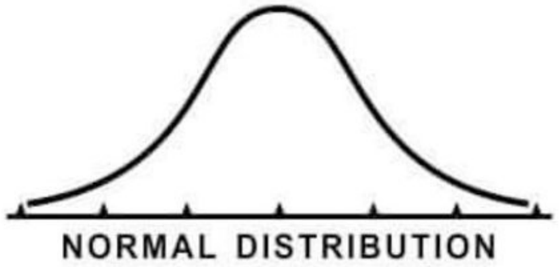
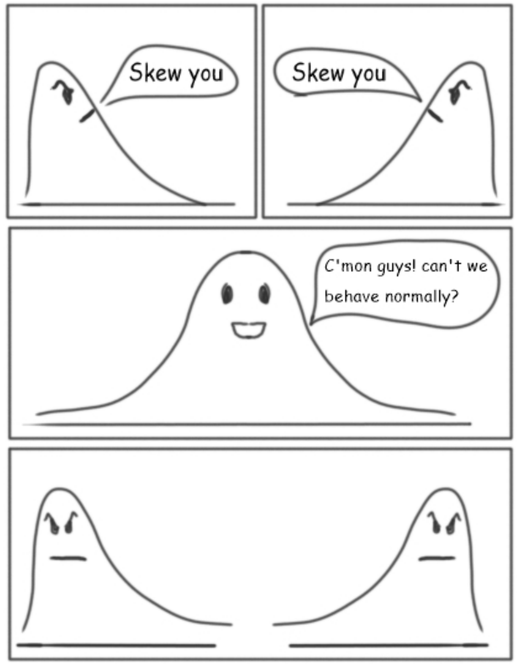




실험심리언어학도를 위한 기초 확률과 통계

2. 확률

2022. 07. 14.

박기효

Cornell University

KU KONKUK UNIVERSITY

Brain & Cognition Research Center

오늘 수업 내용

- 확률
- 확률분포
- 표본분포와 중심극한정리

  ➡ (여러분께) 바라는 것: 통계학의 근간이 되는 분야 중 하나인 확률론과 친해지기

오늘 수업 내용

- 오늘은 수학을 많이 다룰 겁니다.
- 하지만 계산을 여러분이 직접하실 필요는 없습니다.
- 왜냐하면 R이라는 공학용 계산기가 다 해줄 거거든요!
- 중요한 건 **개념**을 이해하는 겁니다.

우리(실험심리언어학자)가 다루는 데이터의 모습

- 어떤 게 있을까요?

우리(실험심리언어학자)가 다루는 데이터의 모습

- 어떤 게 있을까요?
- 빈칸 채우기(Cloze task), 수용성 판단(O/X 혹은 1점~7점 사이로 응답) …
- 반응시간, 안구운동, 뇌파(EEG 등) …

우리(실험심리언어학자)가 다루는 데이터의 모습

- 어떤 게 있을까요?
- 빈칸 채우기(Cloze task), 수용성 판단(O/X 혹은 1점~7점 사이로 응답) …

  예) 수용할 수 있다(=O)라고 응답한 경우가 몇 번 나왔는가?

- 반응시간, 안구운동, 뇌파(EEG 등) …

  예) 문법적이지 않은 문장에 대한 반응시간이 얼마만큼 나왔는가?

우리(실험심리언어학자)가 다루는 데이터의 모습

- 어떤 게 있을까요?
- 빈칸 채우기(Cloze task), 수용성 판단(O/X 혹은 1점~7점 사이로 응답) …

  예) 수용할 수 있다(=O)라고 응답한 경우가 몇 번 나왔는가?

  → **이산형(discrete)**
- 반응시간, 안구운동, 뇌파(EEG 등) …

  예) 문법적이지 않은 문장에 대한 반응시간이 얼마만큼 나왔는가?

  → **연속형(continuous)**

우리(실험심리언어학자)가 다루는 데이터의 모습

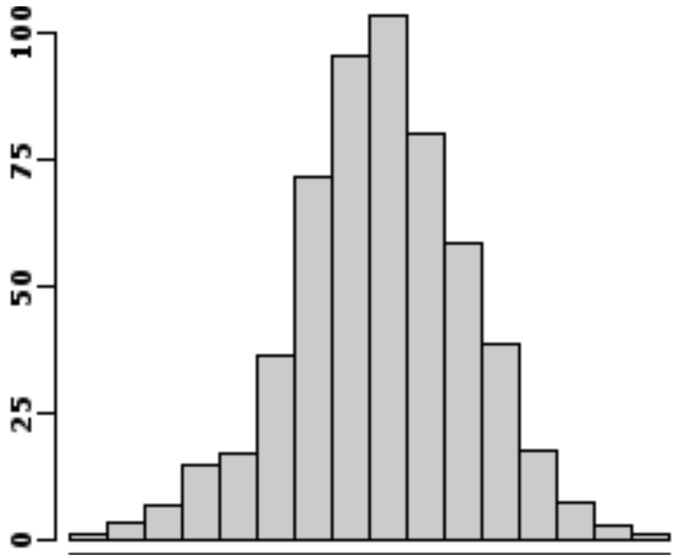
- 우리는 실험 데이터가 어떤 모습을 나타낼지 (암묵적으로) 가정함.

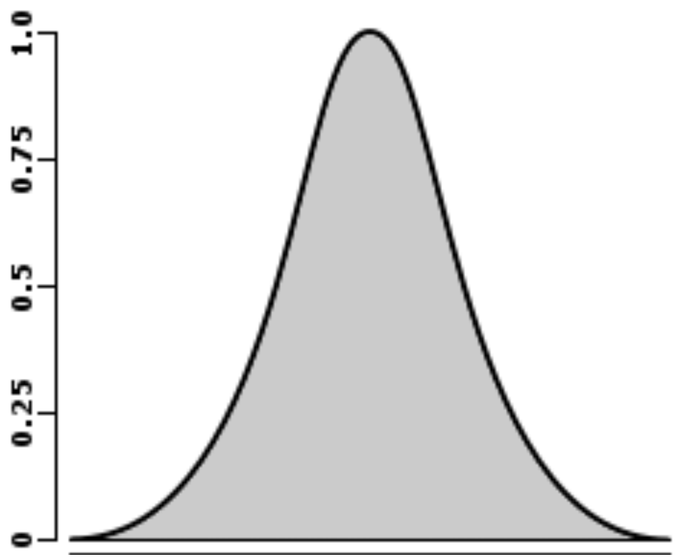
우리(실험심리언어학자)가 다루는 데이터의 모습

- 우리는 실험 데이터가 어떤 모습을 나타낼지 (암묵적으로) 가정함.
- 다른 말로 하면, 우리는 관찰값, 즉, 종속변수가 어떤 분포를 따른다고 가정함.

우리(실험심리언어학자)가 다루는 데이터의 모습

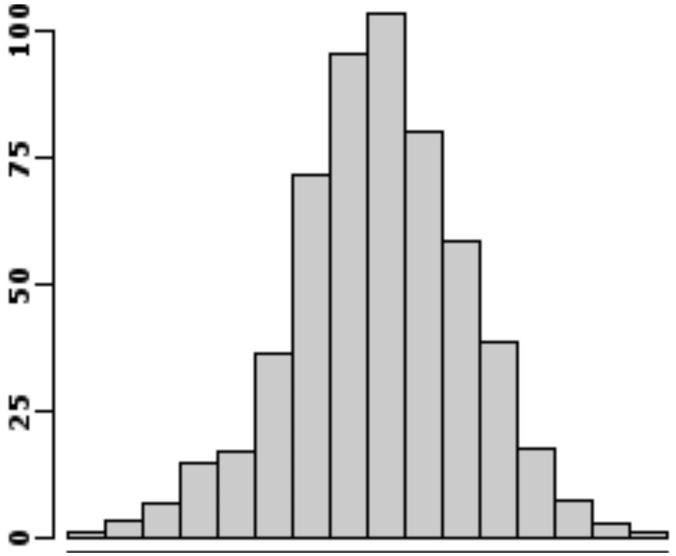
- 우리는 실험 데이터가 어떤 모습을 나타낼지 (암묵적으로) 가정함.
- 다른 말로 하면, 우리는 관찰값, 즉, 종속변수가 어떤 분포를 따른다고 가정함.



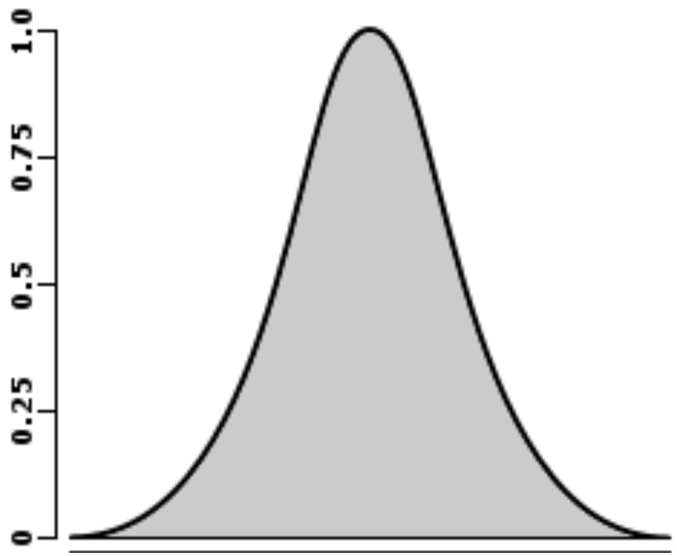


우리(실험심리언어학자)가 다루는 데이터의 모습

- 우리는 실험 데이터가 어떤 모습을 나타낼지 (암묵적으로) 가정함.
- 다른 말로 하면, 우리는 관찰값이 어떤 분포를 따른다고 가정함.



이산형

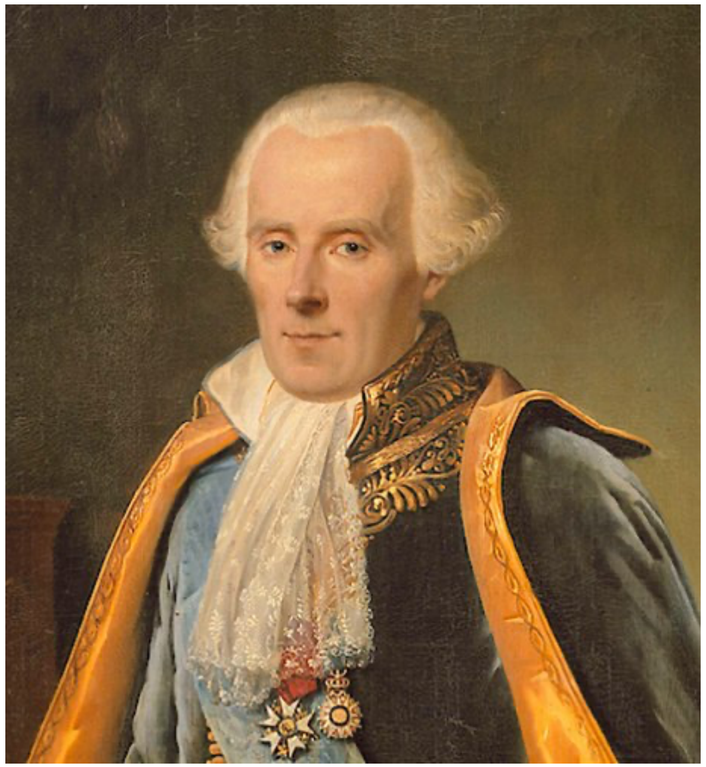


연속형

오늘 같이 얘기할 것들

- 실제 실험심리언어학에서 데이터를 수집할 때 활용되는 분포들에 대한 가정들.
- 하지만 이를 알기 위해선 기초적인 확률론에 대해서 알아야함.
- 특히나 데이터를 생성해내는 기능으로서 확률 변수에 담긴 개념들을 살펴볼 것!
- 귀찮지만 안 할 수 없으며, 무엇보다 알아두면 앞으로 매우 유용함!

좀 더 엄격하게 확률 바라보기



“확률론이란 상식을 산수로 간추린 것에 불과하다.”

- 피에르시몽 드 라플라스 (1749 ~ 1827) -

↑재수없다.

확률이란?

- **빈도주의(Frequentist) 관점**

  ✓ 비율(상대빈도) = 특정 사건이 일어난 빈도 / 모든 사건이 일어난 빈도

  ✓ 무한한 데이터를 가지고 있을 때 나타나는 상대빈도

확률이란?

- **빈도주의(Frequentist) 관점**

  ✓ 비율(상대빈도) = 특정 사건이 일어난 빈도 / 모든 사건이 일어난 빈도

  ✓ 무한한 데이터를 가지고 있을 때 나타나는 상대빈도

- **베이즈(Bayesian) 관점**

  ✓ 불확실성에 대한 주관적 측도 (혹은 '주관적 확률', '믿음의 정도')

  ✓ 베이즈 정리(Bayes Theorem): $P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$

확률이란?

- **빈도주의(Frequentist) 관점**

  ✓ 비율(상대빈도) = 특정 사건이 일어난 빈도 / 모든 사건이 일어난 빈도

  ✓ 무한한 데이터를 가지고 있을 때 나타나는 상대빈도

- **베이즈(Bayesian) 관점**

  ✓ 불확실성에 대한 주관적 측도 (혹은 '주관적 확률', '믿음의 정도')

  ✓ 베이즈 정리(Bayes Theorem): $P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$

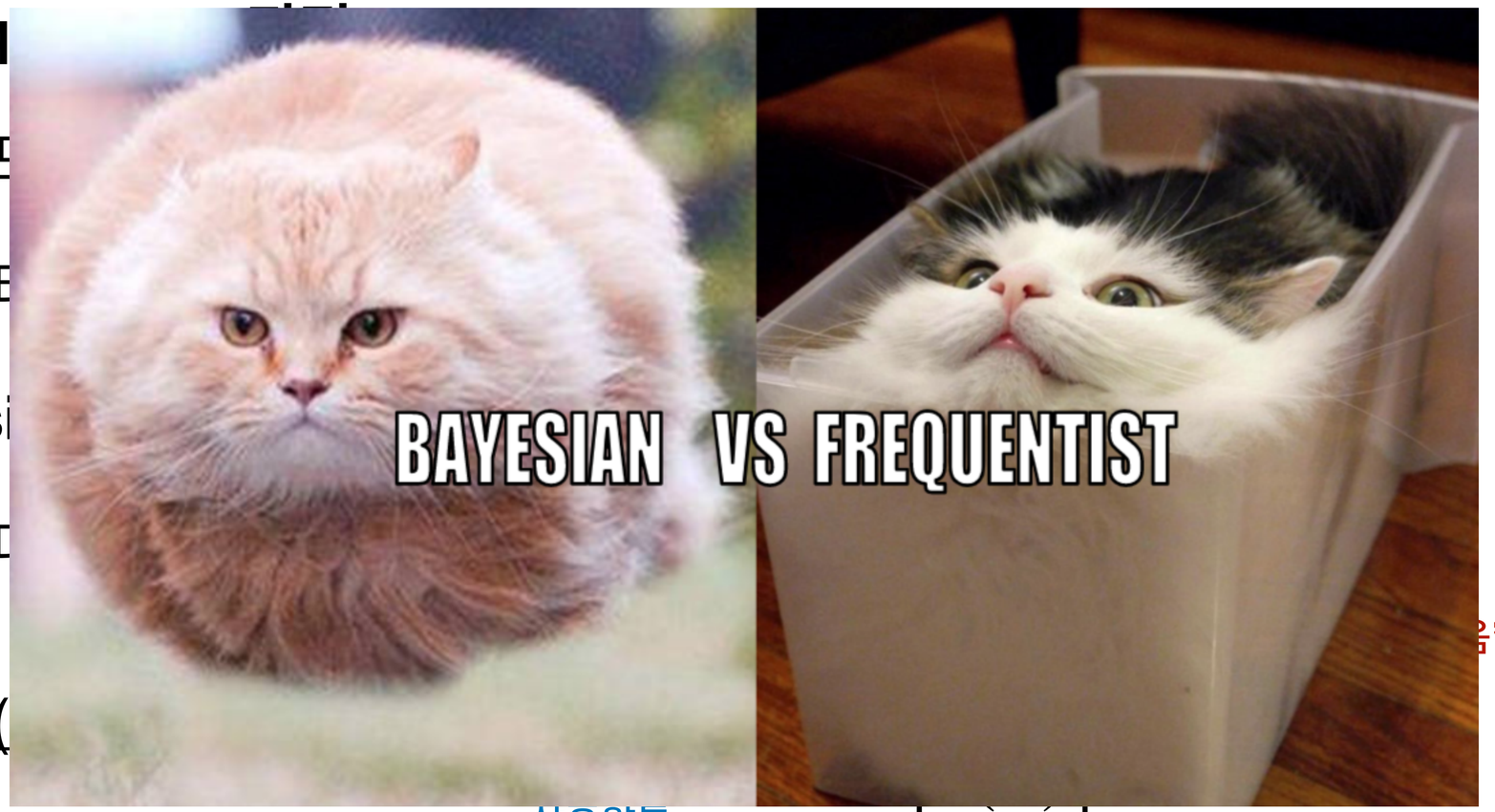
  사후확률

  사전확률 = '믿음' 혹은 '직관'

  가능도 = 관찰을 통해 얻은 확률

확률이란?

- **빈도주의(Freq**
  - ✓ 비율(상대빈도
  - ✓ 무한한 데이터
- **베이즈(Bayesi**
  - ✓ 불확실성에 대
  - ✓ 베이즈 정리(



움’ 혹은 ‘직관’

사후확률

가능도 = 관찰을 통해 얻은 확률

확률이란?

