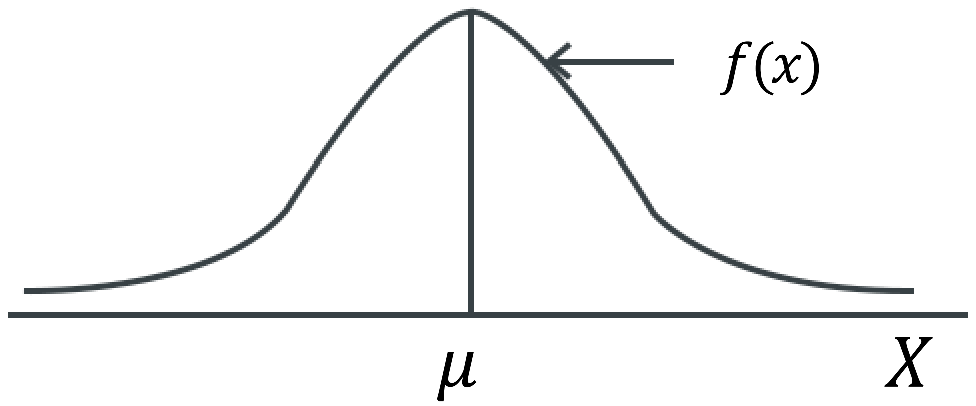
- $F(-\infty) = 0,\ F(\infty) = 1$
-x가 증가할 때 $F(x)$도 증가하며, $F(x)$는 음의 값을 가질 수 없음.

확률변수와 확률분포

확률분포의 특성치

- 기대값 : 분포의 무게중심, 중심 위치를 나타냄.

$$E[X] = \mu = \begin{cases} \sum_{all\ x} xf(x), & X\text{는 이산형} \\ \int_{-\infty}^{\infty} xf(x)\,dx, & X\text{는 연속형} \end{cases}$$



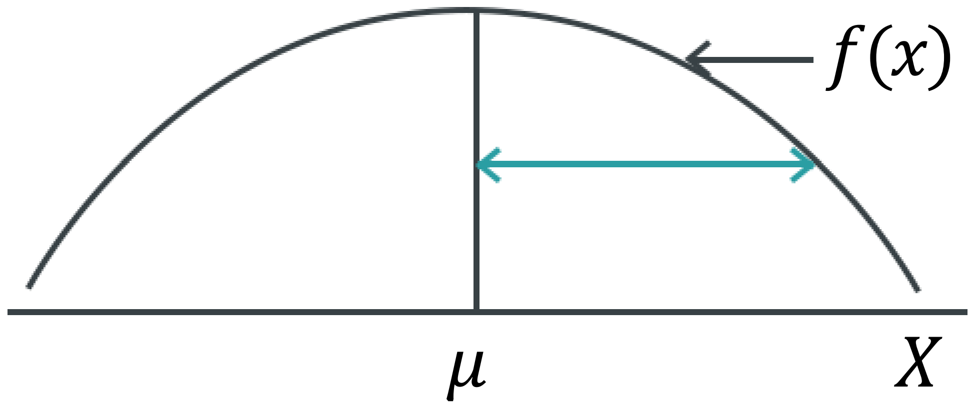
확률변수와 확률분포

확률분포의 특성치

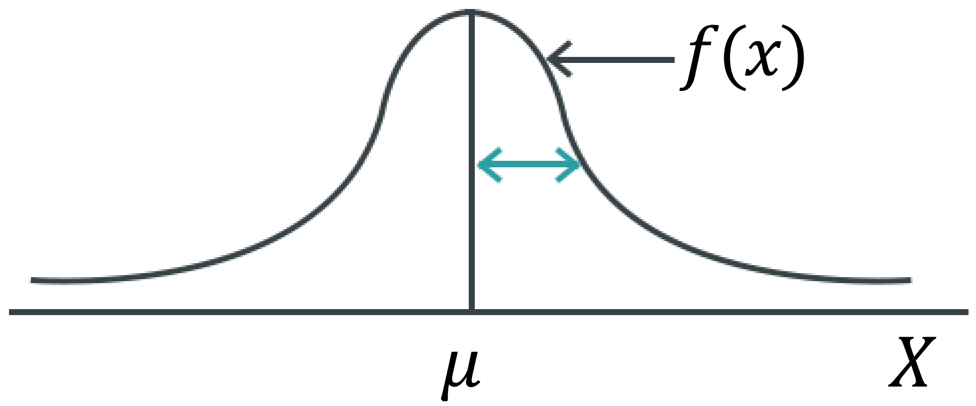
- 분산 : 분포의 산포를 나타냄.

$$V[X] = \sigma^2 = E[(X-\mu)^2]$$

1) σ^2 이 큰 경우



2) σ^2 이 작은 경우



확률변수와 확률분포

확률분포의 특성치

- 표준편차 :분산의 제곱근. 단위가 보정됨.

$$S[X] = \sigma = \sqrt{V[X]}$$

주요 확률분포:
이항분포, 포아송분포, 지수분포, 감마분포

Key words

#베르누이시행 #이항분포
#포아송분포 #지수분포 #감마분포

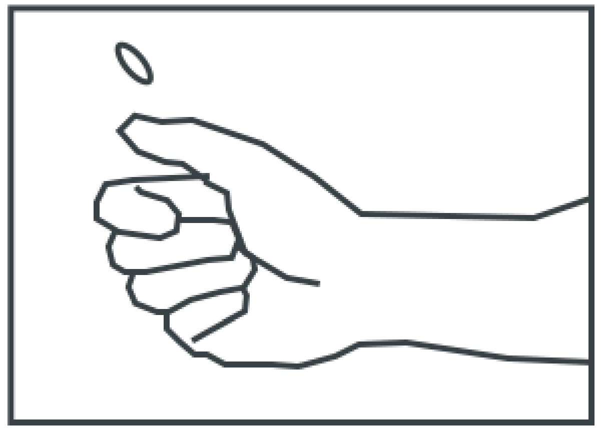
이항분포

베르누이 시행

- 매 시행 마다,

  - '성공' 또는 '실패'의 오직 두 가지 가능한 결과만 가짐
  - '성공'의 확률이 p로 일정함.

  의 조건을 만족하는 실험.



이항분포

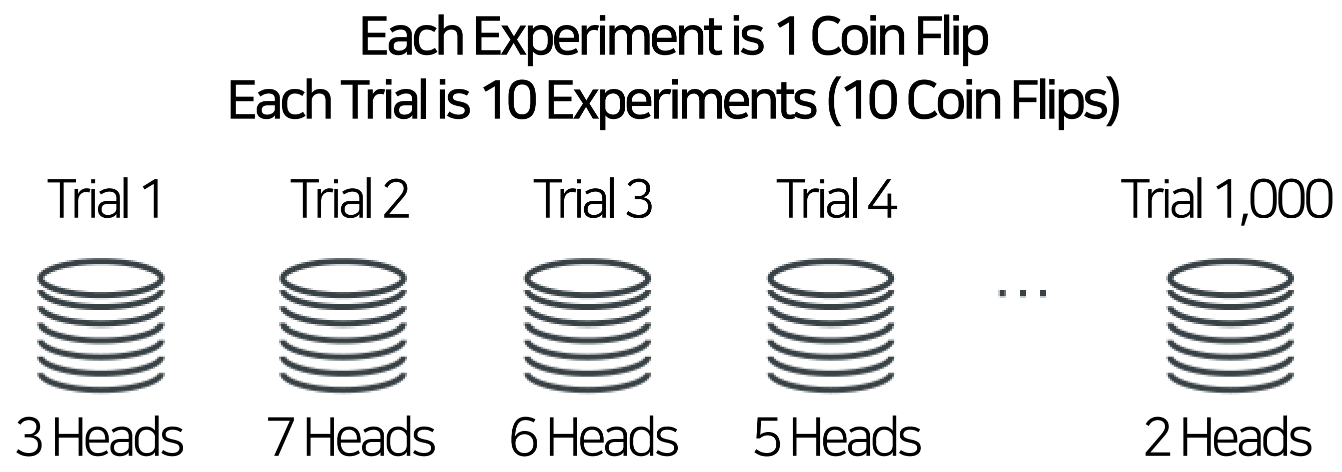
이항 확률변수가 고려되는 실험

- 매 시행 마다,

  - ‘성공’ 또는 ‘실패’의 오직 두 가지 가능한 결과만 가짐
  - ‘성공’의 확률이 p로 일정함.

