강의 소개

데이터와 의사결정 | 정종빈

강사 소개: 정종빈

연세대학교 경영학과 학부/석사

스탠포드 Management Science & Engineering 박사
Decision Analysis Working Group
Computational Policy Lab

구글 Content recommendation/Ads metrics 계량 분석

세일즈포스IQ NLP 머신러닝 개발팀

우버 잇츠 Courier pricing 데이터사이언티스트

[현] 볼트 데이터사이언티스트

강의 목적

데이터를 활용한 의사결정의 다양한 접근 방법

불확실성(확률), 계량, 머신러닝, 인과관계분석

예제를 중심으로 실무에 도움이 되게

직관적인 이해와 활용에 초점

1부: 불확실성과 데이터

- 의사결정 모형
- 불확실성 계량화 전략
  - 빈도주의(Frequentist)
  - 베이지안
- 최적 의사결정

2부: ML 예측과 인과관계 분석

- ML 예측모형 기초
- 의사결정과 인과관계 분석

강의 개요

선수 지식

확률 분포에 대한 기본적인 이해

이항분포, 정규분포, 베타 분포 … ?

수리적 코딩(e.g., R, numpy, julia)에 대한 기초 지식

강의에서는 python/numpy로 실습

호기심 😅

속 시원한 해법 보다는 답 없는 문제 제기

참고자료

- Foundations of Decision Analysis
  by Ron Howard and Ali Abbas
- Data Analysis Using Regression
  and Multilevel/Hierarchical Models
  by Andrew Gelman and Jennifer Hill
- Causal Inference
  for Statistics, Social, and Biomedical Sciences
  by Guido Imbens and Donald Rubin
- MS&E 125: Introduction to Applied statistics
  https://github.com/stanford-policylab/mse125
- MS&E 226: Fundamentals of Data Science
  https://web.stanford.edu/class/msande226/

1부: 불확실성과 데이터

예제 소개 및 강의 개요

데이터와 의사결정 | 정종빈

1부: 불확실성과 데이터

- 의사결정 모형
- 불확실성 계량화 전략
  - 빈도주의(Frequentist)
  - 베이지안
- 최적 의사결정

2부: ML 예측과 인과관계 분석

- ML 예측모형 기초
- 의사결정과 인과관계 분석

강의 개요

예제 소개

아이폰일까 안드로이드일까?

친구와 지하철 타고 이동 중

친구와 지하철 타고 이동 중

다른 승객 중 10명은 핸드폰을 들여다 보고 있음

친구와 지하철 타고 이동 중

다른 승객 중 10명은 핸드폰을 들여다 보고 있음

10명 중 4명은 아이폰, 6명은 안드로이드 유저

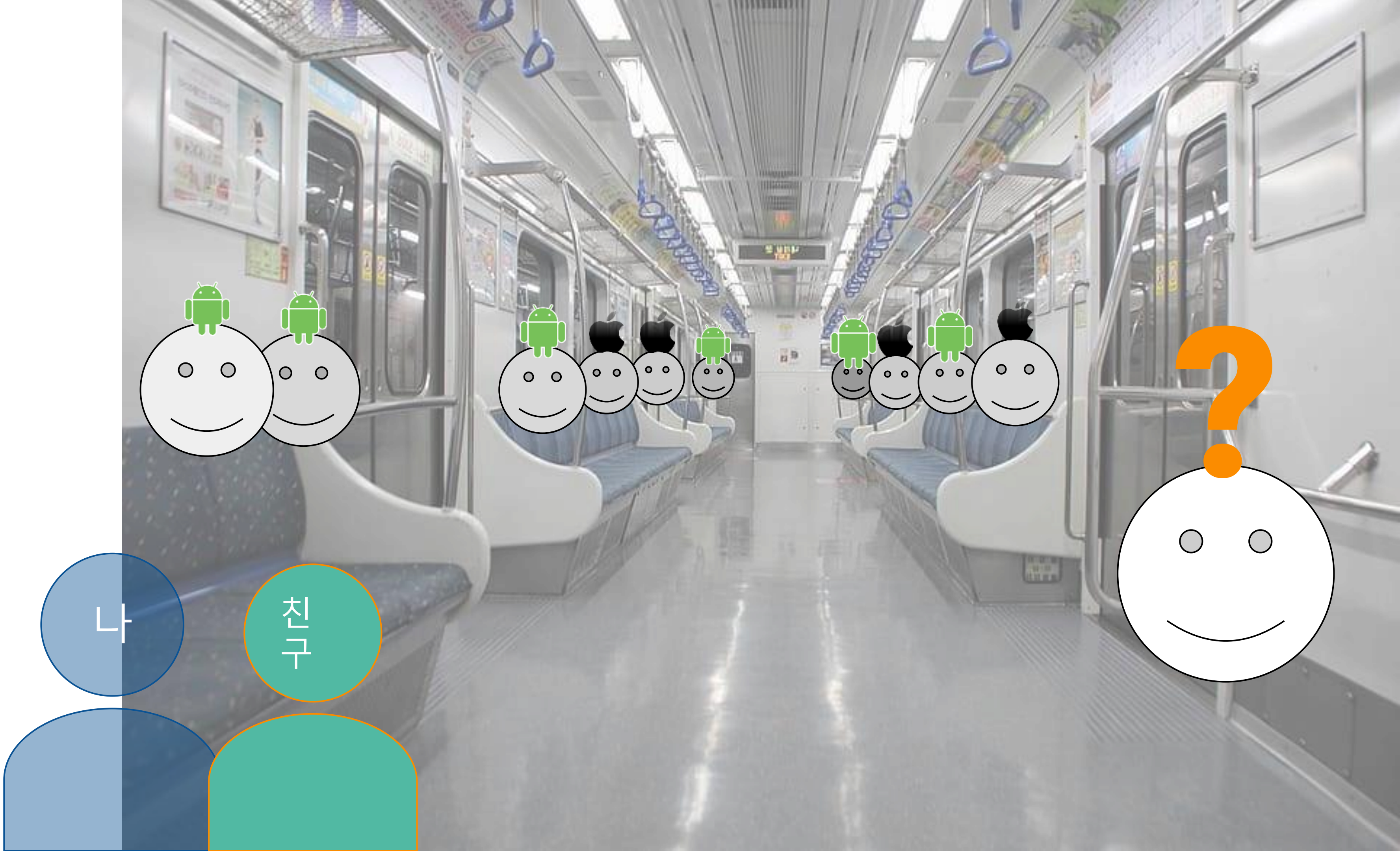
친구와 지하철 타고 이동 중

다른 승객 중 10명은 핸드폰을 들여다 보고 있음

10명 중 4명은 아이폰, 6명은 안드로이드 유저

마침, 폰을 아직 꺼내지 않은 한 승객

친구와 지하철 타고 이동 중
다른 승객 중 10명은 핸드폰을 들여다 보고 있음
10명 중 4명은 아이폰, 6명은 안드로이드 유저
마침, 폰을 아직 꺼내지 않은 한 승객


나
친구

여러분은 어떻게 하시겠습니까?

여러분은 어떻게 하시겠습니까?

“두 가지 경우의 수 = 50% 확률?”

“안드로이드 유저일 확률이
6/10 = 60% 니까 … ”

여러분은 어떻게 하시겠습니까?

“두 가지 경우의 수 = 50% 확률?”

“안드로이드 유저일 확률이
6/10 = 60% 니까 … ”

“핸드폰 시장 점유율을
생각해보면 … ”

여러분은 어떻게 하시겠습니까?

“두 가지 경우의 수 = 50% 확률?”

“안드로이드 유저일 확률이
6/10 = 60% 니까 … ”

“핸드폰 시장 점유율을
생각해보면 … ”

여러분은 어떻게 하시겠습니까?

“만원은 좀 많고 … 오백원만?”

“두 가지 경우의 수 = 50% 확률?”

"지금 점심 시간인데, 이 시간에 지하철 타는 사람이면 … "

"안드로이드 유저일 확률이 6/10 = 60% 니까 … "

"랩탑은 애플 쓰는거 같은데?"

"핸드폰 시장 점유율을 생각해보면 … "

여러분은 어떻게 하시겠습니까?

"신촌역에서 탔으니까 … "

"만원은 좀 많고 … 오백원만?"

"두 가지 경우의 수 = 50% 확률?"

"핸드폰 안 꺼낸 다른 승객들 … "

1부 강의 목적 및 개요

체계적이고 일반적인 <u>의사결정 모형</u> 소개

불확실성을 계량화하는 두 가지 접근법:

- 빈도주의(frequentist) 통계
  Estimand, estimator, estimate, 신뢰구간, bootstrap
- 베이지안(Bayesian) 확률
  Prior, likelihood, posterior (손계산 → 복잡한 모델링)

다음: 의사결정 모형

1부: 불확실성과 데이터

의사결정 모형

1부 강의 목적 및 개요

체계적이고 일반적인 <u>의사결정 모형</u> 소개

불확실성을 계량화하는 두 가지 접근법:

- 빈도주의(frequentist) 통계
  - Estimand, estimator, estimate, 신뢰구간, bootstrap
- 베이지안(Bayesian) 확률
  - Prior, likelihood, posterior (손계산 → 복잡한 모델링)

의사결정 모형

기초 모델링

- 의사결정의 3요소
- 의사결정 나무 만들기
- 의사결정 나무 계산

의사결정 모형

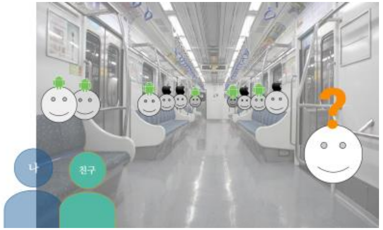
기초 모델링

- 의사결정의 3요소
- 의사결정 나무 만들기
- 의사결정 나무 계산

의사결정의 3요소

의사결정의 3요소

선택지
Alternatives



의사결정의 3요소

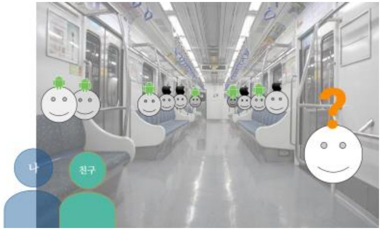
선택지
Alternatives

1. 내기를 할까?
   한다 / 안한다
2. 어디에 걸까?
   안드로이드 / 아이폰

의사결정의 3요소

선택지
Alternatives

불확실성/정보
Uncertainty/information



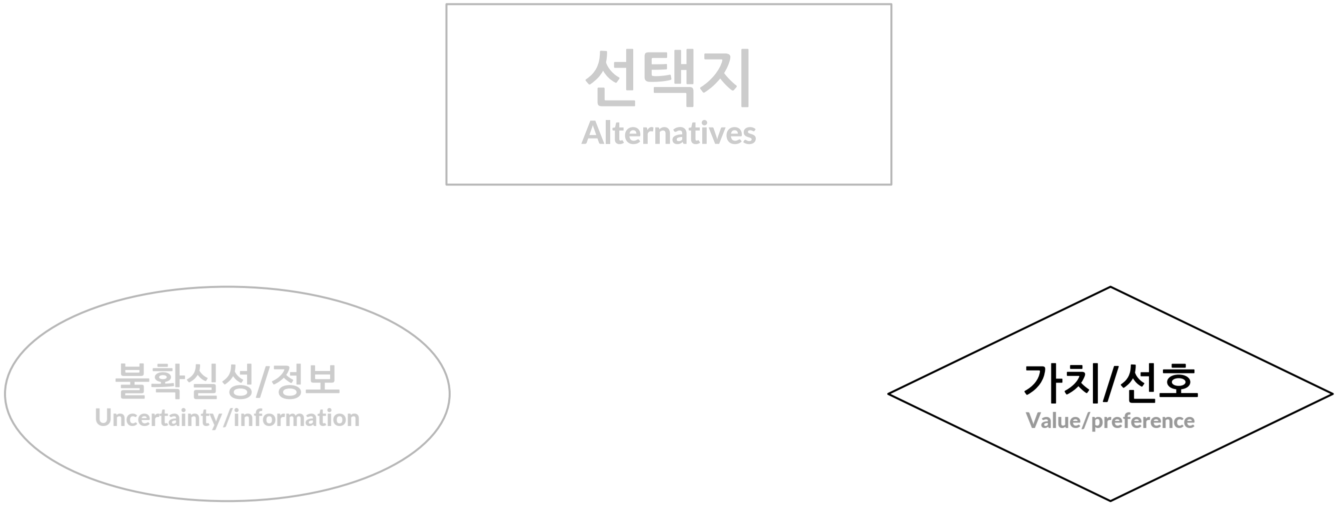
의사결정의 3요소

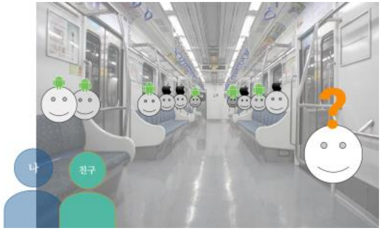
선택지
Alternatives

불확실성/정보
Uncertainty/information

1. 불확실성: 안드로이드? 아이폰?
2. 정보: 10명 중 6명은 안드로이드 …

의사결정의 3요소


선택지
Alternatives
불확실성/정보
Uncertainty/information
가치/선호
Value/preference



의사결정의 3요소

선택지
Alternatives

불확실성/정보
Uncertainty/information

1. 안해: 0
2. 맞춰: + 만원
3. 틀려: - 만원

가치/선호
Value/preference

의사결정의 3요소

선택지
Alternatives

불확실성/정보
Uncertainty/information

가치/선호
Value/preference

의사결정의 3요소 → 의사결정 나무

의사결정 모형

기초 모델링

- 의사결정의 3요소
- 의사결정 나무 만들기
- 의사결정 나무 계산

선택지

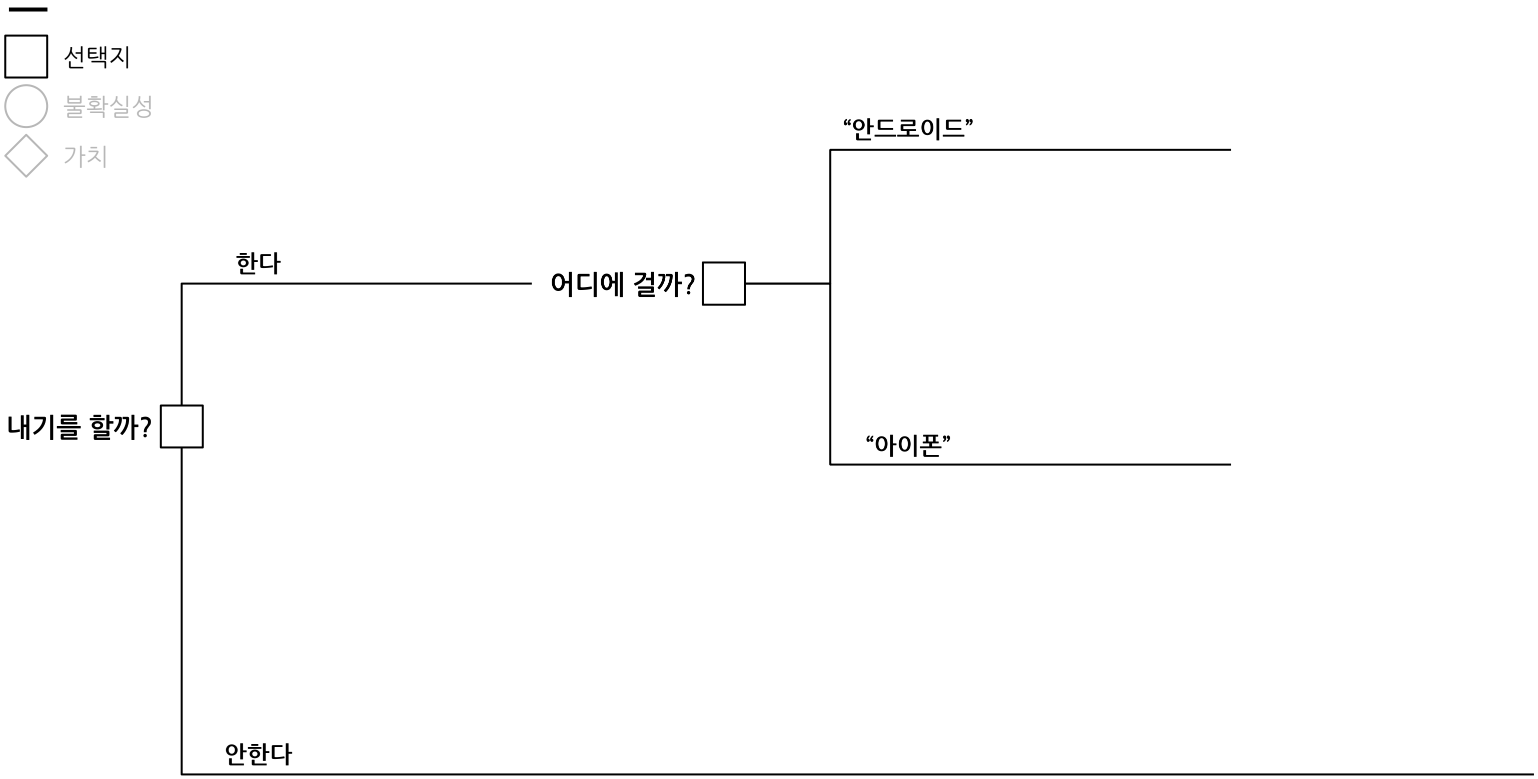
불확실성

가치

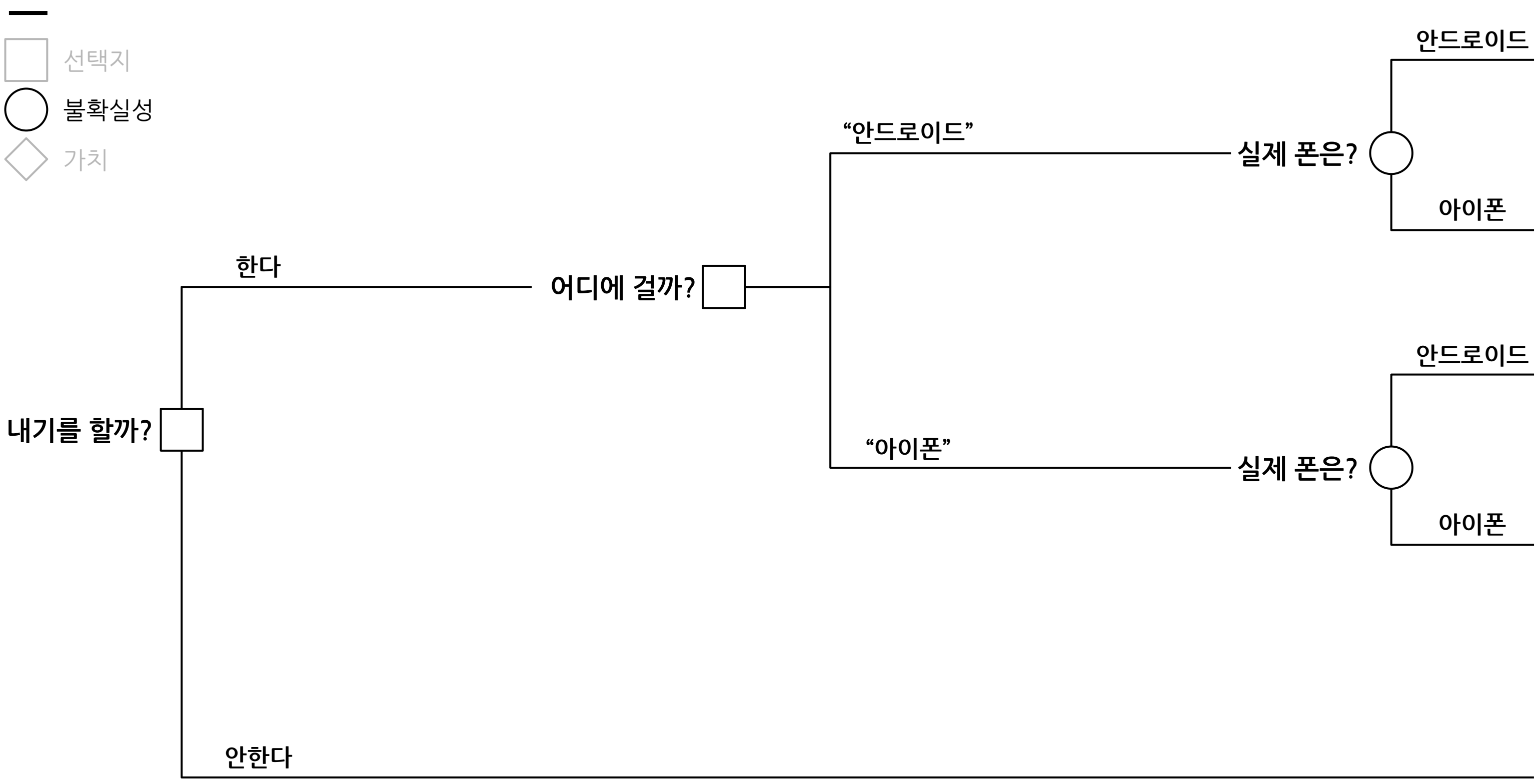
의사결정 나무 (Decision tree)

선택지
불확실성
가치

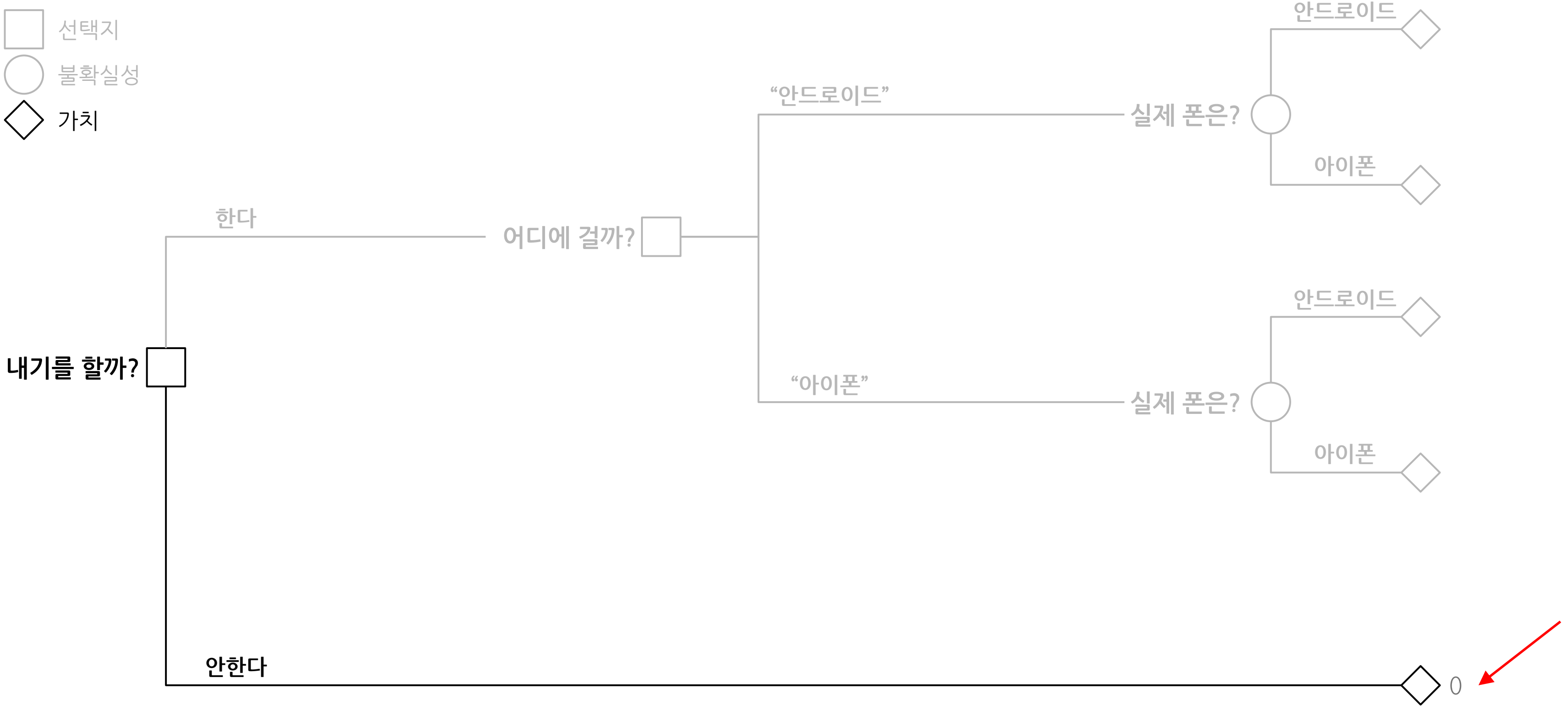
의사결정 나무 (Decision tree)



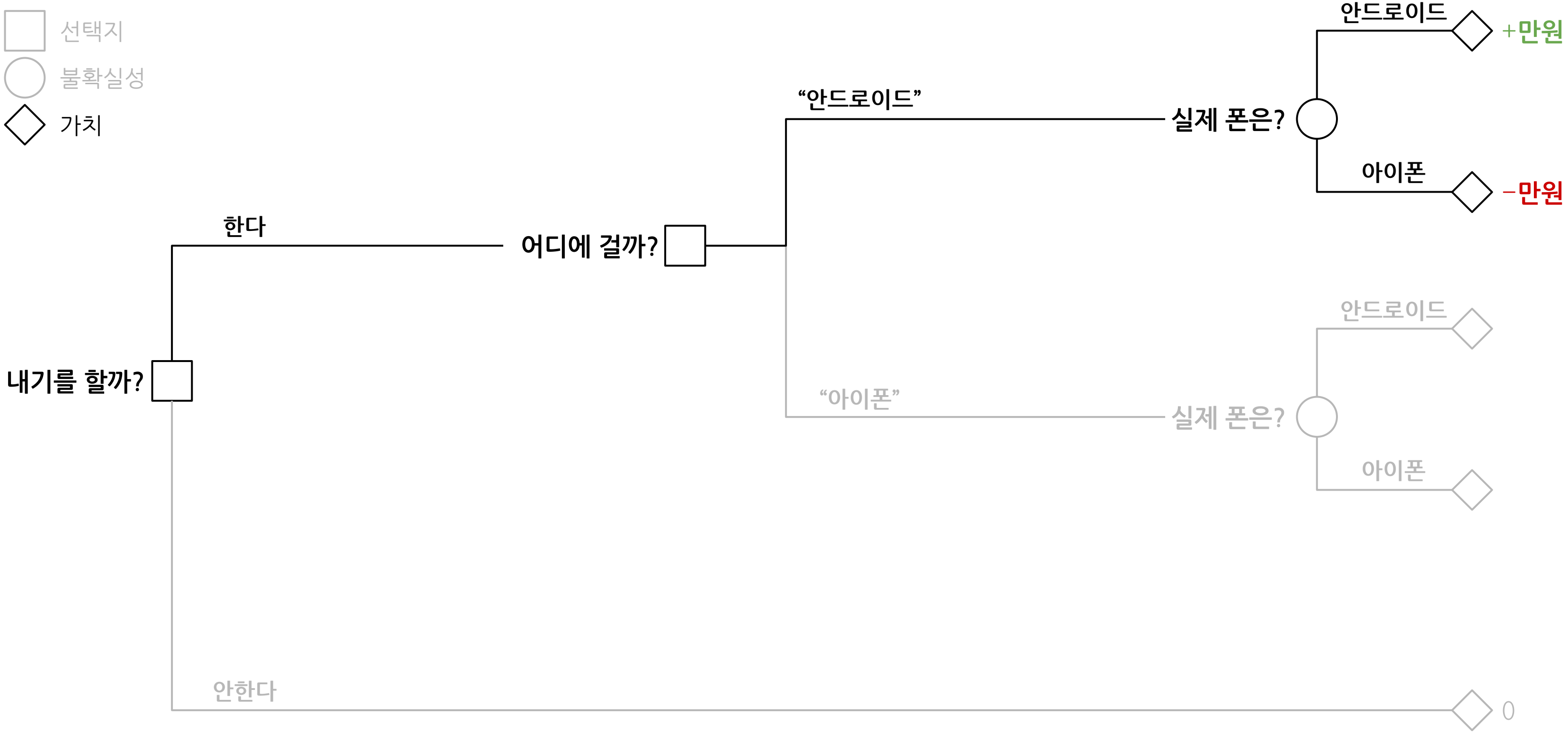
의사결정 나무 (Decision tree)



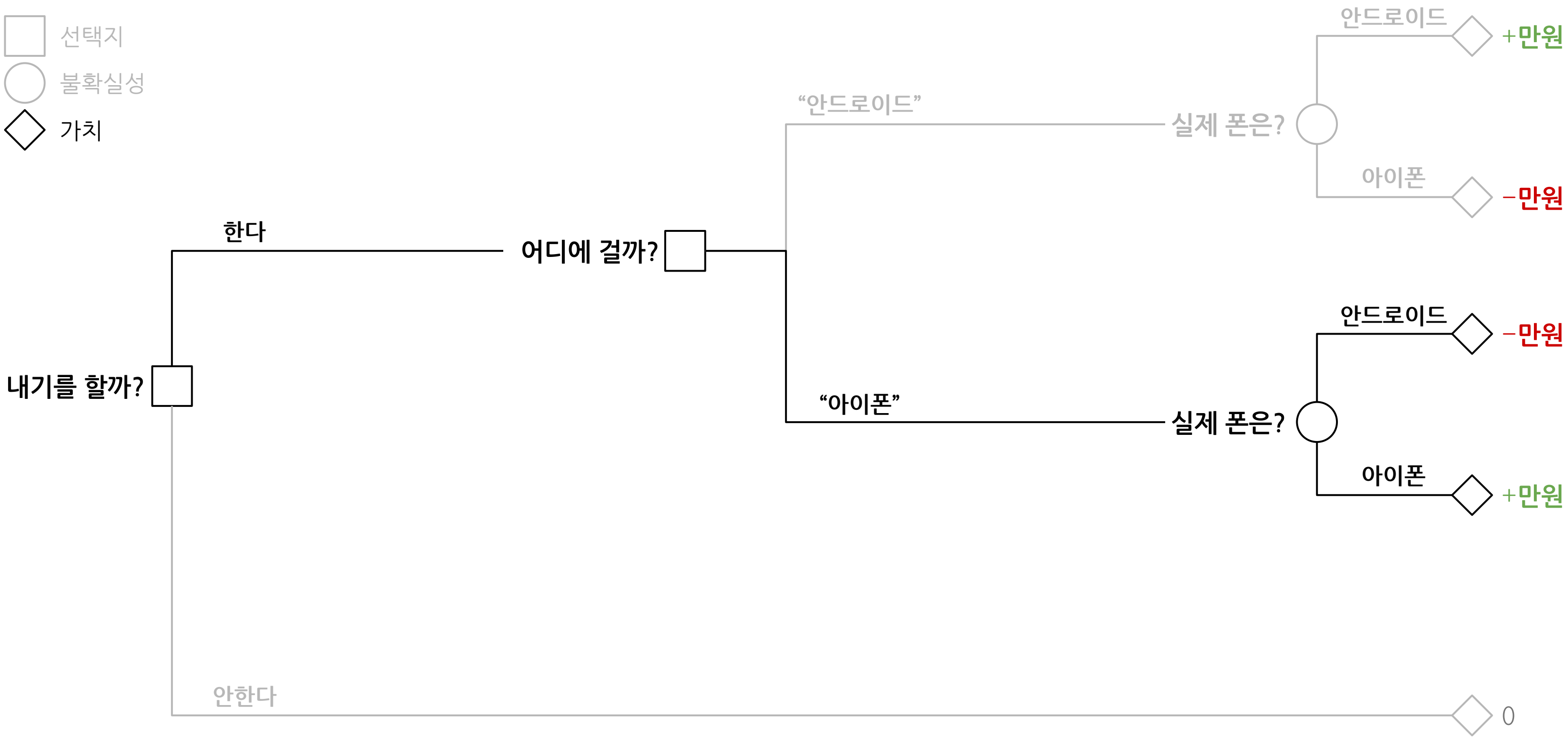
의사결정 나무 (Decision tree)