- **확률 변수** X

  ✓ 무작위 실험을 통해 얻은 결과값들과 어떤 성질을 담고 있는 집합을 사상(mapping)하는 함수 (뒤에서 더 얘기할 예정)

- **확률 분포** $p(X = x)$

  ✓ 어떤 성질 x를 담고 있는 결과값들이 발생할 확률을 측도하는 함수

확률 공간

- **표본 공간 Ω**

  ✓ 어떤 실험에서 얻어낼 수 있는 모든 가능한 결과들의 집합

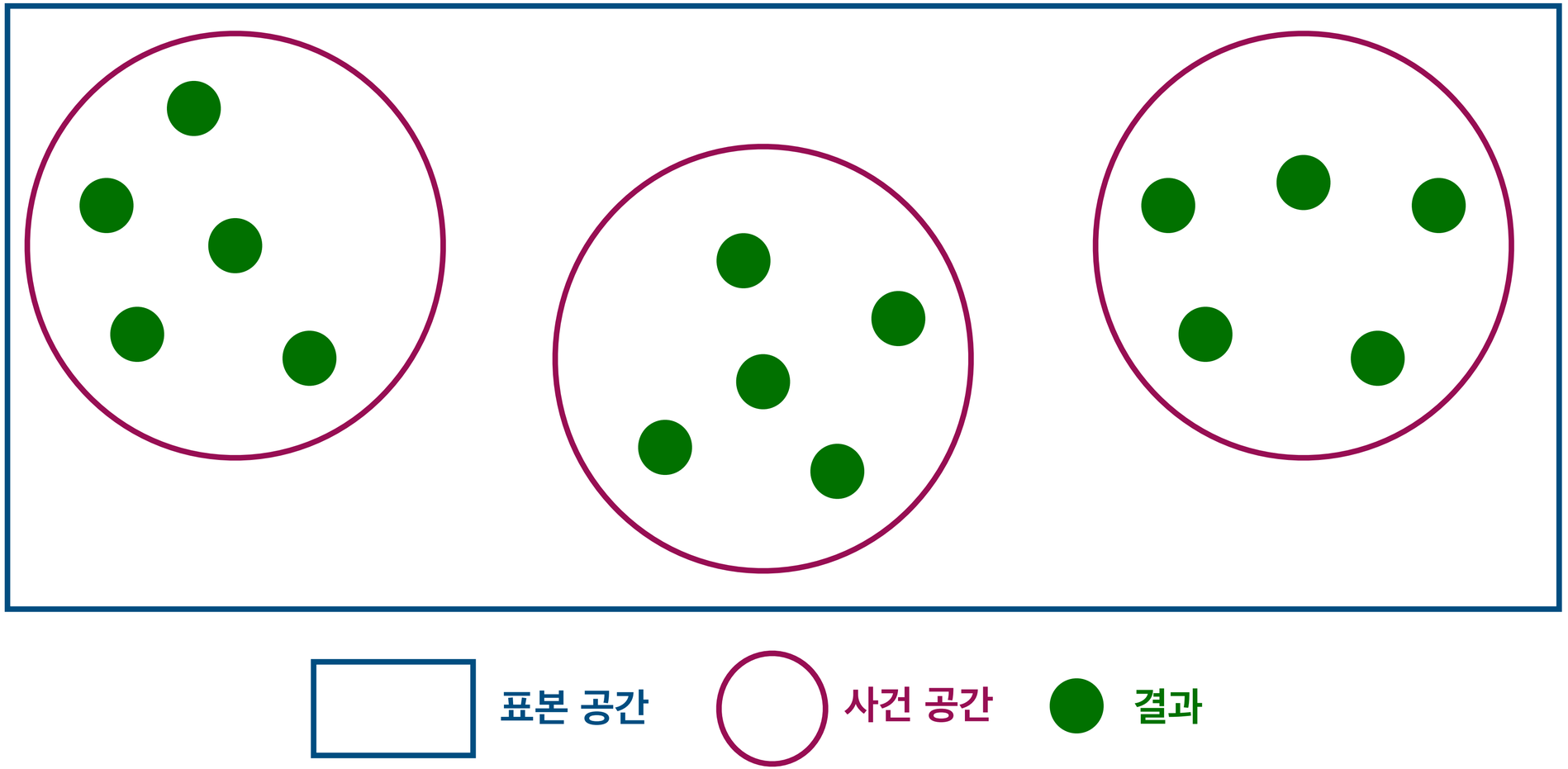
  ✓ 예) O/X 수용성 판단이 가능한 임의의 2개의 문장에 대한 결과가 들어있는 표본공간 Ω

  ➡ {OO, XX, OX, XO}

- **사건 공간 E**

  ✓ 어떤 실험에서 얻을 수 있는 잠재적 결과들이 담긴 공간

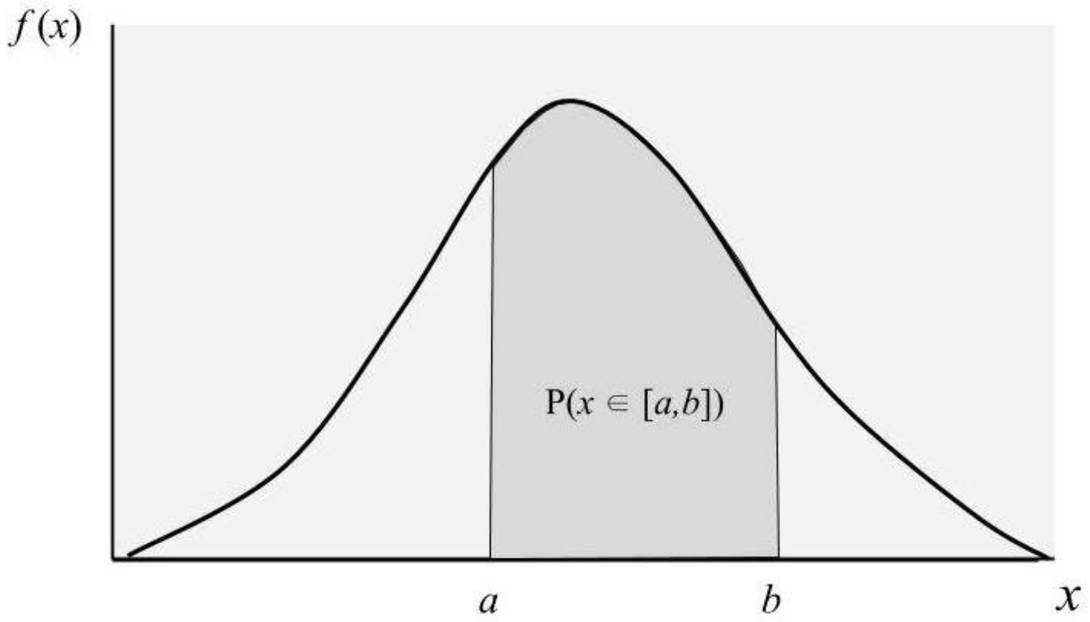
확률 공간



확률 공간

- **확률함수 p**

  ✓ 어떤 사건 $e \in E$ 이 존재할 때, 확률함수 $p(e) \in [0,1]$ (0 이상 1 이하)는 실험 결과의 확률 혹은 믿음의 정도를 측도한다.



확률론의 공리 (=확률을 정의하는 필요충분조건)

1. **음수가 아님!**

   ✓ 임의의 사건 A가 일어날 확률을 $P(A)$라 할 때, $P(A) \geq 0$ 이다.

2. **합하면 1!**

   ✓ $P(\Omega) = 1$이다.

3. **가산성(=더하기가 된다)!**

   ✓ 표본공간 Ω에 정의된 연속된 사건 $A_1, A_2, \cdots$가 있다고 하자. 이 때, 모든 $i \neq j$에 대하여 $A_i \cap A_j = \varnothing$이면, $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$ 이다.

확률 변수

- 어떤 확률 공간 (Ω, E, p)가 주어졌을 때, 그 결과를 나타내고 있는 공간 Ω을 목표 공간 T라는 것으로 변환하고자 한다고 하자.

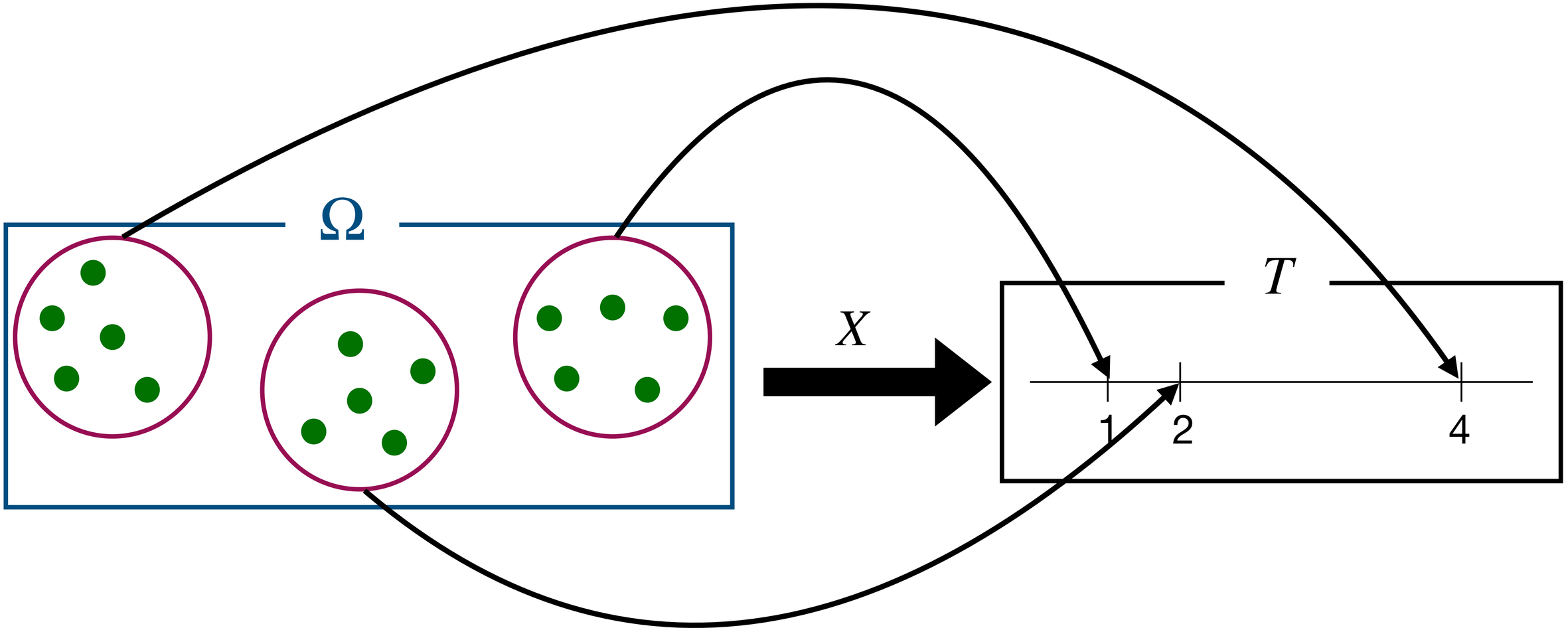
- 따라서 우리가 관심 있는 건 목표 공간 T !

- **확률 변수(random variable)**

  ✓ $X : \Omega \rightarrow T$

  ✓ 즉, 표본 공간의 원소(결과)를 목표 공간의 **값**으로 반환해주는 함수가 바로 확률변수! (아마 여러분 머릿속에 들어있는 기존의 그 '변수'가 아닐 것!)

확률 변수



확률 변수

- **예제**
  - ✓ 확률변수 $X : \Omega \rightarrow T$
  - ✓ O/X 수용성 판단이 가능한 임의의 2개 문장의 결과가 들어 있는 표본 공간
    - ➡ $\Omega : \{OO, XX, OX, XO\}$
  - ✓ '수용 가능하다(=O)'가 나올 수 있는 결과값을 반환해주는 확률변수 X는 다음과 같다.
  - ✓ $X(OO) = 2$, $X(OX) = 1$, $X(XO) = 1$, $X(XX) = 0$
    - ➡ $T : \{0, 1, 2\}$
- **이 때 사건 공간 E가 성립하기 위해선 확률 변수 X는 측도 가능한(=좋은) 함수여야 함!**
  - ✓ 이는 결국 측도 가능한 함수에 속하는 모든 집합은 말 그대로 측도 가능함(=좋음)을 의미!

확률 변수가 왜 필요하죠?

“왜 굳이 이렇게 해요? 그냥 수용하거나 안 할 확률 2분의 1씩 해서 하면 되는 거 아닌가요?”

확률 변수가 왜 필요하죠?

- **새로운 확률 공간을 한 번 정의해봅시다.**

  ✓ $\Omega =$ 한국어를 제1언어로 습득하는 24개월 아기

  ✓ $E = \mathscr{P}\{\Omega\}$

  ✓ 모든 $x \in \Omega$에 대하여 $p(x) = 1/|\Omega|$

- **하루에 명사(구체명사 혹은 추상명사)를 산출할 확률 변수**

  ✓ $N(x) = x$가 명사를 산출하면 **참** (그렇지 않으면 **거짓**)

  ✓ $ConcreteN(x) = x$가 구체명사를 산출하면 **참** (그렇지 않으면 **거짓**)

  ✓ $AbstractN(x) = x$가 추상명사를 산출하면 **참** (그렇지 않으면 **거짓**)

확률 변수가 왜 필요하죠?

- **단일 무작위 공간 Ω에서 물어볼 수 있는 질문들 몇 개**

✓ $p(x$가 구체명사와 추상명사 둘 다 산출하는 경우)

$= p(ConcreteN = \text{참}, AbstractN = \text{참})$

✓ $p(x$가 명사를 산출한다는 조건 하에 x가 구체명사를 산출하는 경우)

$= p(ConcreteN = \text{참} \mid N = \text{참})$

✓ $p(x$가 명사를 산출한다는 조건 하에 x가 추상명사를 산출하지 않는 경우)

$= p(AbstractN = \text{ 거짓} \mid N = \text{참})$

확률 변수가 왜 필요하죠?

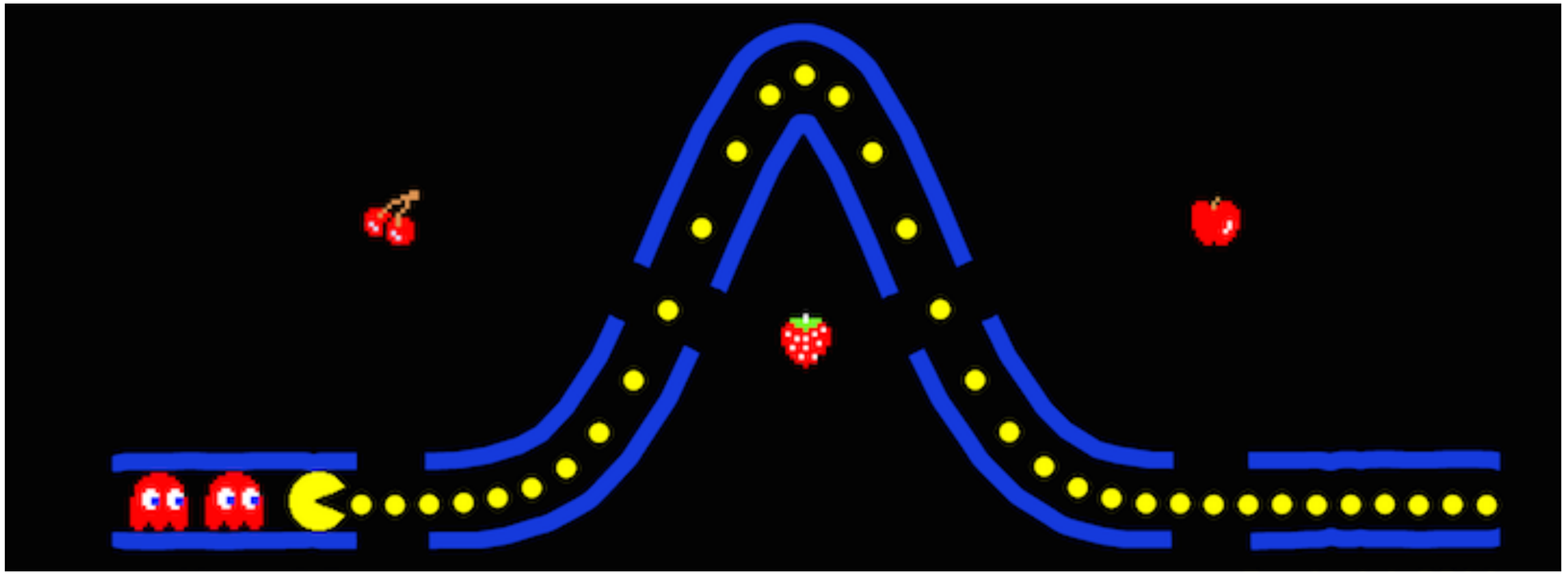
- **확률 변수가 필요한 주요 이유 2개**

  ✓ 표본 공간에 있는 각 원소가 지닌 서로 다른 **성질들** 간의 상호 관계를 볼 수 있게 함.

  ✓ 확률 변수를 사용하지 않으면 독립, 상관성 등을 이야기하기가 어려워짐.

- **중요:** 독립과 상관성은 확률 공간의 성질이 아니라, **확률 변수**의 성질!

2. 이산형 확률변수와 연속형 확률변수



“통계학에서 확률변수는 밥줄입니다.”