  의 조건을 만족하는 베르누이 시행을

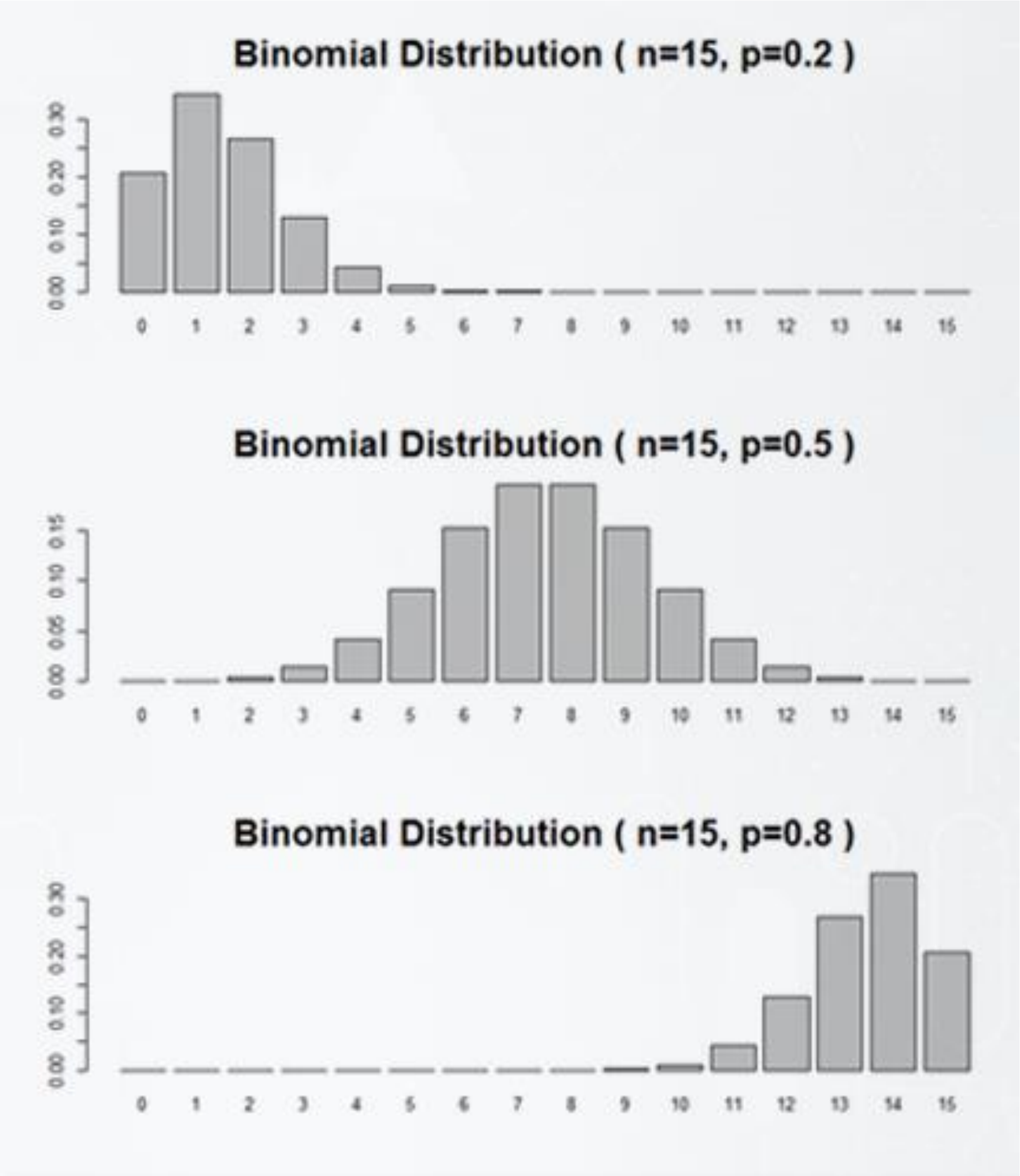
  - 독립적으로
  - n번 반복하는 실험



이항분포

이항확률변수와 확률질량함수

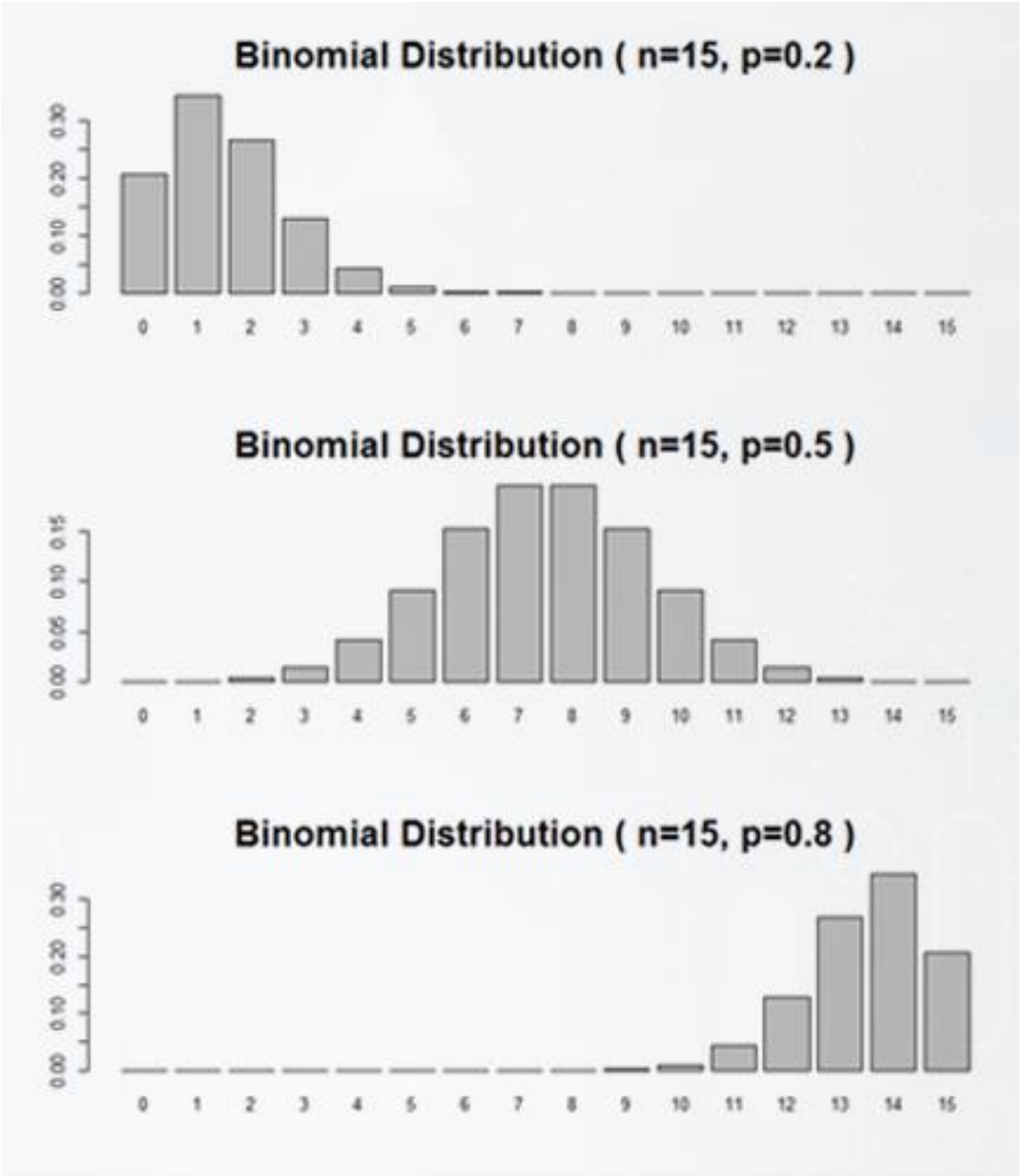
- X : n번 시행 중 '성공'의 횟수로 정의
- $x = 0,1, \ldots, n$의 값을 가짐
- $f(x) = P(X = x) = \binom{n}{x} \; p^x(1-p)^{n-x}$
- 이 경우 $X \sim Bin[n, p]$라고 함.