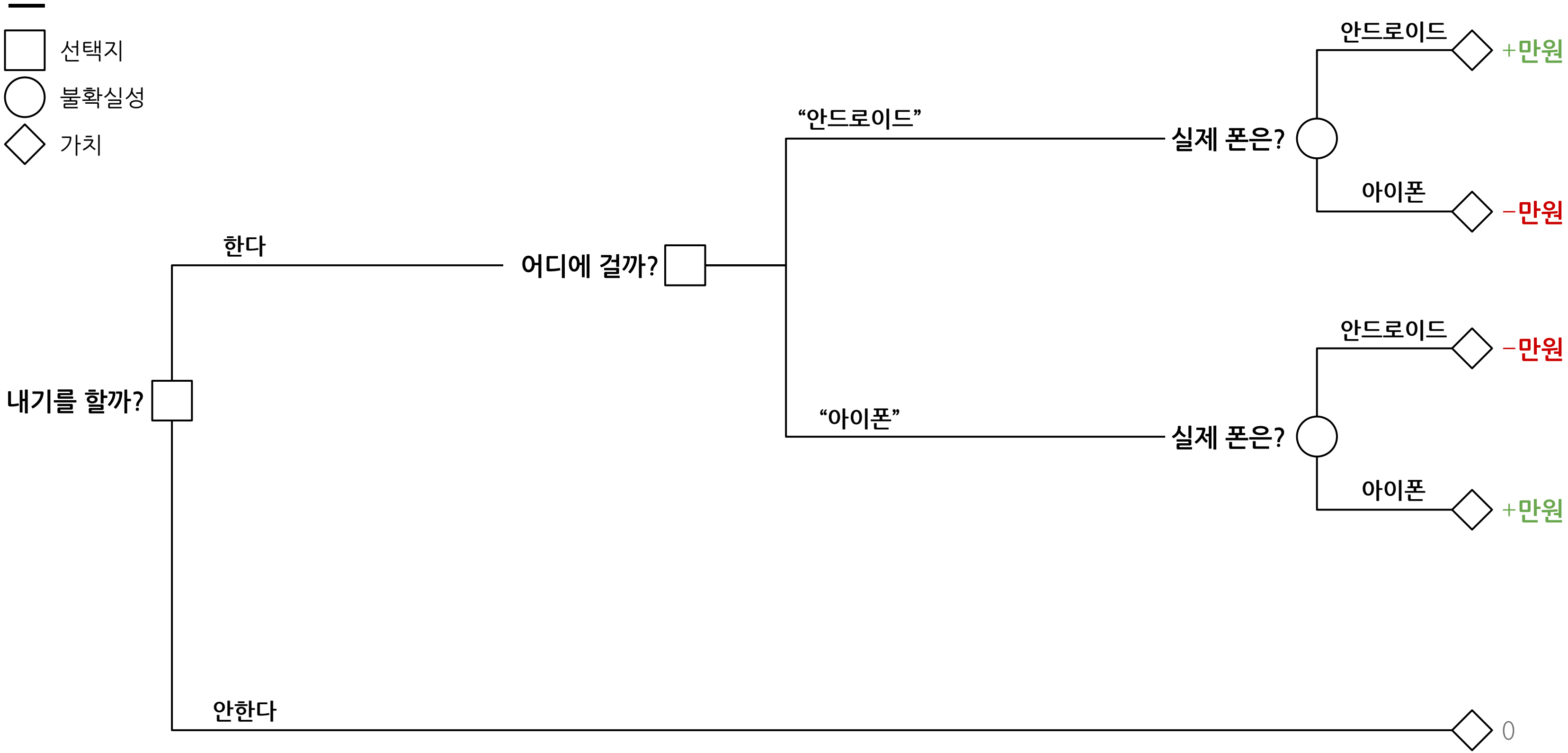
의사결정 나무 (Decision tree)



의사결정 나무 (Decision tree)



의사결정 나무 (Decision tree)

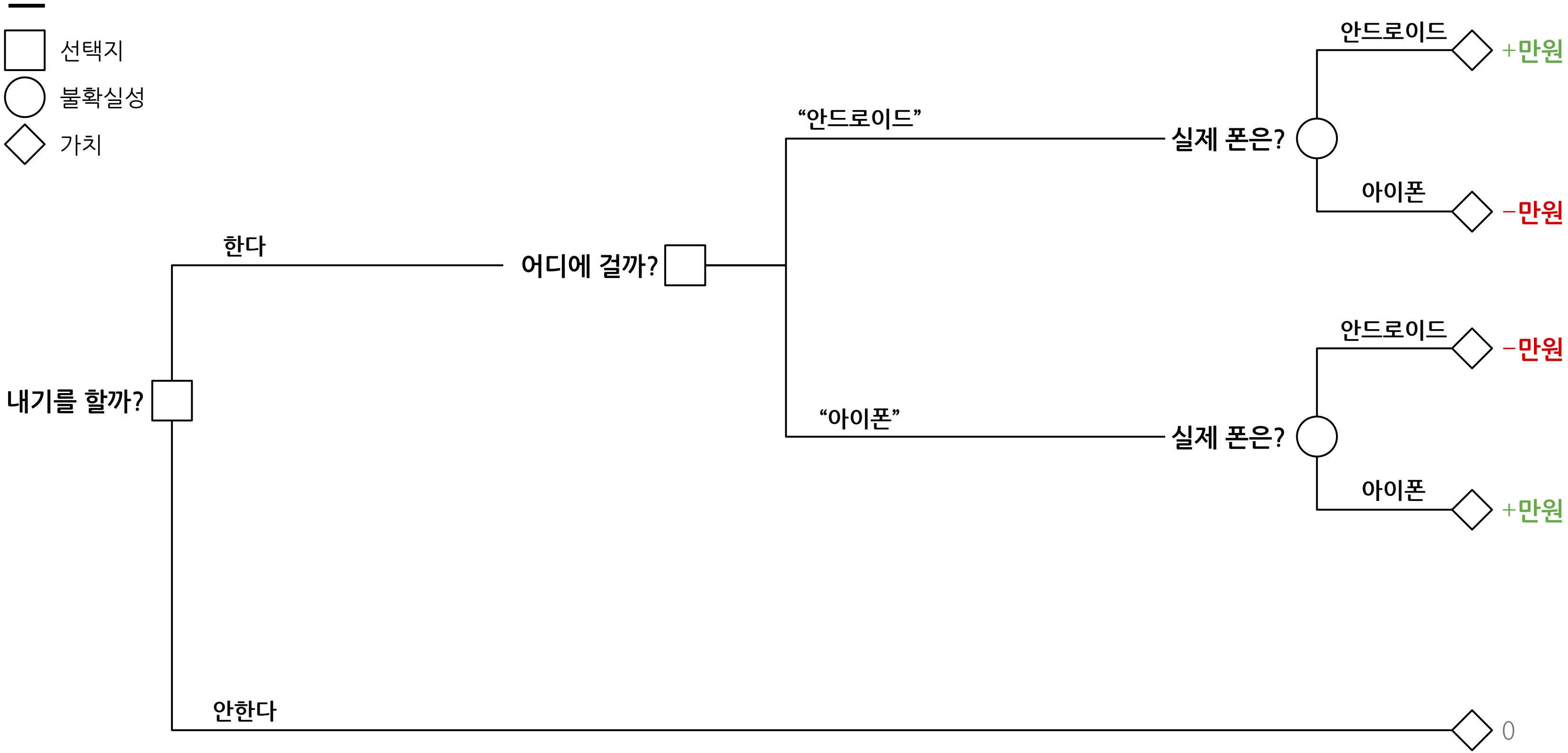


의사결정 나무 (Decision tree)

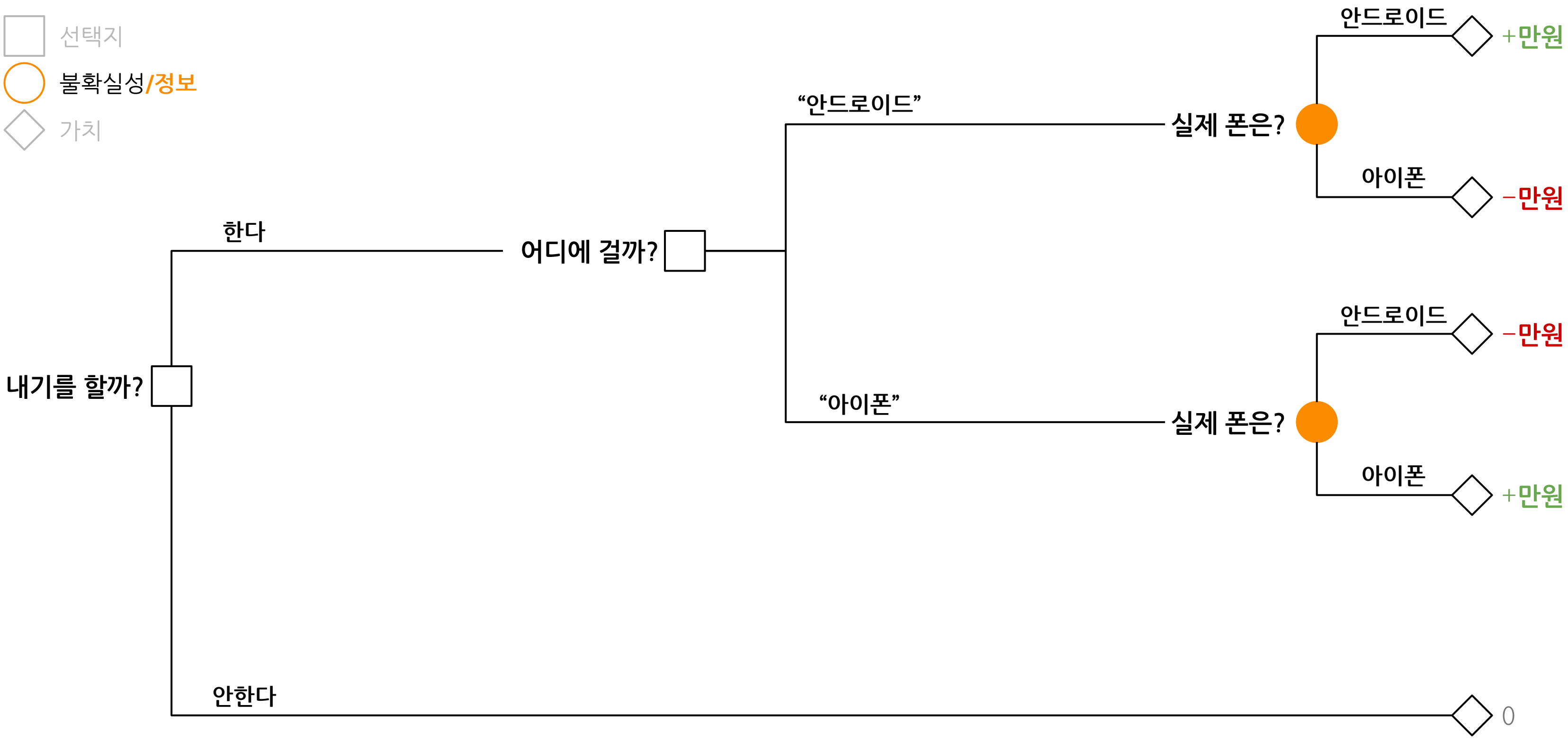
의사결정 모형

기초 모델링

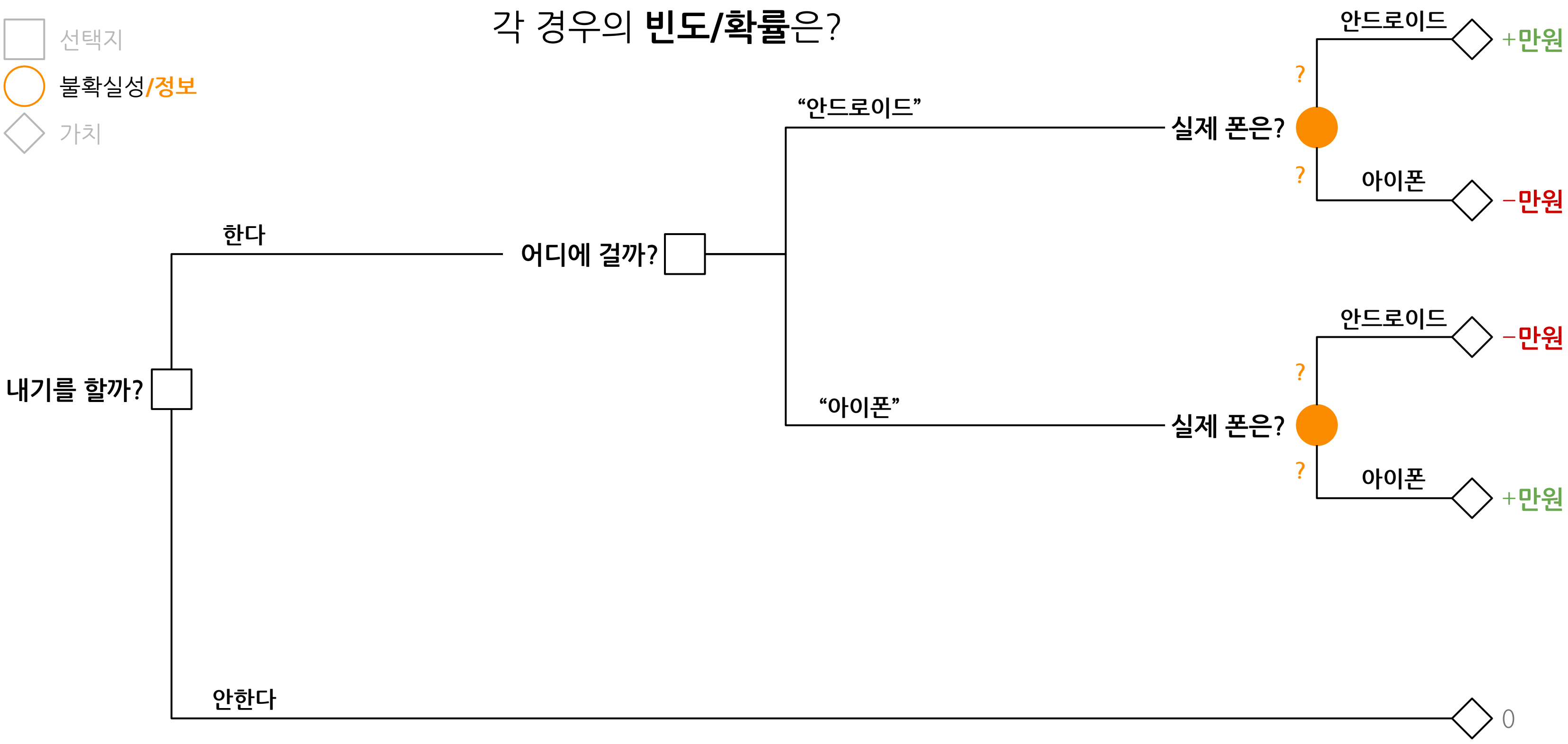
- 의사결정의 3요소
- 의사결정 나무 만들기
- 의사결정 나무 계산


선택지
불확실성
가치
내기를 할까?
한다
안한다
어디에 걸까?
“안드로이드”
“아이폰”
실제 폰은?
안드로이드
아이폰
+만원
−만원
−만원
+만원
0

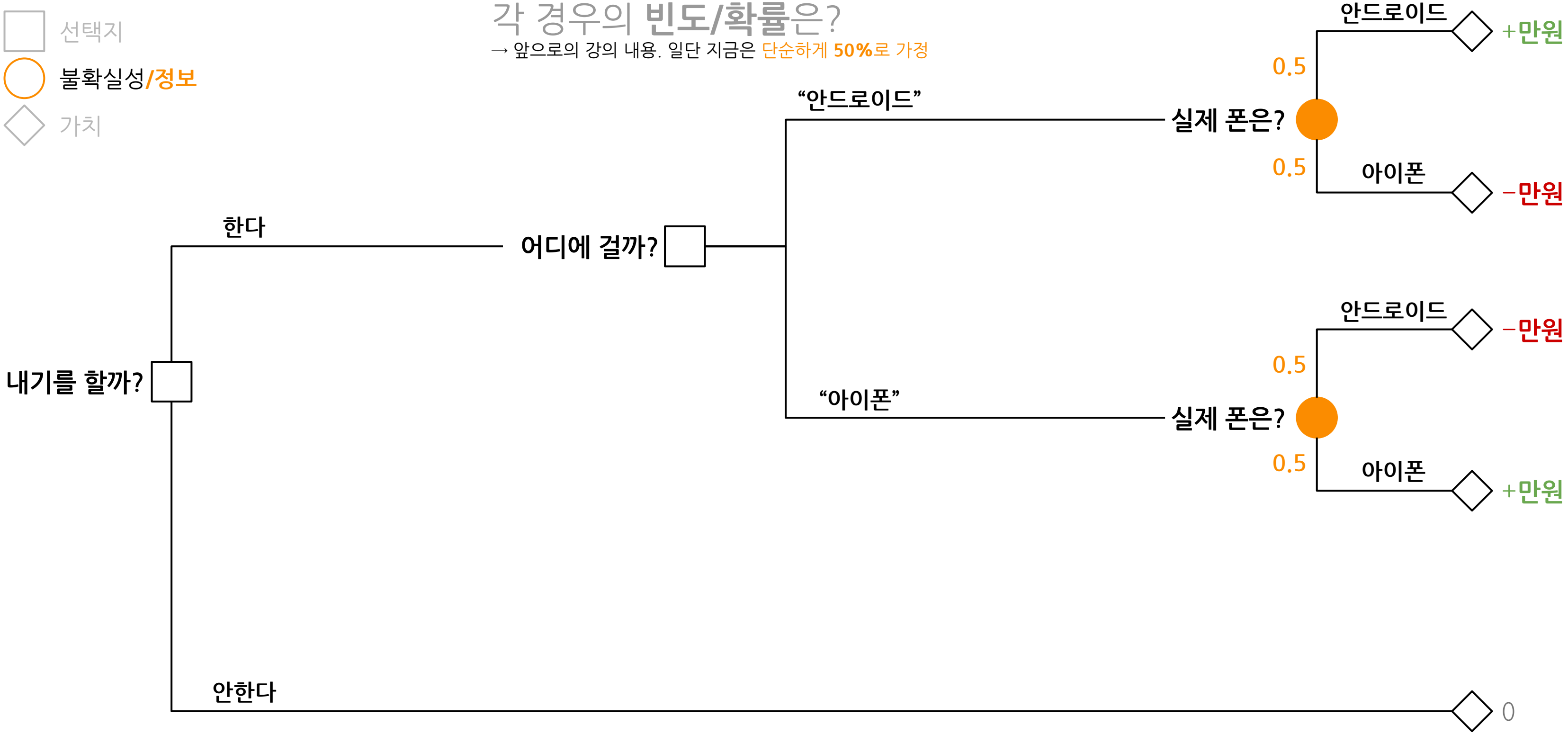
의사결정 나무 계산


선택지
불확실성/정보
가치
내기를 할까?
한다
안한다
어디에 걸까?
“안드로이드”
“아이폰”
실제 폰은?
안드로이드
아이폰
+만원
−만원
실제 폰은?
안드로이드
아이폰
−만원
+만원
0

의사결정 나무 계산


선택지
불확실성/정보
가치
각 경우의 빈도/확률은?
내기를 할까?
한다
안한다
어디에 걸까?
"안드로이드"
"아이폰"
실제 폰은?
?
안드로이드
아이폰
+만원
−만원
0

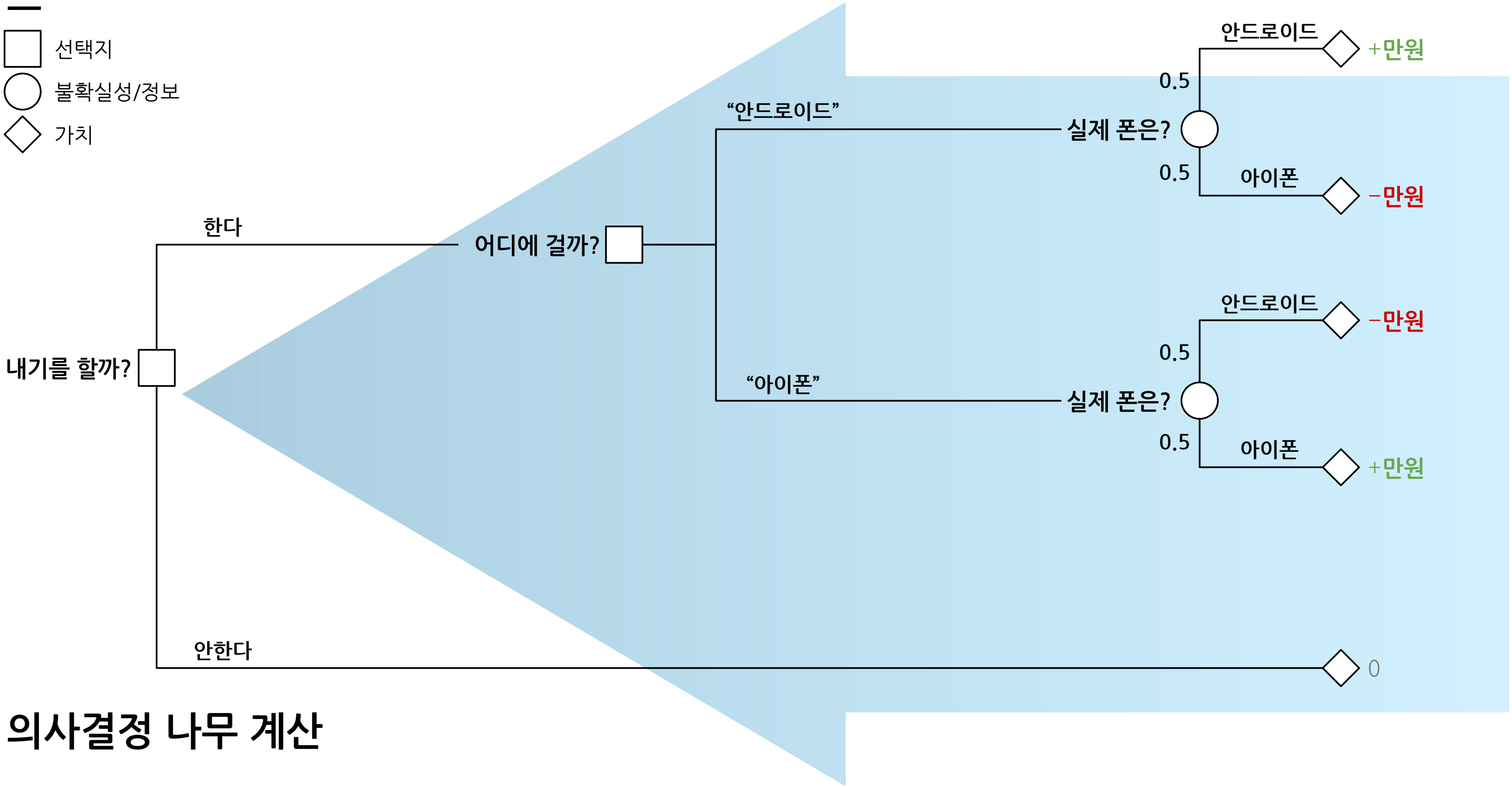
의사결정 나무 계산



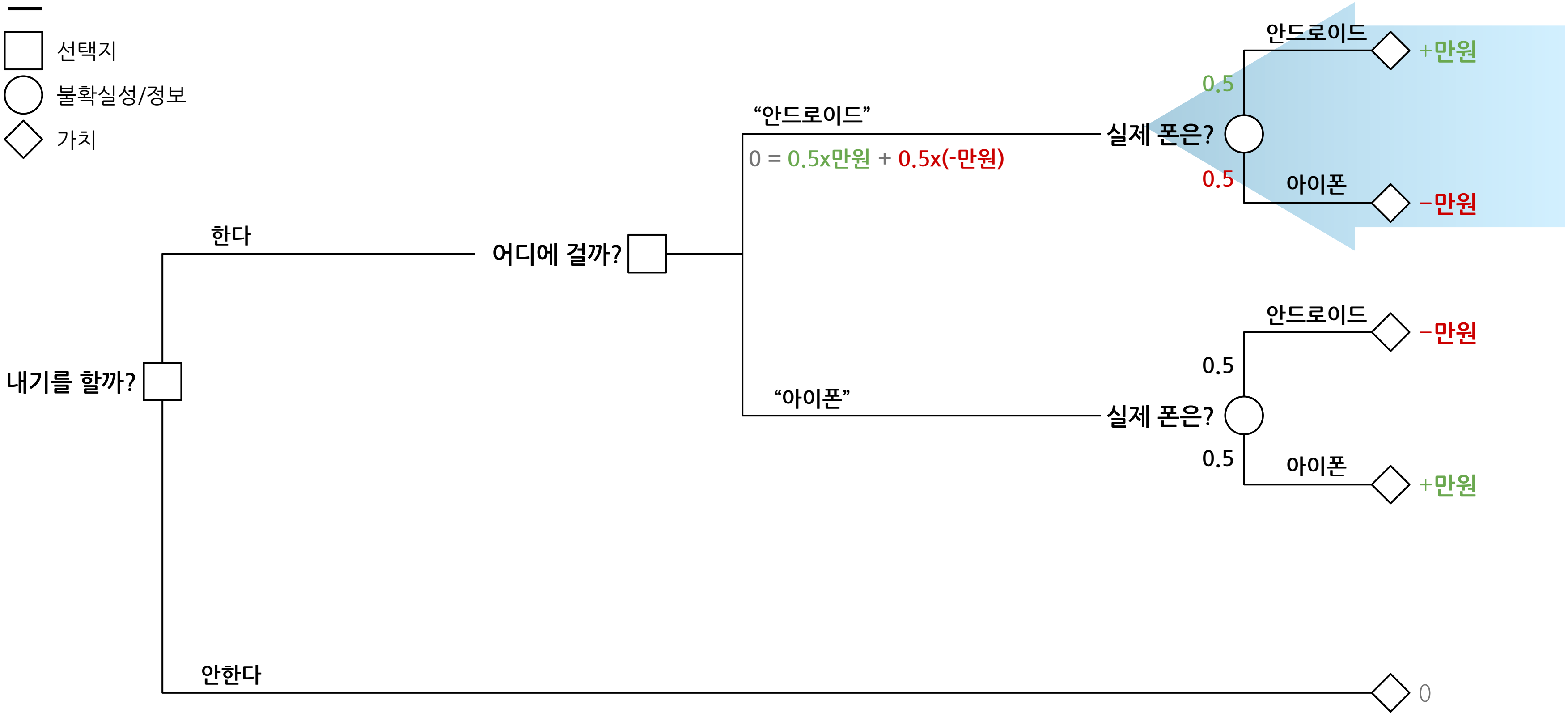
각 경우의 **빈도/확률**은?