- 조 블리츠스테인 (하버드 대학교 확률론 개론 중) -

이항분포로 살펴보는 이산형 확률 변수

어떤 문장들에 대한 문법성을 판단한 실험 데이터가 있다고 생각해봅시다.

(예: Tomorrow I will have ice cream. vs. *Tomorrow I had ice cream.)

Cornell University

KU KONKUK UNIVERSITY

Brain & Cognition Research Center

이항분포로 살펴보는 이산형 확률 변수

- 각 실험참여자들의 응답은 다음과 같이 간주됨(혹은 코딩됨).

  ✓ “문법적이다.” = 1

  ✓ “비문법적이다.” = 0

- 총 20명의 참여자들에 대해서 각 실험 마다 10개의 문장에 대한 응답을 얻어냄.
- 이 때, “문법적이다”(=1)이라고 응답한 결과를 R 코드(뒤에서 설명)는 아래와 같이 생성함.

  [1] 8 5 4 6 7 5 6 4 7 6 4 2 7 3 7 8 3 5 4 5

- 그리고 위 결과는 어떤 확률 분포 $p(Y)$를 따름.

이항분포로 살펴보는 이산형 확률 변수

- **이항분포(Binomial distribution)**

  ✓ 성공(=1 혹은 “문법적이다”) 혹은 실패(=0 혹은 “비문법적이다”)만을 갖는 이산형(=혹은 베르누이) 확률 변수 $X_1, \cdots, X_n$가 있을 때, 이들의 합으로 나타내는 $X = \sum_{i=1}^{n} X_i$를 이항 확률 변수라 한다.

  ✓ 이항 확률 변수가 따르는 분포는 이항분포라 한다.

  ✓ 이산형 확률 변수가 따른 확률 분포는 확률 질량 함수(Probability Mass Function, 혹은 PMF)다.

이항분포로 살펴보는 이산형 확률 변수

- **이항분포(Binomial distribution)**

  ✓ 이항 확률 변수의 확률 질량 함수는 다음과 같다.

$$Binomial(k \mid n, \theta) = \binom{n}{k} \theta^k (1-\theta)^{n-k} \text{ (단, } x = 0,1,2,\cdots,n)$$

  ✓ 이 때, n은 총 시행 횟수, k는 성공 횟수, θ는 성공 확률을 지칭한다. (퀴즈: $1-\theta$ = ?)

  ✓ 위 함수에서 θ는 성공 확률을 가리키는 모수(parameter)라고 한다.

  ✓ 하지만 위 식을 다 쓰는 건 귀찮으니깐 보통 책이나 논문에선 $X \sim B(n, \theta)$로 표기한다.

cf. $\binom{n}{k} = {}_nC_k = \frac{n!}{k!(n-k)!}$ = 고딩 때 경우의 수 시간에 배운 '조합'

이항분포로 살펴보는 이산형 확률 변수

- 앞에서 본 문법성 판단 실험을 다시 생각해보자.

  ✓ 각 실험참여자는 10개의 문장에 대해 10번의 판단을 함.

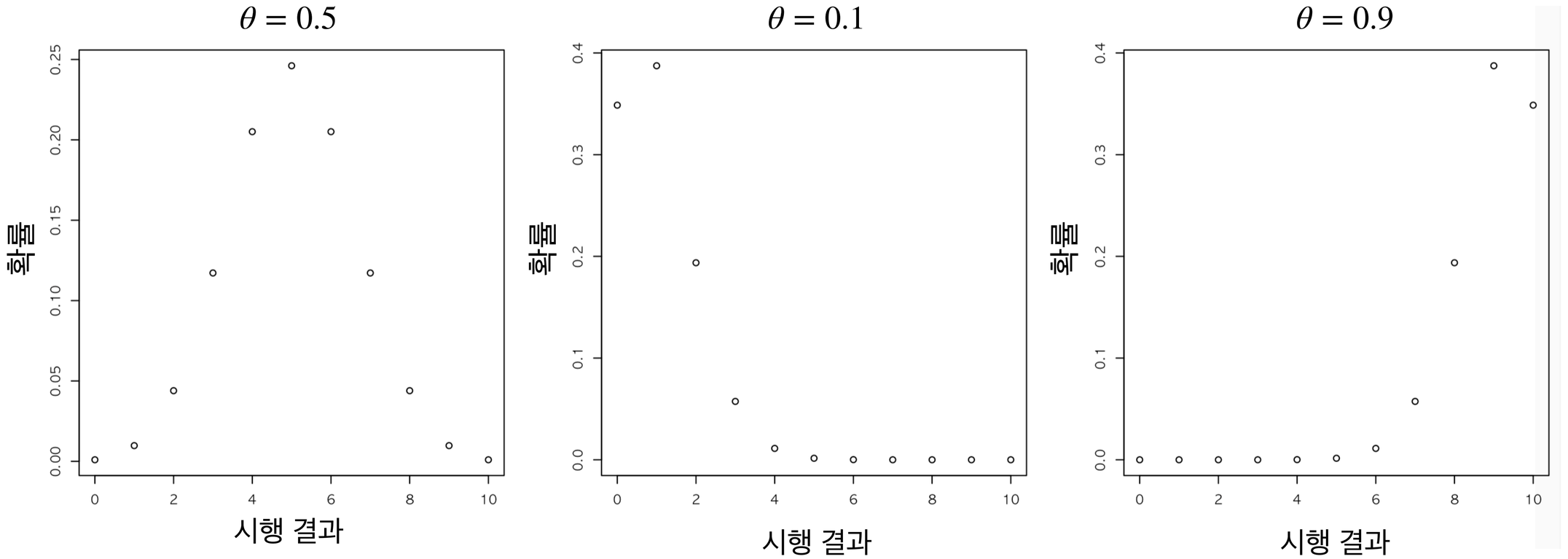
  ✓ 이 때 “문법적이다”라고 판단한 결과들은 총 11개의 경우의 수가 있음.

  ✓ 왜 10개가 아니라 11개? 한 번도 판단을 하지 않는 경우도 있기 때문.

- (실험 문장의 실제 문법성 여부는 잠시 접어두고) 문제는 “문법적이다”라고 응답할 확률, 즉, 성공확률 θ의 참값은 아무도 모름.

이항분포로 살펴보는 이산형 확률 변수

- 물론, (현실성을 떠나) 우리가 임의대로 θ 값을 아래처럼 설정해 볼 수는 있음.



이항분포로 살펴보는 이산형 확률 변수

- 하지만 θ의 참값은 여전히 알 수 없음.
- 따라서 실제 통계분석 진행시 주된 목표는 수집한 데이터를 갖고서 위와 같은 모수를 **추정**하는 것!

이항분포로 살펴보는 이산형 확률 변수

- **이항분포의 평균과 분산**

  ✓ …이 어떻게 수식으로 도출되는지에 대한 구체적인 설명은 생략!

  ✓ 한편, 이제부터 기대값(Expectation) 혹은 평균은 $E[X]$, 분산(Variance)은 $Var(X)$이라 하겠음.

  ✓ 이항분포의 평균과 분산은 다음과 같음.

  ➡$E[X] = n\theta,\ Var(X) = n\theta(1-\theta)$

이항분포로 살펴보는 이산형 확률 변수

- **최대 가능도 추정(Maximum Likelihood Estimation, MLE)**

  ✓ θ는 데이터를 통해 추정할 수 있다고 했음. 이 때, 데이터는 '관찰'된 것임.

  ✓ 이렇게 관찰된 데이터를 통해 추정한 모수를 $\hat{\theta}$ ("theta-hat")이라 하자.

  ✓ 어떻게 추정할까? 앞서 문법성 판단 실험 데이터를 갖고 생각해보자.

  ✓ 여기서 $\hat{\theta}$ 는 어떤 문장에 대해서 "문법적이다"라고 판단할 확률!

이항분포로 살펴보는 이산형 확률 변수

- **최대 가능도 추정(Maximum Likelihood Estimation, MLE)**

  ✓ 어렵게 생각할 필요 없다. 직관적으로 생각해보자.

이항분포로 살펴보는 이산형 확률 변수

- **최대 가능도 추정(Maximum Likelihood Estimation, MLE)**

  ✓ 어렵게 생각할 필요 없다. 직관적으로 생각해보자.

  ✓ n = 총 문법성 판단 횟수, k = “문법적이다”(=1 혹은 성공)라고 판단한 횟수

이항분포로 살펴보는 이산형 확률 변수

- **최대 가능도 추정(Maximum Likelihood Estimation, MLE)**

  ✓ 어렵게 생각할 필요 없다. 직관적으로 생각해보자.

  ✓ n = 총 문법성 판단 횟수, k = “문법적이다”(=1 혹은 성공)라고 판단한 횟수

  ✓ $\hat{\theta}$ = 어떤 문장에 대해서 “문법적이다”라고 판단할 확률 = ?

이항분포로 살펴보는 이산형 확률 변수

- **최대 가능도 추정(Maximum Likelihood Estimation, MLE)**

  ✓ 어렵게 생각할 필요 없다. 직관적으로 생각해보자.

  ✓ n = 총 문법성 판단 횟수, k = “문법적이다”(=1 혹은 성공)라고 판단한 횟수

  ✓ $\hat{\theta}$ = 어떤 문장에 대해서 “문법적이다”라고 판단할 확률 = k/n

이항분포로 살펴보는 이산형 확률 변수

- **최대 가능도 추정(Maximum Likelihood Estimation, MLE)**

  ✓ 이렇게 추정한 $\hat{\theta}$를 성공 관찰 비율(the observed proportion of successes)라고 하며, 이를 (평균은 모르지만) 참에 대한 **최대 가능도 추정**이라 함.

  ✓ 유사한 방식으로 분산에 대한 최대 가능도 추정은 $n\hat{\theta}(1-\hat{\theta})$이 됨.

  ✓ 이러한 추정들을 사용해서 통계적 추론(inference)를 함.

이항분포로 살펴보는 이산형 확률 변수

"최대 가능도 추정? 몰라..뭐야, 그거…무서워.."

이항분포로 살펴보는 이산형 확률 변수

- **최대 가능도 추정(Maximum Likelihood Estimation, MLE)**

✓ **가능도** = 관찰값들이 있을 때, 이항분포 함수를 통해 얻을 수 있는 특정한 θ의 값.

✓ $n = 10, k = 7$ 일 때, 10번의 시도 중 7번의 성공을 관찰할 확률은 얼마일까?

✓ $Binomial(k = 7 \mid n = 10, \theta) = \binom{10}{7} \theta^7 (1 - \theta)^{10-7}$

✓ 위 함수는 다른 관찰값들이 실험을 통해 수집이 되었을 때, θ에 의해 최종 성공 확률값이 달라진다는 것을 지칭.

✓ 이 경우 위 함수는 **가능도 함수(likelihood function)**라 함.

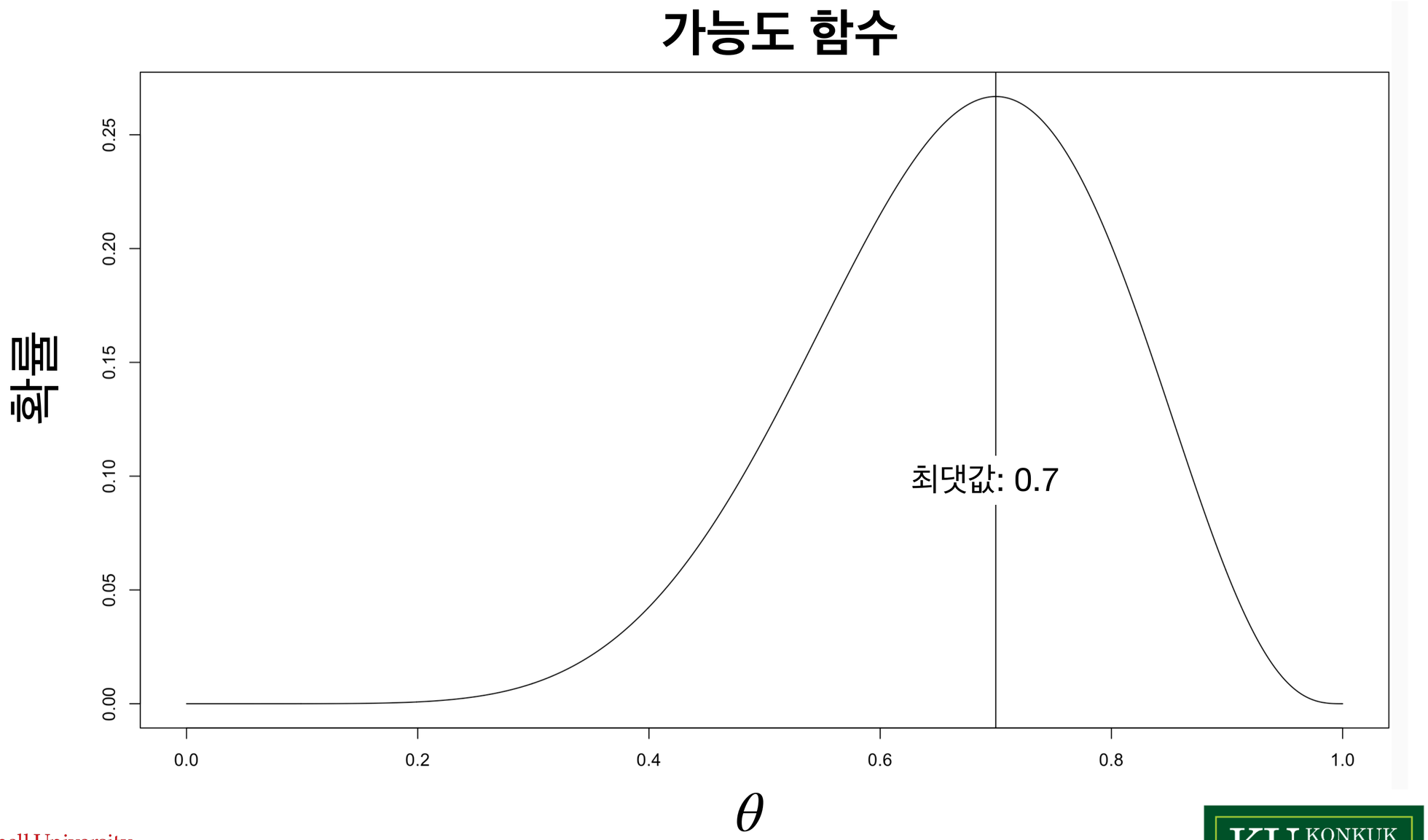
이항분포로 살펴보는 이산형 확률 변수

- **최대 가능도 추정(Maximum Likelihood Estimation, MLE)**

  ✓ 따라서 가능도 함수 (혹은 θ에 대한 함수) 다음과 같이 종종 표기됨.

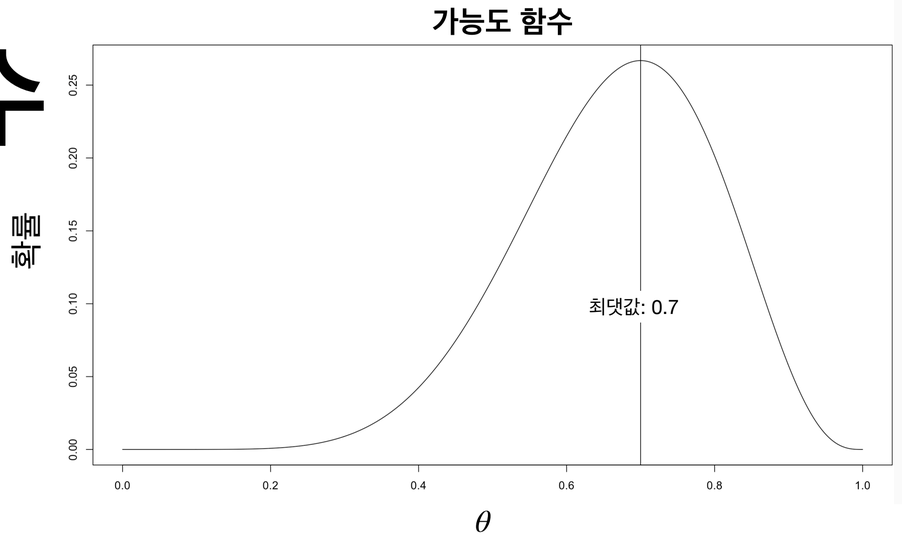
  ✓ $p(y \mid \theta) = p(k = 7, n = 10 \mid \theta) = \mathscr{L}(\theta)$

이항분포로 살펴보는 이산형 확률 변수

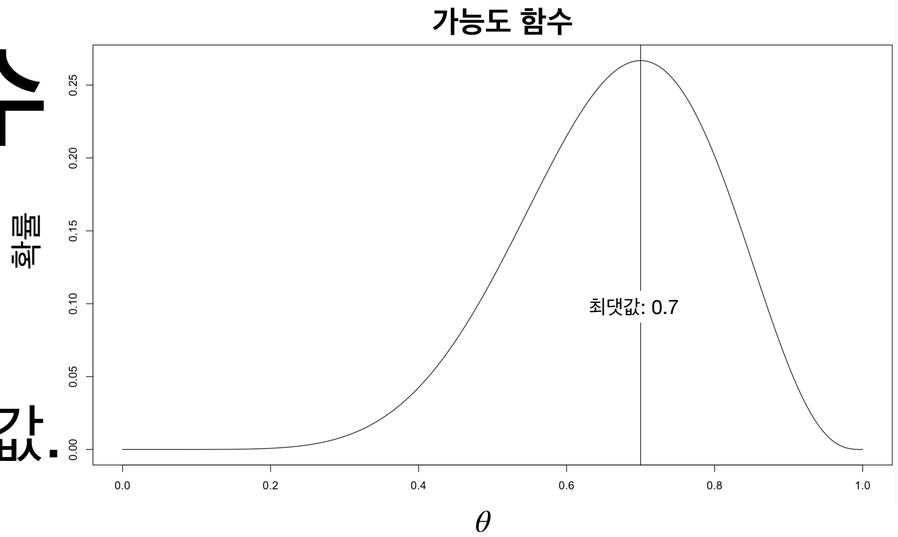


이항분포로 살펴보는 이산형 확률 변수

- 최대 가능도 추정



이항분포로 살펴보는 이산형 확률 변수



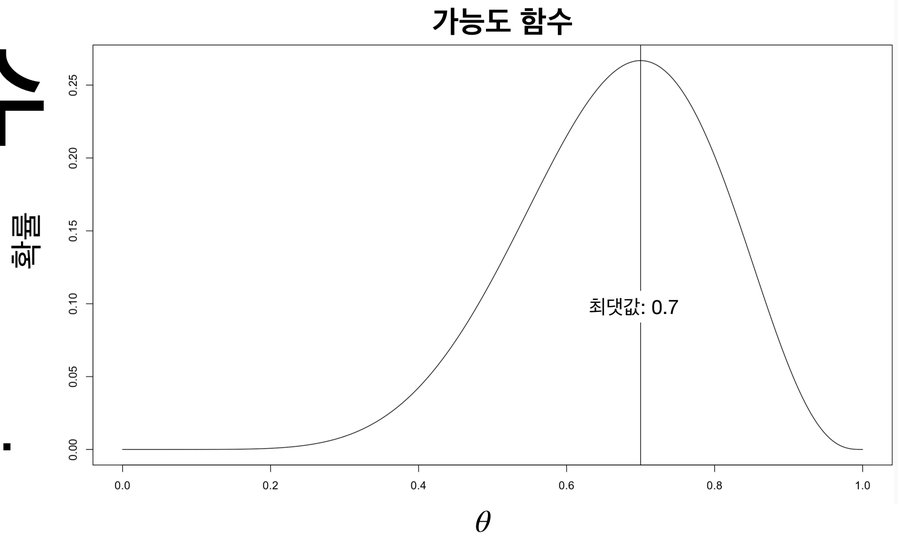
- **최대 가능도 추정**

  ✓ 어떤 데이터가 주어졌을 때, $\theta = 0.7$에서 최대 추정 확률값.