이항분포

이항분포의 특성치

- $X \sim Bin[n, p]$인 경우,
  - $E[X] = np$
  - $V[X] = np(1-p)$



포아송 분포

포아송 확률변수와 확률질량함수

- 단위시간에(t = 1), 포아송 확률과정을 따르는 사건 A 가 발생하는 횟수를 X 로 정의하면,

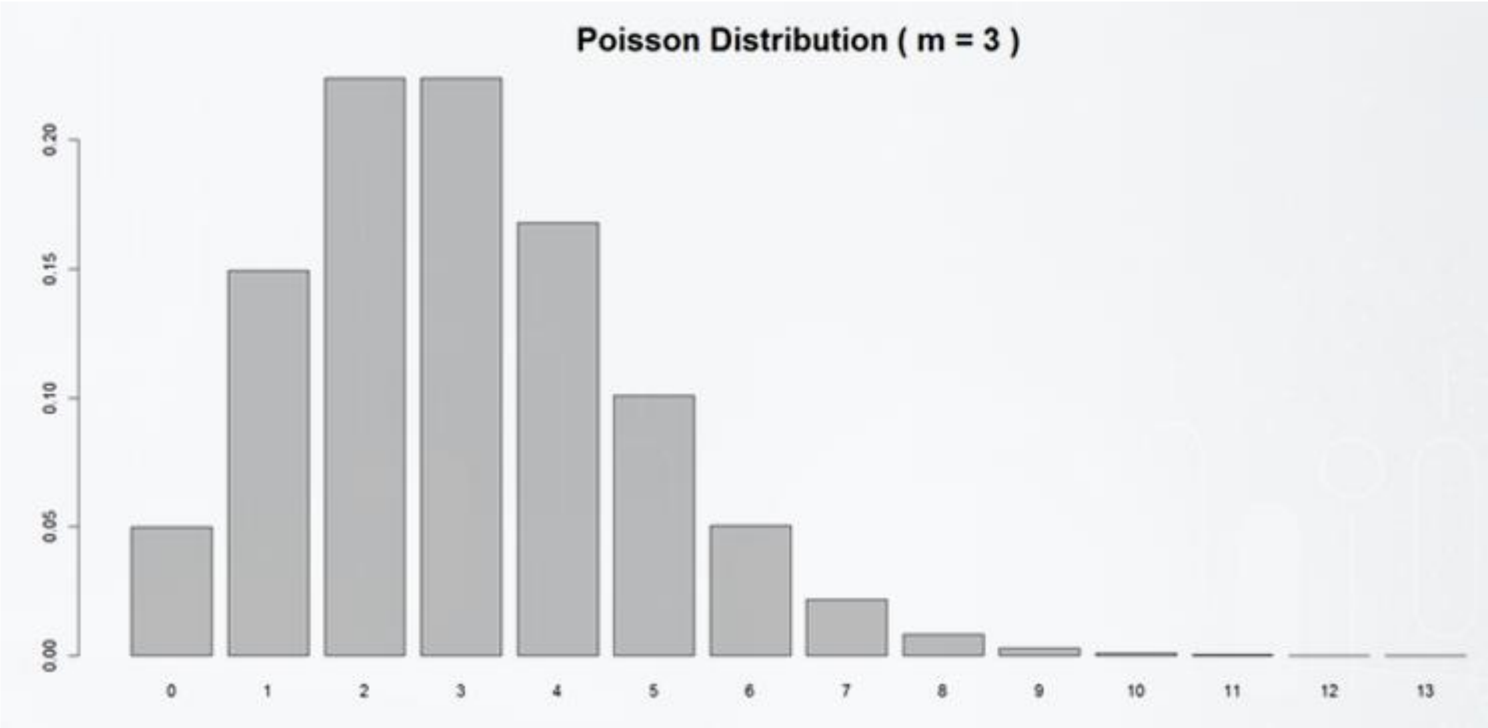
$$f(x) = P(X = x) = \frac{\exp(-m)\, m^x}{x!}, \quad x = 0,1,2,\ldots$$

- 이 경우 $X \sim POI[m]$라고 함.

포아송 분포

포아송 분포의 특성치

- $X \sim POI[m]$인 경우,
  - $E[X] = V[X] = m$



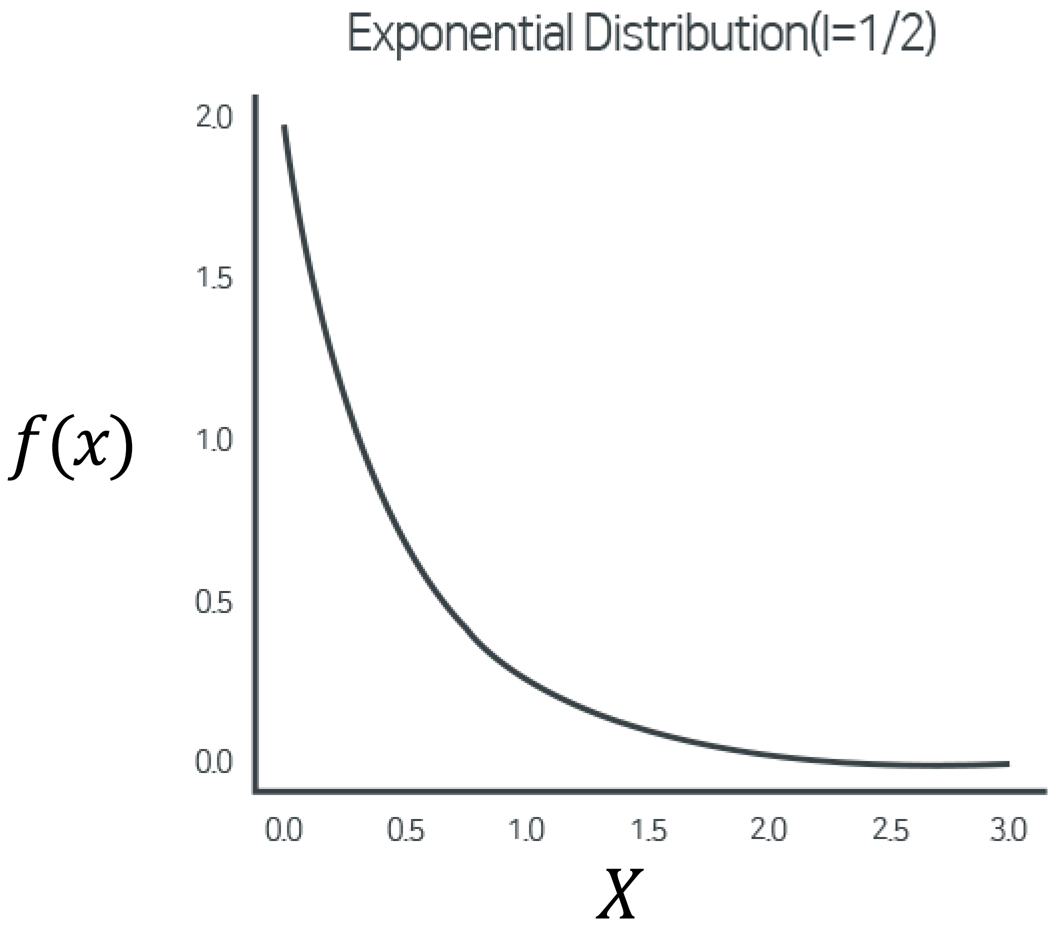
지수분포

지수확률변수와 확률밀도함수

- 단위구간에서 평균발행횟수가 m인 포아송 확률과정을 따르는 사건 A가 한번 일어난 뒤 그 다음 또 일어날 때까지 걸리는 시간 X로 정의됨.

$$f(x) = \frac{1}{\lambda}\exp\left(-\frac{x}{\lambda}\right), \qquad x > 0\text{임.} \quad \left(\text{단}, m = \frac{1}{\lambda}\right)$$

- 이 경우 $X \sim EXP[\lambda]$라고 함.