→ 앞으로의 강의 내용. 일단 지금은 단순하게 **50%**로 가정

의사결정 나무 계산

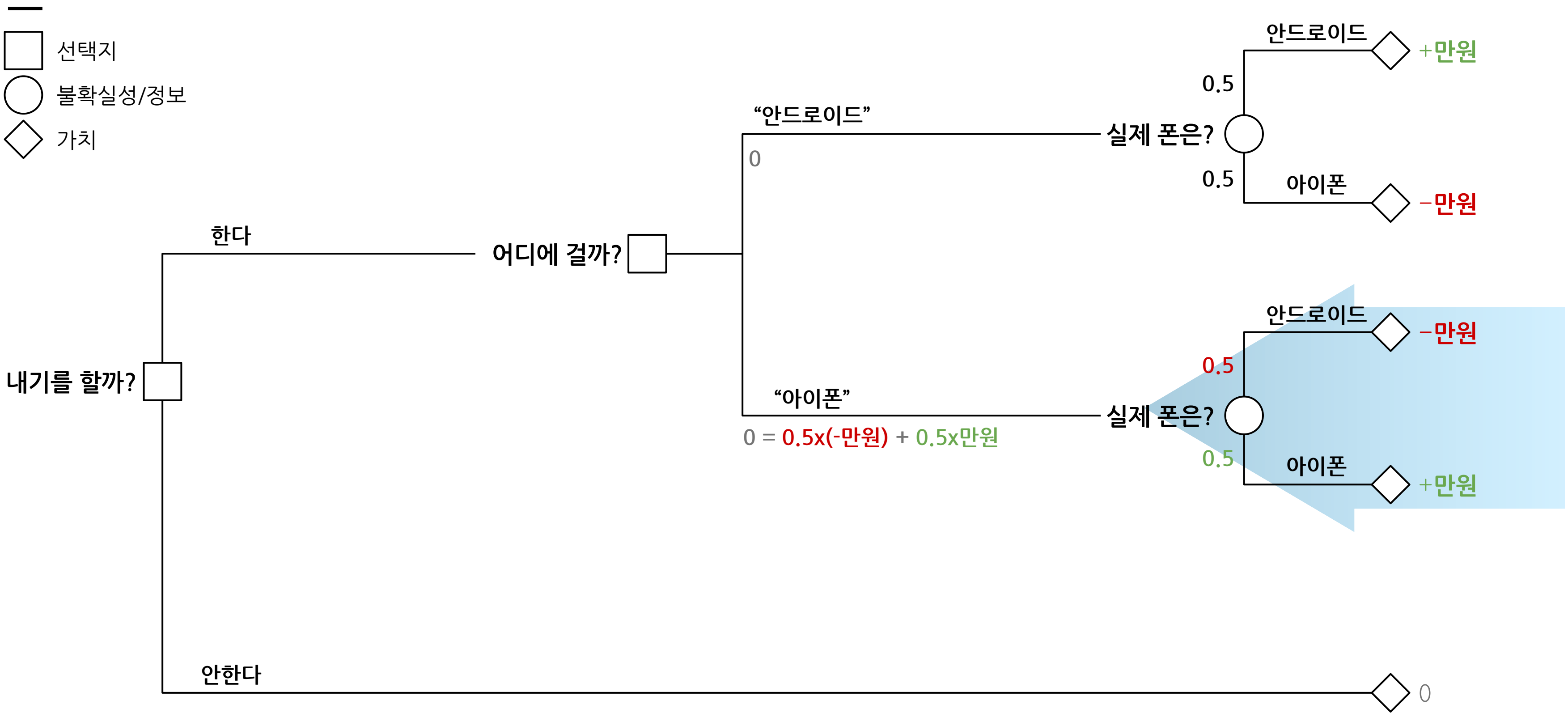

선택지
불확실성/정보
가치
내기를 할까?
한다
어디에 걸까?
"안드로이드"
실제 폰은?
0.5
안드로이드
+만원
0.5
아이폰
−만원
"아이폰"
실제 폰은?
0.5
안드로이드
−만원
0.5
아이폰
+만원
안한다
0

의사결정 나무 계산



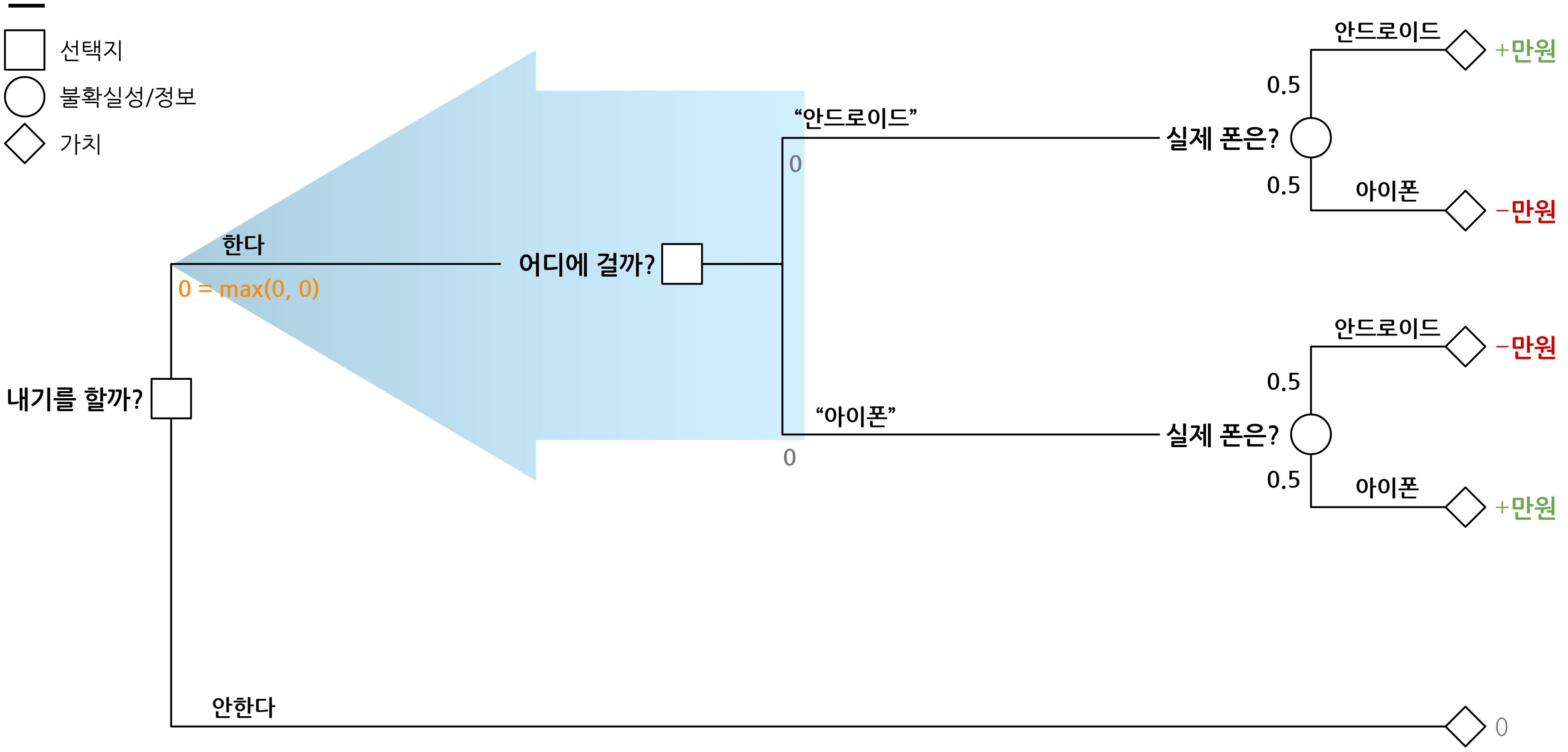
선택지
불확실성/정보
가치
내기를 할까?
한다
안한다
어디에 걸까?
"안드로이드"
0 = 0.5x만원 + 0.5x(-만원)
"아이폰"
실제 폰은?
0.5
0.5
안드로이드
아이폰
+만원
−만원
실제 폰은?
0.5
0.5
안드로이드
아이폰
−만원
+만원
0

의사결정 나무 계산



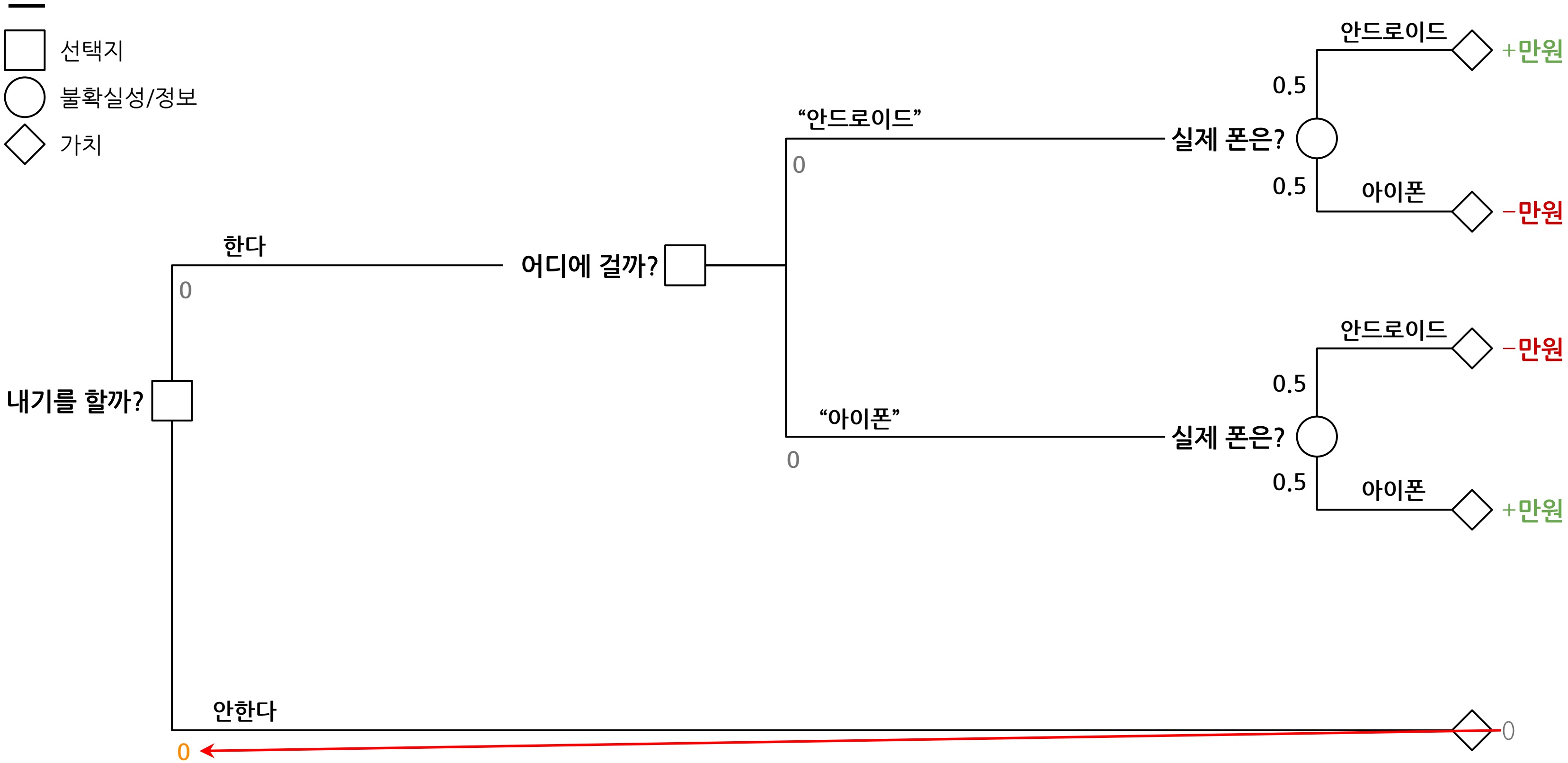
선택지
불확실성/정보
가치
내기를 할까?
한다
안한다
어디에 걸까?
"안드로이드"
0
"아이폰"
0 = 0.5x(-만원) + 0.5x만원
실제 폰은?
0.5
0.5
안드로이드
아이폰
+만원
−만원
실제 폰은?
0.5
0.5
안드로이드
아이폰
−만원
+만원
0

의사결정 나무 계산



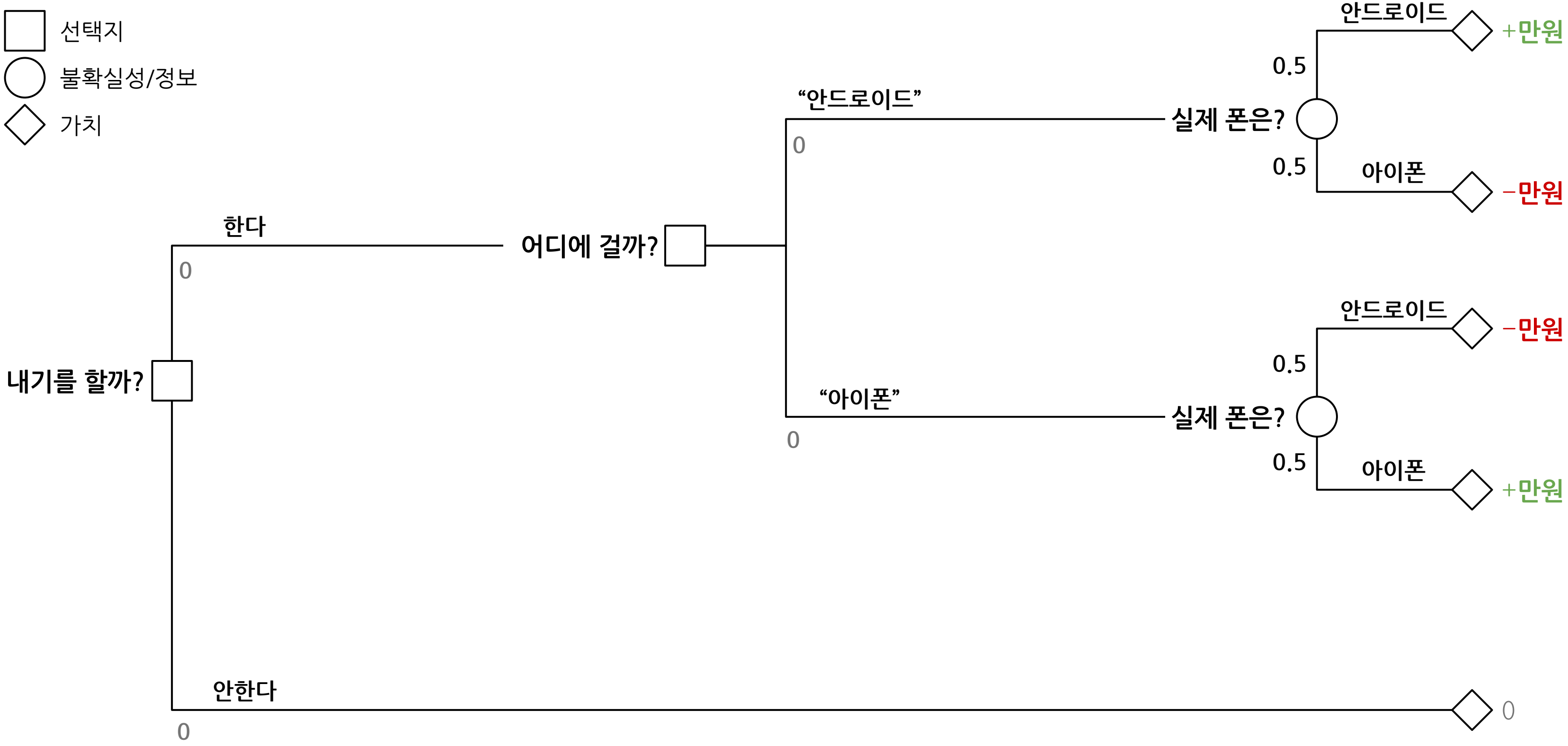
선택지
불확실성/정보
가치
내기를 할까?
한다
0 = max(0, 0)
어디에 걸까?
"안드로이드"
0
"아이폰"
0
실제 폰은?
0.5
안드로이드
+만원
0.5
아이폰
−만원
실제 폰은?
0.5
안드로이드
−만원
0.5
아이폰
+만원
안한다
0

의사결정 나무 계산



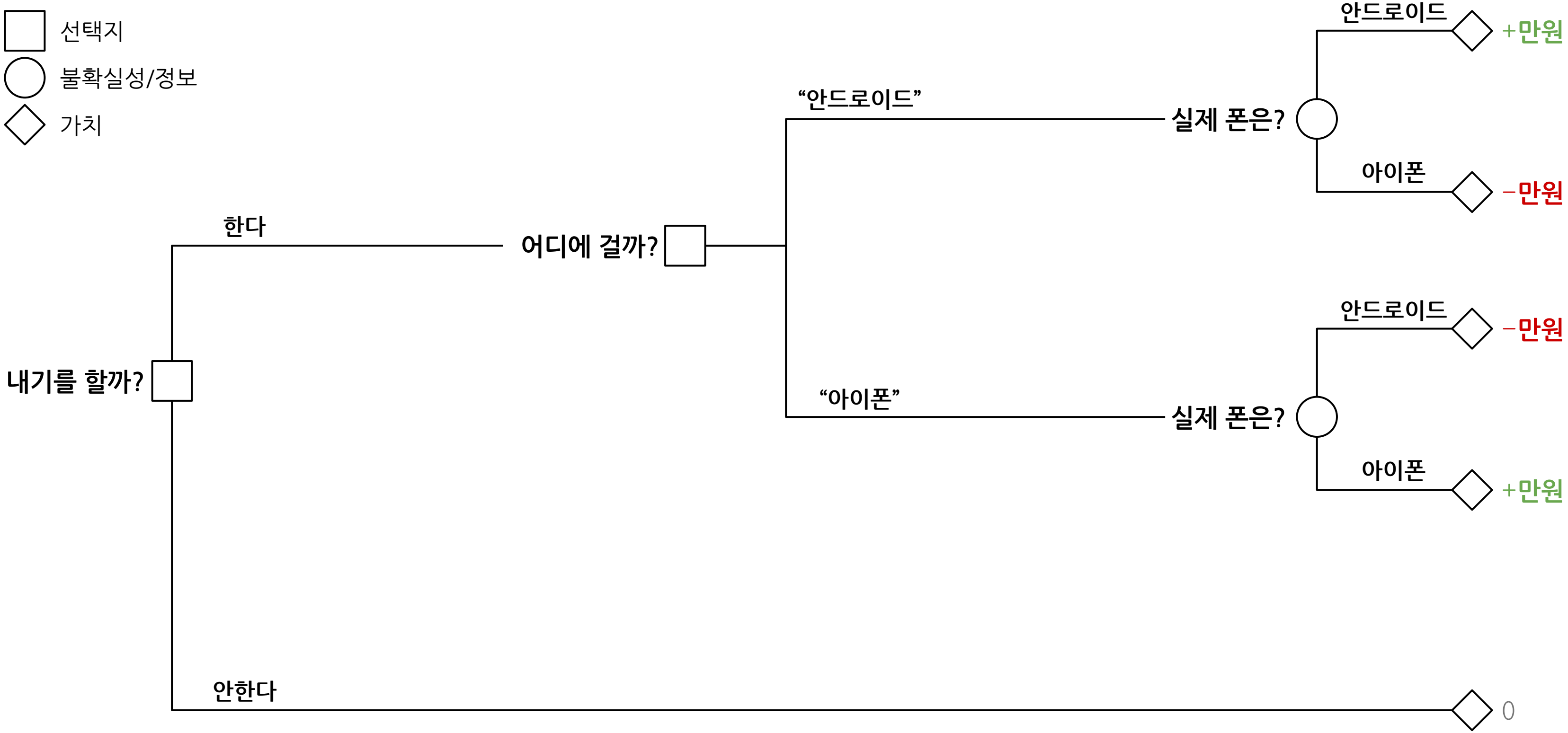
선택지
불확실성/정보
가치
한다
0
어디에 걸까?
"안드로이드"
0
"아이폰"
0
실제 폰은?
0.5
안드로이드
+만원
0.5
아이폰
−만원
실제 폰은?
0.5
안드로이드
−만원
0.5
아이폰
+만원
내기를 할까?
안한다
0
0

의사결정 나무 계산

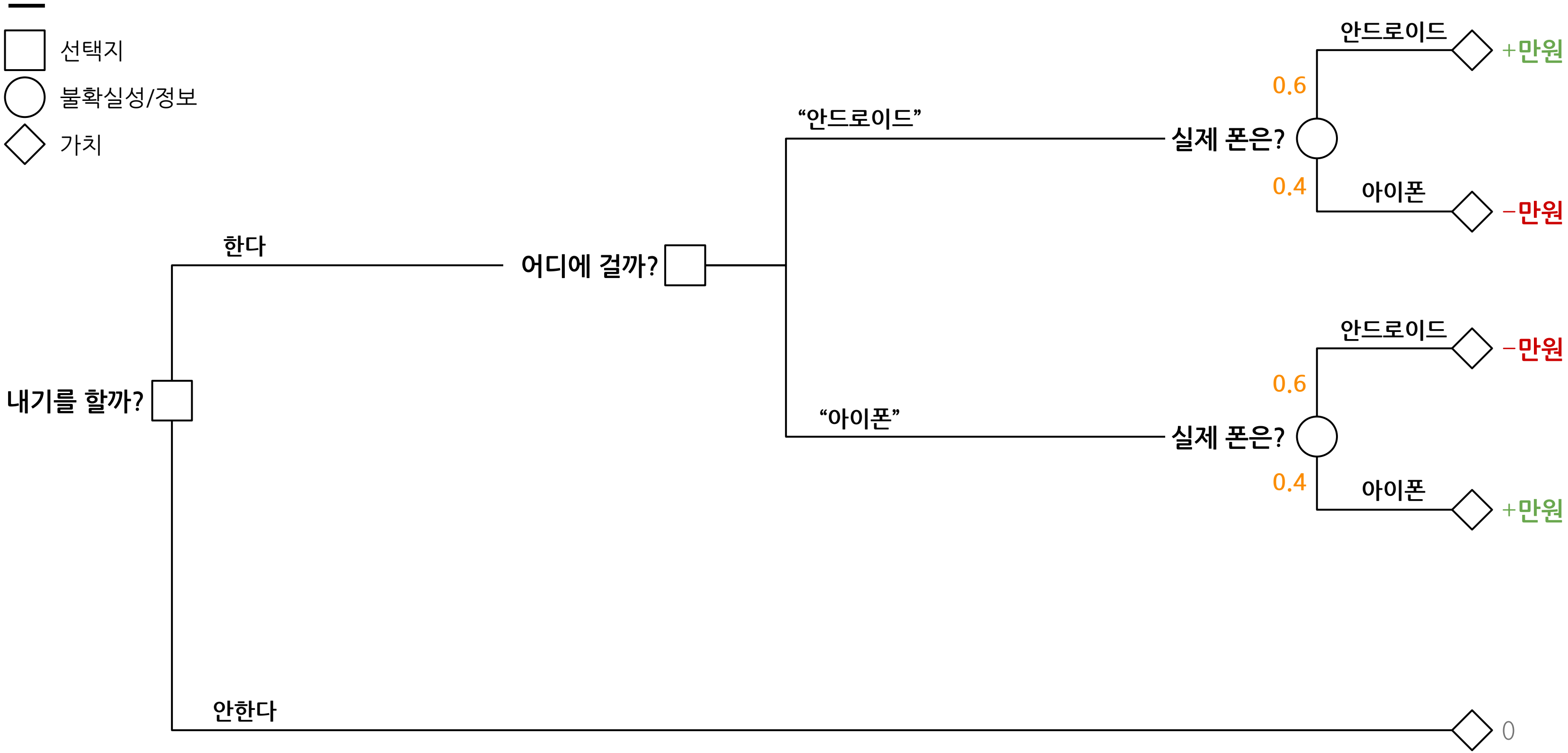

선택지
불확실성/정보
가치
내기를 할까?
한다
0
어디에 걸까?
"안드로이드"
0
실제 폰은?
0.5
안드로이드
+만원
0.5
아이폰
−만원
"아이폰"
0
실제 폰은?
0.5
안드로이드
−만원
0.5
아이폰
+만원
안한다
0
0

의사결정 나무 계산

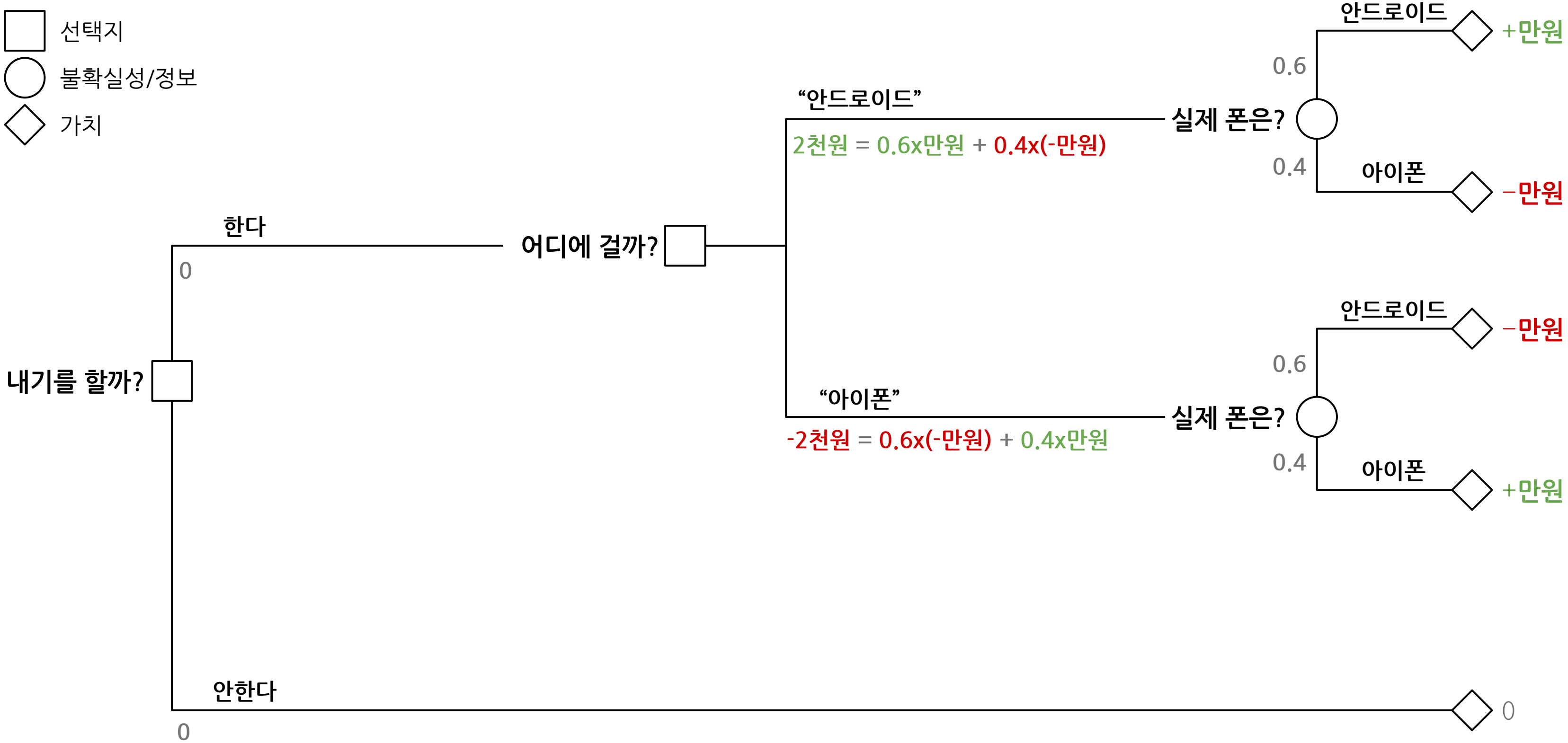
연습문제



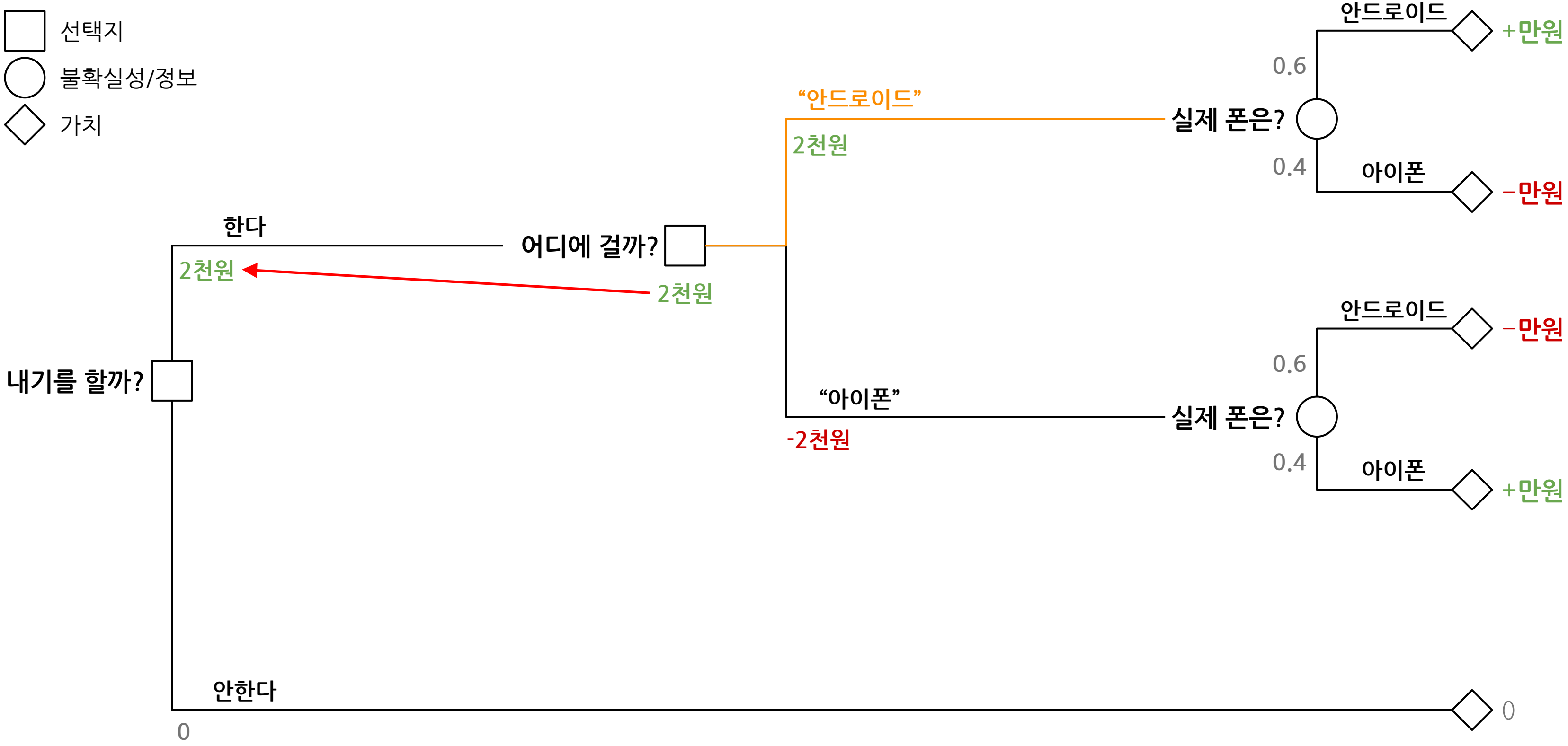
실제폰이 안드로이드일 확률을 6/10 = 60%로 본다면?



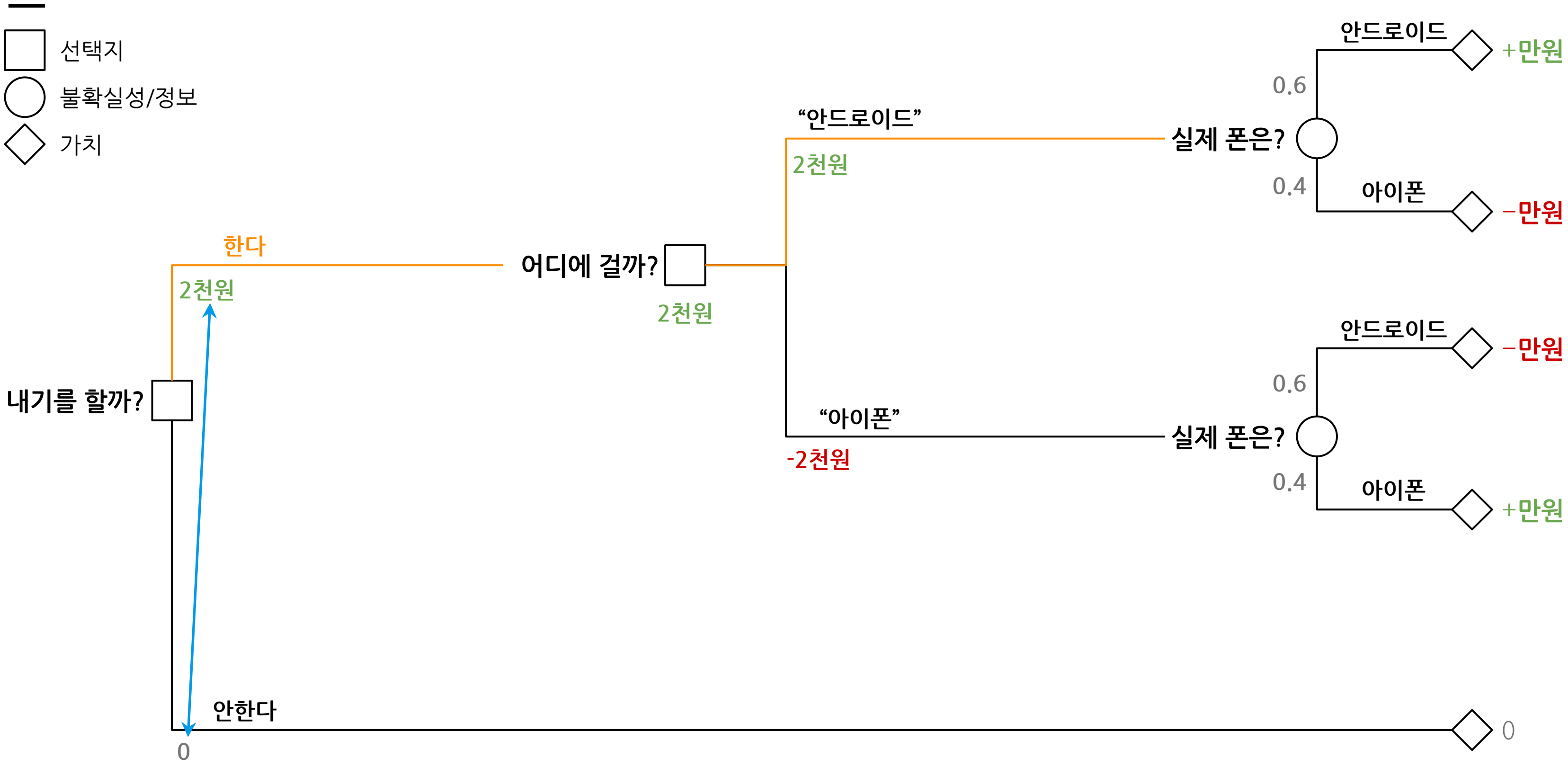
실제폰이 안드로이드일 확률을 6/10 = 60%로 본다면?