  ✓ 결국 최대 가능도 추정이란 어떤 데이터가 있을 때 모수 θ가 지닐 수 있는 가장 그럴 듯한 값을 의미.

  ✓ “가장 그럴 듯 한 값”은 늘 변할 수 있으며, 고정된 게 아님!

이항분포로 살펴보는 이산형 확률 변수



- **최대 가능도 추정**

✓ 어떤 데이터가 주어졌을 때, $\theta = 0.7$에서 최대 추정 확률값.

✓ 결국 최대 가능도 추정이란 어떤 데이터가 있을 때 매개변수 θ가 지닐 수 있는 가장 그럴 듯한 값을 의미.

✓ "가장 그럴 듯 한 값"은 늘 변할 수 있으며, 고정된 게 아님!

✓ 10번의 시행에서 1번의 성공을 얻었다면, 최대 가능도는 0.1!

✓ 심지어 $\theta = 0.5$더라도 10번의 시행에서 1번의 성공만 얻을 수도 있음! (cf. 동전 던지기)

✓ 물론, 시행이 무한에 가까울수록 θ에 대한 추정값은 점점 더 참에 가까워 질 수도 있음!

이항분포로 살펴보는 이산형 확률 변수

- **확률 분포의 유용성 - R을 통해 이항분포를 이루는 특정 결과에 대한 확률 계산하기**

$$Binomial(k \mid n, \theta) = \binom{n}{k} \theta^k (1-\theta)^{n-k} \text{ (단, } x = 0,1,2,\cdots,n)$$

✓ $k = 1, n = 5, \theta = 0.5$

✓ $Binomial(1 \mid 5, 0.5) = \binom{5}{1} 0.5^1 (1-0.5)^{5-1} = 0.15625$

✓ R에서는 dbinom() 함수 사용해서 얻을 수 있음.

이항분포로 살펴보는 이산형 확률 변수

- **확률 분포의 유용성 - R을 통해 이항분포를 이루는 특정 결과에 대한 확률 계산하기**

✓ $Binomial(1 \mid 5, 0.5) = \binom{5}{1} 0.5^1 (1 - 0.5)^{5-1} = 0.15625$

```
> dbinom(1, size = 5, prob = 0.5)
[1] 0.15625
```

이항분포로 살펴보는 이산형 확률 변수

- **확률 분포의 유용성 - R을 통해 이항분포를 이루는 특정 결과에 대한 확률 계산하기**

```
> dbinom(7, size = 10, prob = c(0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1))
```

이항분포로 살펴보는 이산형 확률 변수

- **확률 분포의 유용성 - R을 통해 이항분포를 이루는 특정 결과에 대한 확률 계산하기**

```
> dbinom(7, size = 10, prob = c(0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1))

[1] 0.000000000 0.000008748 0.000786432 0.009001692 0.042467328 0.117187500 0.214990848
    0.266827932
[9] 0.201326592 0.057395628 0.000000000
```

이항분포로 살펴보는 이산형 확률 변수

- **확률 분포의 유용성 - R을 통해 이항분포를 이루는 특정 결과에 대한 확률 계산하기**

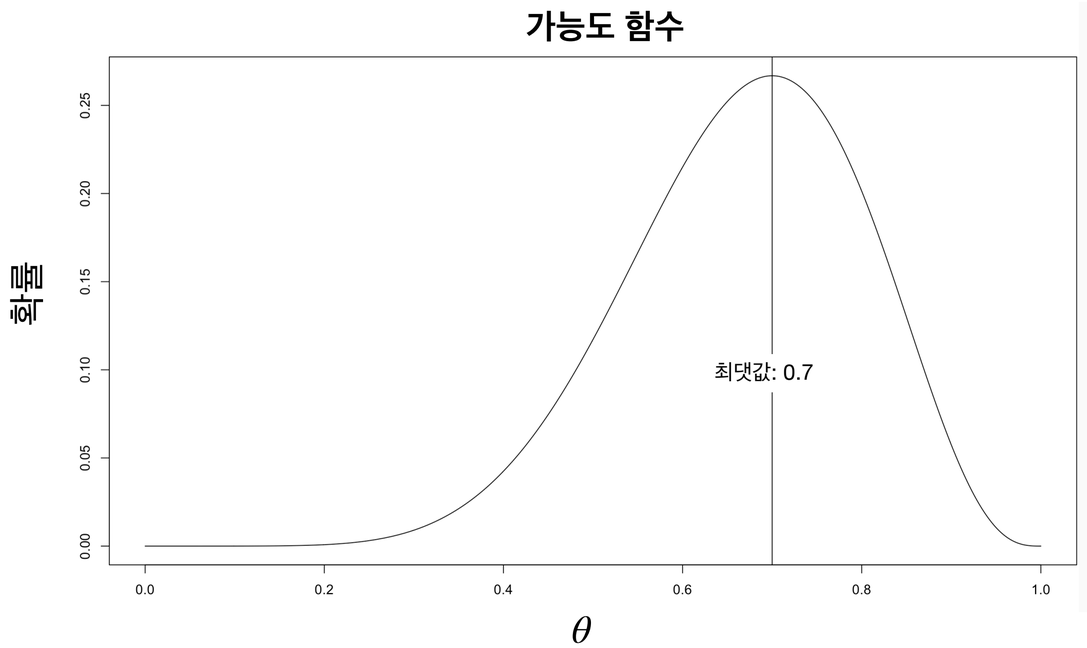
```
> dbinom(7, size = 10, prob = c(0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1))

[1] 0.000000000 0.000008748 0.000786432 0.009001692 0.042467328 0.117187500 0.214990848
    0.266827932
[9] 0.201326592 0.057395628 0.000000000
```

이항분포로 살펴보는 이산형 확률 변수

- **확률 분포의 유용성 - R을 통해 이항분포를 이루는 특정 결과에 대한 확률 계산하기**

```
> dbinom(7, size = 10, prob = c(0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1))
```

```
[1] 0.000000000 0.000008748 0.000786432 0.009001692 0.042467328 0.117187500 0.214990848
    0.266827932
[9] 0.201326592 0.057395628 0.000000000
```



이항분포로 살펴보는 이산형 확률 변수

- **확률 분포의 유용성 - R을 통해 이항분포를 이루는 특정 결과에 대한 확률 계산하기**

  ✓ 개개의 결과보다는 특정 결과 내에서의 확률 값을 구하려면?

  ✓ 예) $k \leq 2$인 경우의 확률값?

  ➡ 누적 확률 (Cumulative probability)

  ➡ $F(k) = P(Y \leq k)$

  ✓ “확률을 누적 시킨다는 게 뭔 소리죠…어떻게 구해요? ㅠㅠ”

이항분포로 살펴보는 이산형 확률 변수

- **확률 분포의 유용성 - R을 통해 이항분포를 이루는 특정 결과에 대한 확률 계산하기**

  ✓ 누적 = 더하기 (cf. 확률론 공리 3번째: 가산성)

  ✓ 따라서 이항 분포 함수에서 얻은 값들을 더하면 됨.

  ✓ $k \leq 2$인 경우의 누적 확률값?

  ➡ $\sum_{k=0}^{2} \binom{n}{k} \theta^k (1-\theta)^{n-k}$

이항분포로 살펴보는 이산형 확률 변수

- **확률 분포의 유용성 - R을 통해 이항분포를 이루는 특정 결과에 대한 확률 계산하기**

✓ $k \leq 2, n = 10, \theta = 0.5$일 때는…

➡ $\sum_{k=0}^{2} \binom{10}{k} 0.5^k (1 - 0.5)^{10-k}$

이항분포로 살펴보는 이산형 확률 변수

- **확률 분포의 유용성 - R을 통해 이항분포를 이루는 특정 결과에 대한 확률 계산하기**

✓ $k \leq 2, n = 10, \theta = 0.5$일 때는…

➡ $\sum_{k=0}^{2} \binom{10}{k} 0.5^k (1 - 0.5)^{10-k}$

```
> dbinom(0, size = 10, prob = 0.5) + dbinom(1, size = 10, prob = 0.5) + dbinom(2, size = 10, prob = 0.5)
[1] 0.0546875
```

이항분포로 살펴보는 이산형 확률 변수

- **확률 분포의 유용성 - R을 통해 이항분포를 이루는 특정 결과에 대한 확률 계산하기**

✓ $k \leq 2, n = 10, \theta = 0.5$일 때는…

➡ $\sum_{k=0}^{2} \binom{10}{k} 0.5^k (1 - 0.5)^{10-k}$

```
> dbinom(0, size = 10, prob = 0.5) + dbinom(1, size = 10, prob = 0.5) + dbinom(2, size = 10, prob = 0.5)
[1] 0.0546875

> sum(dbinom(0:2, size = 10, prob = 0.5))
[1] 0.0546875
```

이항분포로 살펴보는 이산형 확률 변수

- **확률 분포의 유용성 - R을 통해 이항분포를 이루는 특정 결과에 대한 확률 계산하기**

  ✓ 이것도 귀찮으면 pbinom() 함수를 쓰면 된다.

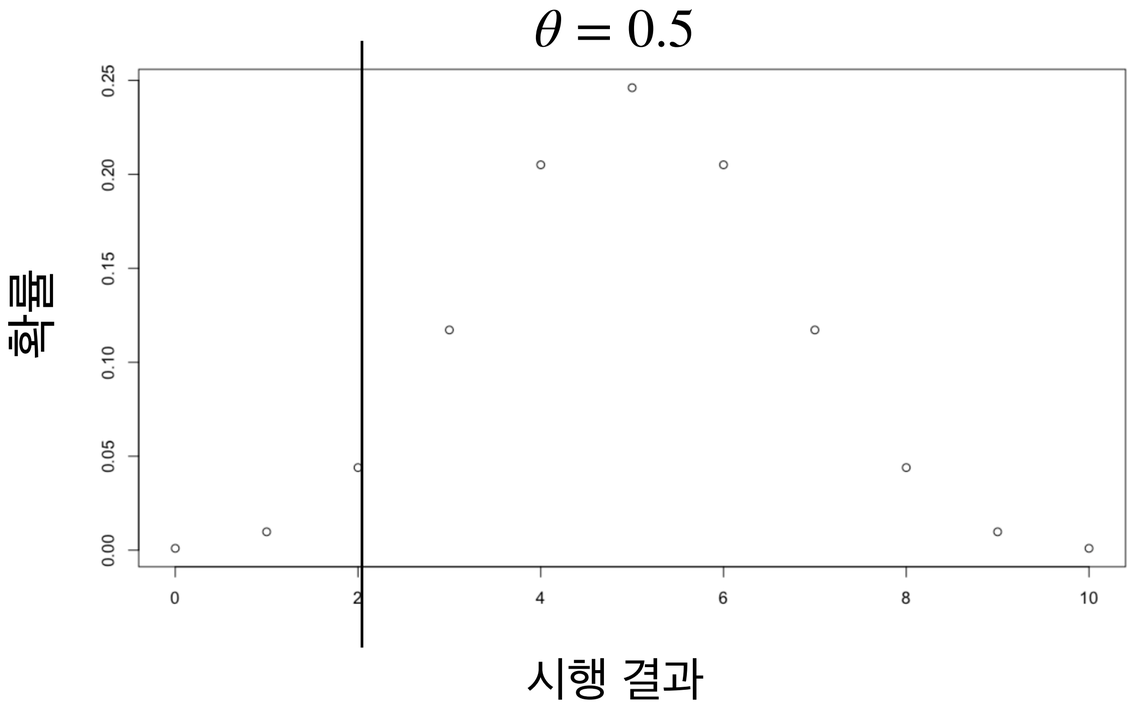
```
> pbinom(2, size = 10, prob = 0.5, lower.tail = TRUE)
[1] 0.0546875
```

이항분포로 살펴보는 이산형 확률 변수

- **확률 분포의 유용성 - R을 통해 이항분포를 이루는 특정 결과에 대한 확률 계산하기**

  ✓ 이것도 귀찮으면 pbinom() 함수를 쓰면 된다.

```
> pbinom(2, size = 10, prob = 0.5, lower.tail = TRUE)
[1] 0.0546875
```