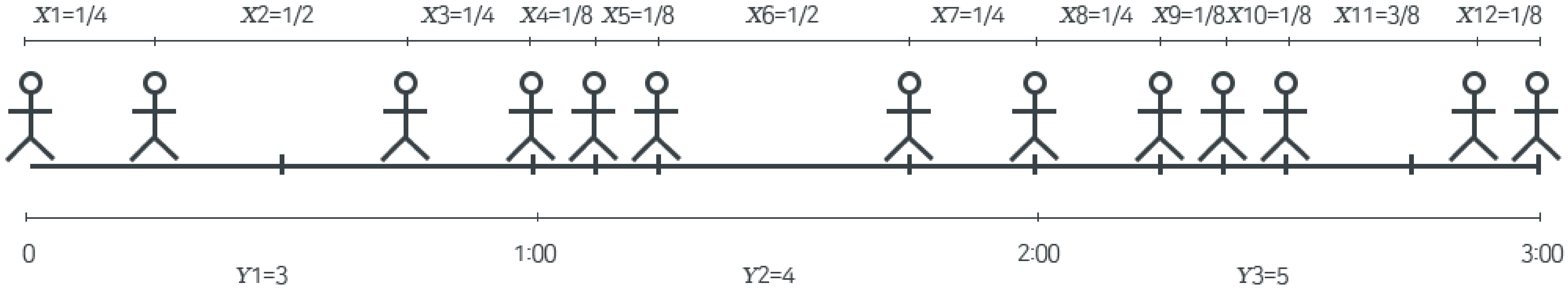


지수분포

지수 분포의 특성치

- 포아송 모수와 지수 모수는 역의 관계
  - 단위구간 내 평균발생횟수가 m인 포아송 과정을 따르는 사건은 사건 사이 소요시간의 평균이 $\lambda = \frac{1}{m}$ 임.

$X \sim EXP[1/4]$



$Y \sim Poisson[4]$

지수분포

지수 분포의 특성치

- $X \sim EXP[\lambda]$인 경우

  - $E[X] = \lambda$

감마분포

감마확률변수와 확률밀도함수

- 감마확률변수 X의 확률밀도함수는 양수인 θ와 k에 대하여 다음과 같이 정의됨.

$$f(x) = \frac{1}{\theta^k \Gamma(k)} x^{k-1} \exp(-x/\theta), \qquad x > 0$$

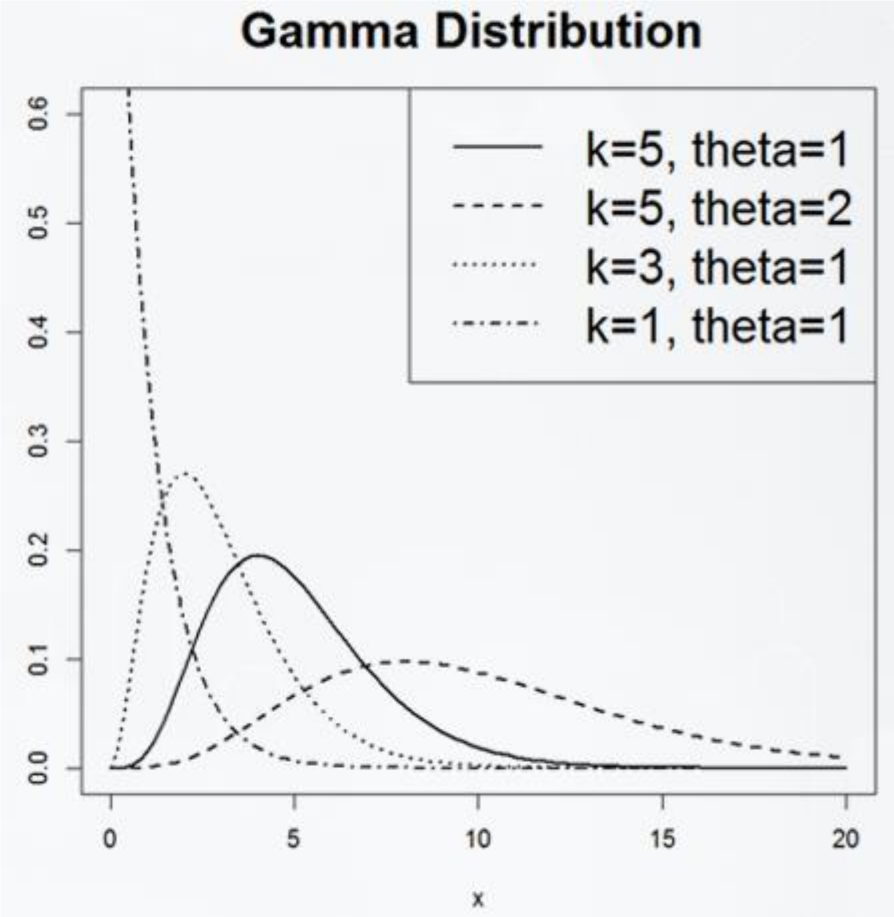
- 이 경우 $X \sim GAMMA[k, \theta]$라고 함.

감마분포

감마분포의 특성치

- $X \sim GAMMA[k, \theta]$ 인 경우에,

-$E[X] = k\theta$
-$V[X] = k\theta^2$



주요 확률분포:
정규분포, 표준정규분포

Key words

#정규분포 #표준정규분포 #표준화

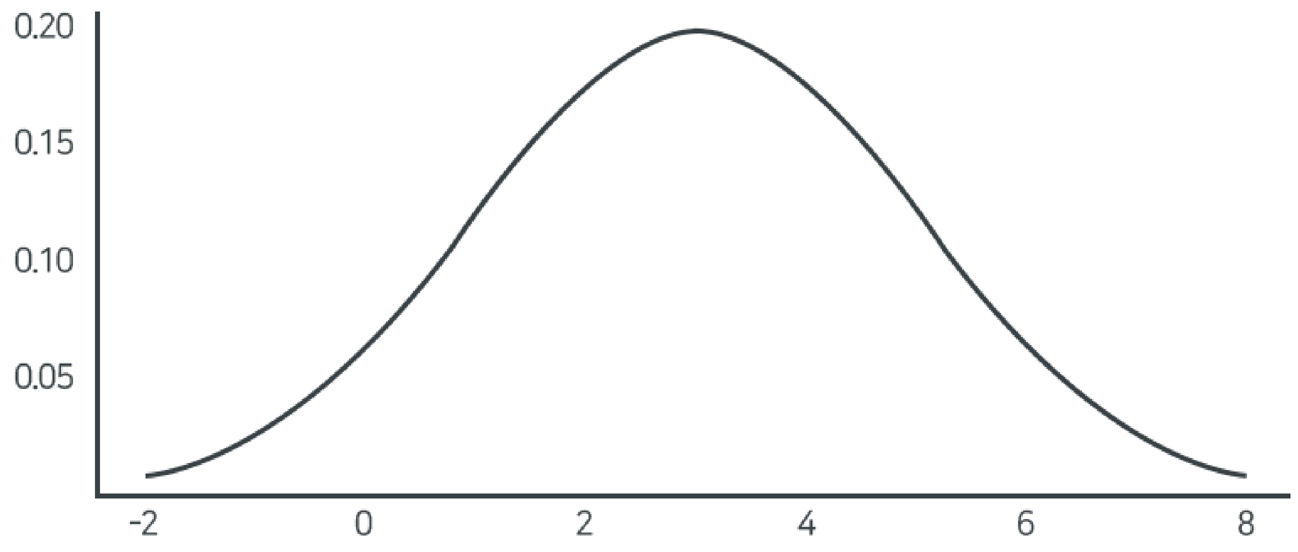
정규분포

정의

- 확률변수 X가 평균이 μ, 분산이 σ^2이고,
  다음의 확률밀도함수를 가질 때, X는 정규분포를 따른다고 함.