실제폰이 안드로이드일 확률을 6/10 = 60%로 본다면?



실제폰이 안드로이드일 확률을 6/10 = 60%로 본다면?



실제폰이 안드로이드일 확률을 6/10 = 60%로 본다면?

붙임

의사결정의 3요소*

* 원래는 decision maker, frame, logic을 포함한 6요소로 보는게 정석
Foundations of Decision Analysis by Ronald Howard and Ali Abbas

의사결정의 3요소*

* 원래는 decision maker, frame, logic을 포함한 6요소로 보는게 정석
Foundations of Decision Analysis by Ronald Howard and Ali Abbas

Decision maker = “나"

Frame
- 지금 이 지하철
- 새로 탄 저 사람
- 핸드폰 OS
- 내기의 규칙
- …

Logic: 수학적 “기대값의 극대화” 논리
(Von Neumann-Morgenstern utility theorem)

선택지
Alternatives

불확실성/정보
Uncertainty/information

가치/선호
Value/preference

심화연습문제

“몬티 홀 문제" 의사결정 모델링

몬티 홀 문제 Monty Hall problem

- 세 개의 문 중에 하나를 선택하여 문 뒤에 있는 선물을 가질 수 있는 게임쇼에 참가
- 한 문 뒤에는 자동차가 있고, 나머지 두 문 뒤에는 꽝
- 세 개 문 중 처음 하나를 선택
- 게임쇼 진행자가 나머지 두 개의 문 중 하나를 열어 꽝임을 보여준 후 “처음 선택을 바꿀 기회를 주겠다”고 함

→ 이 상황을 의사결정 모형으로 표현해 보시오.

다음: 심화연습문제 풀이

1부: 불확실성과 데이터

심화연습문제 풀이

데이터와 의사결정 | 정종빈

몬티 홀 문제 Monty Hall problem

- 세 개의 문 중에 하나를 선택하여 문 뒤에 있는 선물을 가질 수 있는 게임쇼에 참가
- 한 문 뒤에는 자동차가 있고, 나머지 두 문 뒤에는 꽝
- 세 개 문 중 처음 하나를 선택
- 게임쇼 진행자가 나머지 두 개의 문 중 하나를 열어 꽝임을 보여준 후 “처음 선택을 바꿀 기회를 주겠다”고 함

→ 이 상황을 의사결정 모형으로 표현해 보시오.

의사결정의 3요소

선택지
Alternatives

불확실성/정보
Uncertainty/information

가치/선호
Value/preference

의사결정의 3요소

선택지
Alternatives

의사결정의 3요소

선택한 문을 바꿀까?
바꾼다 / 안 바꾼다

선택지
Alternatives

의사결정의 3요소

선택한 문을 바꿀까?
바꾼다 / 안 바꾼다

선택지
Alternatives

불확실성/정보
Uncertainty/information

의사결정의 3요소

선택한 문을 바꿀까?
바꾼다 / 안 바꾼다

선택지
Alternatives

처음 선택이 꽝?
보다 구체적인 불확실성 모델링은 이후
[베이지안 확률] 파트에서 강의

불확실성/정보
Uncertainty/information

의사결정의 3요소

선택한 문을 바꿀까?
바꾼다 / 안 바꾼다

선택지
Alternatives

처음 선택이 꽝?

불확실성/정보
Uncertainty/information

가치/선호
Value/preference

의사결정의 3요소

선택한 문을 바꿀까?
바꾼다 / 안 바꾼다

선택지
Alternatives

처음 선택이 꽝?

불확실성/정보
Uncertainty/information

1. 자동차: +
2. 꽝: 0

가치/선호
Value/preference

의사결정의 3요소

선택한 문을 바꿀까?

바꾼다 / 안 바꾼다

선택지

Alternatives

처음 선택이 꽝?

불확실성/정보

Uncertainty/information

1. 자동차: +
2. 꽝: 0

가치/선호

Value/preference

□ 선택지: 선택한 문을 바꿀까?

○ 불확실성: 처음 선택은?

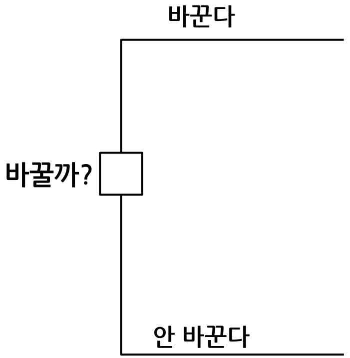
◇ 가치: 자동차 = +, 꽝 = 0

의사결정 나무 (Decision tree)

선택지: 선택한 문을 바꿀까?

불확실성: 처음 선택은?

가치: 자동차 = +, 꽝 = 0

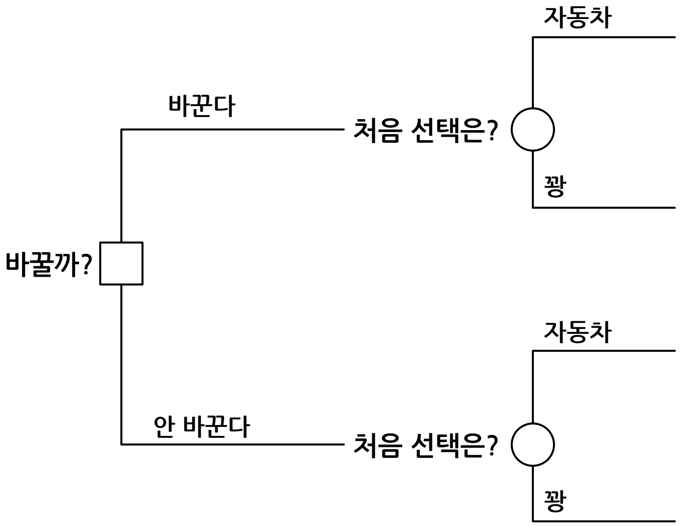


의사결정 나무 (Decision tree)

선택지: 선택한 문을 바꿀까?

불확실성: 처음 선택은?

가치: 자동차 = +, 꽝 = 0

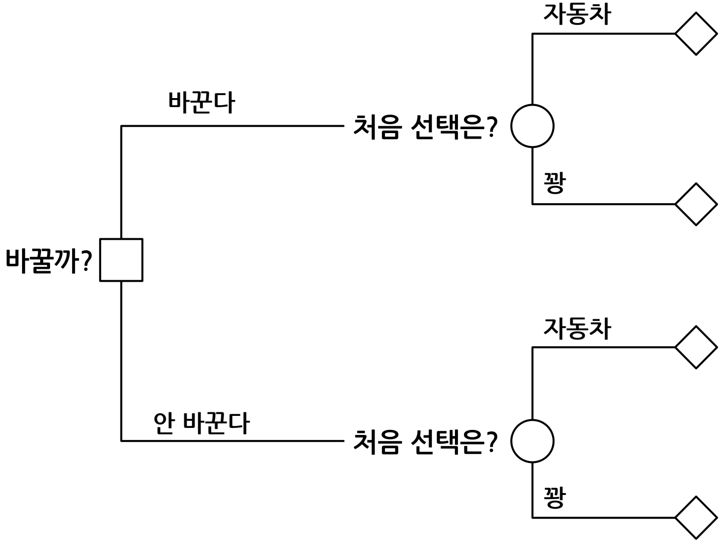


의사결정 나무 (Decision tree)

선택지: 선택한 문을 바꿀까?

불확실성: 처음 선택은?

가치: 자동차 = +, 꽝 = 0

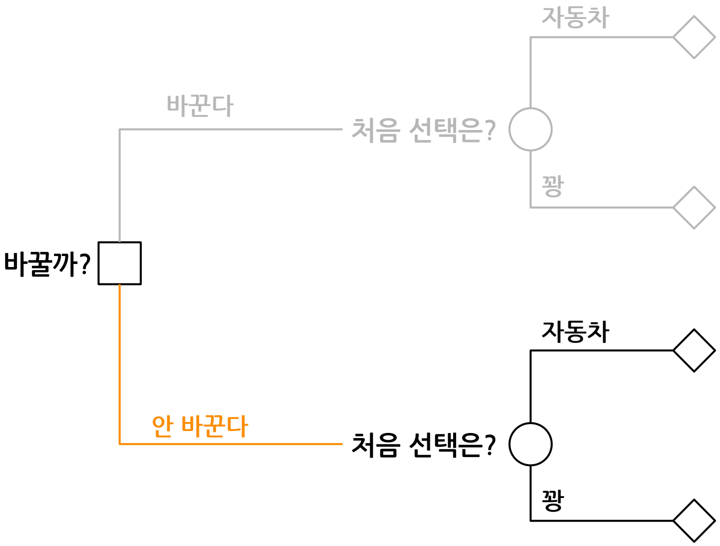


의사결정 나무 (Decision tree)

□ 선택지: 선택한 문을 바꿀까?

○ 불확실성: 처음 선택은?

◇ 가치: 자동차 = +, 꽝 = 0

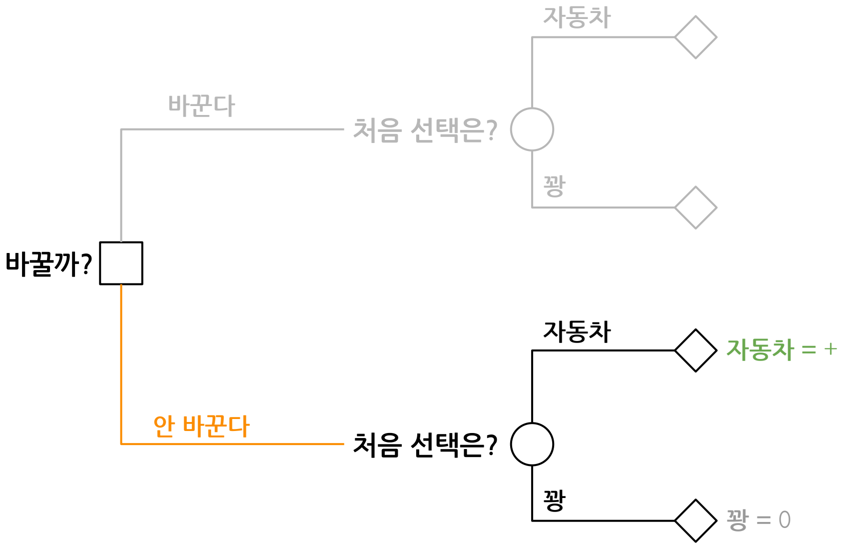


의사결정 나무 (Decision tree)

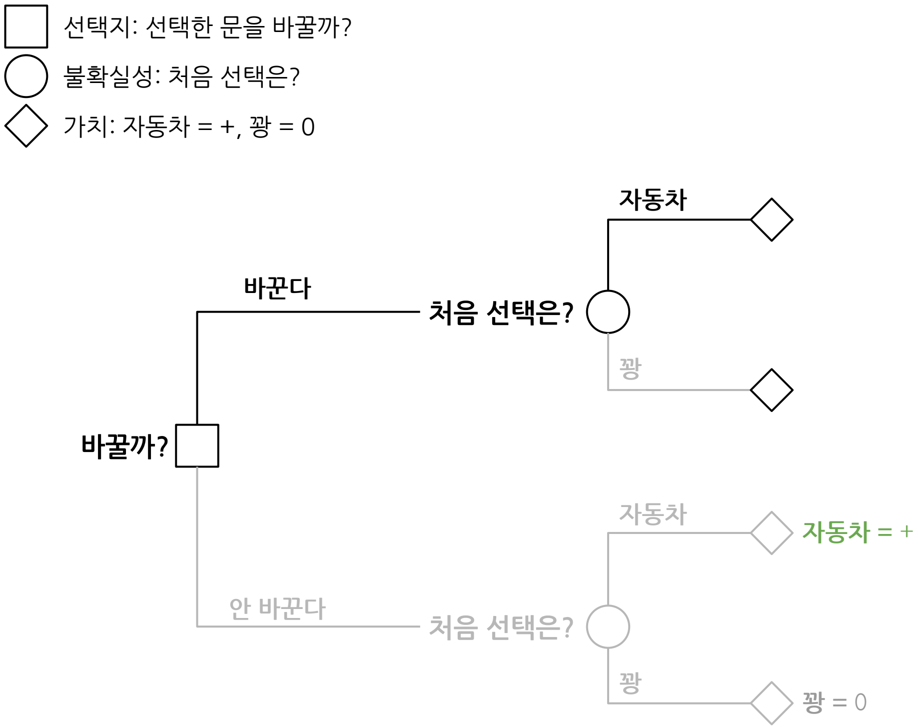
선택지: 선택한 문을 바꿀까?

불확실성: 처음 선택은?

가치: 자동차 = +, 꽝 = 0

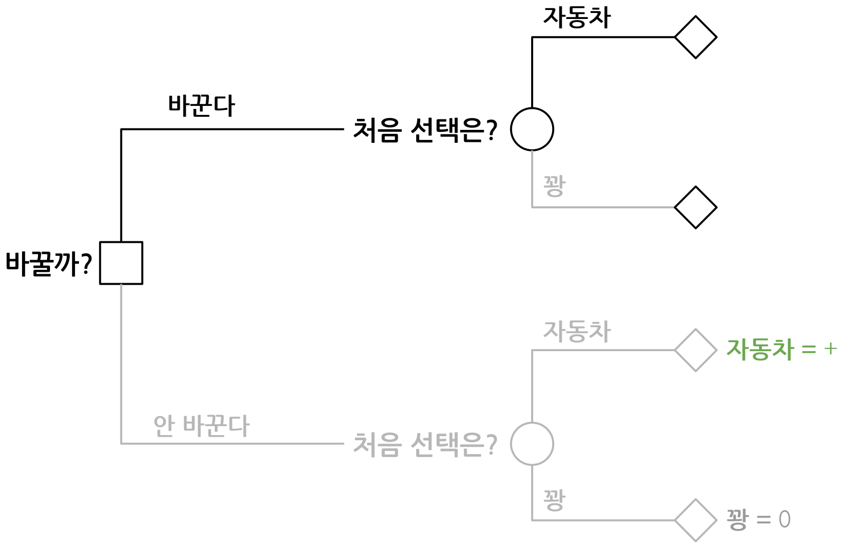


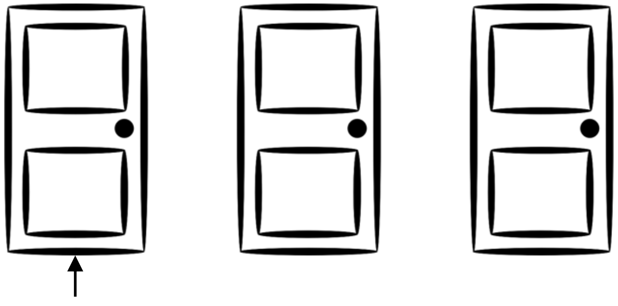
의사결정 나무 (Decision tree)



의사결정 나무 (Decision tree)

선택지: 선택한 문을 바꿀까?
불확실성: 처음 선택은?
가치: 자동차 = +, 꽝 = 0



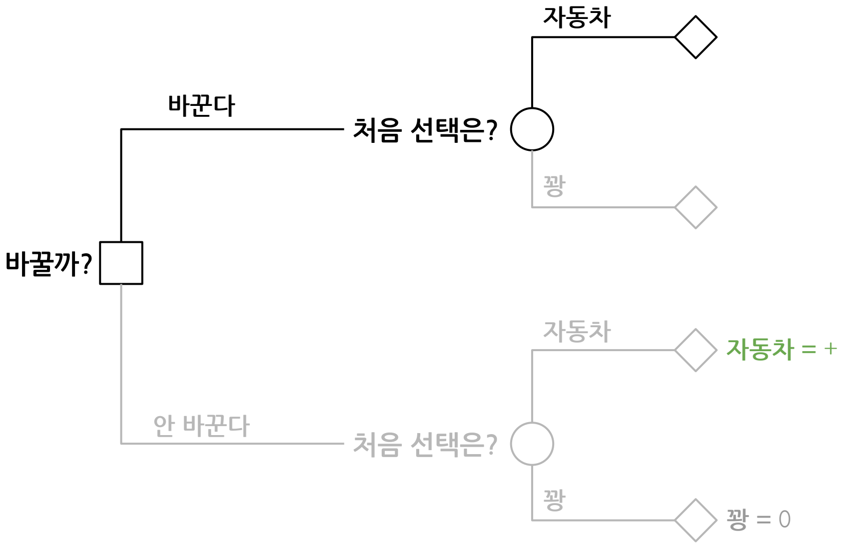


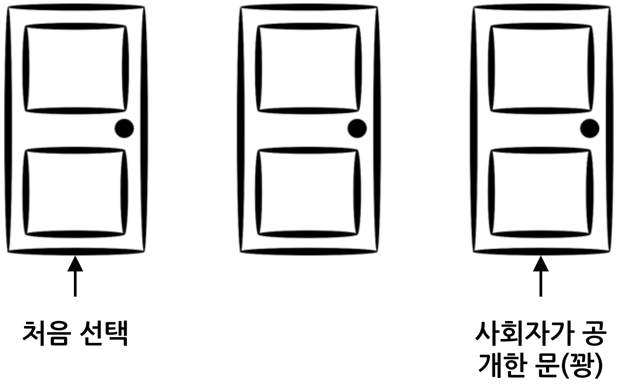
의사결정 나무 (Decision tree)

선택지: 선택한 문을 바꿀까?

불확실성: 처음 선택은?

가치: 자동차 = +, 꽝 = 0



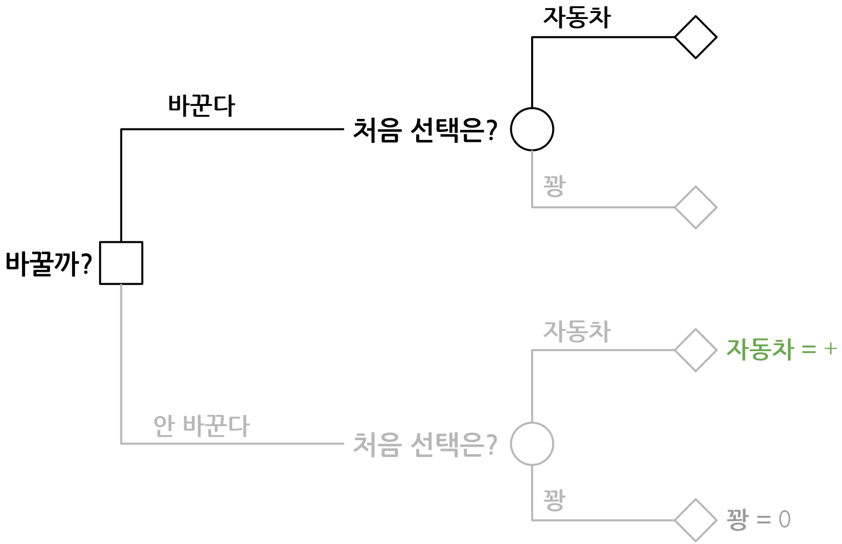


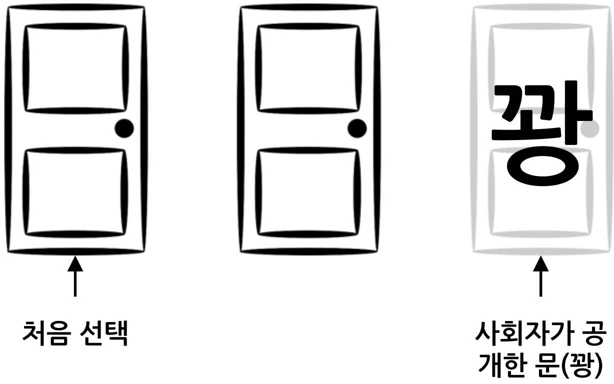
의사결정 나무 (Decision tree)

선택지: 선택한 문을 바꿀까?

불확실성: 처음 선택은?

가치: 자동차 = +, 꽝 = 0



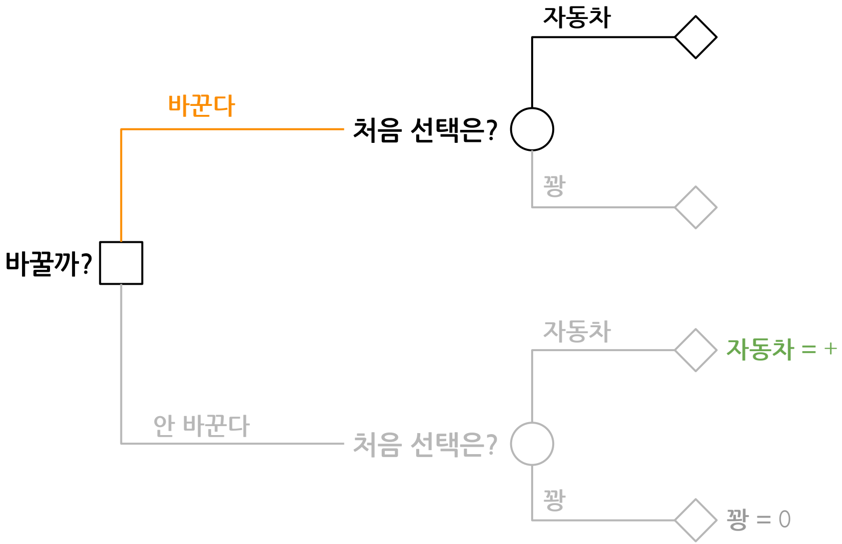


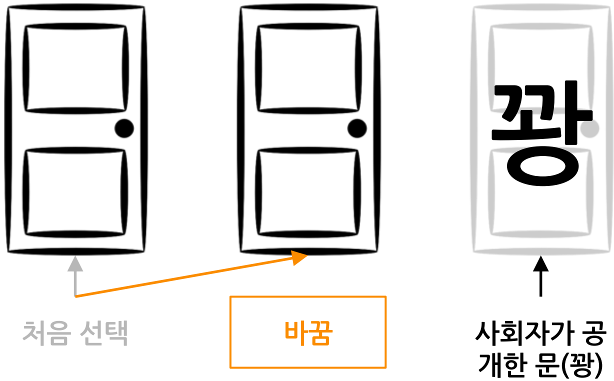
의사결정 나무 (Decision tree)

선택지: 선택한 문을 바꿀까?

불확실성: 처음 선택은?

가치: 자동차 = +, 꽝 = 0



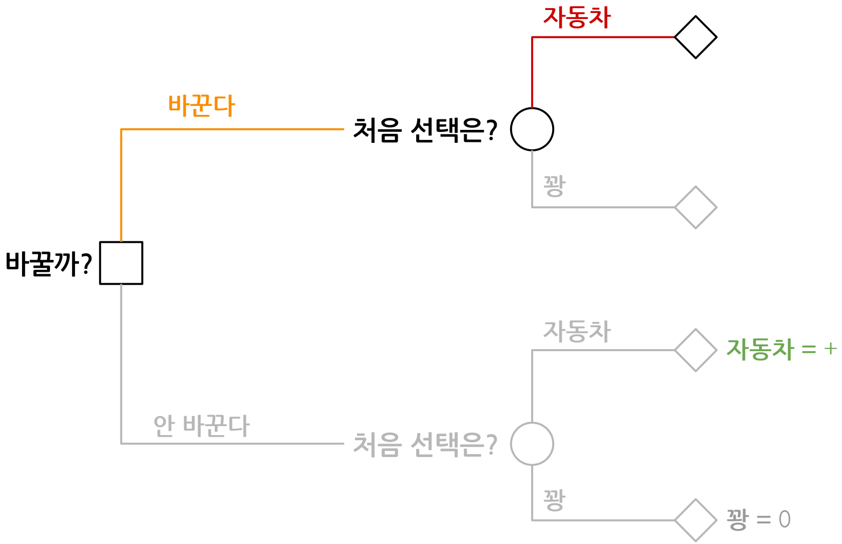


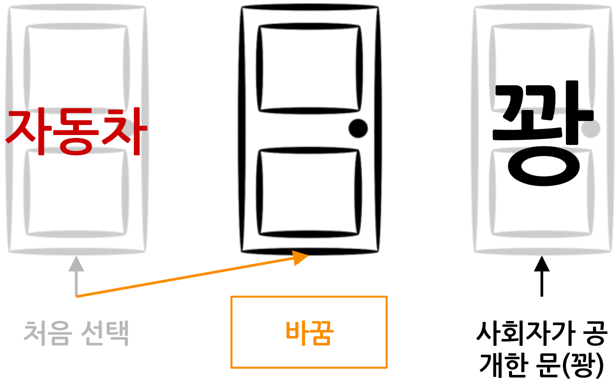
의사결정 나무 (Decision tree)

선택지: 선택한 문을 바꿀까?

불확실성: 처음 선택은?

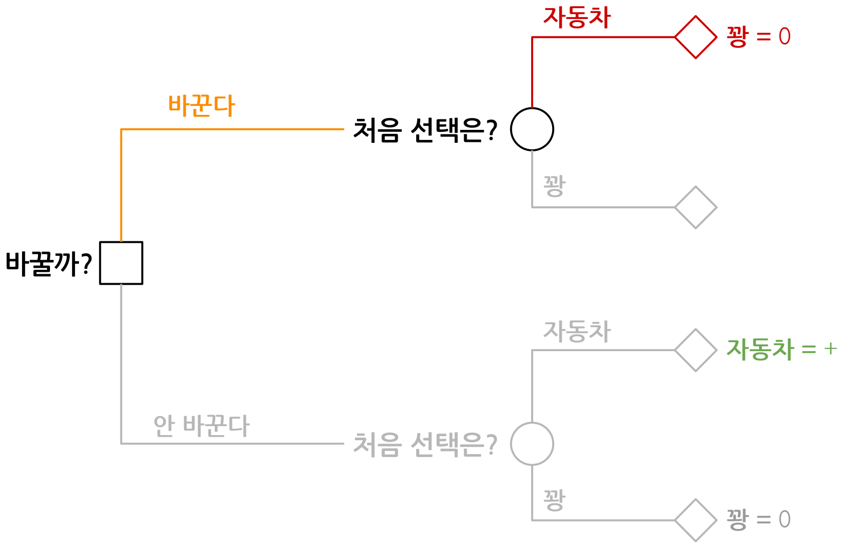
가치: 자동차 = +, 꽝 = 0





의사결정 나무 (Decision tree)

선택지: 선택한 문을 바꿀까?
불확실성: 처음 선택은?
가치: 자동차 = +, 꽝 = 0

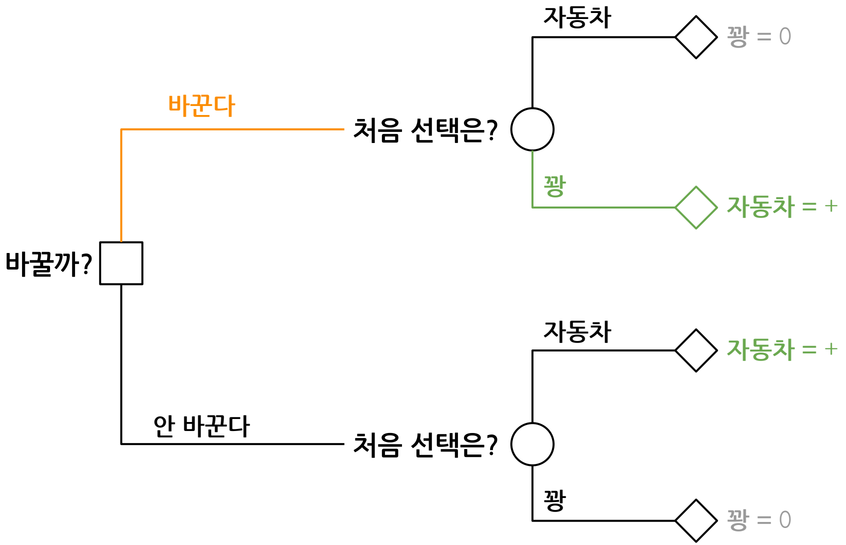


의사결정 나무 (Decision tree)

선택지: 선택한 문을 바꿀까?

불확실성: 처음 선택은?

가치: 자동차 = +, 꽝 = 0

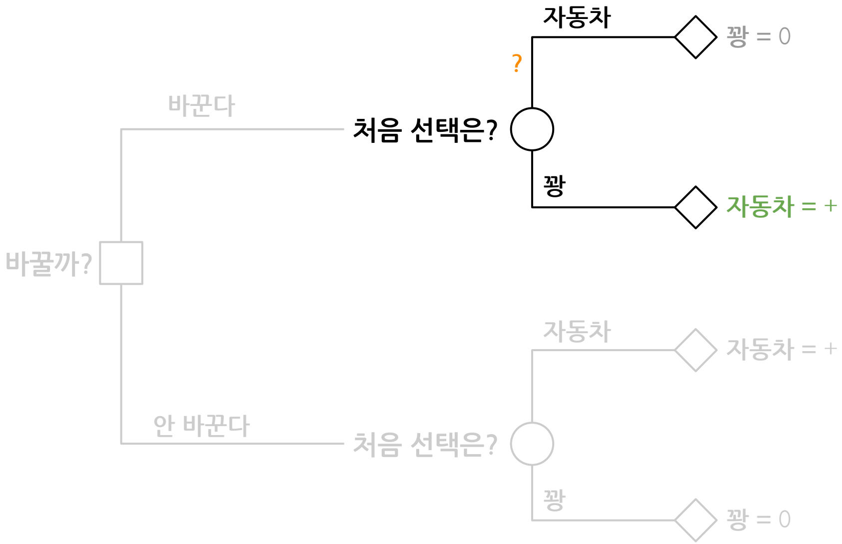


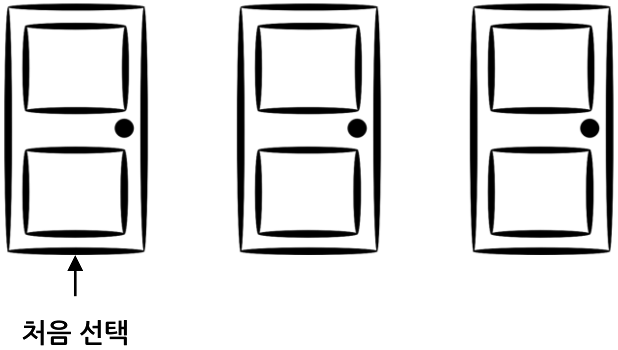
의사결정 나무 (Decision tree)

□ 선택지: 선택한 문을 바꿀까?