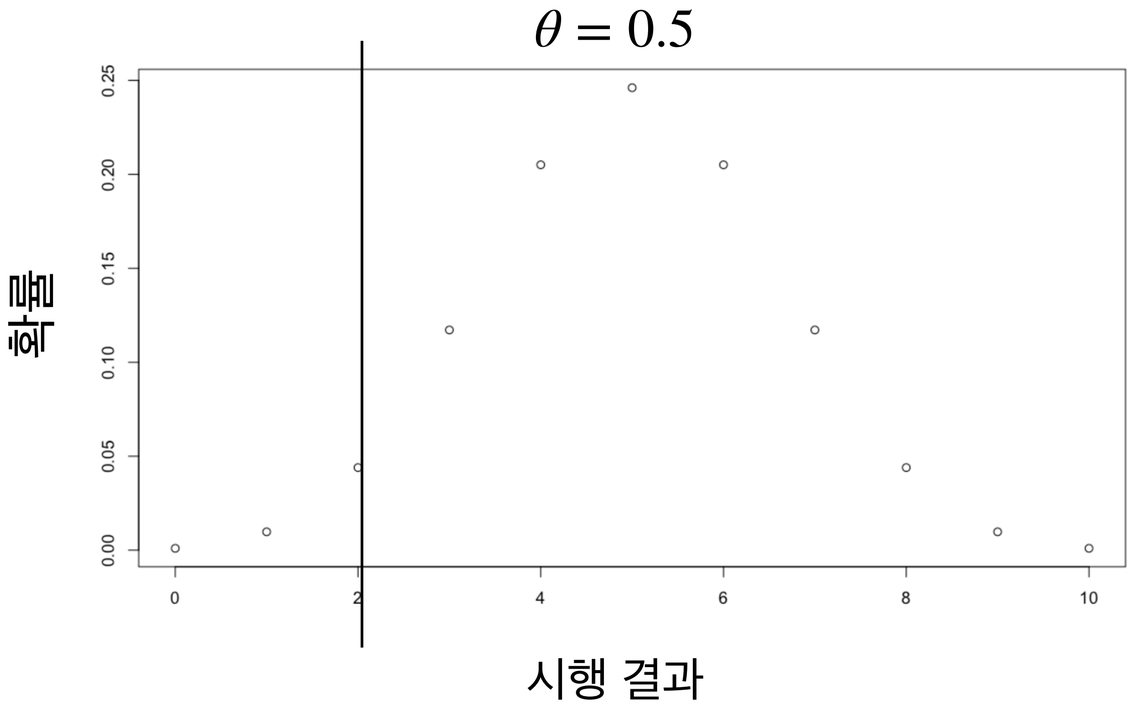


이항분포로 살펴보는 이산형 확률 변수

- **확률 분포의 유용성 - R을 통해 이항분포를 이루는 특정 결과에 대한 확률 계산하기**

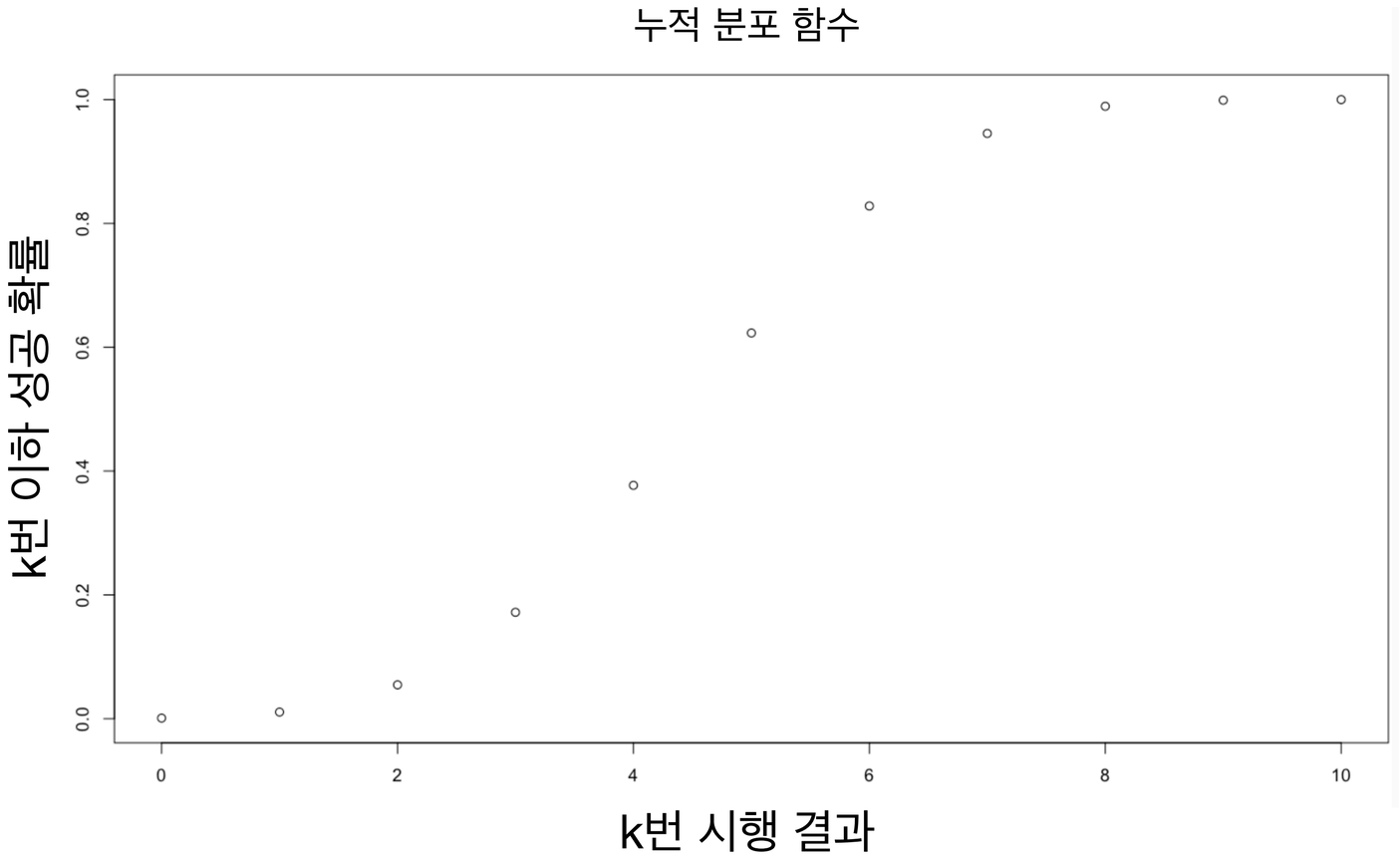
  ✓ 이것도 귀찮으면 pbinom() 함수를 쓰면 된다.

```
> pbinom(2, size = 10, prob = 0.5, lower.tail = FALSE)
[1] 0.9453125
```



이항분포로 살펴보는 이산형 확률 변수

- **확률 분포의 유용성 - R을 통해 이항분포를 이루는 특정 결과에 대한 확률 계산하기**



이항분포로 살펴보는 이산형 확률 변수

- **확률 분포의 유용성 - R을 통해 이항분포를 이루는 특정 결과에 대한 확률 계산하기**

  ✓ 반대로 어떤 성공확률을 알고 있으면, 이 때의 k 값이 언제인지 알 수 있음.

  ✓ 이 값은 qbinom() 함수가 구해준다.

  ✓ 그리고 이건 사실 누적 분포 함수의 역!

```
> qbinom(0.37, size = 10, prob = 0.5)
[1] 4
```

이항분포로 살펴보는 이산형 확률 변수

- **확률 분포의 유용성 - R을 통해 이항분포를 이루는 특정 결과에 대한 확률 계산하기**

  ✓ 한편, 이항 분포를 따르는 난수(random number)를 생성하려면 rbinom()을 쓰면 됨.

```
> rbinom(n = 10, size = 1, prob = 0.5)
 [1] 0 1 1 0 1 1 1 1 0 1
```

정규분포로 살펴보는 연속형 확률 변수

- **정규분포(Normal distribution)와 연속형 확률 변수**

  ✓ 오도문(garden-path sentence)에 대한 반응시간 데이터를 수집했다고 생각해봅시다.

  ✓ 이 데이터를 y라 합시다.

  ✓ 그리고 y는 정규분포를 따르는 확률밀도함수를 보이는 확률변수 Y에서 얻어졌다고 가정해 봅시다 (TMI: 사실 현실에서 반응시간 데이터는 정규분포를 따르지 않습니다).

  ✓ 이러한 데이터는 다음과 같이 표기됩니다.

  ➡ $Y \sim Normal(\mu, \sigma)$

정규분포로 살펴보는 연속형 확률 변수

- **정규분포(Normal distribution)와 연속형 확률 변수**

  ✓ $Y \sim Normal(\mu, \sigma)$ (cf. R은 표준편차로 정규분포를 정의)

  ➡ "확률변수 Y는 확률밀도함수 $Normal(\mu, \sigma)$을 갖고 있다."

  ✓ 실제 관찰된 (혹은 인위적으로 만든) 데이터는 y로 표기.

정규분포로 살펴보는 연속형 확률 변수

- **정규분포(Normal distribution)와 연속형 확률 변수**

  ✓ $Y \sim Normal(\mu, \sigma)$ (cf. R은 표준편차로 정규분포를 정의)

  ➡ "확률변수 Y는 확률밀도함수 $Normal(\mu, \sigma)$을 갖고 있다."

  ✓ 실제 관찰된 (혹은 인위적으로 만든) 데이터는 y로 표기.

  ✓ **중요 가정:** y는 독립항등분포(independent and identically distributed, i.i.d.)를 따른다.

  ➡ 즉, 데이터 y를 구성하는 각 관찰값들이 서로 영향을 주지 않고, 동일한 확률 분포를 따른다는 뜻! (각 실험참여자의 데이터를 개별적인 데이터로 여기는 이유!)

  ➡ $Y \overset{i.i.d}{\sim} Normal(\mu, \sigma)$

정규분포로 살펴보는 연속형 확률 변수

- **정규분포(Normal distribution)와 연속형 확률 변수**

  ✓ 정규 분포의 확률 밀도 함수는 다음과 같다.

$$Normal(y \mid \mu, \sigma) = f(y) = \frac{1}{\sqrt{2\mu\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

  ✓ 위 함수 계산은 R이 다 해주므로 걱정 안해도 됨!

정규분포로 살펴보는 연속형 확률 변수

- 정규분포(Normal distribution)와 연속형 확률 변수 - R에서 정규 분포 함수 정의하기

정규분포로 살펴보는 연속형 확률 변수

- **정규분포(Normal distribution)와 연속형 확률 변수 - R에서 정규 분포 함수 정의하기**
- 사실 대부분의 확률 분포 함수는 R에 이미 내장되어 있다.
- 그러나 이미 쓰여있는 것(=내장되어 있는 것)을 쓰는 것보단 실제로 이것이 어떻게 '쓰이게 되는지' 알아보면 R이 어떤 방식으로 확률을 계산하는지 어느 정도 감이 온다.
- 그 예시 중 하나인 정규 분포 함수를 직접 정의해보자

  cf. 이런 함수를 '사용자정의 함수'라 한다.

정규분포로 살펴보는 연속형 확률 변수

- **정규분포(Normal distribution)와 연속형 확률 변수 - R에서 정규 분포 함수 정의하기**

$$Normal(y \mid \mu, \sigma) = f(y) = \frac{1}{\sqrt{2\mu\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

정규분포로 살펴보는 연속형 확률 변수

- **정규분포(Normal distribution)와 연속형 확률 변수 - R에서 정규 분포 함수 정의하기**

$$Normal(y \mid \mu, \sigma) = f(y) = \frac{1}{\sqrt{2\mu\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

```
function(y = NULL, mu = 500, sigma = 100)
```

정규분포로 살펴보는 연속형 확률 변수

- **정규분포(Normal distribution)와 연속형 확률 변수 - R에서 정규 분포 함수 정의하기**

$$Normal(y \mid \mu, \sigma) = f(y) = \frac{1}{\sqrt{2\mu\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

function(y = NULL, mu = 500, sigma = 100)

정규분포로 살펴보는 연속형 확률 변수

- **정규분포(Normal distribution)와 연속형 확률 변수 - R에서 정규 분포 함수 정의하기**

$$Normal(y \mid \mu, \sigma) = f(y) = \frac{1}{\sqrt{2\mu\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

(1 / sqrt(2 * pi * sigma^2)) * exp(-(y - mu)^2 / (2 * sigma^2))

정규분포로 살펴보는 연속형 확률 변수

- **정규분포(Normal distribution)와 연속형 확률 변수 - R에서 정규 분포 함수 정의하기**

$$Normal(y \mid \mu, \sigma) = f(y) = \frac{1}{\sqrt{2\mu\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

```
(1 / sqrt(2 * pi * sigma^2)) * exp(-(y - mu)^2 / (2 * sigma^2))
```

정규분포로 살펴보는 연속형 확률 변수

- **정규분포(Normal distribution)와 연속형 확률 변수 - R에서 정규 분포 함수 정의하기**

$$Normal(y \mid \mu, \sigma) = f(y) = \frac{1}{\sqrt{2\mu\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

```
(1 / sqrt(2 * pi * sigma^2)) * exp(-(y - mu)^2 / (2 * sigma^2))
```

정규분포로 살펴보는 연속형 확률 변수

- **정규분포(Normal distribution)와 연속형 확률 변수 - R에서 정규 분포 함수 정의하기**

$$Normal(y \mid \mu, \sigma) = f(y) = \frac{1}{\sqrt{2\mu\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

```
(1 / sqrt(2 * pi * sigma^2)) * exp(-(y - mu)^2 / (2 * sigma^2))
```

정규분포로 살펴보는 연속형 확률 변수

- **정규분포(Normal distribution)와 연속형 확률 변수 - R에서 정규 분포 함수 정의하기**

$$Normal(y \mid \mu, \sigma) = f(y) = \frac{1}{\sqrt{2\mu\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

```
> my_normal
```

정규분포로 살펴보는 연속형 확률 변수

- **정규분포(Normal distribution)와 연속형 확률 변수 - R에서 정규 분포 함수 정의하기**

$$Normal(y \mid \mu, \sigma) = f(y) = \frac{1}{\sqrt{2\mu\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

```
> my_normal <- function(y = NULL, mu = 500, sigma = 100)
```

정규분포로 살펴보는 연속형 확률 변수

- **정규분포(Normal distribution)와 연속형 확률 변수 - R에서 정규 분포 함수 정의하기**

$$Normal(y \mid \mu, \sigma) = f(y) = \frac{1}{\sqrt{2\mu\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

```
> my_normal <- function(y = NULL, mu = 500, sigma = 100) {

}
```

정규분포로 살펴보는 연속형 확률 변수

- **정규분포(Normal distribution)와 연속형 확률 변수 - R에서 정규 분포 함수 정의하기**

$$Normal(y \mid \mu, \sigma) = f(y) = \frac{1}{\sqrt{2\mu\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

```
> my_normal <- function(y = NULL, mu = 500, sigma = 100) {
    (1 / sqrt(2 * pi * sigma^2))
}
```

정규분포로 살펴보는 연속형 확률 변수

- **정규분포(Normal distribution)와 연속형 확률 변수 - R에서 정규 분포 함수 정의하기**

$$Normal(y \mid \mu, \sigma) = f(y) = \frac{1}{\sqrt{2\mu\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

```
> my_normal <- function(y = NULL, mu = 500, sigma = 100) {
    (1 / sqrt(2 * pi * sigma^2)) * exp(-(y - mu)^2 / (2 * sigma^2))
}
```

정규분포로 살펴보는 연속형 확률 변수

- **정규분포(Normal distribution)와 연속형 확률 변수 - R에서 정규 분포 함수 정의하기**

$$Normal(y \mid \mu, \sigma) = f(y) = \frac{1}{\sqrt{2\mu\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

```
> my_normal <- function(y = NULL, mu = 500, sigma = 100) {
    (1 / sqrt(2 * pi * sigma^2)) * exp(-(y - mu)^2 / (2 * sigma^2))
}
```

✓ y는 밀리세컨드 단위 반응시간 데이터!

정규분포로 살펴보는 연속형 확률 변수

- **정규분포(Normal distribution)와 연속형 확률 변수 - R에서 정규 분포 함수 정의하기**

$$Normal(y \mid \mu, \sigma) = f(y) = \frac{1}{\sqrt{2\mu\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

```
> my_normal <- function(y = NULL, mu = 500, sigma = 100) {
    (1 / sqrt(2 * pi * sigma^2)) * exp(-(y - mu)^2 / (2 * sigma^2))
}
```

✓ y는 밀리세컨드 단위 반응시간 데이터!

```
> (y <- seq(100, 900, by = 0.01))
  [1] 100.00 100.01 100.02 100.03 100.04 100.05 …
```

정규분포로 살펴보는 연속형 확률 변수

- **정규분포(Normal distribution)와 연속형 확률 변수 - R에서 정규 분포 함수 정의하기**

$$Normal(y \mid \mu, \sigma) = f(y) = \frac{1}{\sqrt{2\mu\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

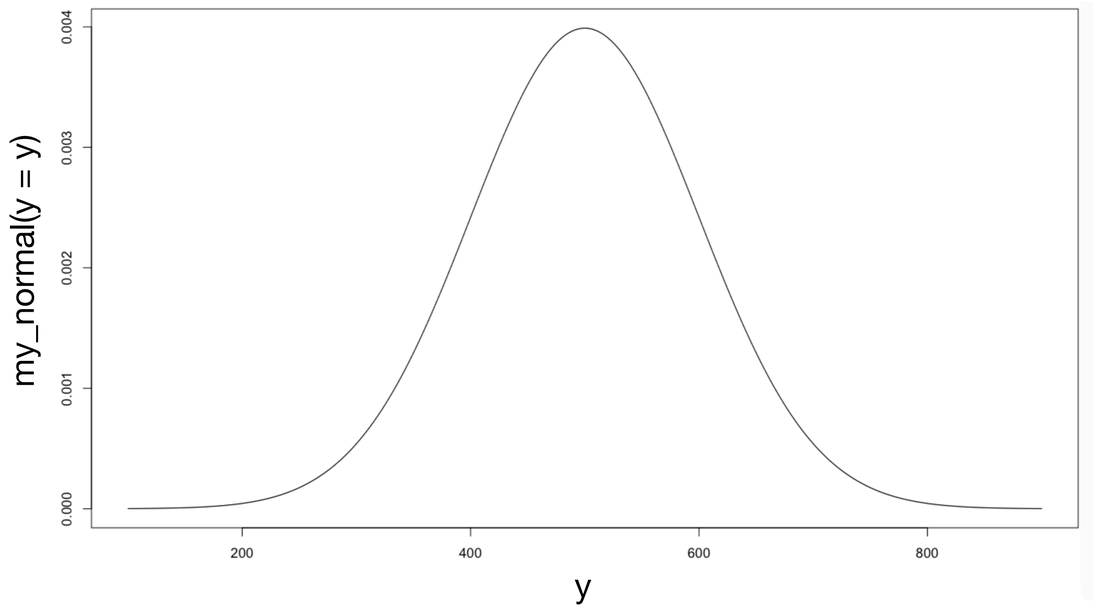
```
> my_normal(y = y)
  [1] 0.000001338302 0.000001338838 0.000001339373 0.000001339909 …
```

정규분포로 살펴보는 연속형 확률 변수

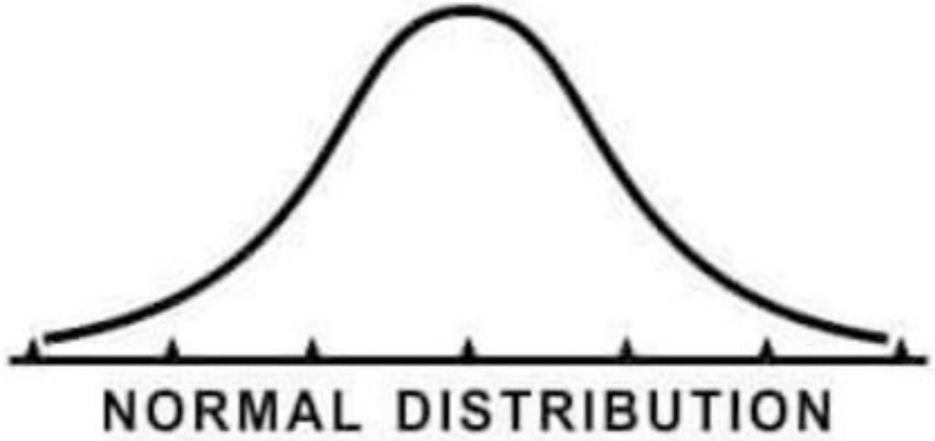
- **정규분포(Normal distribution)와 연속형 확률 변수 - R에서 정규 분포 함수 정의하기**

$$Normal(y \mid \mu, \sigma) = f(y) = \frac{1}{\sqrt{2\mu\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

```
> plot(y, my_normal(y = y), type = "l")
```



정규분포로 살펴보는 연속형 확률 변수



Cornell University

KU KONKUK UNIVERSITY

Brain & Cognition Research Center

정규분포로 살펴보는 연속형 확률 변수

- **정규분포(Normal distribution)와 연속형 확률 변수 - R에서 정규 분포 함수 정의하기**

  ✓ 이항 분포 함수를 계산하는데 사용했던 dbinom(), pbinom(), qbinom() 함수와 마찬가지로 R은 정규분포를 계산하는 함수가 이미 만들어져있음.