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

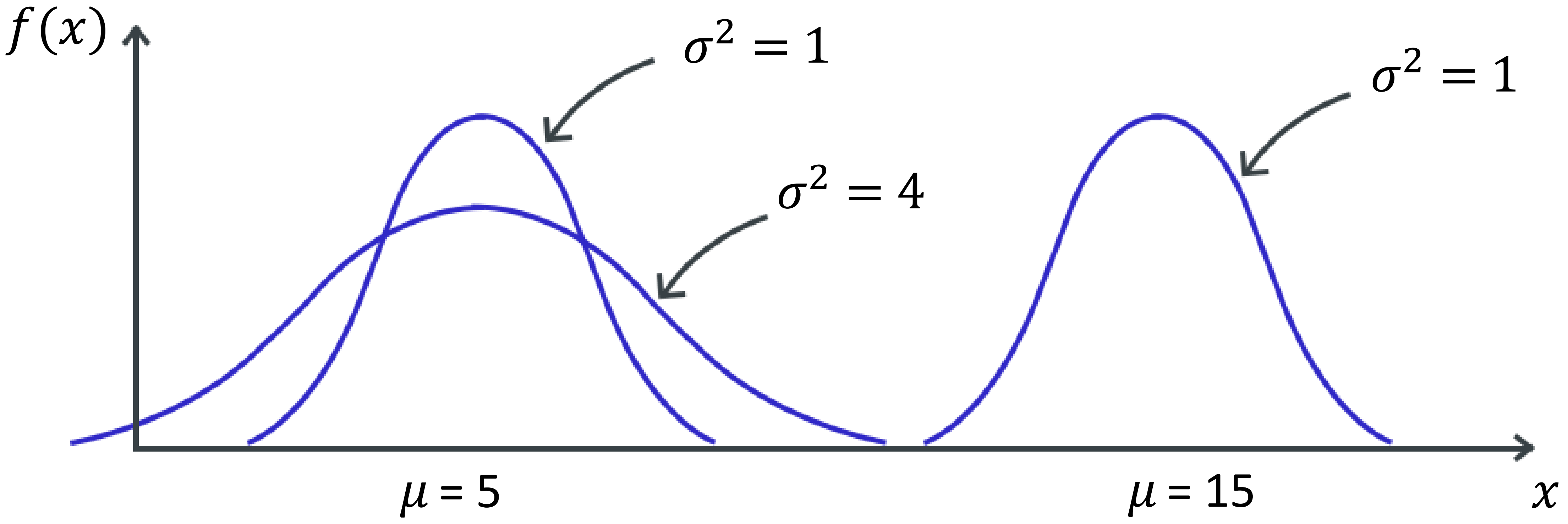
▶ 이 경우 $X \sim N\ [\mu, \sigma^2]$라고 함.



정규분포

정규분포 확률밀도함수의 개형

- μ는 분포의 중심.
- μ를 중심으로 대칭이고, μ에서 가장 큰 값이 되는 하나의 봉우리만 가짐.
- σ^2이 크면 분포의 산포가 커지고, σ^2이 작으면 분포의 산포가 작아짐.



정규분포

정규분포의 특성치

- $X \sim N[\mu, \ \sigma^2]$인 경우
  - $E[X] = \mu$
  - $V[X] = \sigma^2$

표준정규분포

표준정규분포와 정규확률변수의 표준화

- 표준정규분포

  - $X \sim N[\mu, \sigma^2]$일 때, 정규분포의 선형불변성에 의해

  $Z = \frac{X-\mu}{\sigma} \sim N[0,1]$ 이 되며,

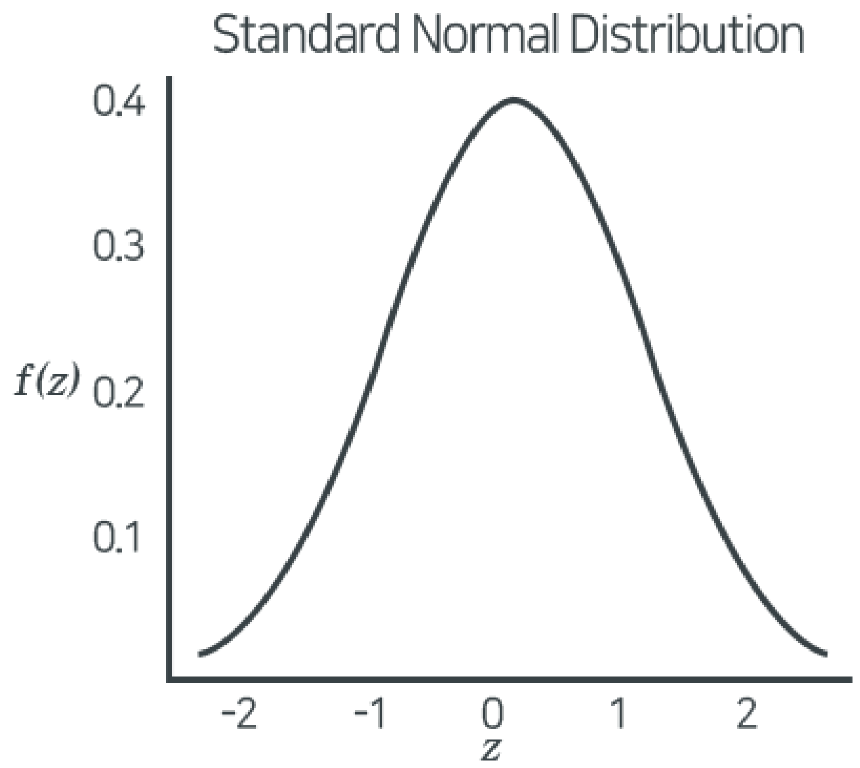
  이 때의 평균이 0 분산이 1인 정규분포를 표준정규분포라 정의함.

표준정규분포

표준정규분포와 정규확률변수의 표준화

- 표준정규분포

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



표준정규분포

표준정규분포와 정규확률변수의 표준화

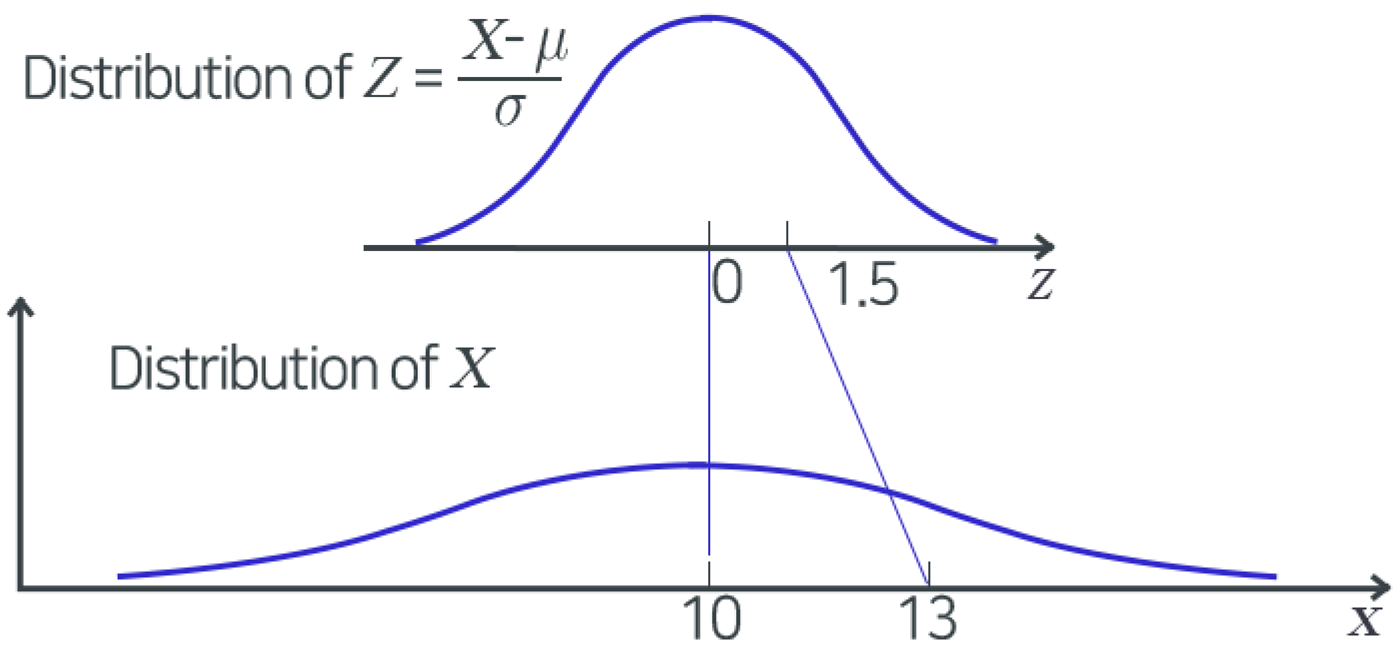
- $X \sim N[10, 2^2]$ 일 때, $P(10 < X < 13)$