○ 불확실성: 처음 선택은?

◇ 가치: 자동차 = +, 꽝 = 0



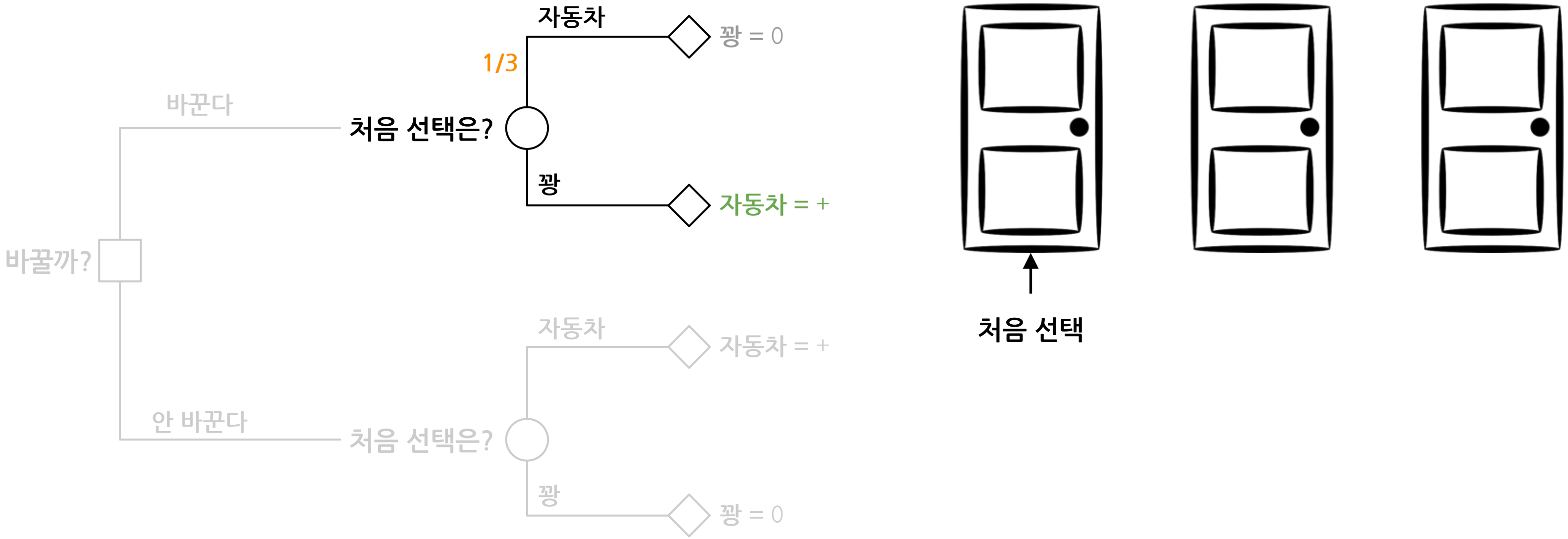


의사결정 나무 (Decision tree)

선택지: 선택한 문을 바꿀까?

불확실성: 처음 선택은?

가치: 자동차 = +, 꽝 = 0

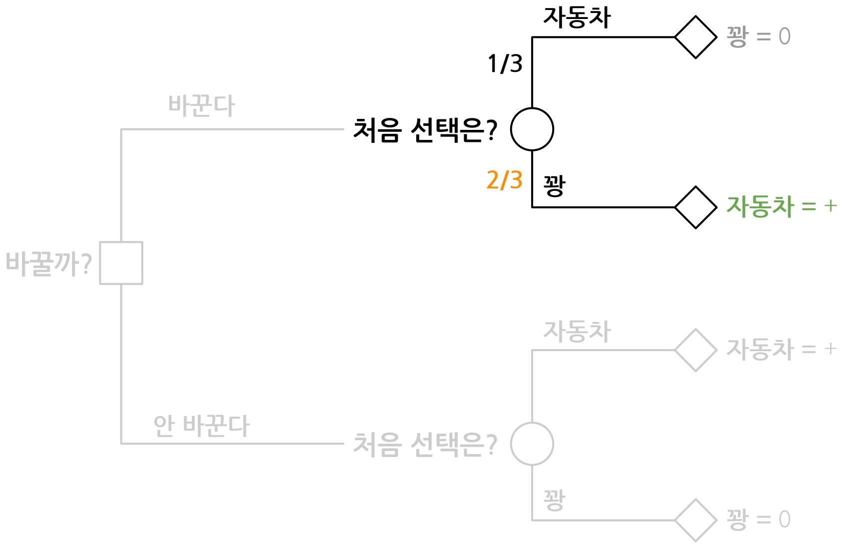


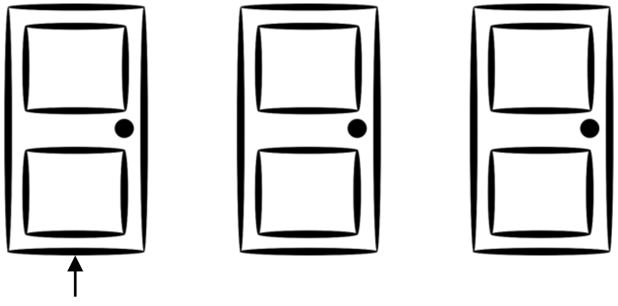
의사결정 나무 (Decision tree)

선택지: 선택한 문을 바꿀까?

불확실성: 처음 선택은?

가치: 자동차 = +, 꽝 = 0



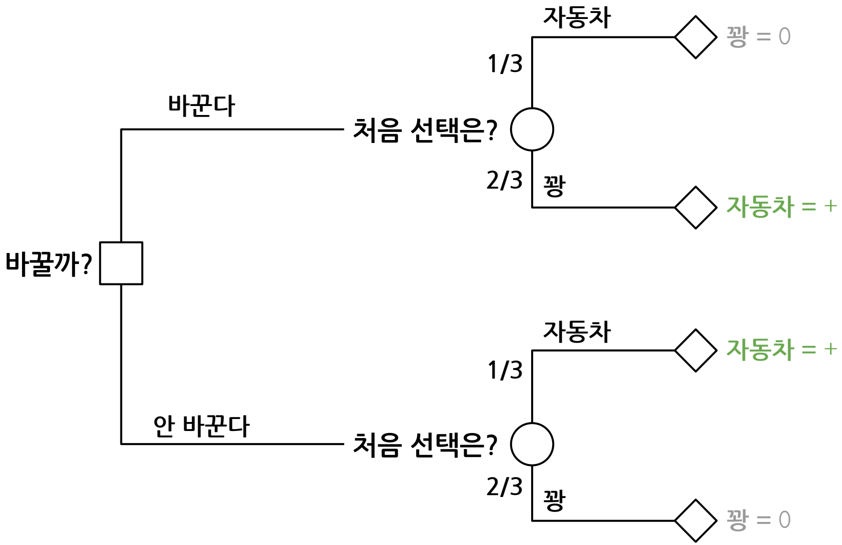


의사결정 나무 (Decision tree)

선택지: 선택한 문을 바꿀까?

불확실성: 처음 선택이 꽝?

가치: 자동차 = +, 꽝 = 0

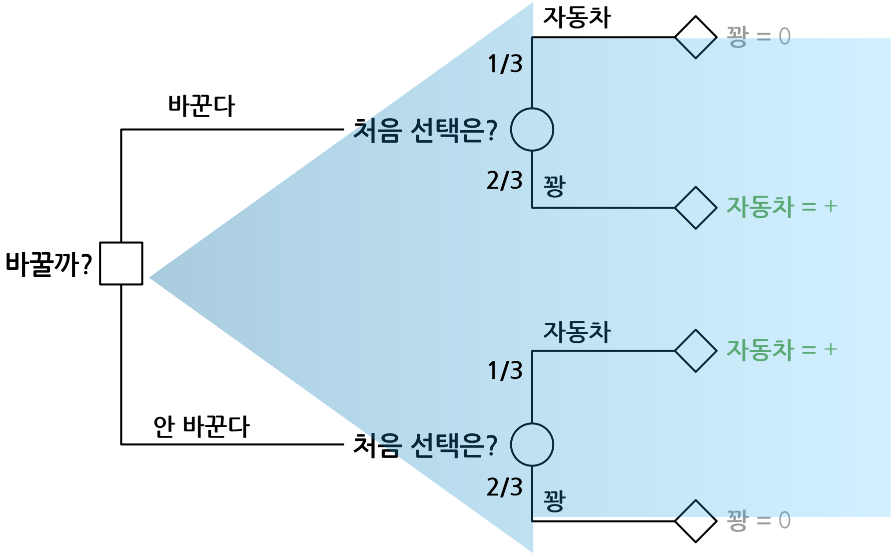


의사결정 나무 (Decision tree)

선택지: 선택한 문을 바꿀까?

불확실성: 처음 선택이 꽝?

가치: 자동차 = +, 꽝 = 0

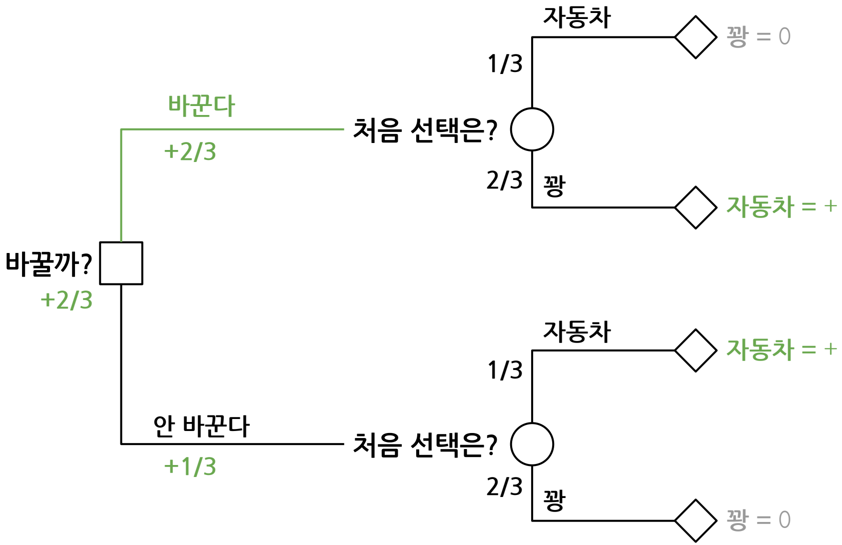


의사결정 나무 (Decision tree): 계산

선택지: 선택한 문을 바꿀까?

불확실성: 처음 선택이 꽝?

가치: 자동차 = +, 꽝 = 0



의사결정 나무 (Decision tree): 계산

다음: 빈도주의 통계 확률모형

1부: 불확실성과 데이터

빈도주의 통계 확률모형

데이터와 의사결정 | 정종빈

1부 강의 목적 및 개요

체계적이고 일반적인 <u>의사결정 모형</u> 소개

불확실성을 계량화하는 두 가지 접근법:

- **빈도주의(frequentist) 통계**
  Estimand, estimator, estimate, 신뢰구간, bootstrap
- **베이지안(Bayesian) 확률**
  Prior, likelihood, posterior (손계산 → 복잡한 모델링)

의사결정의 3요소

선택지
Alternatives

불확실성/정보
Uncertainty/information

가치/선호
Value/preference

불확실성을 계량적으로 표현하는 전략 → 확률모형

절대적 진리

불확실성 無, 관찰이 불가능 할 수도 있음

- 대한민국 아이폰 유저 비중
- 만 31세 국민 키의 평균
- 동전을 무한히 던졌을 때 뒷면 대비 앞면이 나오는 빈도

빈도주의 확률모형

관심 변수

표현하고자하는 불확실성

절대적 진리

불확실성 無, 관찰이 불가능 할 수도 있음

- 대한민국 아이폰 유저 비중
- 만 31세 국민 키의 평균
- 동전을 무한히 던졌을 때 뒷면 대비 앞면이 나오는 빈도

빈도주의 확률모형

관심 변수

표현하고자하는 불확실성

절대적 진리

불확실성 無, 관찰이 불가능 할 수도 있음

Data generating process

관찰된 정보

“절대적 진리"와 연관된 현상

- 대한민국 아이폰 유저 비중
- 만 31세 국민 키의 평균
- 동전을 무한히 던졌을 때 뒷면 대비 앞면이 나오는 빈도

- 지하철 같은 칸에서 폰을 꺼낸 사람 /아이폰을 꺼낸 사람의 수
- 우리팀 31세 8 명의 키
- 동전 10번 던져 앞면이 나온 횟수

빈도주의 확률모형

관심 변수

표현하고자하는 불확실성

절대적 진리

불확실성 無, 관찰이 불가능 할 수도 있음

- 대한민국 아이폰 유저 비중
- 만 31세 국민 키의 평균
- 동전을 무한히 던졌을 때 뒷면 대비 앞면이 나오는 빈도

Data generating process

관찰된 정보

“절대적 진리"와 연관된 현상

- 지하철 같은 칸에서 폰을 꺼낸 사람 /아이폰을 꺼낸 사람의 수
- 우리팀 31세 8 명의 키
- 동전 10번 던져 앞면이 나온 횟수

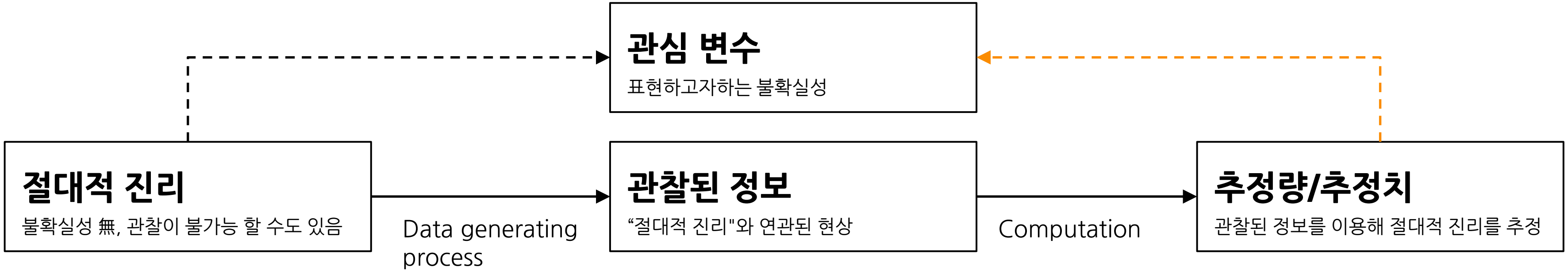
Computation

추정량/추정치

관찰된 정보를 이용해 절대적 진리를 추정

- 아이폰 꺼낸 사람 수 / 폰 꺼낸 사람 수
- 8명 키의 평균
- 앞면 나온 횟수 / 10

빈도주의 확률모형



- 대한민국 아이폰 유저 비중
- 만 31세 국민 키의 평균
- 동전을 무한히 던졌을 때 뒷면 대비 앞면이 나오는 빈도

- 지하철 같은 칸에서 폰을 꺼낸 사람 /아이폰을 꺼낸 사람의 수
- 우리팀 31세 8 명의 키
- 동전 10번 던져 앞면이 나온 횟수

- 아이폰 꺼낸 사람 수 / 폰 꺼낸 사람 수
- 8명 키의 평균
- 앞면 나온 횟수 / 10

빈도주의 확률모형

새로 관찰된 폰이 안드로이드 일 확률(p)?

관심 변수 Estimand
표현하고자하는 불확실성

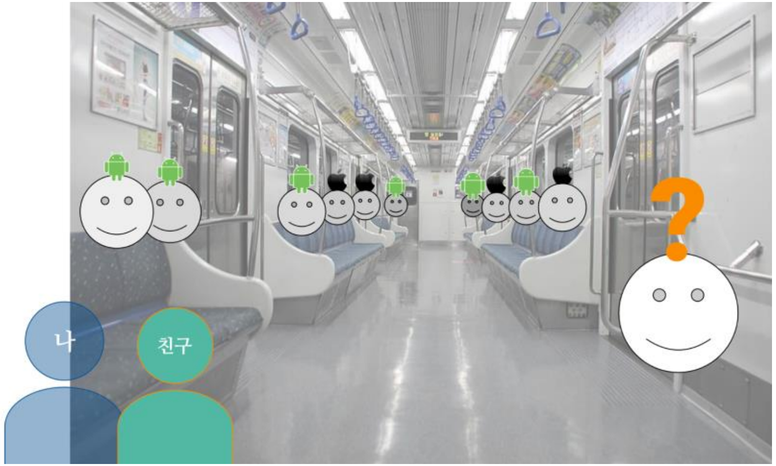
절대적 진리
불확실성 無, 관찰이 불가능 할 수도 있음

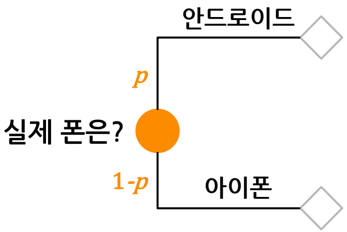
Data generating process

관찰된 정보
“절대적 진리"와 연관된 현상

Computation

추정량/추정치
관찰된 정보를 이용해 절대적 진리를 추정





빈도주의 확률모형

새로 관찰된 폰이 안드로이드 일 확률(p)?

관심 변수
표현하고자하는 불확실성

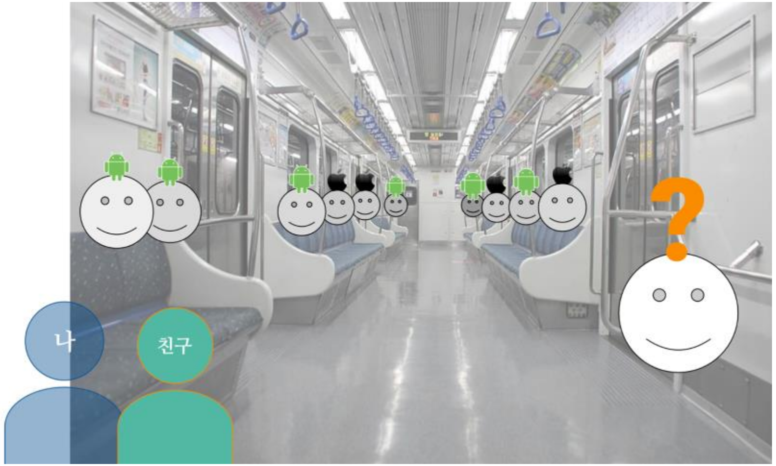
절대적 진리
불확실성 無, 관찰이 불가능 할 수도 있음

Data generating process

관찰된 정보
“절대적 진리"와 연관된 현상

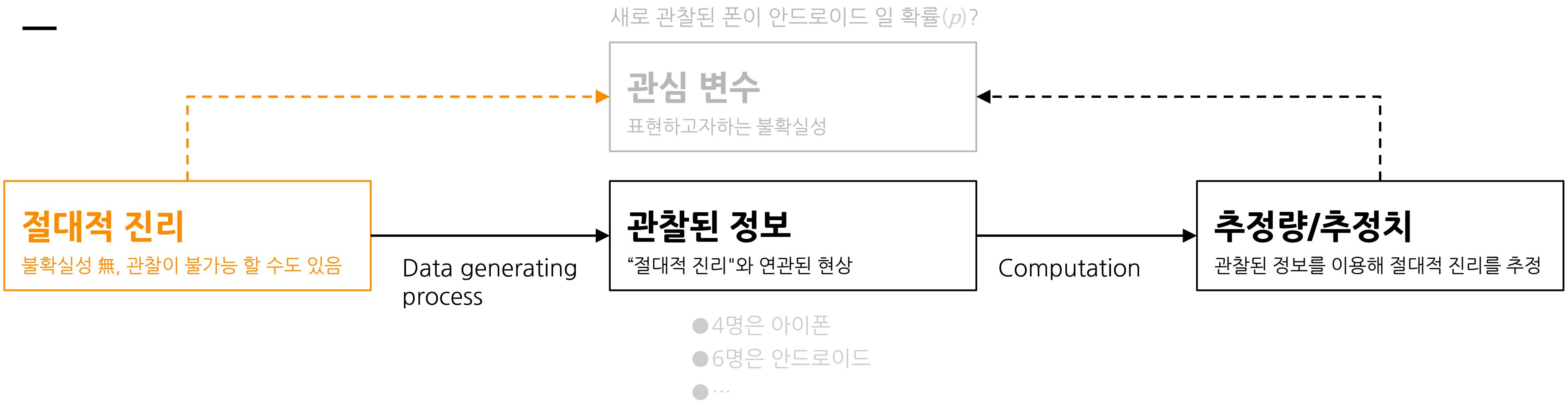
Computation

추정량/추정치
관찰된 정보를 이용해 절대적 진리를 추정



- 4명은 아이폰
- 6명은 안드로이드
- …

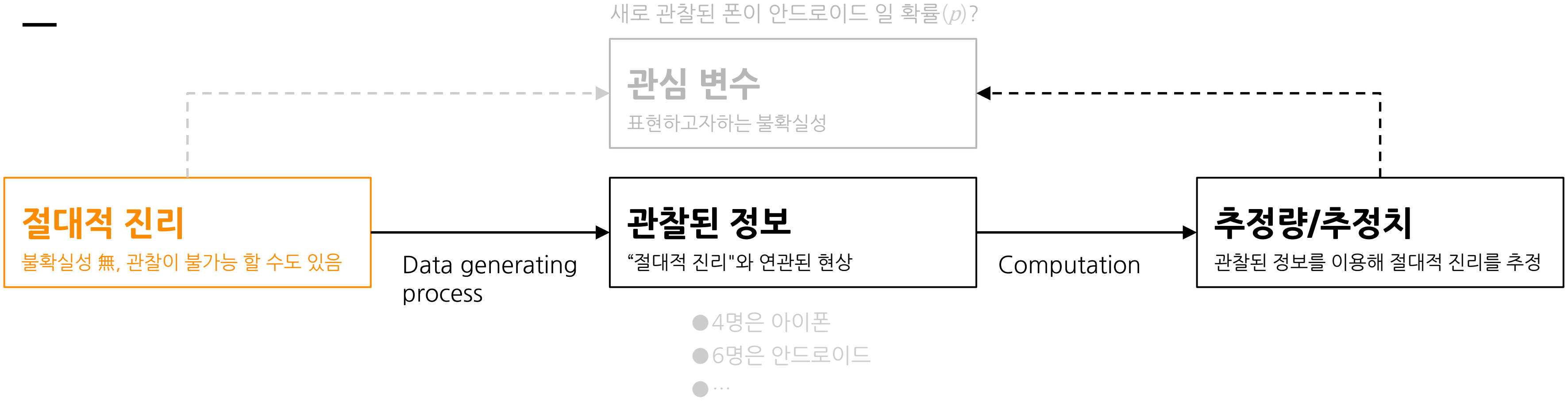
빈도주의 확률모형



“어쩌다 이 데이터를 보게 됐을까?” → 새롭게 볼 데이터에 대한 이해/기대

- 절대적 진리에는 **불확실성 無**, “확률”과 “불확실성”은 우리가 관찰하게 된 데이터를 통해 유입
- 주로 현실적인 제약(데이터의 부재, 문제의 복잡도 등)으로 인해 상당히 단순화
- 대체로, 무리한 가정을 많이 요구하는 단순한 모형이 실증적으로 유용한 경우가 많음

빈도주의 확률모형

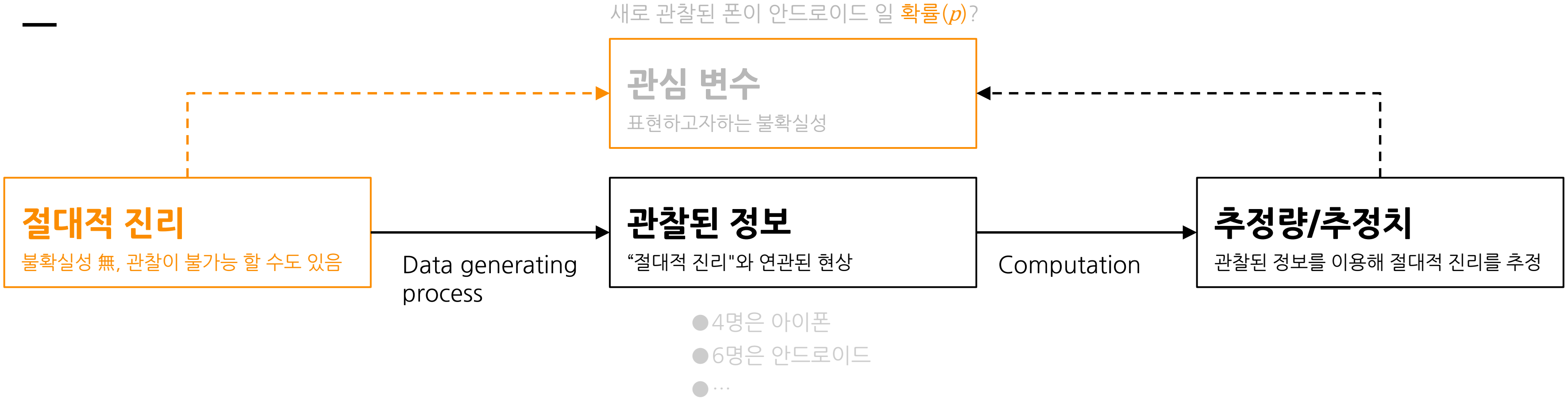


현실적 단순모형 ●——● 사실적 복잡모형

- 이 지하철(칸)을 탄 전체 인구
- 그 중 안드로이드 유저의 수

- 각 역에서 이 지하철을 탄 인구 수
- 그 중 나이, 성별, … 에 따른 안드로이드 유저의 수
- 안드로이드 유저가 아이폰 유저 대비 지하철에서 폰을 꺼내는 경우의 수
- …

빈도주의 확률모형



현실적 단순모형

- 이 지하철(칸)을 탄 전체 인구
- 그 중 안드로이드 유저의 수

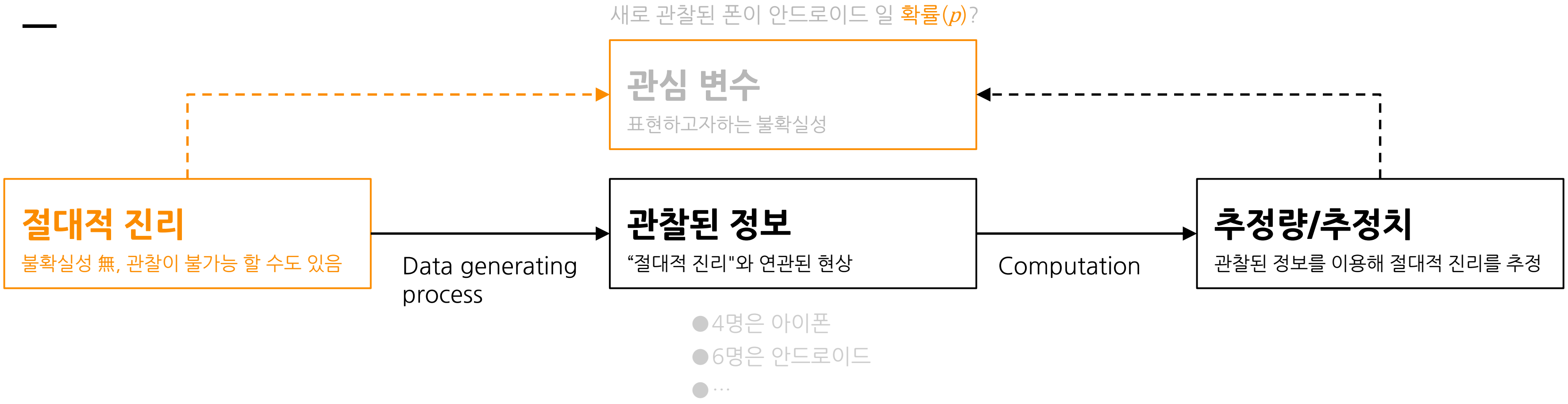
빈도주의자의

확률 = 여러번 반복했을 때, 발생하는 빈도의 비중

동전을 던져 앞면이 나올 “확률”

주사위를 던져 6이 나올 “확률”

빈도주의 확률모형



현실적 단순모형

- 이 지하철(칸)을 탄 전체 인구
- 그 중 안드로이드 유저의 수
- p = (안드로이드 유저의 수) / (전체 인구)

빈도주의 확률모형

새로 관찰된 폰이 안드로이드 일 확률(p)?

관심 변수

표현하고자하는 불확실성

절대적 진리

불확실성 無, 관찰이 불가능 할 수도 있음

Data generating process

관찰된 정보

“절대적 진리"와 연관된 현상

Computation

추정량/추정치

관찰된 정보를 이용해 절대적 진리를 추정

- 4명은 아이폰
- 6명은 안드로이드
- …

현실적 단순모형

- 이 지하철(칸)을 탄 전체 인구
- 그 중 안드로이드 유저의 수
- p = (안드로이드 유저의 수) / (전체 인구)

?

빈도주의 확률모형 → 불확실성의 근원은 “데이터”

새로 관찰된 폰이 안드로이드 일 확률(p)?

관심 변수

표현하고자하는 불확실성

절대적 진리

불확실성 無, 관찰이 불가능 할 수도 있음

Data generating process

관찰된 정보

“절대적 진리"와 연관된 현상

Computation

추정량/추정치

관찰된 정보를 이용해 절대적 진리를 추정

- 4명은 아이폰
- 6명은 안드로이드
- …

현실적 단순모형

- 이 지하철(칸)을 탄 전체 인구
- 그 중 안드로이드 유저의 수
- p = (안드로이드 유저의 수) / (전체 인구)

앞면(안드로이드)이 나올 확률이 p인 동전을 **10번** 던졌더니 **앞면이 6번** 나왔다

p의 값은 무엇일까??

빈도주의 확률모형

새로 관찰된 폰이 안드로이드 일 확률(p)?