  ✓확률 밀도 함수: dnorm(n, mean = μ, sd = σ)

  ✓누적 밀도 함수: pnorm(n, mean = μ, sd = σ)

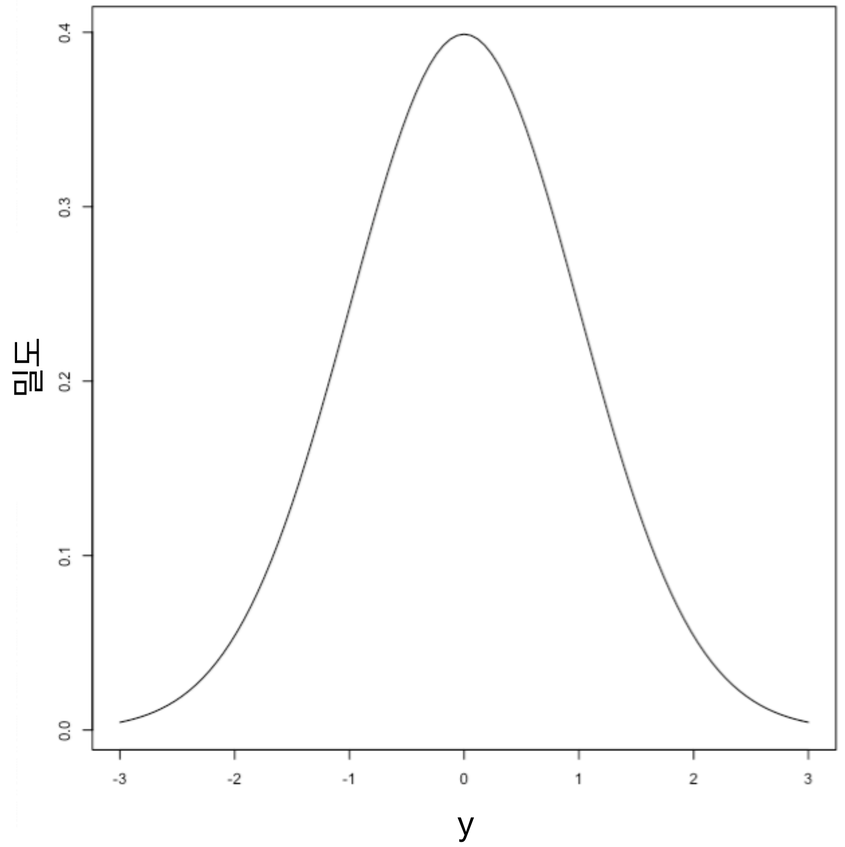
  ✓누적 밀도 함수의 역: qnorm(n, mean = μ, sd = σ)

  ✓난수 생성: rnorm(n, mean = μ, sd = σ)

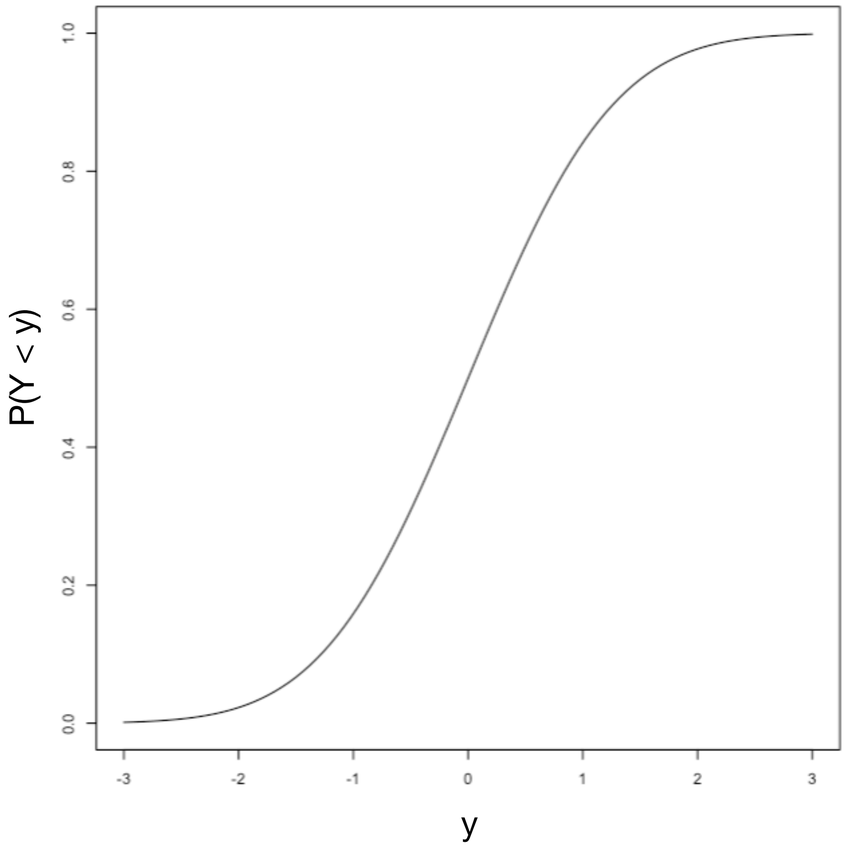
정규분포로 살펴보는 연속형 확률 변수

- **정규분포(Normal distribution)와 연속형 확률 변수 - R에서 정규 분포 함수 정의하기**

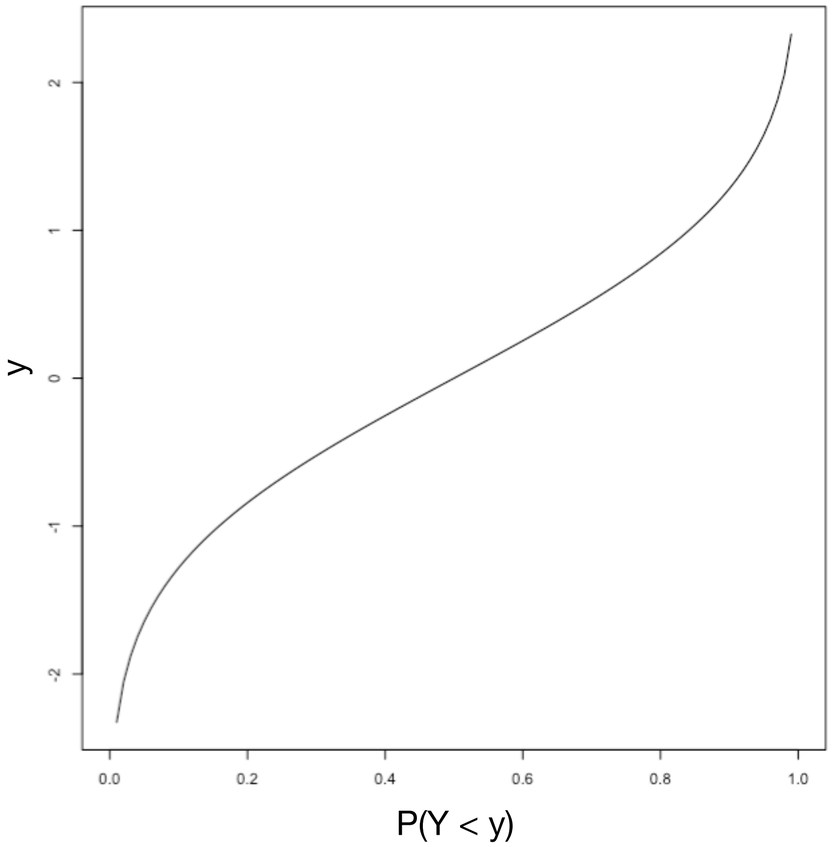
Y의 확률밀도함수 ~ Normal(0, 1)



Y의 누적밀도함수 ~ Normal(0, 1)



Y의 누적밀도함수의 역 ~ Normal(0, 1)



Cornell University

KU KONKUK UNIVERSITY

Brain & Cognition Research Center

정규분포로 살펴보는 연속형 확률 변수

- **정규분포(Normal distribution)와 연속형 확률 변수 - R에서 정규 분포 함수 정의하기**

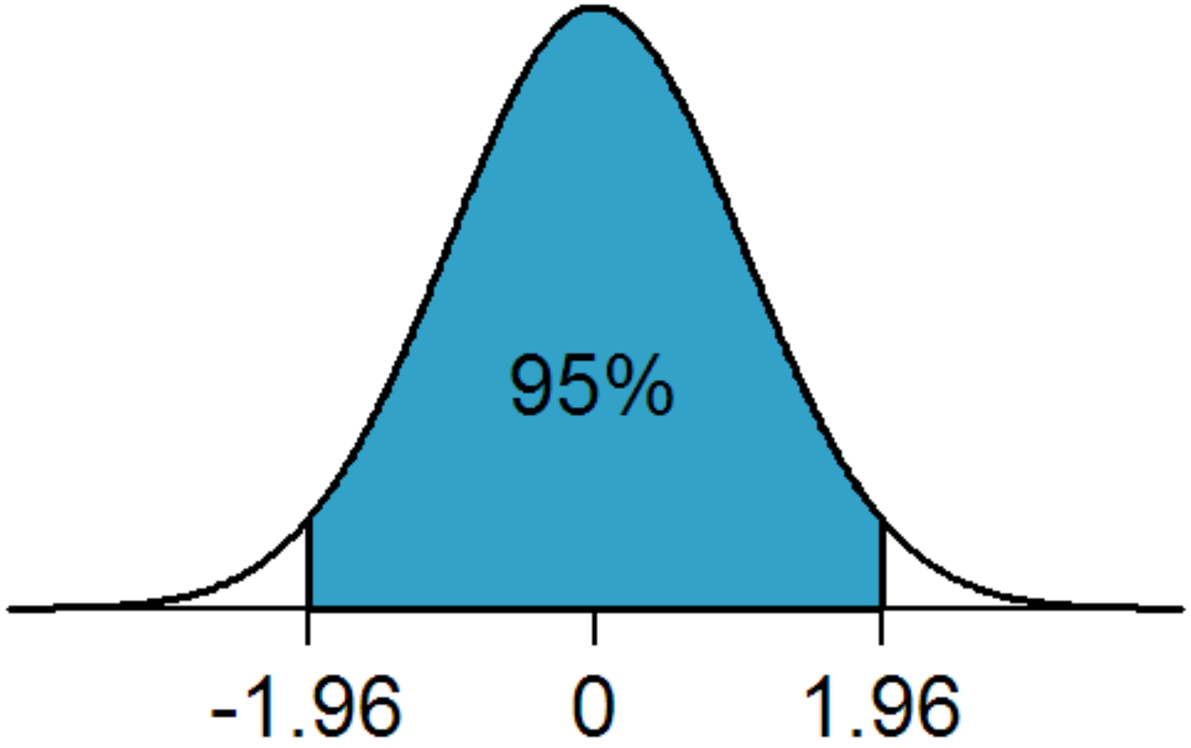
✓ 확률 밀도 함수는 데이터 y가 가질 수 있는 값들에 대해서 '밀도(density)'를 제공함.

✓ 여기서 말하는 밀도란, 확률을 말하는 것이 아니라 $f(y)$가 생성하는 값에 대해서 0을 포함한 양수를 가져다 준다는 것을 의미!

✓ 이산형 확률 변수와 마찬가지로, 누적 밀도 함수 $P(Y < y)$와 그 역 또한 동일하게 성립함.

정규분포로 살펴보는 연속형 확률 변수

- 정규분포(Normal distribution)와 연속형 확률 변수 - R에서 정규 분포 함수 정의하기



“P = .05는 1.96 혹은 2에 근접하다. 유의성 여부는 이 지점을 한계점으로 간주하여 결정하는 것이 편리하다.”

Statistical Methods for Research Workers (1925), p. 47

- 로널드 피셔 (1890 ~ 1962) -

정규분포로 살펴보는 연속형 확률 변수

- **평균 μ와 표준편차 σ를 가진 정규분포를 따르는 관찰 구간 [a, b]의 확률은?**

✓예) $\mu = 0, \sigma = 1, n = [-\infty, 1]$

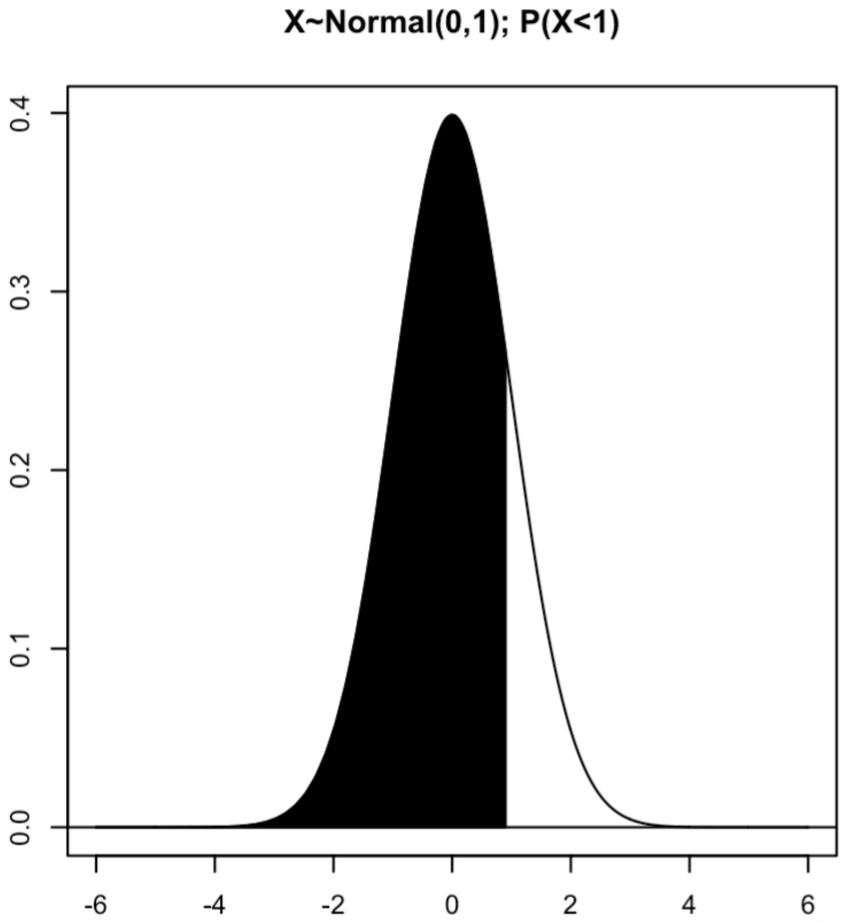
```
> pnorm(1, mean = 0, sd = 1)
[1] 0.8413447
```

정규분포로 살펴보는 연속형 확률 변수

- 평균 μ와 표준편차 σ를 가진 정규분포를 따르는 관찰 구간 [a, b]의 확률은?

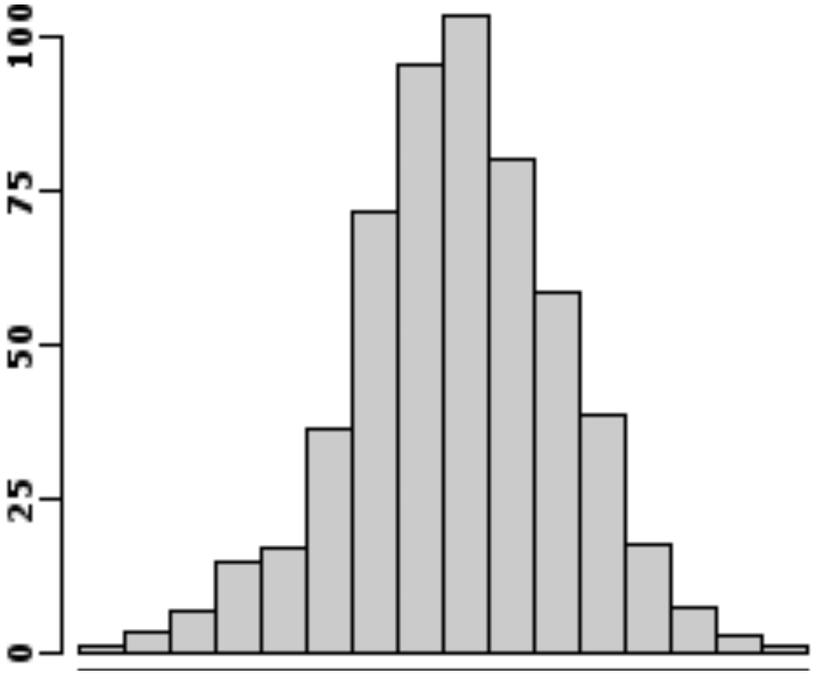
✓예) $\mu = 0, \sigma = 1, n = [-\infty, 1]$

```
> pnorm(1, mean = 0, sd = 1)
[1] 0.8413447
```



멈춰!