$$
\begin{aligned}
&P(10 < X < 13) \\
&= P\left(\frac{10-10}{2} < \frac{\mathrm{X}-10}{2} < \frac{13-10}{2}\right) \\
&= P(0 < Z < 1.5), Z \sim N[0,1]
\end{aligned}
$$

표준정규분포

표준정규분포와 정규확률변수의 표준화

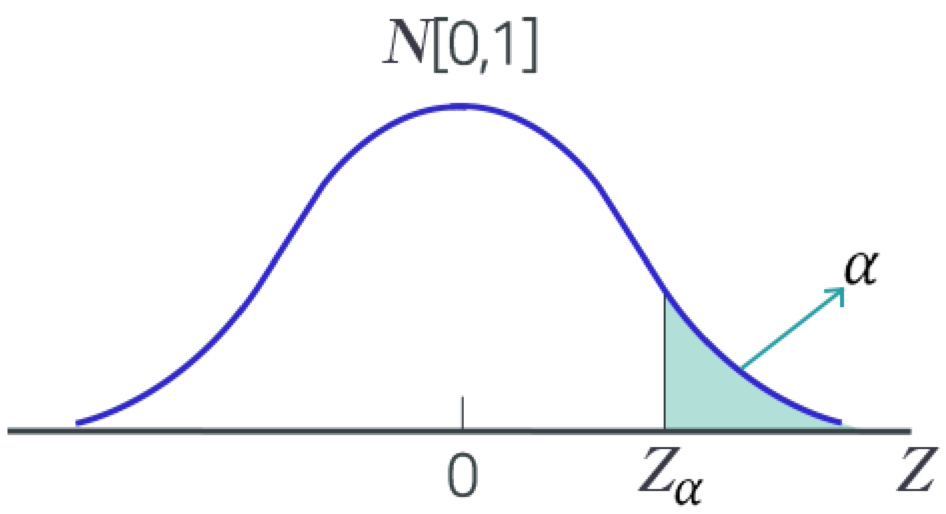
- $X \sim N[10, 2^2]$ 일 때, $P(10 < X < 13)$



표준정규분포

표준정규 확률변수의 $(1-\alpha)$분위수 : Z_α

- $Z \sim N\,[0,1]$일 때, $\mathrm{P}[Z < c] = 1 - \alpha$를
  만족하는 Z의 $(1-\alpha)$분위수 c를 Z_α으로 표기.



주요 확률분포:
카이제곱분포, t분포, f분포

Key words

#카이제곱분포 #t분포 #F분포

카이제곱 분포

정의

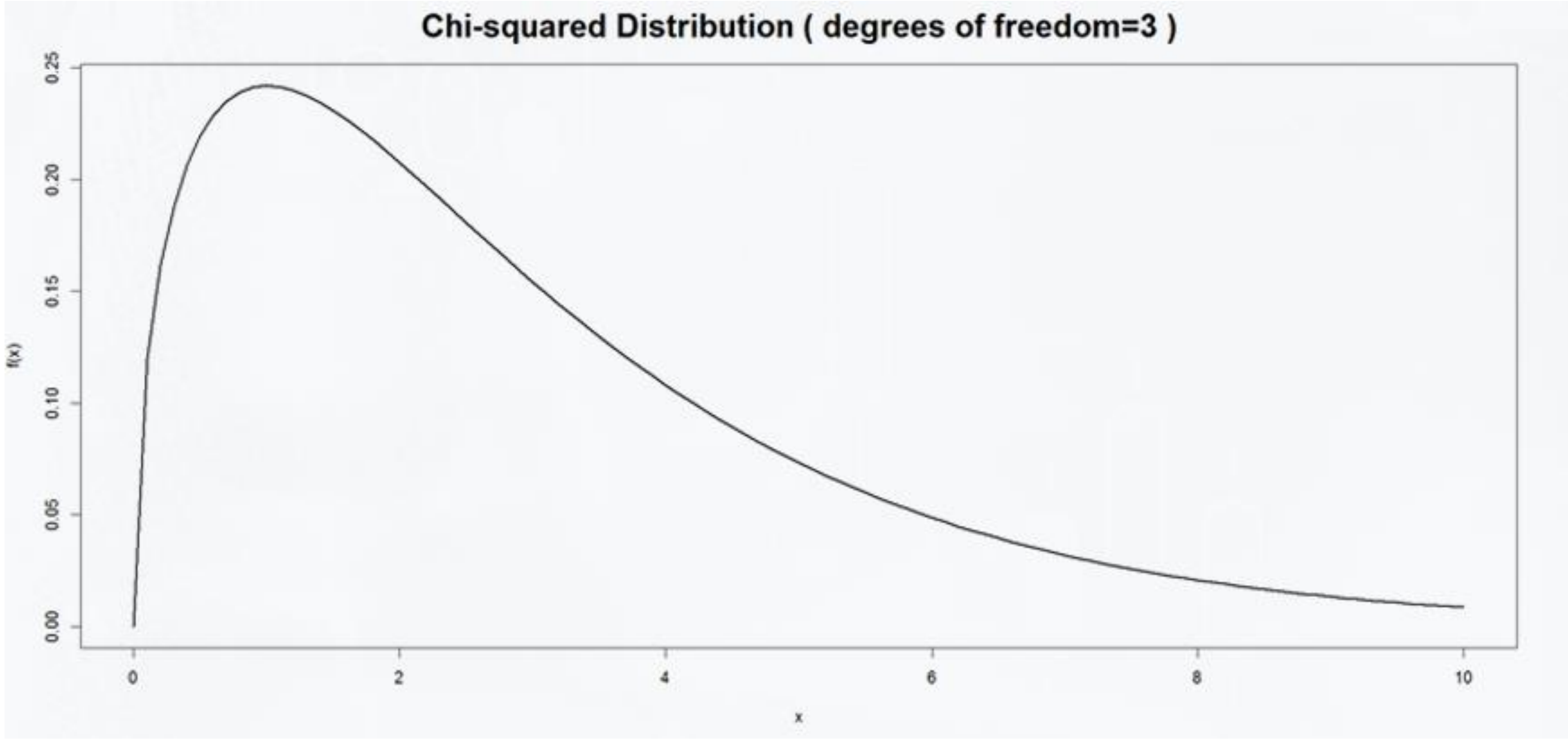
- $Z_1, Z_2, \ldots, Z_k$가 k개의 서로 독립인 표준정규 확률변수 $(Z_i \sim N[0,1],\ i = 1,2,\ldots,k)$ 라고 할 때, $X = {Z_1}^2 + {Z_2}^2 + \cdots + {Z_k}^2$가 따르는 분포를 자유도가 k인 카이제곱 분포라고 정의함.

$$f(x) = \frac{1}{\Gamma\left(\frac{1}{2}\right) 2^{\frac{k}{2}}} x^{\frac{k}{2}-1} e^{-\frac{x}{2}}, 0 < x < \infty$$

- 이 경우 $X \sim \chi^2[k]$라고 함.