관심 변수

표현하고자하는 불확실성

절대적 진리

불확실성 無, 관찰이 불가능 할 수도 있음

Data generating process

관찰된 정보

“절대적 진리"와 연관된 현상

- 4명은 아이폰
- 6명은 안드로이드
- …

Computation

추정량/추정치

관찰된 정보를 이용해 절대적 진리를 추정

$\hat{p}$

현실적 단순모형

- 이 지하철(칸)을 탄 전체 인구
- 그 중 안드로이드 유저의 수
- p = (안드로이드 유저의 수) / (전체 인구)

앞면(안드로이드)이 나올 확률이 p인 동전을 **10번** 던졌더니 **앞면이 6번** 나왔다

p의 값은 무엇일까??

추정량(estimator)

주어진 정보(“10번 중 앞면 6번”)를 이용해 관심변수(“앞면이 나올 확률”)를 추정하는 전략/함수

빈도주의 확률모형

새로 관찰된 폰이 안드로이드 일 확률(p)?

관심 변수

표현하고자하는 불확실성

절대적 진리

불확실성 無, 관찰이 불가능 할 수도 있음

Data generating process

관찰된 정보

“절대적 진리"와 연관된 현상

Computation

추정량/추정치

관찰된 정보를 이용해 절대적 진리를 추정

- 4명은 아이폰
- 6명은 안드로이드
- …

$$\hat{p}$$

현실적 단순모형

- 이 지하철(칸)을 탄 전체 인구
- 그 중 안드로이드 유저의 수
- p = (안드로이드 유저의 수) / (전체 인구)

앞면(안드로이드)이 나올 확률이 p인 동전을
10번 던졌더니 **앞면이 6번** 나왔다
p의 값은 무엇일까??

= (앞면이 나온 수) / (던진 횟수)
= (안드로이드 유저) / (관찰한 승객 수)
= 6 / 10
= 0.6
^^^ 추정치 (estimate)

빈도주의 확률모형

새로 관찰된 폰이 안드로이드 일 확률(p)?

관심 변수

표현하고자하는 불확실성

절대적 진리

불확실성 無, 관찰이 불가능 할 수도 있음

Data generating process

관찰된 정보

“절대적 진리"와 연관된 현상

Computation

추정량/추정치

관찰된 정보를 이용해 절대적 진리를 추정

현실적 단순모형

- 이 지하철(칸)을 탄 전체 인구
- 그 중 안드로이드 유저의 수
- p = (안드로이드 유저의 수) / (전체 인구)

$$\hat{p}$$

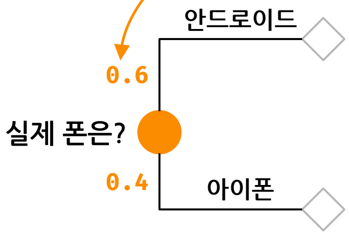
= (앞면이 나온 수) / (던진 횟수)

= (안드로이드 유저) / (관찰한 승객 수)

= 6 / 10

= 0.6

^^^ 추정치 (estimate)



빈도주의 확률모형

새로 관찰된 폰이 안드로이드 일 확률(p)?

관심 변수
표현하고자하는 불확실성

절대적 진리
불확실성 無, 관찰이 불가능 할 수도 있음

Data generating process

관찰된 정보
“절대적 진리"와 연관된 현상

Computation

추정량/추정치
관찰된 정보를 이용해 절대적 진리를 추정

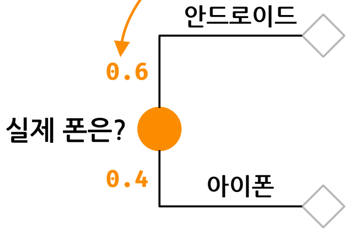
$$\hat{p}$$

= (앞면이 나온 수) / (던진 횟수)
= (안드로이드 유저) / (관찰한 승객 수)
= 6 / 10
= 0.6
^^^ 추정치 (estimate)

현실적 단순모형

- 이 지하철(칸)을 탄 전체 인구
- 그 중 안드로이드 유저의 수
- p = (안드로이드 유저의 수) / (전체 인구)

그런데, 10 명의 데이터로 계산한 이 0.6은 과연 얼마나 정확할까?



빈도주의 확률모형

다음: 빈도주의와 평행우주론

빈도주의 통계에서의 불확실성 계량 전략에 관하여

1부: 불확실성과 데이터

빈도주의와 평행우주론

빈도주의 통계에서의 불확실성 계량 전략에 관하여

데이터와 의사결정 | 정종빈

1부 강의 목적 및 개요

체계적이고 일반적인 의사결정 모형 소개

불확실성을 계량화하는 두 가지 접근법:

- **빈도주의(frequentist) 통계**
  Estimand, estimator, estimate, 신뢰구간, bootstrap
- **베이지안(Bayesian) 확률**
  Prior, likelihood, posterior (손계산 → 복잡한 모델링)

새로 관찰된 폰이 안드로이드 일 확률(p)?

관심 변수
표현하고자하는 불확실성

절대적 진리
불확실성 無, 관찰이 불가능 할 수도 있음

Data generating process

관찰된 정보
“절대적 진리"와 연관된 현상

Computation

추정량/추정치
관찰된 정보를 이용해 절대적 진리를 추정

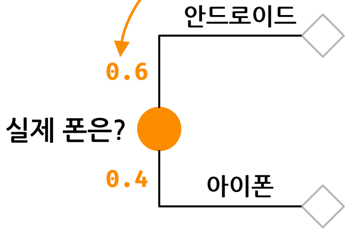
$$\hat{p}$$

현실적 단순모형

- 이 지하철(칸)을 탄 전체 인구
- 그 중 안드로이드 유저의 수
- p = (안드로이드 유저의 수) / (전체 인구)

그런데, 10 명의 데이터로 계산한 이 0.6은 과연 얼마나 정확할까?

= (앞면이 나온 수) / (던진 횟수)
= (안드로이드 유저) / (관찰한 승객 수)
= 6 / 10
= 0.6
^^^ 추정치 (estimate)



빈도주의 확률모형

새로 관찰된 폰이 안드로이드 일 확률(p)?

관심 변수
표현하고자하는 불확실성

절대적 진리
불확실성 無, 관찰이 불가능 할 수도 있음

Data generating process

관찰된 정보
"절대적 진리"와 연관된 현상

Computation

추정량/추정치
관찰된 정보를 이용해 절대적 진리를 추정

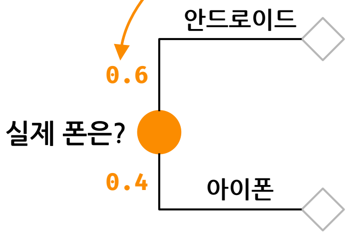
$$\hat{p}$$

= (앞면이 나온 수) / (던진 횟수)
= (안드로이드 유저) / (관찰한 승객 수)
= 6 / 10
= 0.6
^^^ 추정치 (estimate)

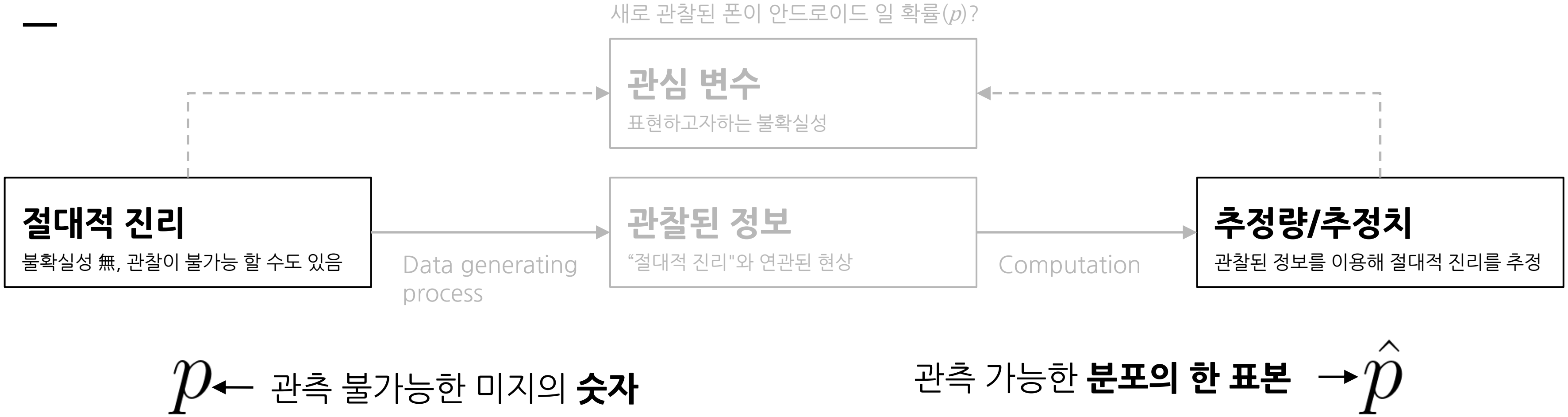
현실적 단순모형

- 이 지하철(칸)을 탄 전체 인구
- 그 중 안드로이드 유저의 수
- p = (안드로이드 유저의 수) / (전체 인구)

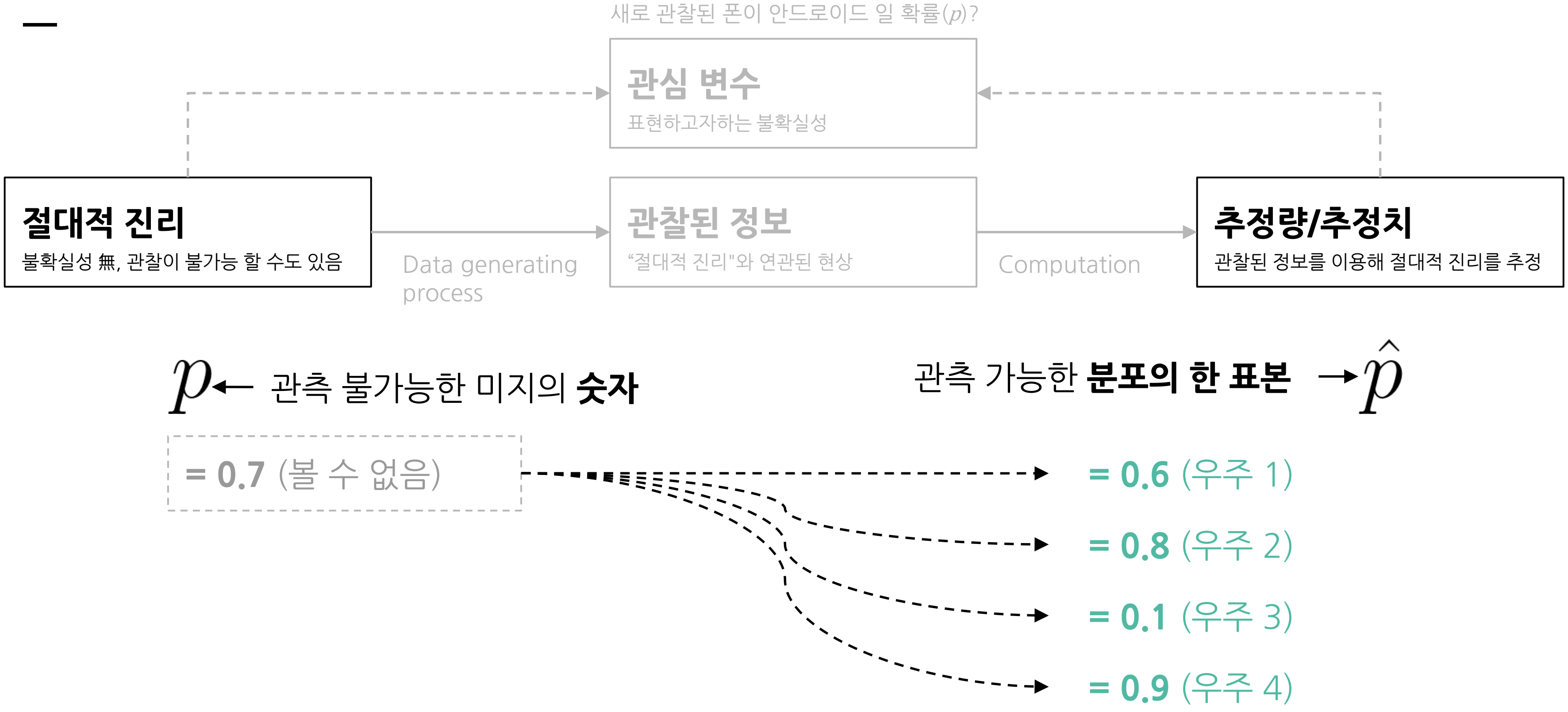
그런데, 10 명의 데이터로 계산한 이 0.6은 과연 얼마나 정확할까?
= 여러번 반복 했을 때의 성과?



빈도주의 확률모형


새로 관찰된 폰이 안드로이드 일 확률(p)?
관심 변수
표현하고자하는 불확실성
절대적 진리
불확실성 無, 관찰이 불가능 할 수도 있음
Data generating process
관찰된 정보
"절대적 진리"와 연관된 현상
Computation
추정량/추정치
관찰된 정보를 이용해 절대적 진리를 추정
p ← 관측 불가능한 미지의 **숫자**
관측 가능한 **분포의 한 표본** → $\hat{p}$

추정량의 분포?



추정량의 분포?: 평행우주의 예

새로 관찰된 폰이 안드로이드 일 확률(p)?

관심 변수
표현하고자하는 불확실성

절대적 진리
불확실성 無, 관찰이 불가능 할 수도 있음

Data generating process

관찰된 정보
"절대적 진리"와 연관된 현상

Computation

추정량/추정치
관찰된 정보를 이용해 절대적 진리를 추정

$p \leftarrow$ 관측 불가능한 미지의 **숫자**

관측 가능한 **분포의 한 표본** $\rightarrow \hat{p}$

그런데, 10 명의 데이터로 계산한 이 0.6은
과연 얼마나 정확할까?
= 여러번 반복 했을 때의 성과?
= 수많은 평행우주에 걸쳐 추정치는 어떻게 달랐을까?

0.6 (우주 1)

0.8 (우주 2)

0.1 (우주 3)

0.9 (우주 4)

빈도주의 확률모형: 평행우주의 예

새로 관찰된 폰이 안드로이드 일 확률(p)?

관심 변수
표현하고자하는 불확실성

절대적 진리
불확실성 無, 관찰이 불가능 할 수도 있음

Data generating process

관찰된 정보
"절대적 진리"와 연관된 현상

Computation

추정량/추정치
관찰된 정보를 이용해 절대적 진리를 추정

$p \leftarrow$ 관측 불가능한 미지의 **숫자**

관측 가능한 **분포의 한 표본** $\rightarrow \hat{p}$

그런데, 10 명의 데이터로 계산한 이 0.6은
과연 얼마나 정확할까?
= 여러번 반복 했을 때의 성과?
= 수많은 평행우주에 걸쳐
추정치는 어떻게 달랐을까?
= 표본분포 (Sampling distribution)

0.6 (우주 1)
0.8 (우주 2)
0.1 (우주 3)
0.9 (우주 4)

빈도주의(frequentist)의 꽃: 표본분포!

다음: 표본분포 코딩 실습

1부: 불확실성과 데이터

표본분포 코딩 실습

데이터와 의사결정 | 정종빈

Google Colab에서 진행

numpy와 seaborn이 설치된 환경 어디든 실습 가능

다음: 표본분포 활용

1부: 불확실성과 데이터

표본분포 활용

데이터와 의사결정 | 정종빈

정리

- 표본분포의 불확실성(분산)은 표본의 크기(n)의 영향
- 표준오차(standard error) = 표본분포의 표준편차

추정치가 실제값에 얼마나 가까울까?

- 실제값(p)은 알 수 없지만, **변치않는 진리**, 불확실 X
  어느 우주에 있든지, 항상 같은 값
- 불확실한건 **추정량**과 **표본분포**
  어느 우주에서 왔느냐에 따라 값이 다름

추정치가 실제값에 얼마나 가까울까?

$$\hat{p} - p$$

(95%) 신뢰구간
95%의 확신을 가지고,
"아, 실제값이 이 범위 안에는 있겠다"
… 싶은 구간

= 실제값이 이론적으로 표본분포의 95% 확률 범위 안에 있도록 설정

같은 신뢰도라면, 짧은 구간을 선호

추정치가 실제값에 얼마나 가까울까?

$$a \leq \hat{p} - p \leq b$$

(95%) 신뢰구간

표본분포에서, 확률밀도가 95%인 지점 a와 b를 찾기

추정치가 실제값에 얼마나 가까울까?

$$a \leq \hat{p} - p \leq b$$

(95%) 신뢰구간

표본분포에서, 확률밀도가 95%인 지점 a와 b를 찾기

1. 표본분포가 정규분포를 따른다고 가정하면?

추정치가 실제값에 얼마나 가까울까?

$$-1.95s \le \hat{p} - p \le 1.95s$$

(95%) 신뢰구간

표본분포에서, 확률밀도가 95%인 지점 a와 b를 찾기

1. 표본분포가 정규분포를 따른다고 가정하면? s는 표본분포의 표준편차(표준오차)

추정치가 실제값에 얼마나 가까울까?

$$-1.95s \le \hat{p} - p \le 1.95s$$

$$\hat{p} - 1.95s \le p \le \hat{p} + 1.95s$$

(95%) 신뢰구간

표본분포에서, 확률밀도가 95%인 지점 a와 b를 찾기

1. 표본분포가 정규분포를 따른다고 가정하면? s는 표본분포의 표준편차(표준오차)

추정치가 실제값에 얼마나 가까울까?

(95%) 신뢰구간

표본분포에서, 확률밀도가 95%인 지점 a와 b를 찾기

1. 표본분포가 정규분포를 따른다고 가정하면? s는 표본분포의 표준편차(표준오차)
2. 표본분포에 대한 특별한 가정 없이, 표본분포 자체를 추정할 수 있다면?

현실적으로 평행우주를 관찰 할 수 없는데?

두 가지 접근법

표본분포에 대한 수학적 분석 및 증명 (전통통계)

- 예: 모든 “평균"값은 관측 수가 증가하면 정규분포로 수렴
- 데이터에 대한 가정과 추정량이 복잡해질수록 비현실적 난이도

두 가지 접근법

표본분포에 대한 수학적 분석 및 증명 (전통통계)

- 예: 모든 “평균"값은 관측 수가 증가하면 정규분포로 수렴
- 데이터에 대한 가정과 추정량이 복잡해질수록 비현실적 난이도

컴퓨터 시뮬레이션 (bootstrap)

- 광범위하게 적용 가능
- 전통적 분석방법에 비해 엄청나게 느림
- 실제 관측한 데이터의 퀄리티에 크게 영향 받음

두 가지 접근법

~~표본분포에 대한 수학적 분석 및 증명 (전통통계)~~

- ~~예: 모든 “평균"값은 관측 수가 증가하면 정규분포로 수렴~~
- ~~데이터에 대한 가정과 추정량이 복잡해질수록 비현실적 난이도~~

컴퓨터 시뮬레이션 (bootstrap)

- 광범위하게 적용 가능
- 전통적 분석방법에 비해 엄청나게 느림
- 실제 관측한 데이터의 퀄리티에 크게 영향 받음

다음: Bootstrap

평행우주 시뮬레이션

1부: 불확실성과 데이터

Bootstrap

현실적인 평행우주 시뮬레이션

데이터와 의사결정 | 정종빈

1부 강의 목적 및 개요

체계적이고 일반적인 의사결정 모형 소개

불확실성을 계량화하는 두 가지 접근법:

- **빈도주의(frequentist) 통계**
  Estimand, estimator, estimate, 신뢰구간, bootstrap
- **베이지안(Bayesian) 확률**
  Prior, likelihood, posterior (손계산 → 복잡한 모델링)

새로 관찰된 폰이 안드로이드 일 확률(p)?

관심 변수

표현하고자하는 불확실성

절대적 진리

불확실성 無, 관찰이 불가능 할 수도 있음

Data generating process

관찰된 정보

"절대적 진리"와 연관된 현상

Computation

추정량/추정치

관찰된 정보를 이용해 절대적 진리를 추정

p ← 관측 불가능한 미지의 **숫자**

관측 가능한 **분포의 한 표본** → $\hat{p}$

= **0.7** (볼 수 없음)

= **0.6** (우주 1)

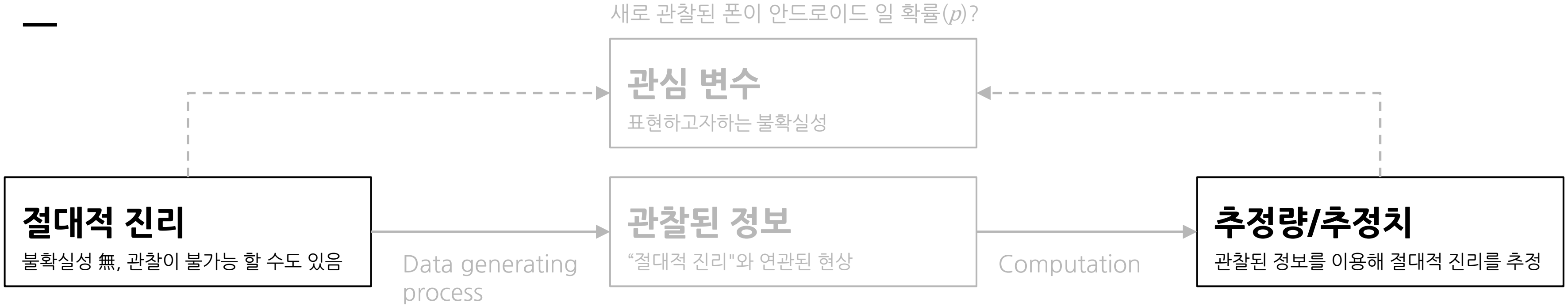
= **0.8** (우주 2)

= **0.1** (우주 3)

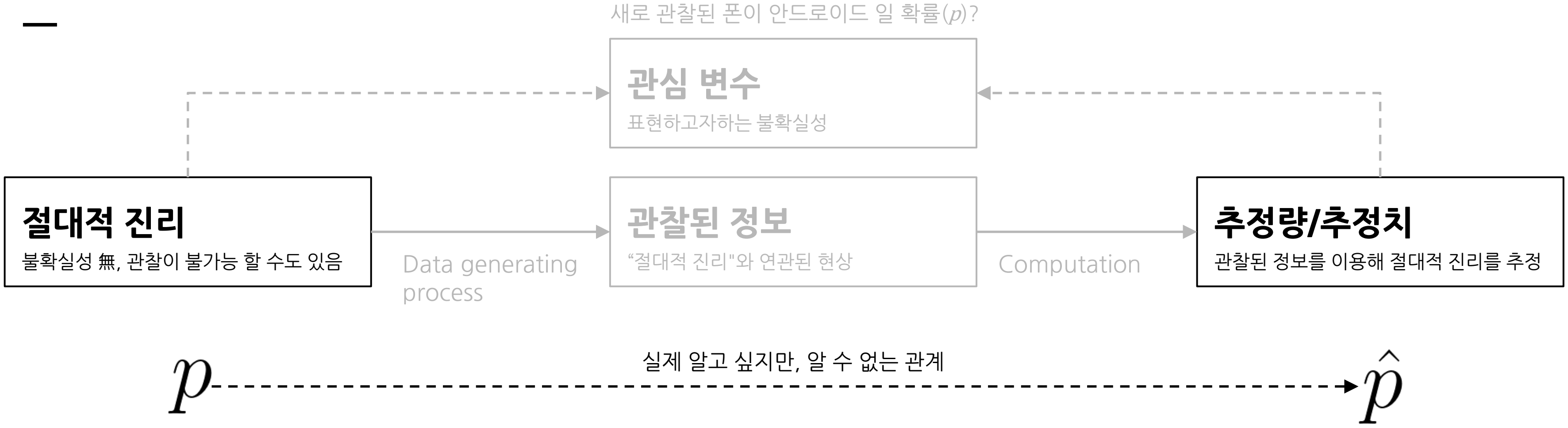
= **0.9** (우주 4)

빈도주의(frequentist)의 꽃: 표본분포!

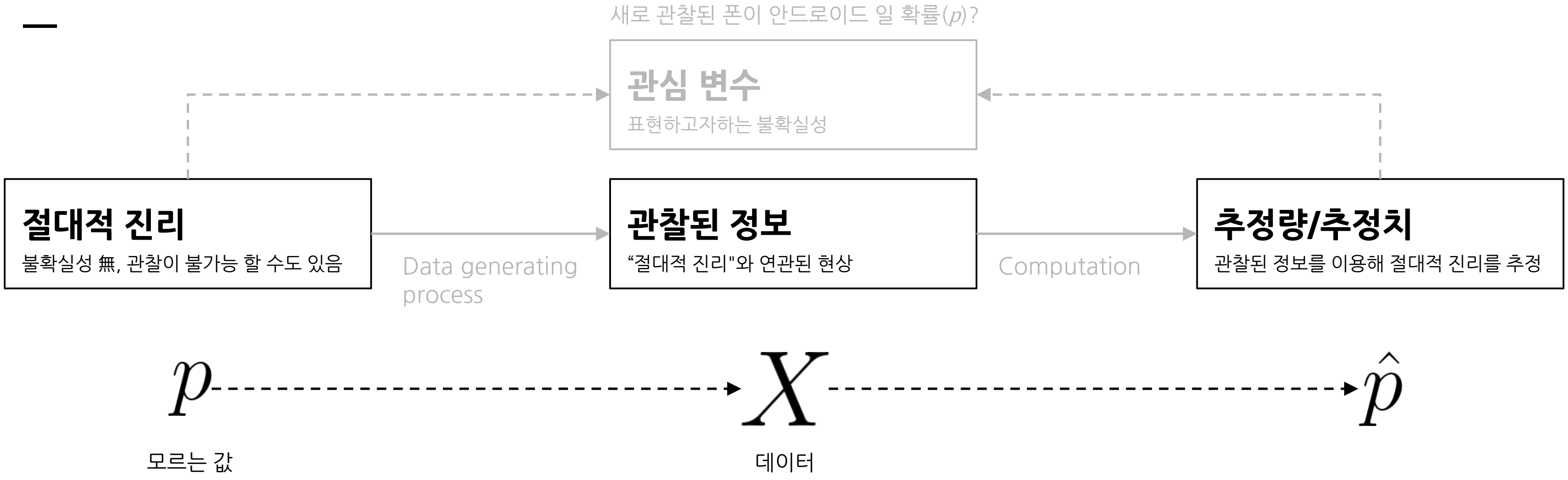
현실적으로 평행우주를 관찰 할 수 없는데?