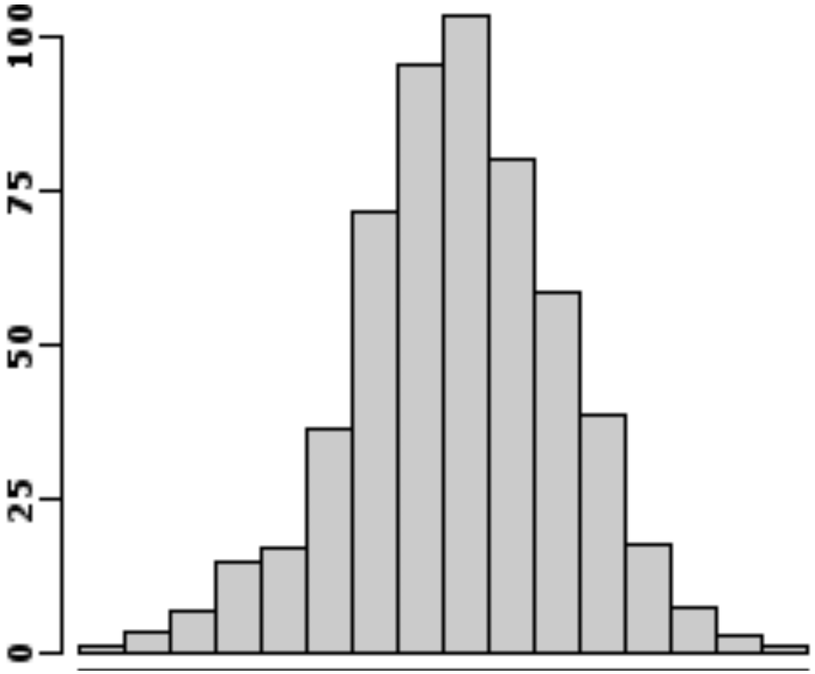
이산형, 연속형 확률 변수 - 확률밀도함수



이산형

연속형

이산형, 연속형 확률 변수 - 확률밀도함수



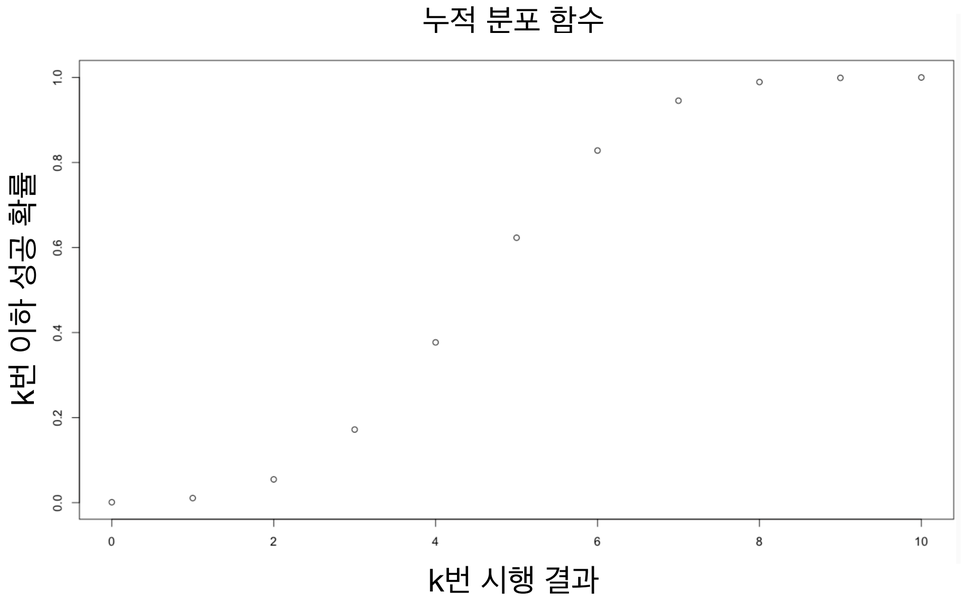
이산형

연속형

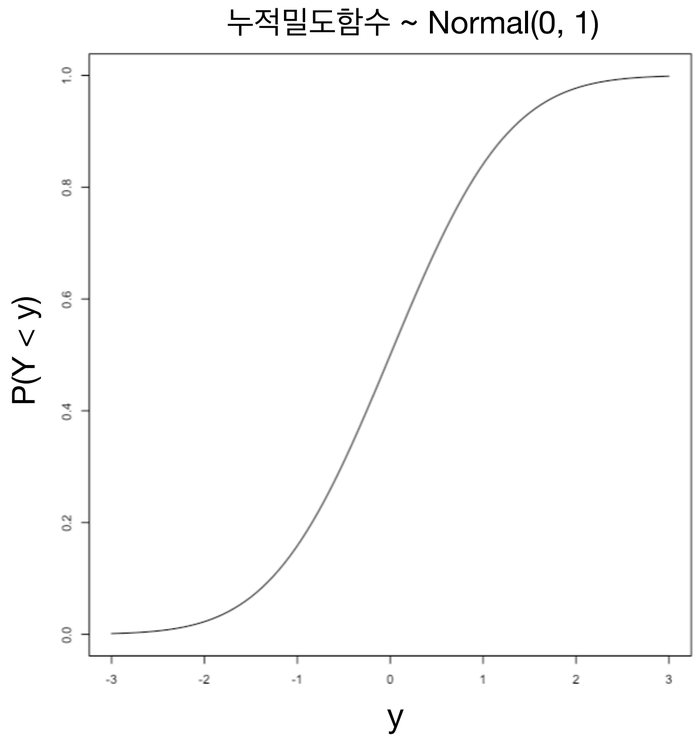
$$P(X \in A) = \sum_{x_i \in A} f(x_i)$$

$$P(X \in A) = \int_A f(x)dx$$

이산형, 연속형 확률 변수 - 누적분포함수

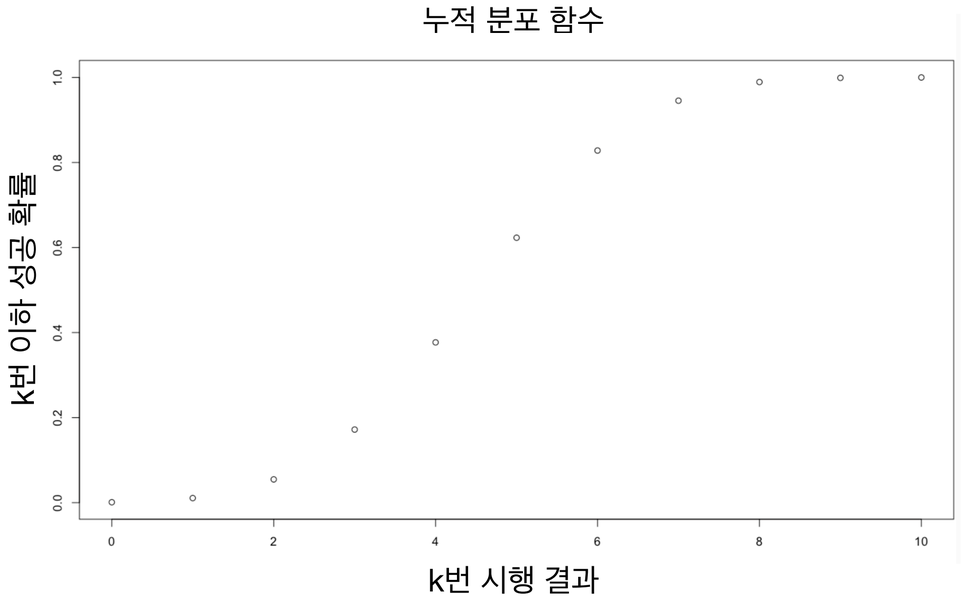


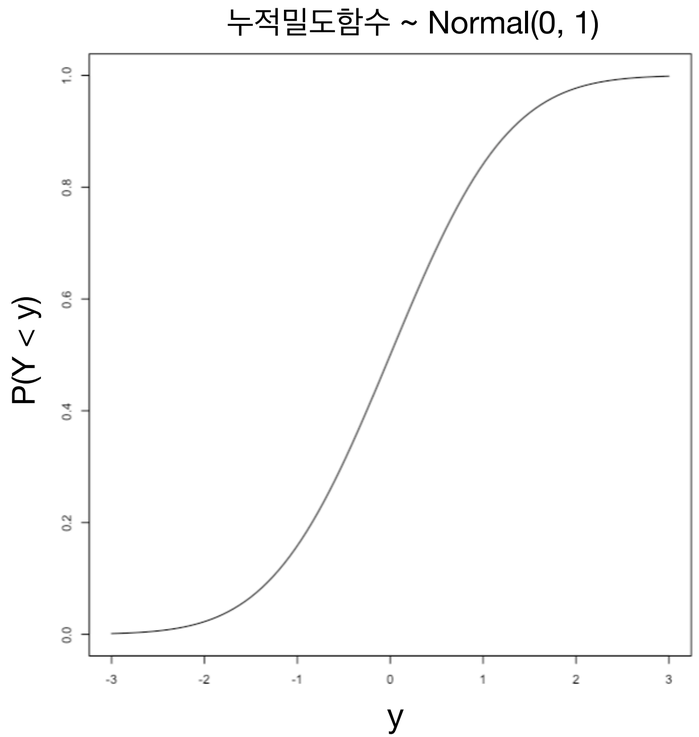
이산형



연속형

이산형, 연속형 확률 변수 - 누적분포함수



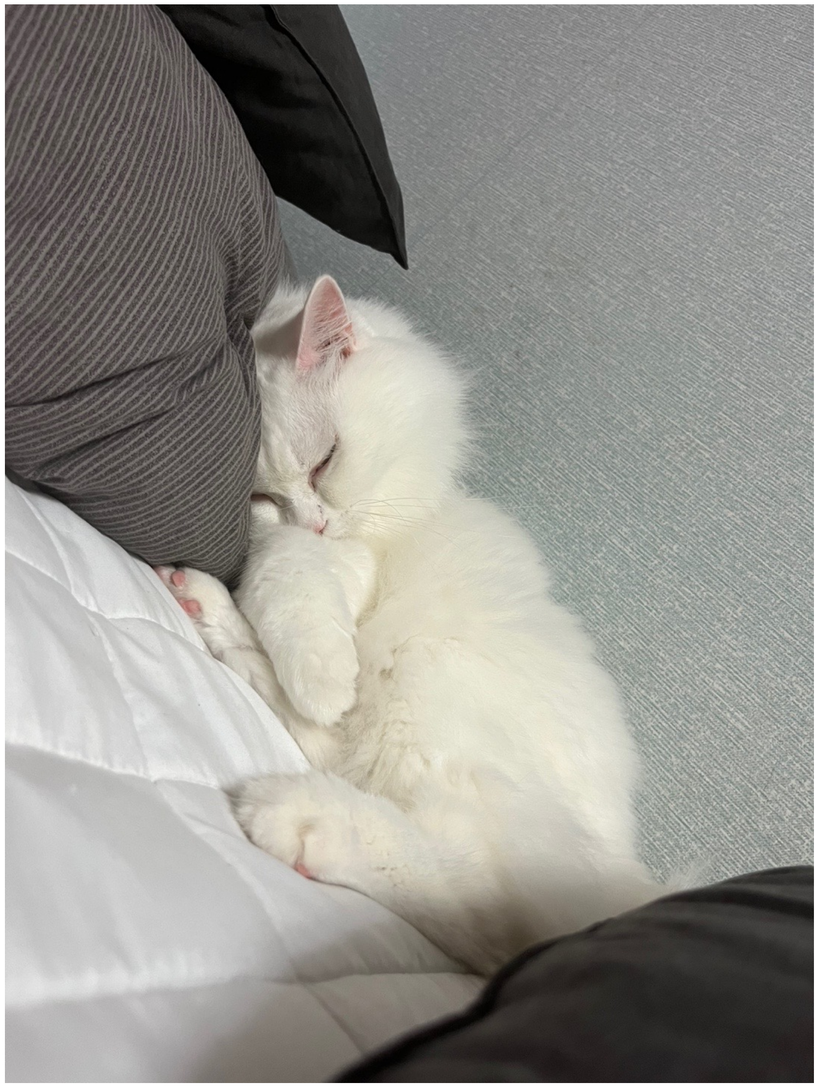


이산형

$$F(x) = \sum_{x_i \le x} f(x_i)$$

연속형

$$F(x) = \int_{-\infty}^{x} f(t)dt$$



쉬는 시간
(특별 게스트: 달리)

3. 알아두면 좋은 분포들

도대체 왜 (확률) 분포를 알아야할까?

- **알아두면 좋은 이유가 있다.**

  ✓ 확률 분포는 현실에서 수집된 데이터를 근사해주기 때문에 통계분석에 적극적으로 사용됨.

  ✓ 적극적으로 사용된다는 것은 다른 말로 하자면 확률 분포가 데이터를 설명하는데 있어 하나의 단위가 될 수 있다는 뜻!

  ➡ “특정 데이터가 t분포를 따를 때 t값이 2 이상이면 어떤 유의미한 것이 있다.”

  ✓ 유의미한 것?

  ➡ “문법적인 문장에 대한 읽기 시간과 비문법적인 문장에 대한 읽기 시간이 차이가 있다.”

  ➡ “특정 통계적 모형이 다른 모형보다 좋다.”

알아두면 좋은 분포들

- **알아두면 좋은 이유가 있다.**

  ✓ 이 파트에서 볼 확률 분포들의 구체적인 성질들을 외울 필요는 없음 (어차피 R이 다 계산)

  ✓ 또한 각 분포 마다 그 기능이 굉장히 다양하기 때문에 기능을 모두 열거할 수 없음.

  ✓ 그럼에도 불구하고, 이제부터 살펴볼 분포들은 수많은 통계분석 관련 해서 대부분의 경우 많고 중요한 역할을 하기 때문에 살펴볼 것임!

표준(Standard) 정규 분포

- **표준인 이유가 있다.**

  ✓ $Normal(\mu = 0, \sigma = 1)$

  ✓ 확률변수 $X_1, \cdots, X_n$이 독립항등분포(=서로 독립)이고 이것이 어떤 평균과 분산을 가진 분포를 따르게 된다면…

  ✓ n이 무한으로 갈수록 우리는 다음과 같은 변환된 확률 변수 Z의 분포를 얻을 수 있다.

표준 정규 분포 (Standard Normal Distribution)

- **표준인 이유가 있다.**

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

✓ $\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n}$ (= 평균)

✓ 그렇습니다. 바로 어디서 들어보셨을 법한 Z값/z-score/표준점수/표준값!

✓ 표준 정규 분포의 확률밀도함수

$$f(z) = \frac{1}{\sqrt{2\mu}} exp\left(-\frac{z^2}{2}\right)$$

t 분포 (t-distribution) - 적은샘플로 유의미한 차이 찾기

- **기네스 맥주 덕후 윌리엄 고셋이 만든 "학생(Student)의 t 분포"**

  ✓ 어떤 확률 변수 X가 있을 때, 아래 확률 질량 함수를 만족하면 이를 t 분포라 함.

$$f_X(x) = \frac{\Gamma(\frac{v+1}{2})}{\sqrt{\pi v}\Gamma(\frac{v}{2})}\left(1+\frac{x^2}{v}\right)^{-\frac{v+1}{2}} \quad \text{for } -\infty < t < \infty$$

  ✓ $X \sim t(v)$로 표기

  ✓ 여기서 v는 자유도(degree of freedom), $\Gamma(n)$는 감마 함수 $\Gamma(n) = (n-1)!$를 지칭 (생긴 것만 무섭지 별 거 아님! 그리고 계산은 R이 다 해준다)

  ✓ "정규분포에서 샘플수가 적을 땐 정규분포 모양이랑 완벽하게 똑같지 않더라"

  ✓ $\mu = 0$ (단, $v > 1$)

  ✓ $\sigma^2 = \frac{v}{v-2}$ (단, $v > 2$)

t 분포 (t-distribution) - 적은샘플로 유의미한 차이 찾기

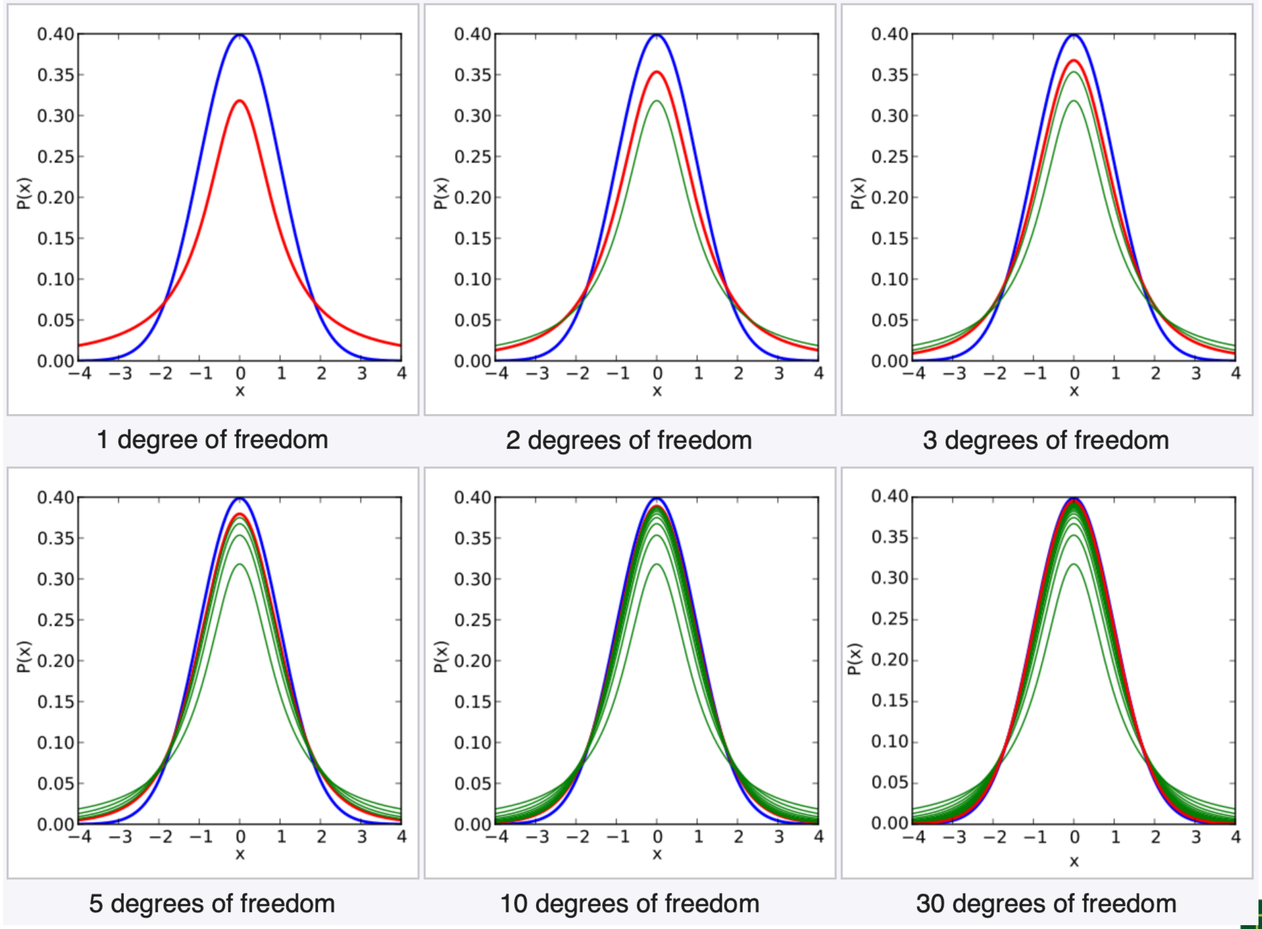
- **기네스 맥주 덕후 윌리엄 고셋이 만든 "학생(Student)의 t 분포"**

  ✓ 자유도 ν가 무한으로 갈수록 (=표본의 크기가 무한으로 갈수록) 표준 정규 분포에 가까워짐.