카이제곱 분포

정의


Chi-squared Distribution (degrees of freedom=3)
f(x)
0.00
0.05
0.10
0.15
0.20
0.25
0
2
4
6
8
10
x

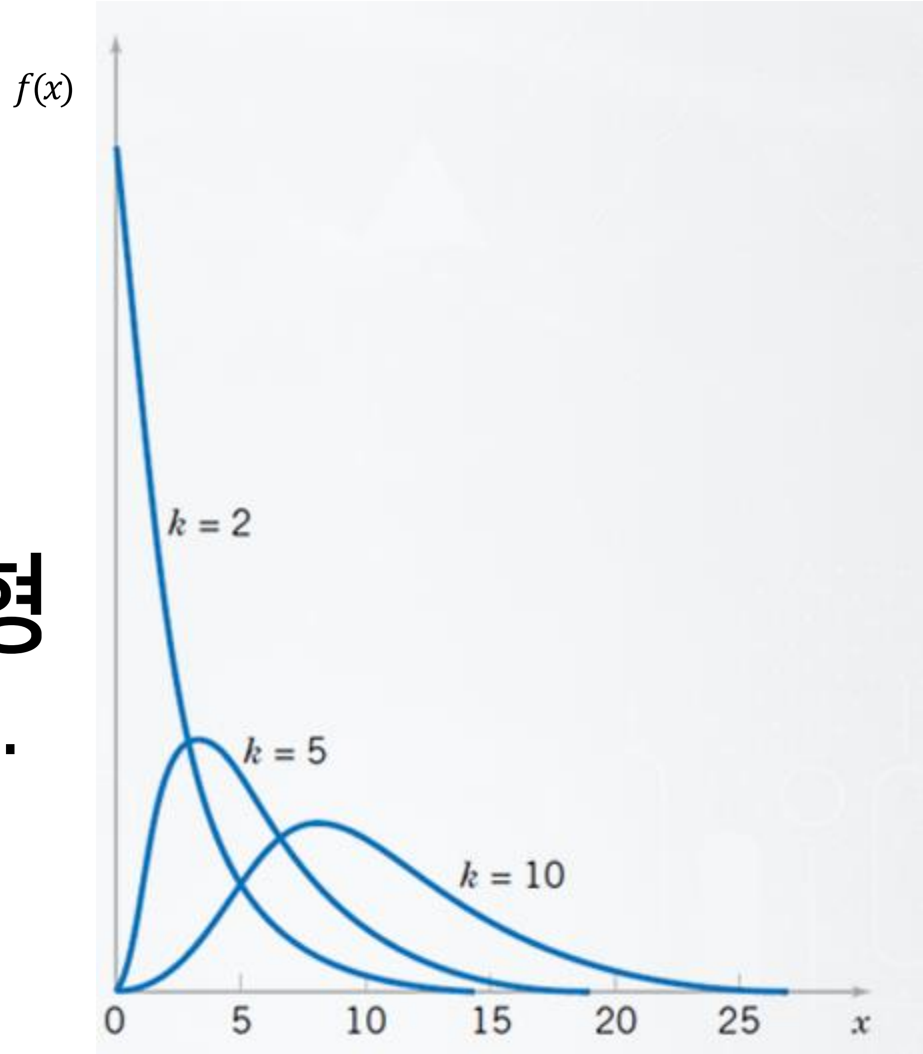
카이제곱 분포

카이제곱 분포의 특성치

- $X \sim \chi^2[k]$인 경우
  - $E[X] = k$
  - $V[X] = 2k$

카이제곱 분포 확률밀도함수 개형

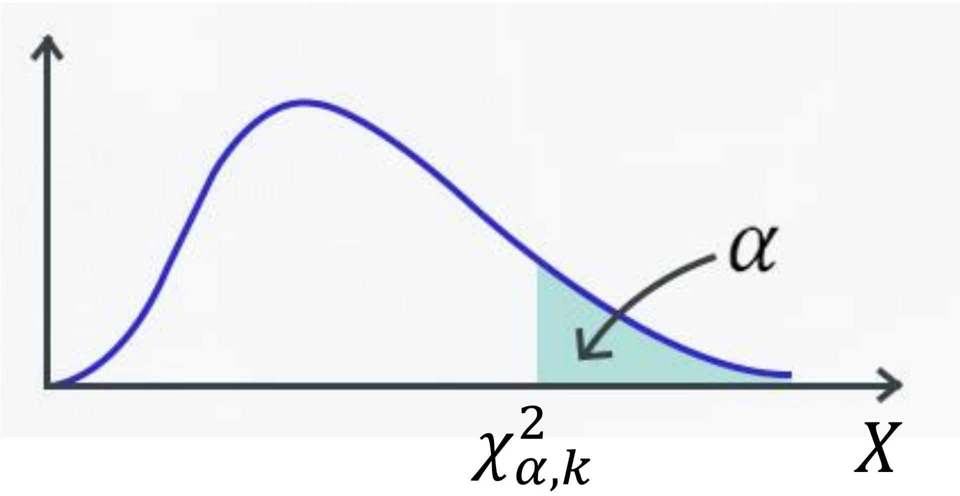
- 오른쪽 꼬리가 길게 늘어진 비대칭 형태임.



카이제곱 분포

카이제곱 확률변수의 $(1-\alpha)$분위수 : $\chi^2{}_{\alpha,k}$

- $X \sim \chi^2[k]$일 때, $\mathrm{P}[X > c] = \alpha$를 만족하는 X의 $(1-\alpha)$분위수 c를 $\chi^2{}_{\alpha,k}$으로 표기함.



t 분포

정의

- Z가 표준정규 확률변수 $Z \sim N[0,1]$이며 할 때, U가 자유도가 k인 카이제곱 확률변수 $U \sim \chi^2[k]$ 이며, Z와 X는 서로 독립이라고 할 때, $X = \frac{Z}{\sqrt{U/k}}$가 따르는 분포를 자유도가 k인 t 분포라고 정의함.

$$f(x) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)} \frac{1}{\sqrt{k\pi}} \frac{1}{\left(1 + \frac{x^2}{k}\right)^{\frac{k+1}{2}}}, \quad -\infty < x < \infty$$

- 이 경우 $X \sim t\,[k]$라고 함.