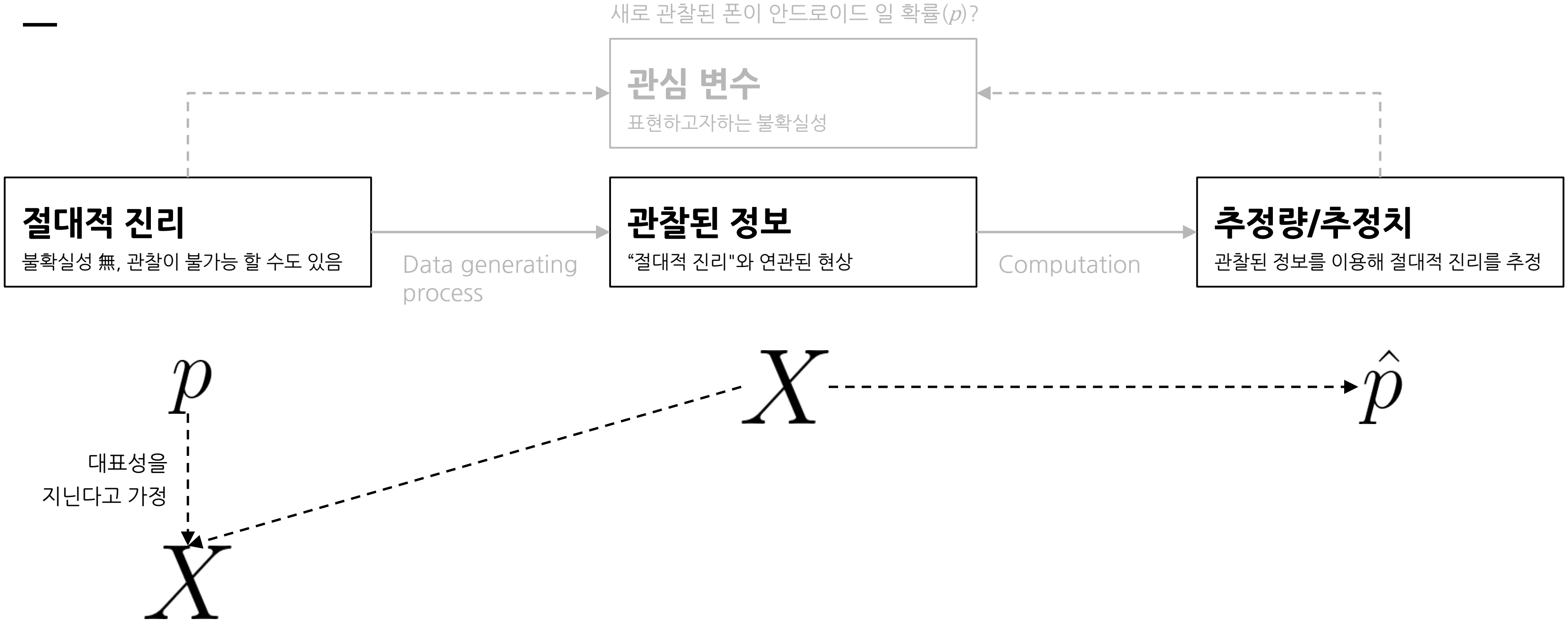
Bootstrap: 짝퉁 표본분포 만들기



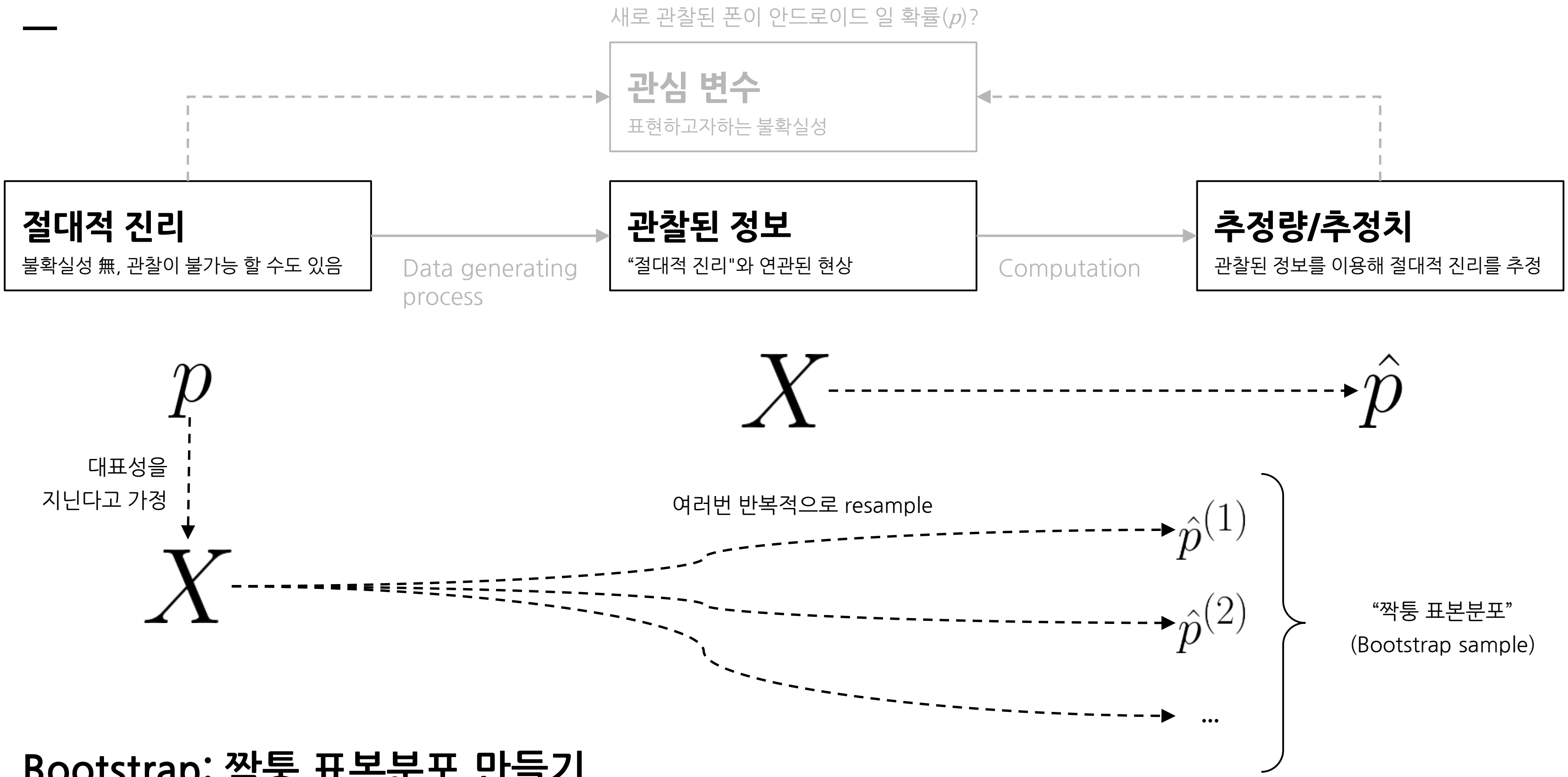
Bootstrap: 짝퉁 표본분포 만들기



Bootstrap: 짝퉁 표본분포 만들기



Bootstrap: 짝퉁 표본분포 만들기



Bootstrap: 짝퉁 표본분포 만들기

다음: Bootstrap 코딩 실습

1부: 불확실성과 데이터

Bootstrap 코딩 실습

데이터와 의사결정 | 정종빈

Google Colab에서 진행

Requires: numpy, scipy, seaborn

1부: 불확실성과 데이터

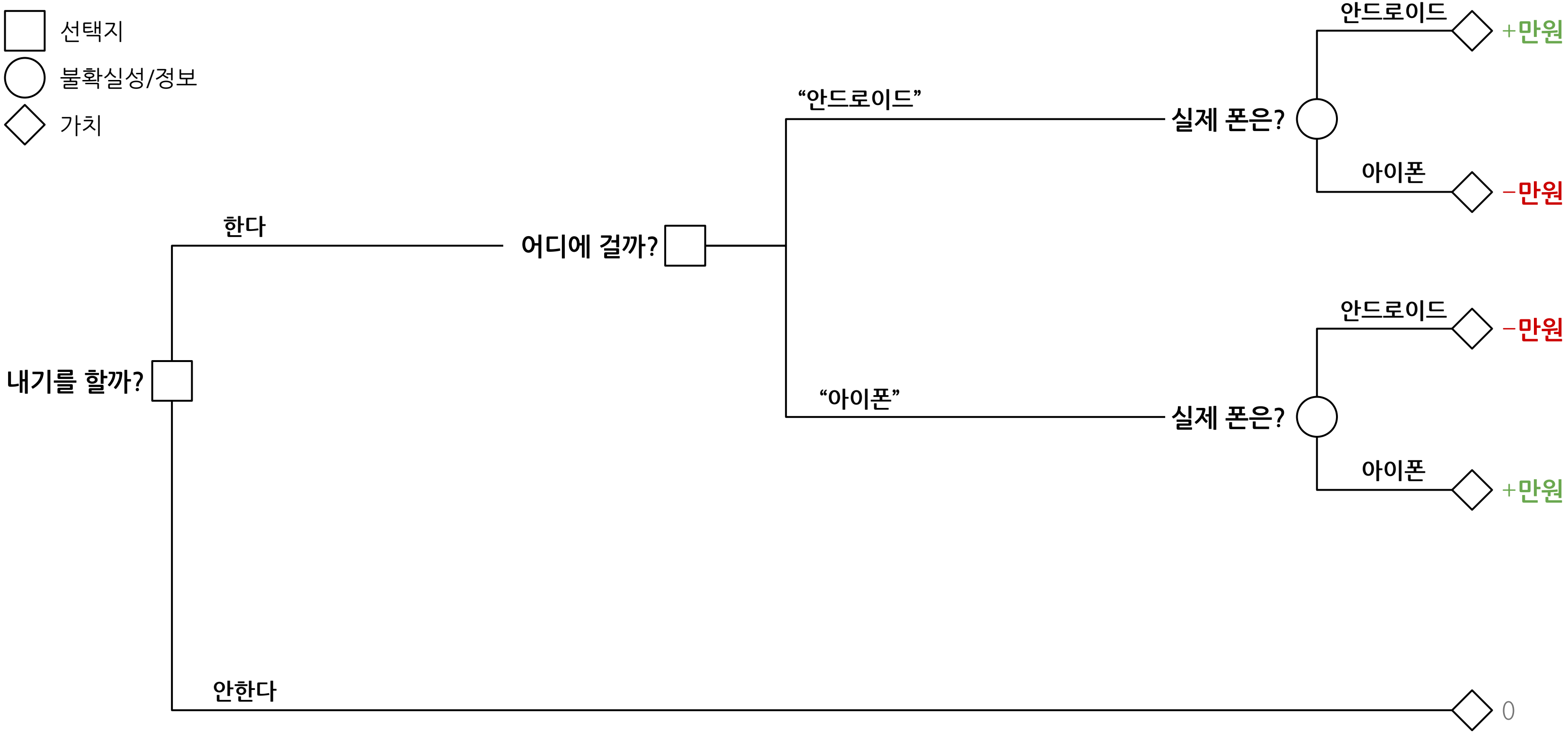
그래서, 내기를 해?

1부 강의 목적 및 개요

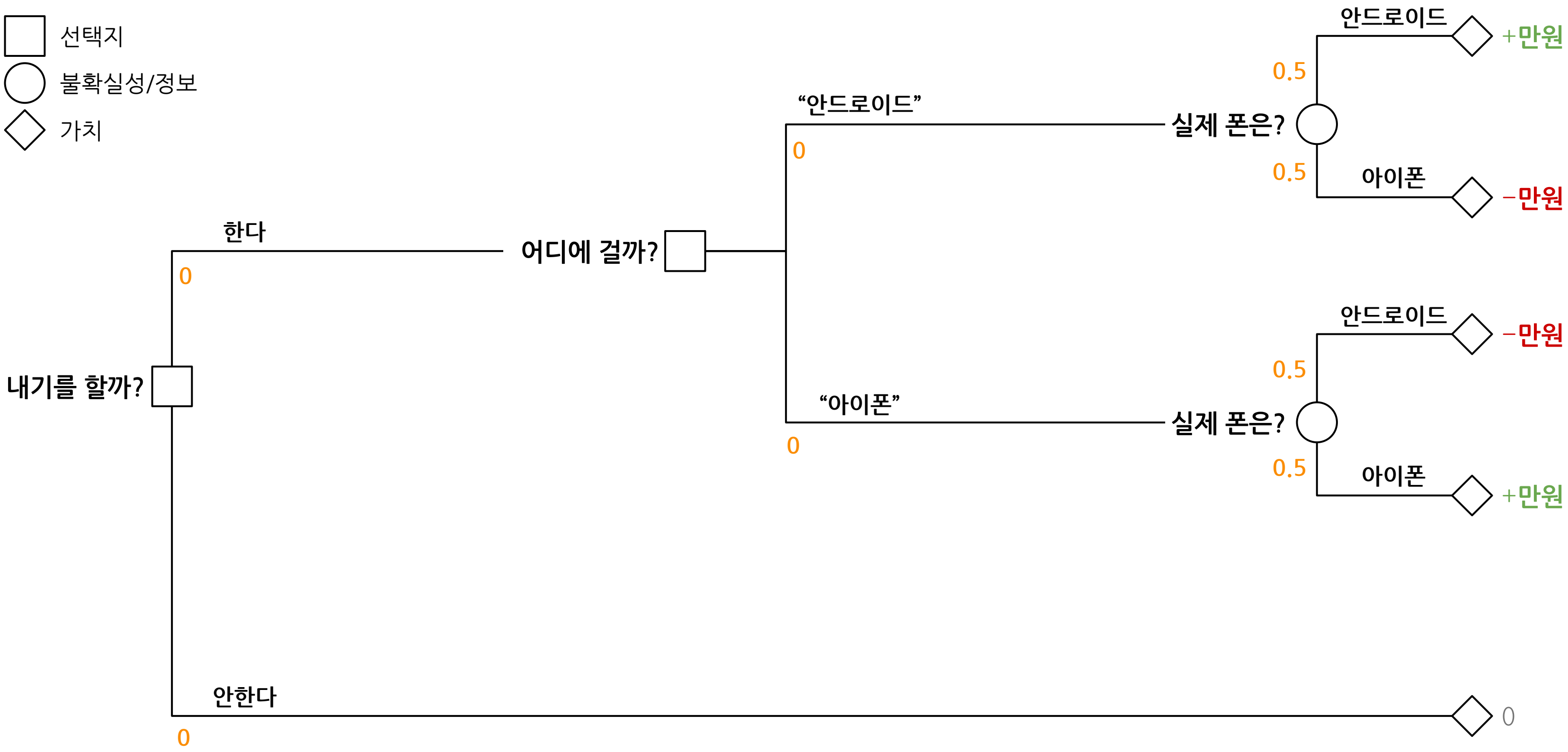
체계적이고 일반적인 의사결정 모형 소개

불확실성을 계량화하는 두 가지 접근법:

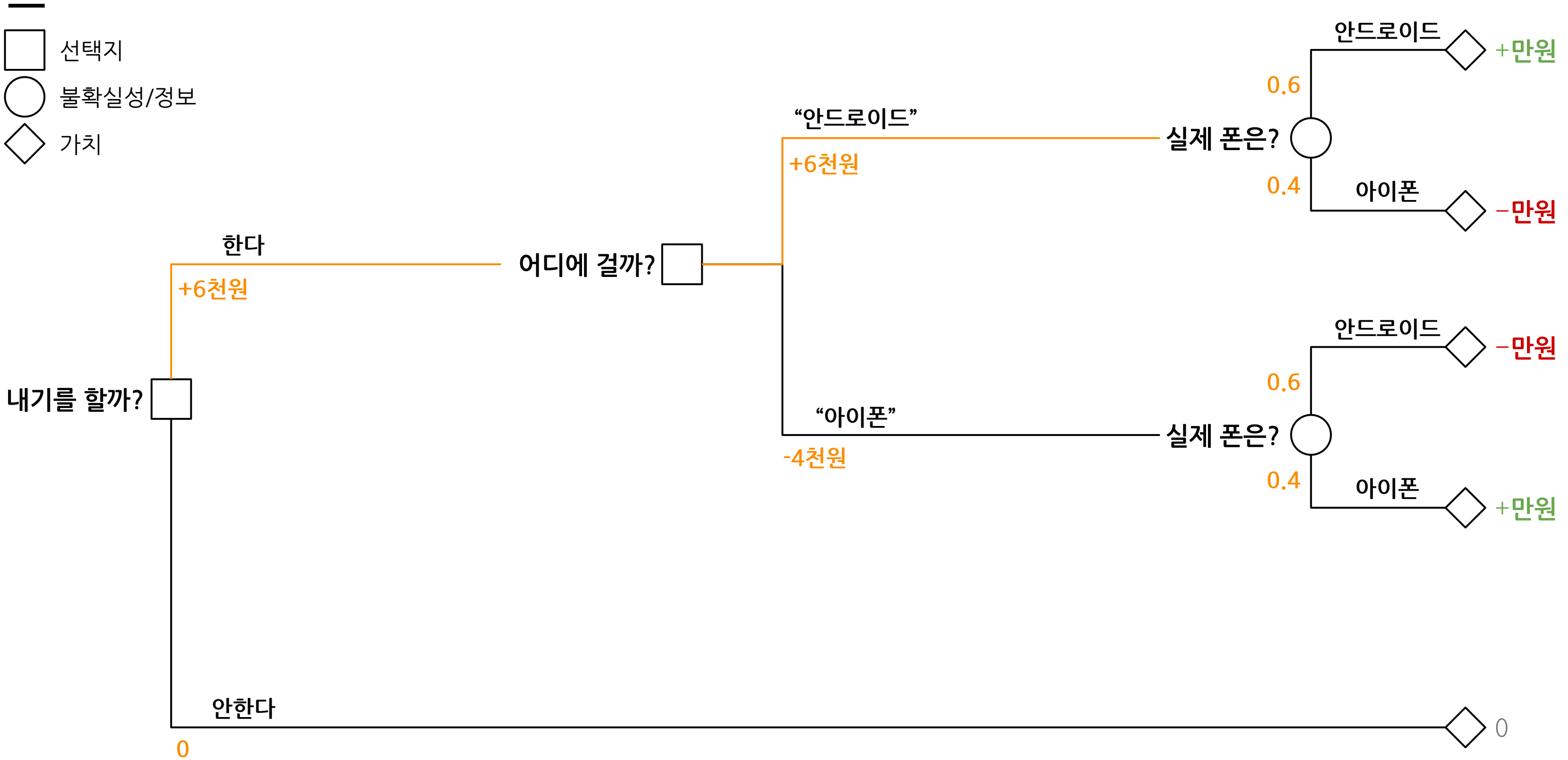
- **빈도주의(frequentist) 통계**
  Estimand, estimator, estimate, 신뢰구간, bootstrap
- **베이지안(Bayesian) 확률**
  Prior, likelihood, posterior (손계산 → 복잡한 모델링)



내기를 해야되냐, 말아야되냐?



내기를 해야되냐, 말아야되냐?



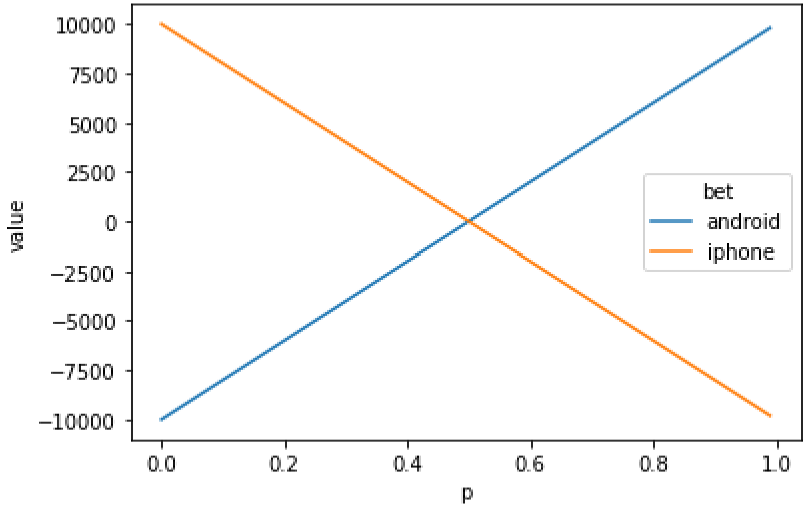
내기를 해야되냐, 말아야되냐?

```
[1] import numpy as np
    import pandas as pd
    import seaborn as sns

    bet = 10000
    p = np.arange(0, 1, .01)
```

```
[2] df = pd.concat([
      pd.DataFrame({
          "p": p,
          "bet": "android",
          "value": p * bet - (1-p) * bet,
      }),
      pd.DataFrame({
          "p": p,
          "bet": "iphone",
          "value": (1-p) * bet - p * bet,
      }),
    ]).reset_index()
    df
```

```
[3] sns.lineplot(data=df, x="p", y="value", hue="bet")
```



Sensitivity to p

궁극적인 질문

안드로이드/아이폰 유저일 확률이

50%

이상/이하라고 믿는가?

통계적 정당성(의 부재)

- 전통적인 “통계적 유의미함"
  - 95% 신뢰구간이 0.5를 포함하지 않는다면 95% 신뢰수준에서 통계적으로 유의미한 결과 (p-value < 0.05)
- 현재 결과는 통계적으로 유의미하지 않음 → 내기 안해
- 객관적인 데이터가 말해주는 정보로서 가치가 있음
- 하지만 …

현실적인 고려 사항

- **데이터의 완전성 (유용한 정보의 부재)**
  - 예: 통계적으로 유의미한 결과는 없지만, 폰을 안꺼낸 승객이 맥북 가방에, 애플워치 차고, 아이패드를 들고 있으면?
- **확률모형의 현실성**
  - 지하철 승객이 모두 “독립시행”?
  - 빈도주의자의 “확률”이 적용되는 상황?

“좋은” 의사결정

Quality of a decision

≠ 의사결정의 결과

의사결정이 “좋은”가는, 결과가 나오기 전에 판단 됨

의사결정의 6요소로 판단

- 가용한 모든 선택지, 정보, 가치를 고려 했는가
- 의사결정의 틀이 명확한가
- 의사결정자가 분명한가
- 의사결정의 논리가 바른가

다음: 베이지안 확률모형

1부: 불확실성과 데이터

베이지안 확률모형

데이터와 의사결정 | 정종빈

1부 강의 목적 및 개요

체계적이고 일반적인 <u>의사결정 모형</u> 소개

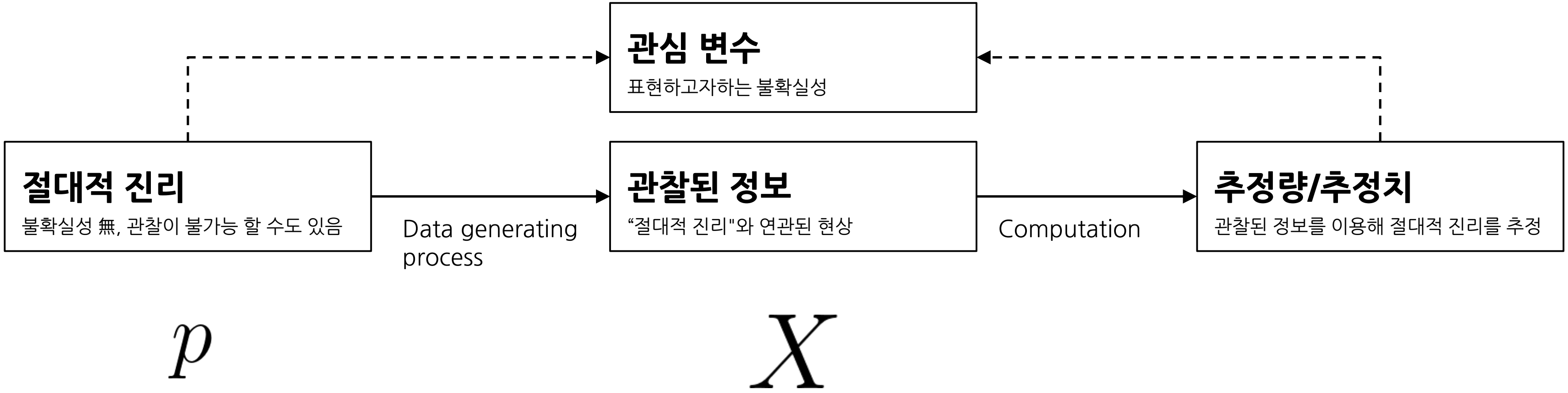
불확실성을 계량화하는 두 가지 접근법:

- **빈도주의(frequentist) 통계**
  Estimand, estimator, estimate, 신뢰구간, bootstrap
- **베이지안(Bayesian) 확률**
  Prior, likelihood, posterior (손계산 → 복잡한 모델링)

의사결정의 3요소


선택지
Alternatives
불확실성/정보
Uncertainty/information
가치/선호
Value/preference

불확실성을 계량적으로 표현하는 전략 → 확률모형


관심 변수
표현하고자하는 불확실성
절대적 진리
불확실성 無, 관찰이 불가능 할 수도 있음
Data generating process
관찰된 정보
“절대적 진리"와 연관된 현상
Computation
추정량/추정치
관찰된 정보를 이용해 절대적 진리를 추정
p
X

빈도주의 확률모형

베이지안 확률모형

관심 변수

표현하고자하는 불확실성

절대적 진리

불확실성 無, 관찰이 불가능 할 수도 있음

p

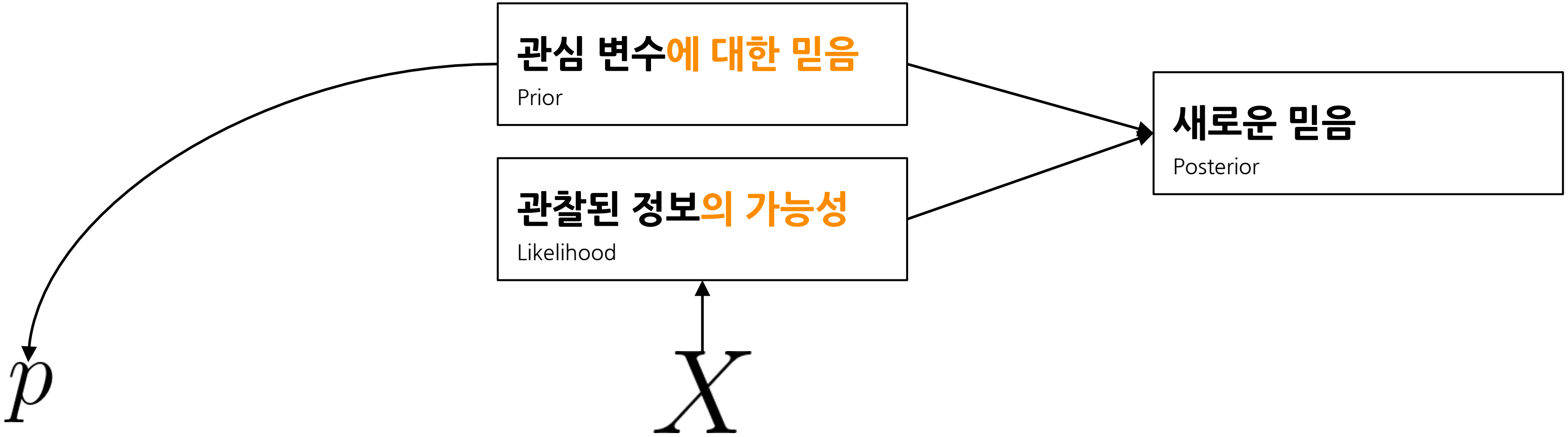
관찰된 정보

“절대적 진리"와 연관된 현상

X

추정량/추정치

관찰된 정보를 이용해 절대적 진리를 추정

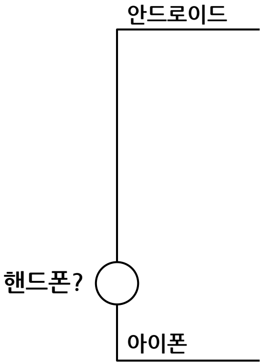


베이지안 확률모형

베이지안 확률모형

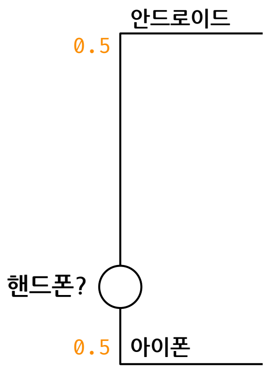
1. (확률적인) 관심변수에 대한 믿음(prior)으로 시작
2. (불확실하지 않은) 데이터의 관측
3. 데이터 관측의 가능성(likelihood) 고려
4. 관심변수에 대한 믿음 업데이트 (→ posterior)

안드로이드나 아이폰 유저일 확률이 같다고 믿음.



베이지안 확률모형: 손계산의 예

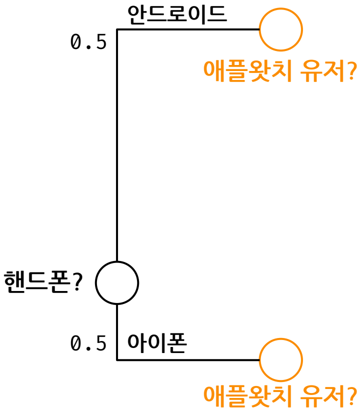
Prior



안드로이드나 아이폰 유저일 **확률이 같다고 믿음**.

1.관심변수에 대한 믿음(prior)으로 시작

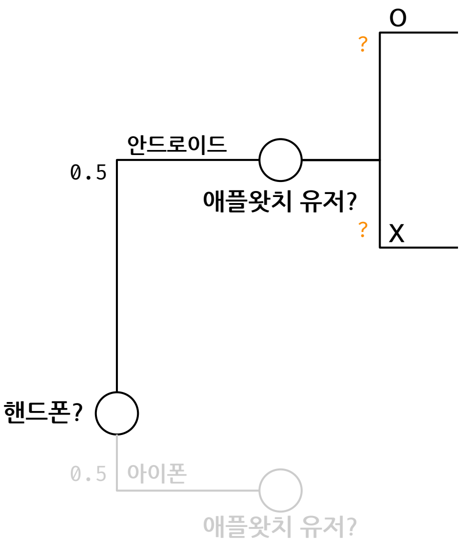
베이지안 확률모형: 손계산의 예



안드로이드나 아이폰 유저일 확률이 같다고 믿음.
그런데, **애플왓치를 차고 있는 것**을 목격.

1. 관심변수에 대한 믿음(prior)으로 시작
2. **데이터의 관측**

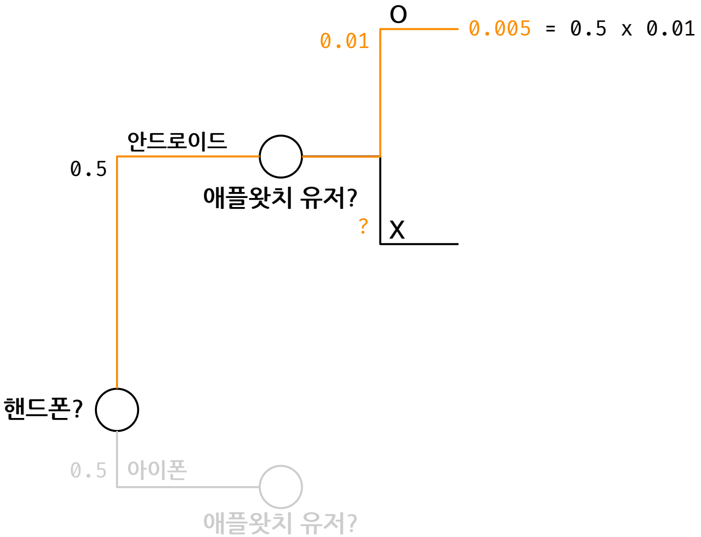
베이지안 확률모형: 손계산의 예



안드로이드나 아이폰 유저일 확률이 같다고 믿음.
그런데, 애플왓치를 차고 있는 것을 목격.

1. 관심변수에 대한 믿음(prior)으로 시작
2. 데이터의 관측
3. 데이터 관측의 가능성(likelihood) 고려

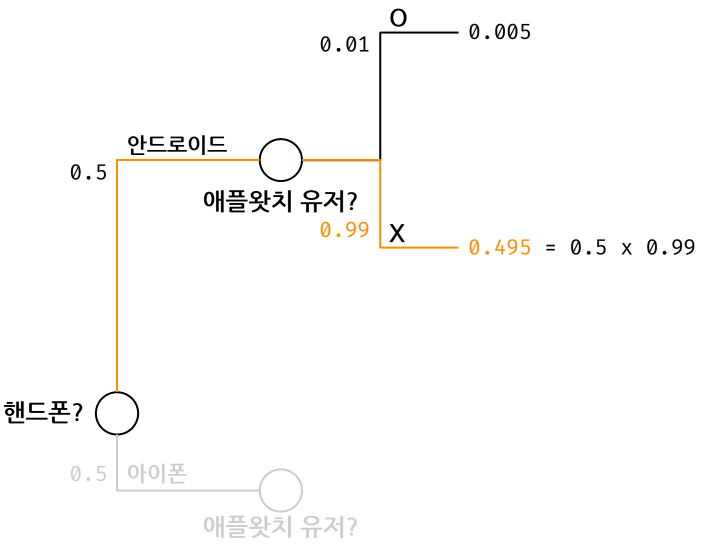
베이지안 확률모형: 손계산의 예



안드로이드나 아이폰 유저일 확률이 같다고 믿음.
그런데, 애플왓치를 차고 있는 것을 목격.

1. 관심변수에 대한 믿음(prior)으로 시작
2. 데이터의 관측
3. **데이터 관측의 가능성(likelihood) 고려**

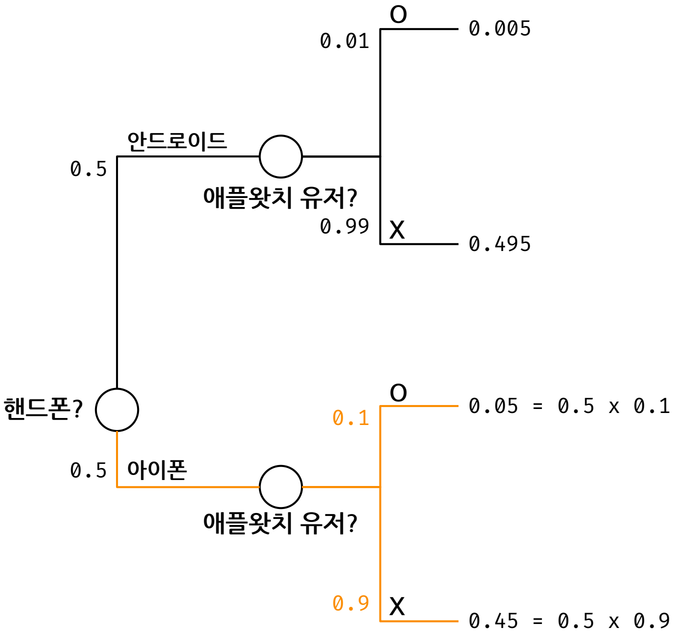
베이지안 확률모형: 손계산의 예



안드로이드나 아이폰 유저일 확률이 같다고 믿음.
그런데, 애플왓치를 차고 있는 것을 목격.