  (cf. $\nu = 1$이 되면 평균과 분산이 정의되지 않는 코시(Cauchy)/로렌츠(Lorentz) 분포가 됨)

t 분포 (t-distribution) - 적은 샘플로 유의미한 차이 찾기

- 기네스 맥주 덕후 윌리엄 고셋이 만든 “학생(Student)의 t 분포”



각 자유도에 따른 t 분포의 변화 (파란색: 표준 정규 분포)

t 분포 (t-distribution) - 적은 샘플로 유의미한 차이 찾기

- **기네스 맥주 덕후 윌리엄 고셋이 만든 "학생(Student)의 t 분포"**

  ✓ 사실 t 분포는 카이 제곱 및 표준 정규 분포와 관련 있음.

  ✓ 이는 나중에 다양한 가설 검정에 사용되는 분포로써도 활용되며, 다음과 같이 정의됨.

  ✓ $Z \sim Normal(0,1)$, $Y \sim \chi^2(v)$이고 서로 독립일 때, t 분포는 아래와 같다.

$$T = \frac{Z}{\sqrt{Y/v}}$$

카이 제곱 분포(Chi-square or χ^2 distribution) - 분산의 크고 작음 비교

- **온갖 검정에 사용되는 인싸 분포 (예: 가능도비 검정, 적합도 검정 등등)**

  ✓ 어떤 확률 변수 X가 있을 때, 아래 확률 질량 함수를 만족하면 이를 카이 제곱 분포라 함.

$$f(x;v) = \begin{cases} \dfrac{x^{\frac{v}{2}-1} exp(-\frac{x}{2})}{2^{\frac{v}{2}}\Gamma\left(\frac{v}{2}\right)}, & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

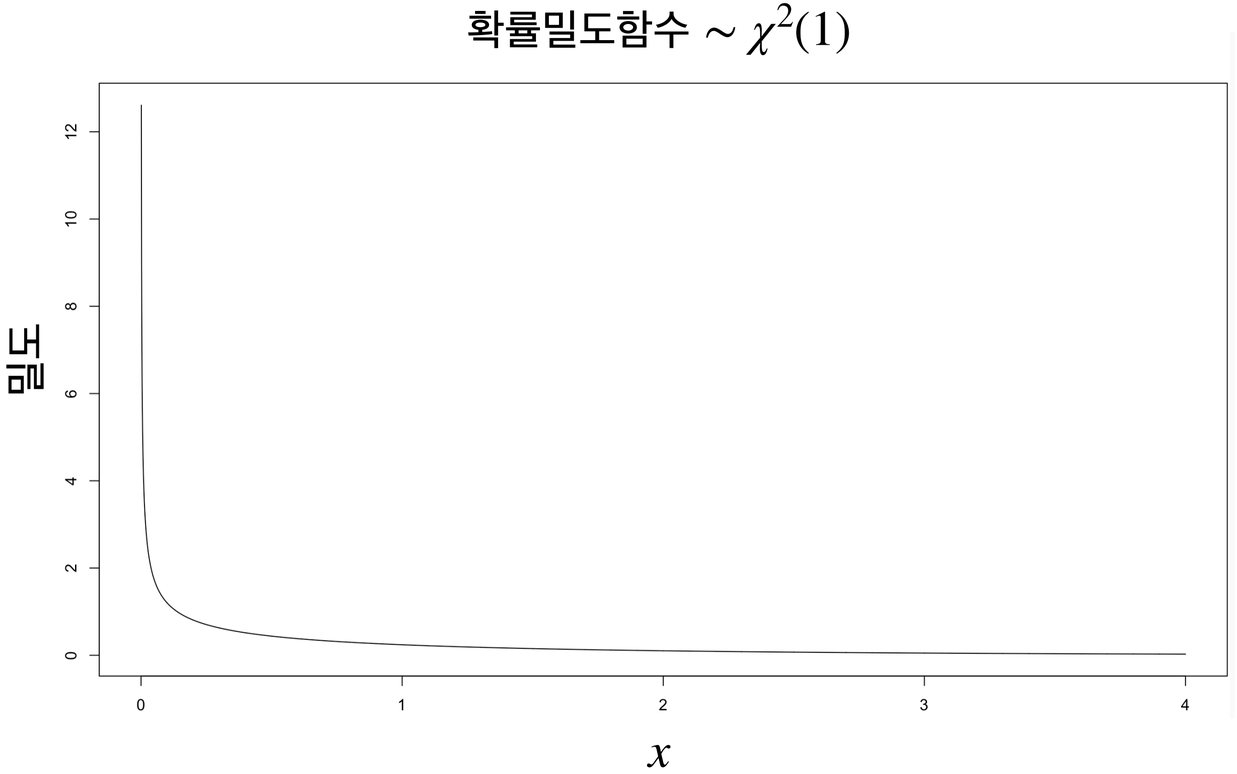
  ✓ $X \sim \chi^2(v)$로 표기

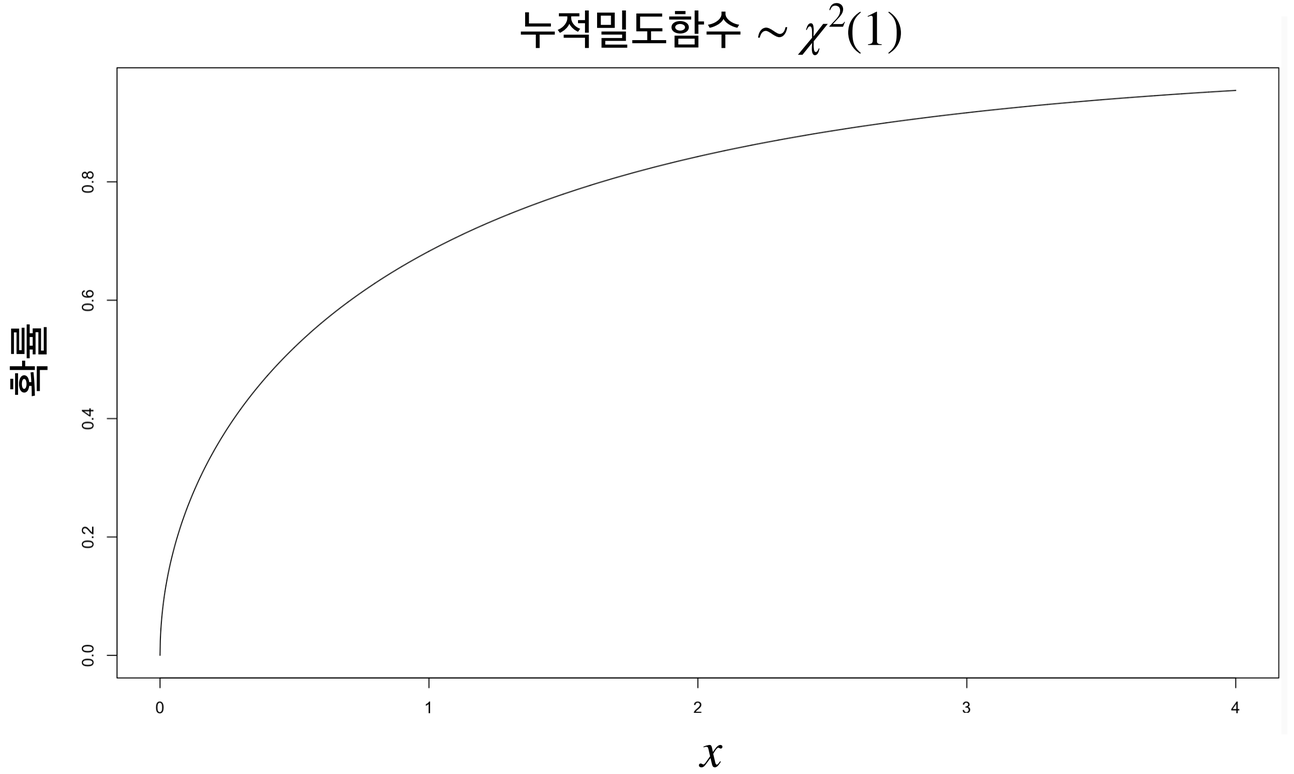
  ✓ $\mu = v$

  ✓ $\sigma^2 = v^2$

카이 제곱 분포(Chi-square or χ^2 distribution)

- 온갖 검정에 사용되는 인싸 분포 (예: 가능도비 검정, 적합도 검정 등등)





F 분포 (F distribution) - 서로 다른 두 개의 분산의 차이를 보기

- **분산분석과 회귀분석에 활용되는 그것!**

✓ 서로 독립인 카이 제곱 확률변수 U, V의 자유도가 각각 n, m이면 아래 F 분포를 따름.

$$X = \frac{U/n}{V/m}$$

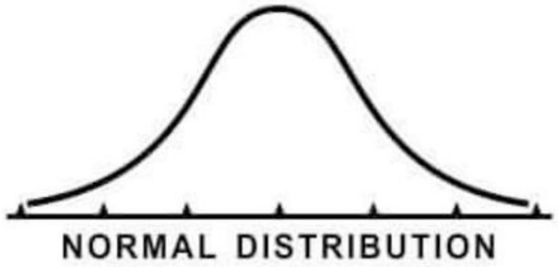
✓ 이 때, 그 확률 밀도 함수는 다음과 같다.

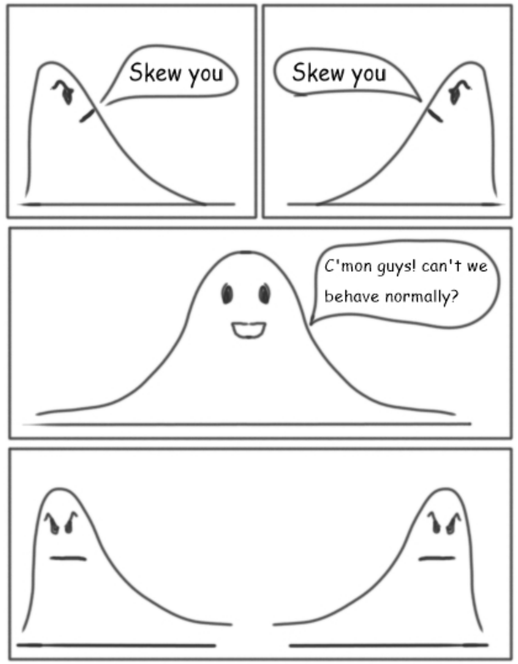
$$f_X(x) = \frac{\Gamma[(n+m)/2]}{\Gamma(n/2)\Gamma(m/2)} \left(\frac{n}{m}\right)^2 \frac{x^{(n-2)/2}}{[1+nx/m]^{(n+m)/2}} (x > 0)$$

✓ $X \sim F(n, m)$으로 표기.

✓ $\mu = \frac{m}{m-2}$ (단, $m > 2$)

✓ $\sigma^2 = \frac{2m^2(m+n-2)}{n(m-2)^2(m-4)}$ (단, $m > 4$)





실험심리언어학도를 위한 기초 확률과 통계

2-3. 표본분포와 중심극한정리

2022. 07. 14.

박기효

Cornell University

KU KONKUK UNIVERSITY

Brain & Cognition Research Center

모집단과 표본

- 우리는 지금까지 모집단과 표본을 종종 얘기했지만, 사실 엄밀하게 정의하지는 않았다.
- 비문법적인 문장에 대한 읽기 시간 데이터를 수집할 때, 현실적으로 지구에 있는 모든 인간들의 데이터를 수집하는 건 불가능에 가깝다.
- 따라서 우리는 일부 데이터, 즉, 표본을 활용해 모집단을 추정해야한다.

모집단과 표본

- **모집단(population)**

  ✓ 관심의 대상이 되는 모든 개체들의 관측값 또는 측정값의 집합

- **표본(sample)**

  ✓ 전체 관측값 대상으로부터 얻어진 일부

표본분포

- 표본을 통해 우리가 추정하고자 하는 정보는 모집단에 대한 정보!

  ✓ 하지만 대부분의 경우 우리는 절대로 모집단에 대한 정보를 온전히 알 수 없다!

  ✓ 그래서 추정한다.

- 이러한 정보를 우리는 모수라고 하며, 이 모수를 추정하기 위해 표본에서 얻은 정보를 통계량이라 한다.

표본분포

- **모수(Parameter)**

  ✓ 모집단에서 얻고자 하는 정보, 모집단에 관한 정보 (예: 모평균, 모분산 등)

- **통계량(Statistic)**

  ✓ 모집단의 일부인 표본에서 얻은 정보, 모수추정을 위한 확률 함수, 표본평균, 표본분산

표본분포

- 통계량은 모수와 일치할 수 없다.

  ✓ 왜? 우리는 추정하니깐!

  ✓ 따라서 통계량은 관찰되는 표본에 따라 달라 질 수 있으며, 각 표본에서 얻은 통계량 개개는 확률변수가 될 수 있으며 이에 따른 확률 분포를 가지게 된다.

  ✓ 그리고 이 통계량은 (암묵적으로) 독립이라고 여긴다!

  ✓ “1번 피험자와 2번 피험자 간 읽기 시간은 서로 다르다(=독립이다).”

  ➡ 이러한 확률 분포를 표본 분포(sampling distribution)이라 한다.

  ➡ 표본분포: 확률변수인 통계량이 가지는 분포 (예: 표본평균의 분포)

표본평균의 분포와 중심극한정리

- 모집단으로부터 크기가 n인 표본 $X_1, X_2, \cdots, X_n$을 추출했다고 하자, 이때 표본 평균은 다음과 같이 정의된다.

✓ $$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- 앞에서 우리는 표본은 서로 다를 수 있다(=독립이다)라고 얘기했으며, 따라서 표본은 사실상 하나의 확률변수다라고 얘기할 수 있다.

- 따라서 표본평균 또한 확률변수 X_i에 따라 값이 달라진다.

- 즉…

표본평균의 분포와 중심극한정리

- X_i가 정해지면 $\bar{X}$도 정해지며, 그 역 또한 성립한다.

  ✓ 좀 더 확률론스럽게 말하면, X_i가 확률변수이고 어떤 확률분포를 가진다면 $\bar{X}$ 또한 확률분포를 가진다.

- 이 때 표본평균 $\bar{X}$의 분포를 표본평균의 분포(sampling distribution of sample mean)이라 한다.

- **표본평균의 분포**

  ✓ 확률변수 X_i에 의하여 정의된 표본평균 $\bar{X}$의 분포

표본평균의 분포와 중심극한정리

- 표본평균의 분포 또한 확률 분포이므로, 여느 분포와 마찬가지로 평균(혹은 기댓값), 분산, 표준편차 등을 가진다.
- 재밌는 사실은 표본평균의 기댓값과 모평균은 동일하며, 표본평균의 분산은 모분산에 비례하며, 표본의 크기에 반비례한다는 것이다.
- 이러한 사실을 일반화한 것이 있는데, 이를 중심극한정리라 한다.
- 표본의 크기 n이 무한대에 접근함에 따라 모집단의 분포와 무관하게 표본평균과 모평균의 차이는 (표준)정규분포로 분포수렴한다.

표본평균의 분포와 중심극한정리

- **중심극한정리(Central limit theorem)**

  ✓ 표본의 크기 n이 무한대에 접근함에 따라 모집단의 분포와 무관하게 표본평균과 모평균의 차이는 (표준)정규분포로 분포수렴한다.

R로 보는 중심극한정리

- 스크립트

끝