t 분포

t 분포의 특성치

- $X \sim t\,[k]$인 경우

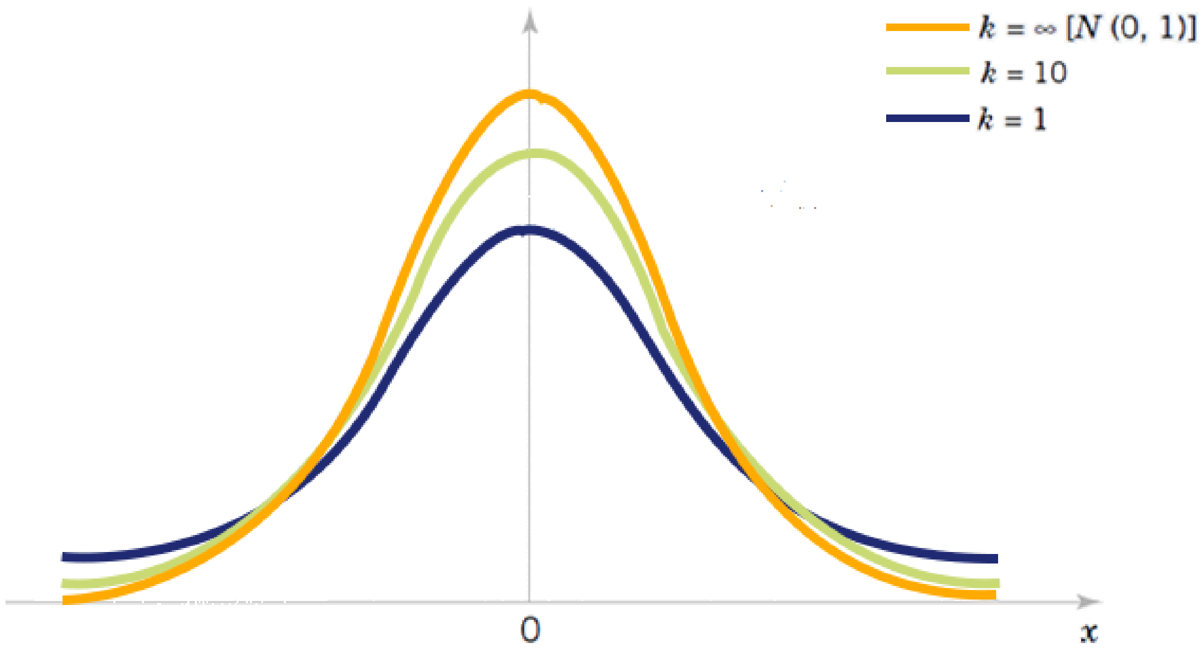
-$E[X] = 0$

-$V[X] = \frac{k}{k-2}$ (단, $k > 2$)

t 분포

t 분포 확률밀도함수 개형

- $X \sim t\,[k]$인 경우
  - 가운데 0을 중심으로 대칭인 종모양의 분포.
  - 표준정규분포 보다 꼬리가 두꺼움.
  - 자유도 k가 커짐에 따라 산포가 줄어들어 표준정규분포로 수렴함.

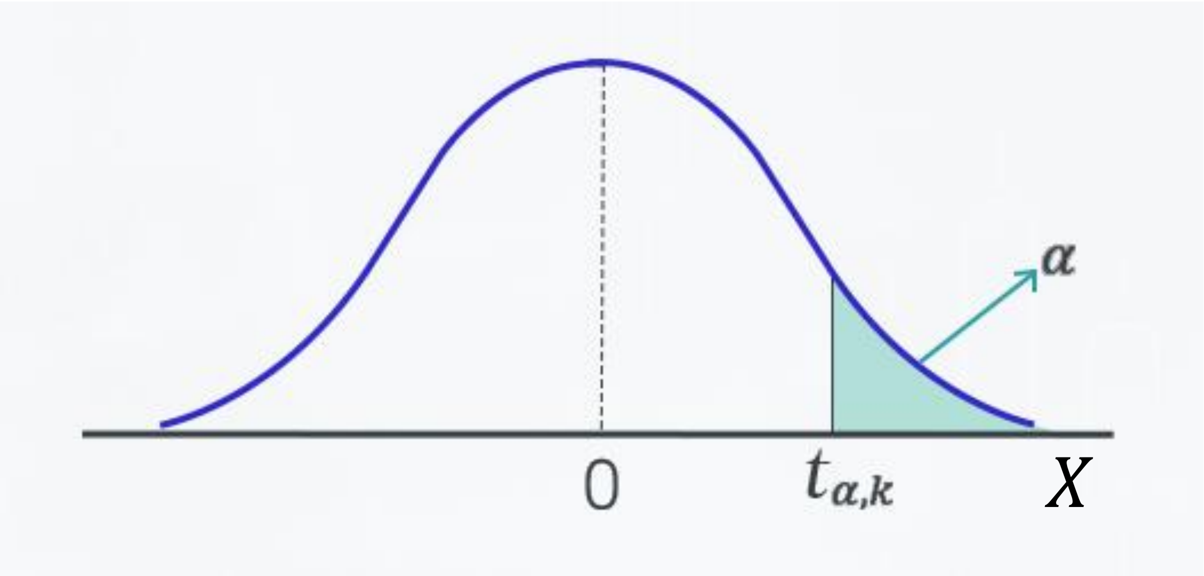


Probability density functions of several t distributions.

t 분포

t 확률변수의 $(1-\alpha)$분위수 : $t_{\alpha,k}$

- $X \sim t\,[k]$일 때, $\mathrm{P}[X > c] = \alpha$를 만족하는 X의 $(1-\alpha)$분위수 c를 $t_{\alpha,k}$으로 표기함.



F 분포

정의

- U가 자유도가 k_1인 카이제곱 확률변수 $U \sim \chi^2[k_1]$이며, V가 자유도가 k_2 인 카이제곱 확률변수 $V \sim \chi^2[k_2]$이고, U와 V는 서로 독립이라고 할 때, $X = \frac{U/k_1}{V/k_2}$ 가 따르는 분포를 자유도가 k_1, k_2 인 F분포라고 정의함.

$$f(x) = \frac{\Gamma\left(\frac{k_1+k_2}{2}\right)}{\Gamma\left(\frac{k_1}{2}\right)\Gamma\left(\frac{k_2}{2}\right)}\left(\frac{k_1}{k_2}\right)^{\frac{k_1}{2}} x^{\frac{k_1}{2}-1}\left(1+\frac{k_1}{k_2}x\right)^{-\frac{1}{2}(k_1+k_2)}, \quad 0 < x < \infty$$

- 이 경우 $X \sim F[k_1, k_2]$ 라고 함.

F 분포

F분포의 특성치

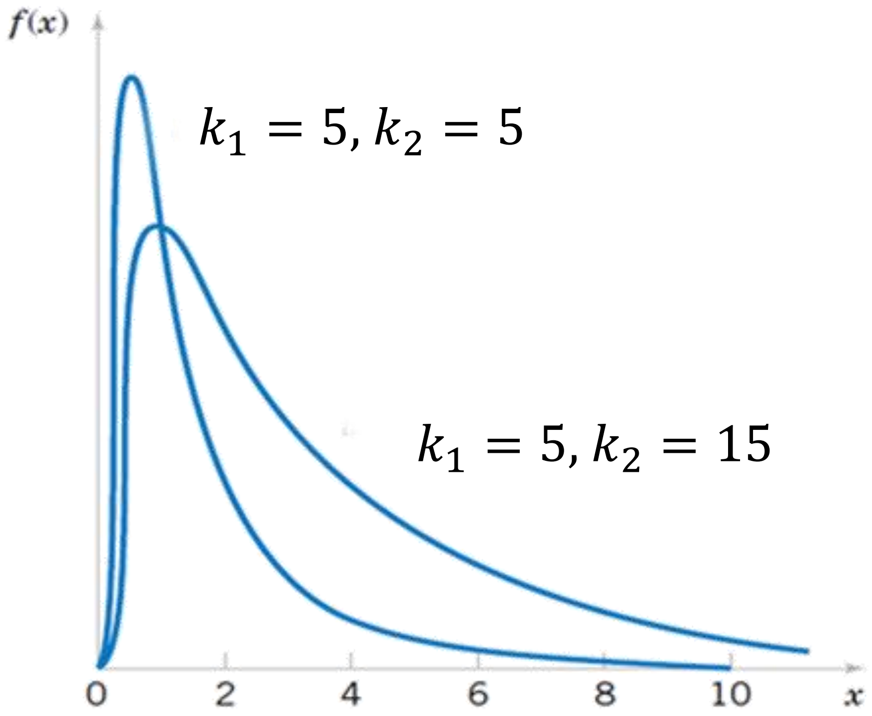
- $X \sim F[k_1, k_2]$ 인 경우

  - $E[X] = \frac{k_2}{k_2 - 2}$

  - $V[X] = \frac{2k_2^2(k_1 + k_2 - 2)}{k_1(k_2 - 2)^2(k_2 - 4)}$

F분포 확률밀도함수 개형

- 카이제곱 분포처럼 오른쪽으로 치우친 비대칭 구조임.



Probability density functions of two F distributions.

F 분포

F 확률변수의 $(1-\alpha)$분위수 : F_{α,k_1,k_2}

- $X \sim F[k_1, k_2]$ 일 때, $\mathrm{P}[X > c] = \alpha$를 만족하는 X의 $(1-\alpha)$분위수 c를 F_{α,k_1,k_2}으로 표기함.