1.관심변수에 대한 믿음(prior)으로 시작
2.데이터의 관측
3.데이터 관측의 가능성(likelihood) 고려

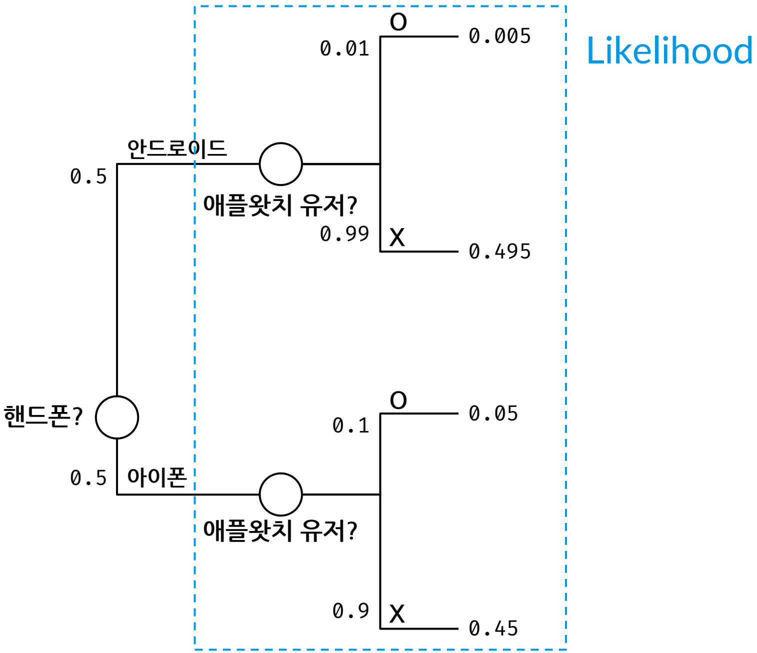
베이지안 확률모형: 손계산의 예



안드로이드나 아이폰 유저일 확률이 같다고 믿음.
그런데, 애플왓치를 차고 있는 것을 목격.

1. 관심변수에 대한 믿음(prior)으로 시작
2. 데이터의 관측
3. 데이터 관측의 가능성(likelihood) 고려

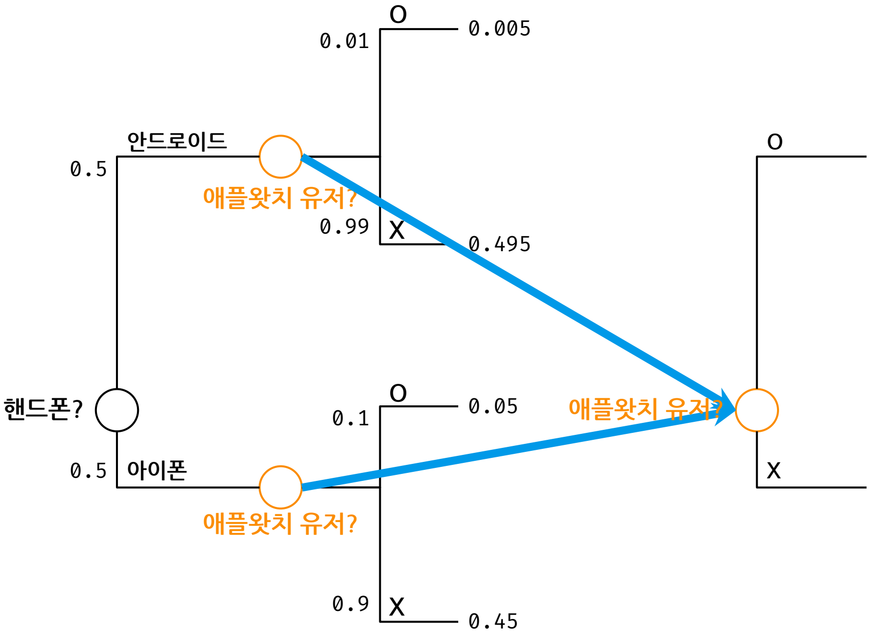
베이지안 확률모형: 손계산의 예



안드로이드나 아이폰 유저일 확률이 같다고 믿음.
그런데, 애플왓치를 차고 있는 것을 목격.
안드로이드 유저일 확률은?

1. 관심변수에 대한 믿음(prior)으로 시작
2. 데이터의 관측
3. 데이터 관측의 가능성(likelihood) 고려

베이지안 확률모형: 손계산의 예

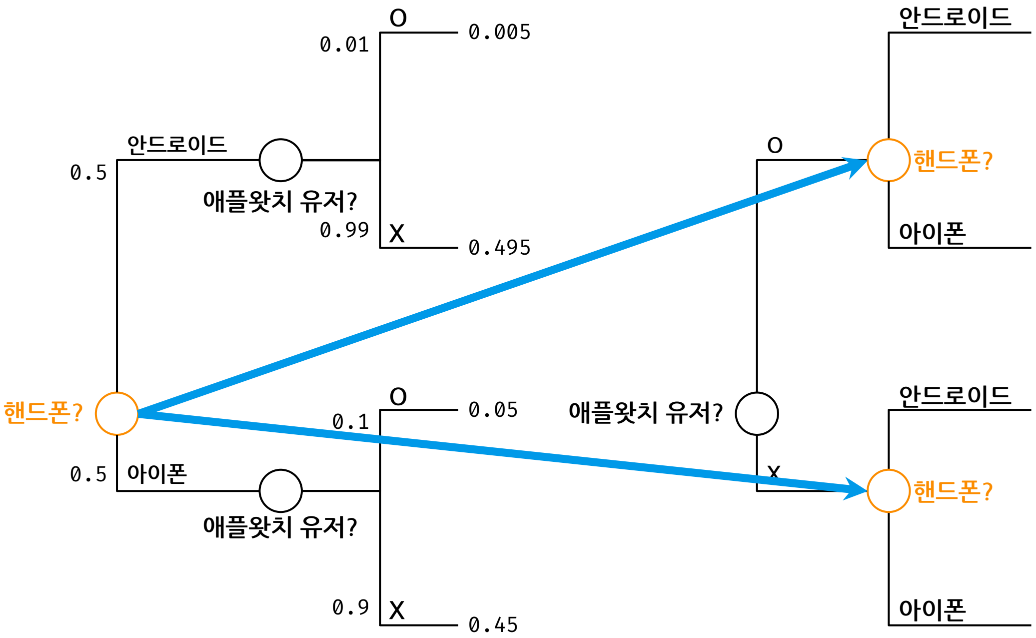


안드로이드나 아이폰 유저일 확률이 같다고 믿음.
그런데, 애플왓치를 차고 있는 것을 목격.
안드로이드 유저일 확률은?

1. 관심변수에 대한 믿음(prior)으로 시작
2. 데이터의 관측
3. 데이터 관측의 가능성(likelihood) 고려
4. **관심변수에 대한 믿음 업데이트 (→ posterior)**

→ **관찰을 했으니**, 이제 불확실성은?

베이지안 확률모형: 손계산의 예

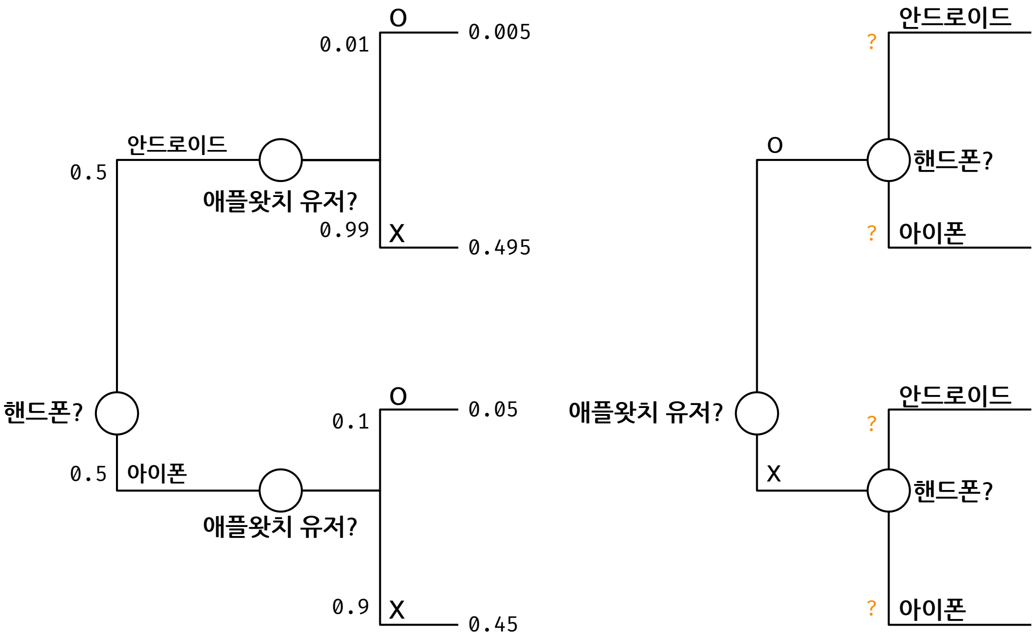


안드로이드나 아이폰 유저일 확률이 같다고 믿음.
그런데, 애플왓치를 차고 있는 것을 목격.
안드로이드 유저일 확률은?

1. 관심변수에 대한 믿음(prior)으로 시작
2. 데이터의 관측
3. 데이터 관측의 가능성(likelihood) 고려
4. **관심변수에 대한 믿음 업데이트 (→ posterior)**

→ 관찰을 했으니, **이제 불확실성**은?

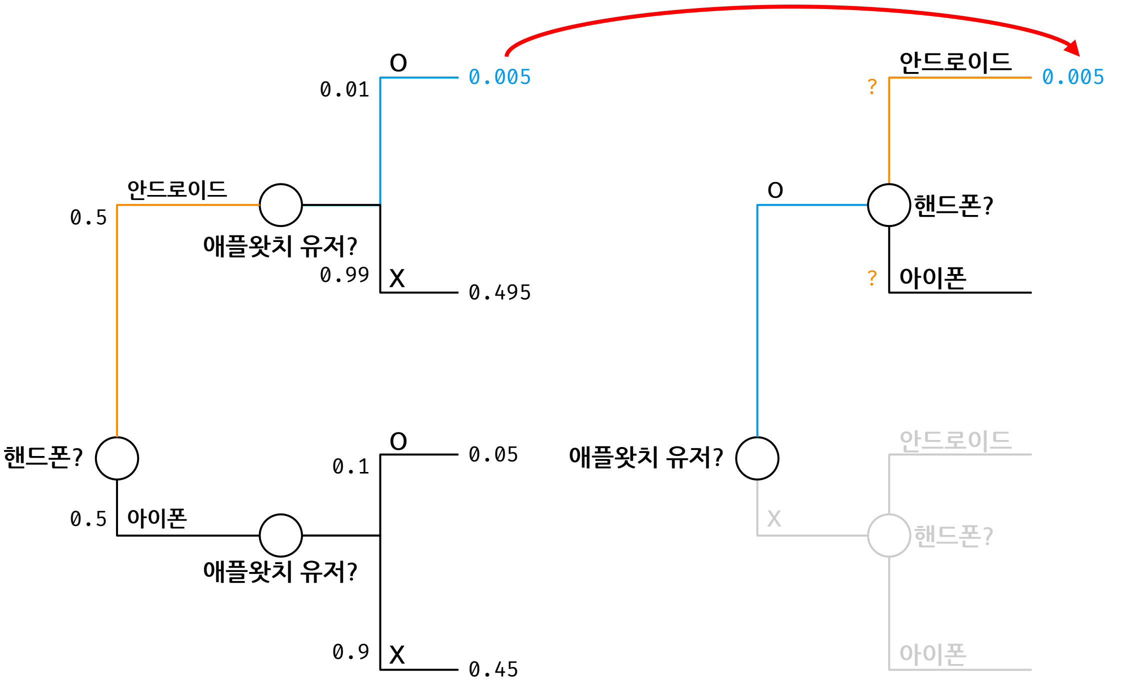
베이지안 확률모형: 손계산의 예



안드로이드나 아이폰 유저일 확률이 같다고 믿음.
그런데, 애플왓치를 차고 있는 것을 목격.
안드로이드 유저일 확률은?

1. 관심변수에 대한 믿음(prior)으로 시작
2. 데이터의 관측
3. 데이터 관측의 가능성(likelihood) 고려
4. **관심변수에 대한 믿음 업데이트 (→ posterior)**

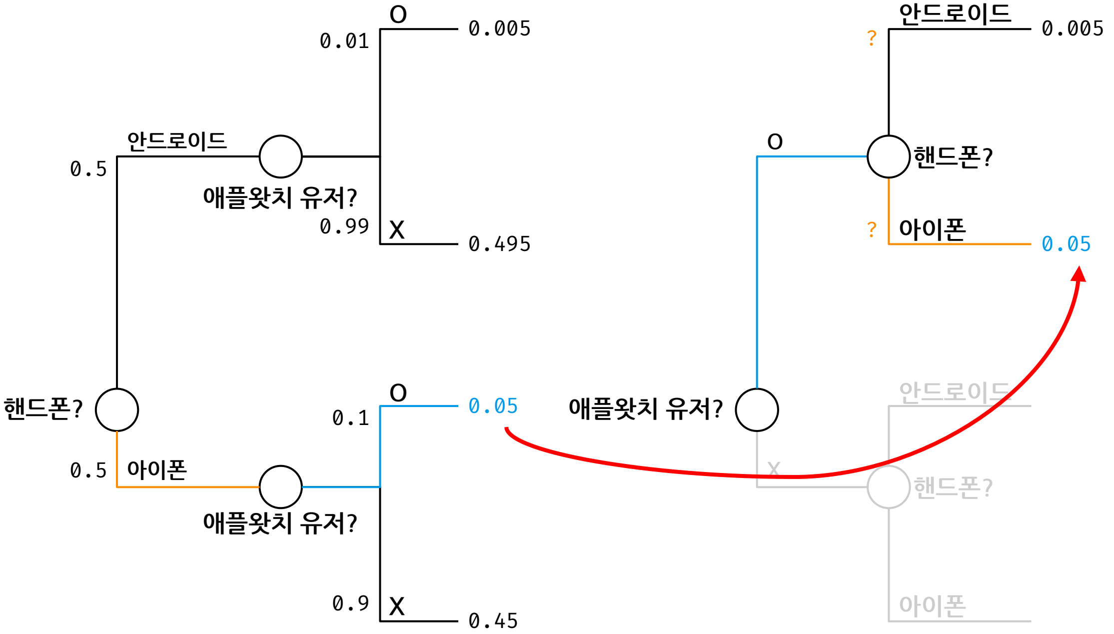
베이지안 확률모형: 손계산의 예



안드로이드나 아이폰 유저일 확률이 같다고 믿음.
그런데, 애플왓치를 차고 있는 것을 목격.
안드로이드 유저일 확률은?

1. 관심변수에 대한 믿음(prior)으로 시작
2. 데이터의 관측
3. 데이터 관측의 가능성(likelihood) 고려
4. **관심변수에 대한 믿음 업데이트 (→ posterior)**

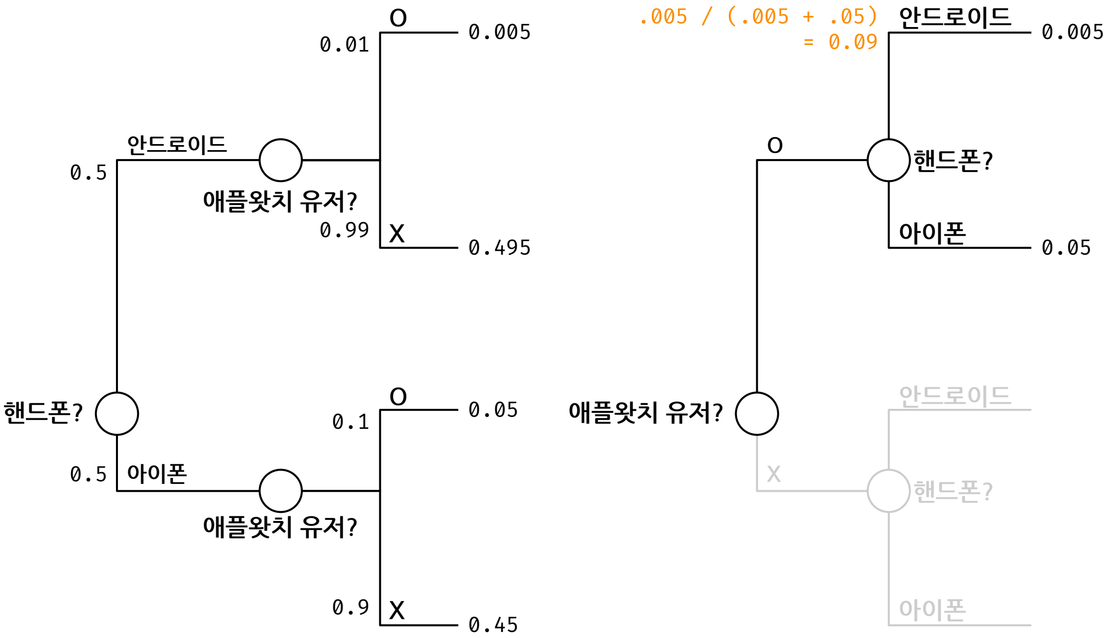
베이지안 확률모형: 손계산의 예



안드로이드나 아이폰 유저일 확률이 같다고 믿음.
그런데, 애플왓치를 차고 있는 것을 목격.
안드로이드 유저일 확률은?

1. 관심변수에 대한 믿음(prior)으로 시작
2. 데이터의 관측
3. 데이터 관측의 가능성(likelihood) 고려
4. **관심변수에 대한 믿음 업데이트 (→ posterior)**

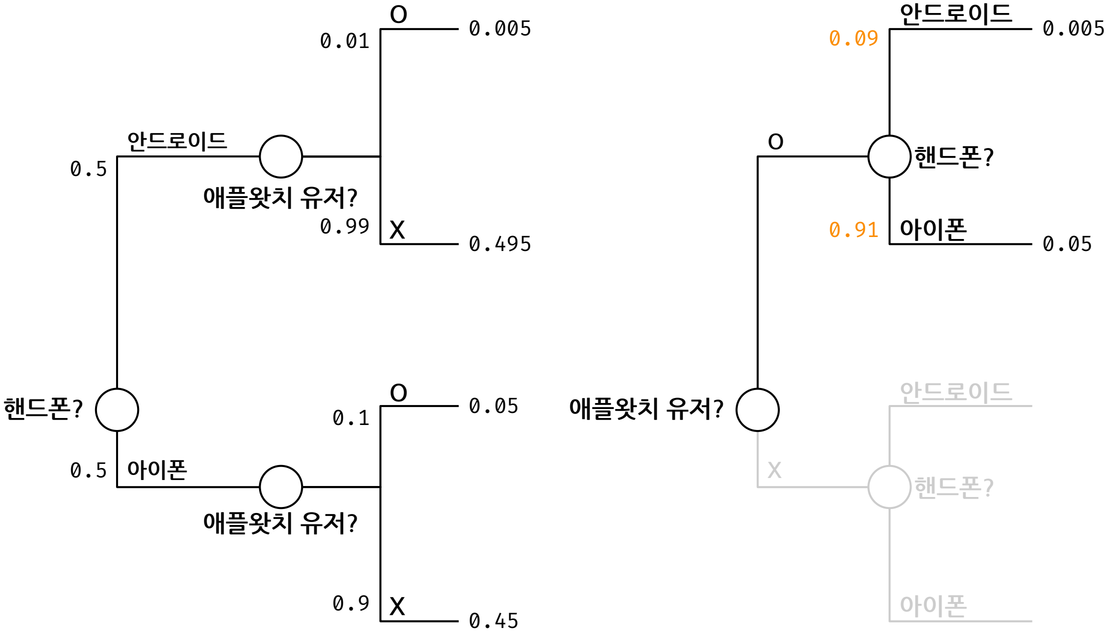
베이지안 확률모형: 손계산의 예



안드로이드나 아이폰 유저일 확률이 같다고 믿음.
그런데, 애플왓치를 차고 있는 것을 목격.
안드로이드 유저일 확률은?

1. 관심변수에 대한 믿음(prior)으로 시작
2. 데이터의 관측
3. 데이터 관측의 가능성(likelihood) 고려
4. 관심변수에 대한 믿음 업데이트 (→ posterior)

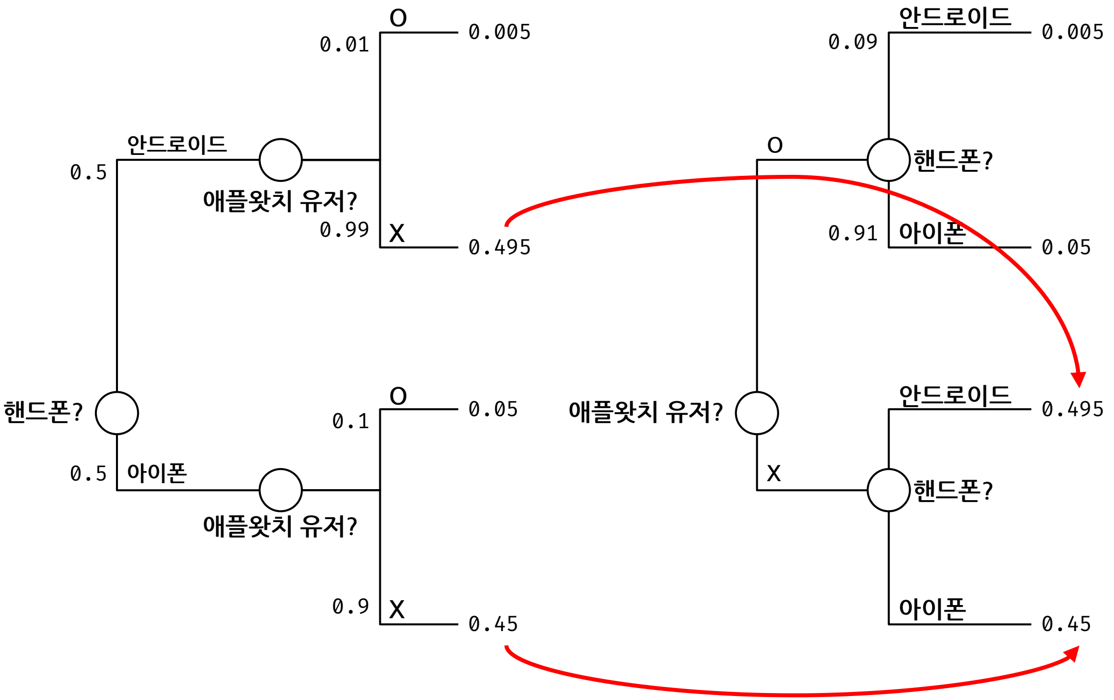
베이지안 확률모형: 손계산의 예



안드로이드나 아이폰 유저일 확률이 같다고 믿음.
그런데, 애플왓치를 차고 있는 것을 목격.
안드로이드 유저일 확률은?

1. 관심변수에 대한 믿음(prior)으로 시작
2. 데이터의 관측
3. 데이터 관측의 가능성(likelihood) 고려
4. **관심변수에 대한 믿음 업데이트 (→ posterior)**

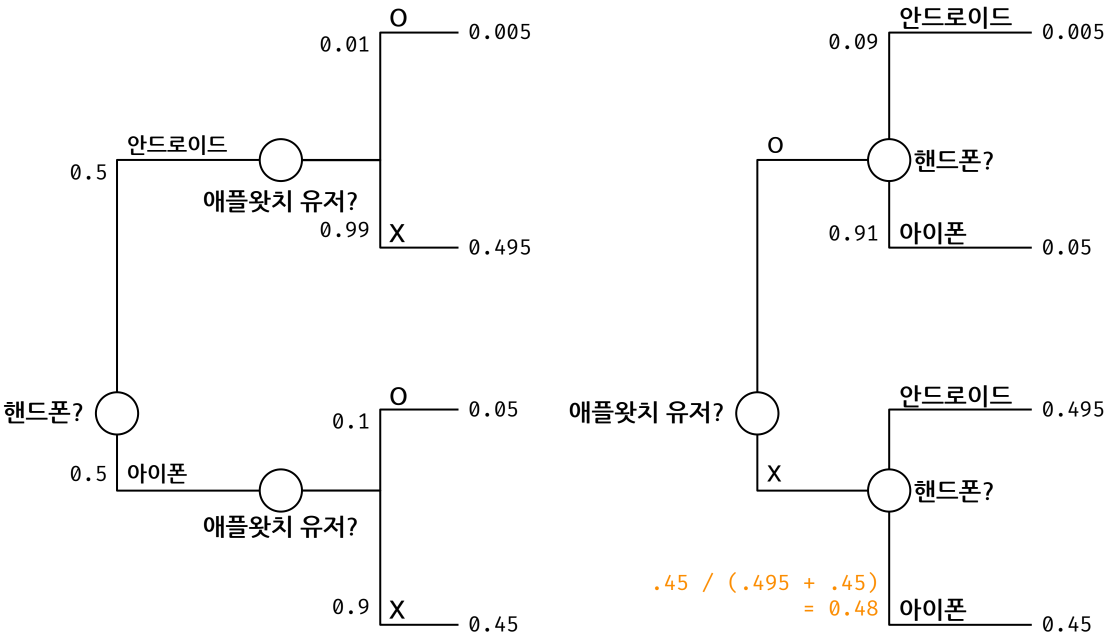
베이지안 확률모형: 손계산의 예



안드로이드나 아이폰 유저일 확률이 같다고 믿음.
그런데, 애플왓치를 차고 있는 것을 목격.
안드로이드 유저일 확률은?

1. 관심변수에 대한 믿음(prior)으로 시작
2. 데이터의 관측
3. 데이터 관측의 가능성(likelihood) 고려
4. **관심변수에 대한 믿음 업데이트 (→ posterior)**

베이지안 확률모형: 손계산의 예

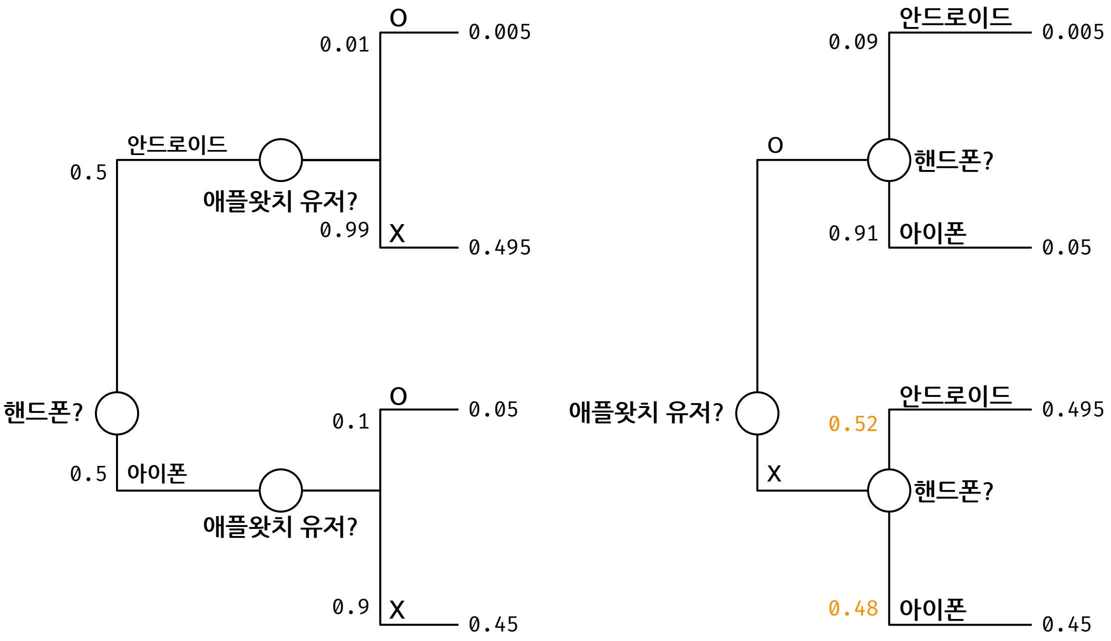


안드로이드나 아이폰 유저일 확률이 같다고 믿음.
그런데, 애플왓치를 차고 있는 것을 목격.
안드로이드 유저일 확률은?

1. 관심변수에 대한 믿음(prior)으로 시작
2. 데이터의 관측
3. 데이터 관측의 가능성(likelihood) 고려
4. **관심변수에 대한 믿음 업데이트 (→ posterior)**

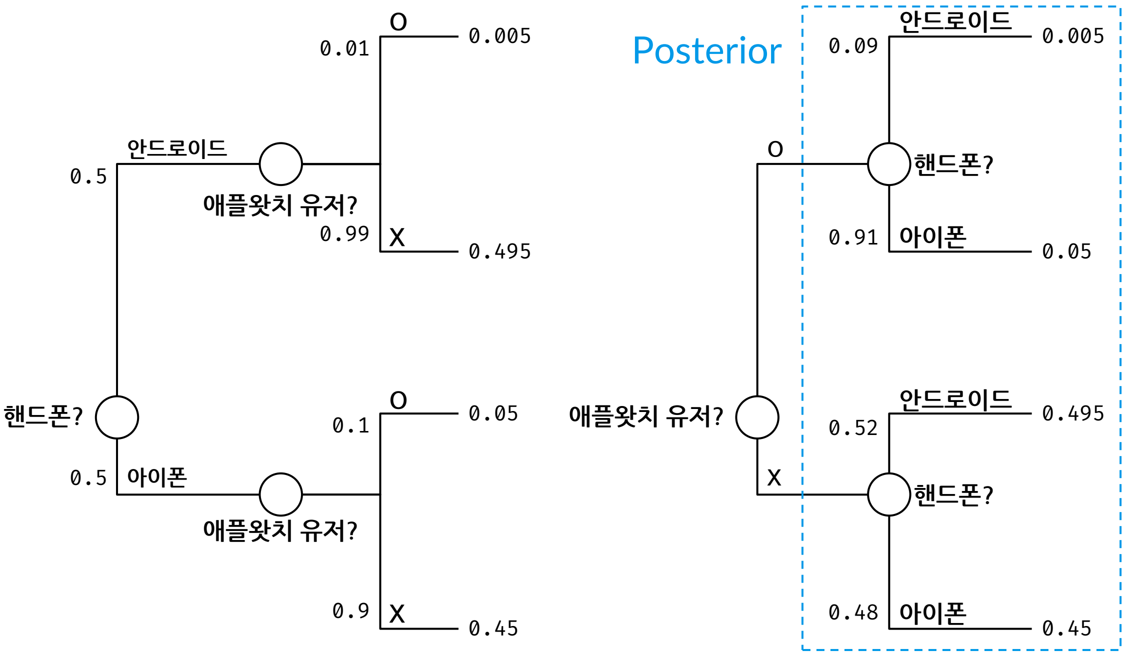
베이지안 확률모형: 손계산의 예

베이지안 확률모형: 손계산의 예



안드로이드나 아이폰 유저일 확률이 같다고 믿음.
그런데, 애플왓치를 차고 있는 것을 목격.
안드로이드 유저일 확률은?

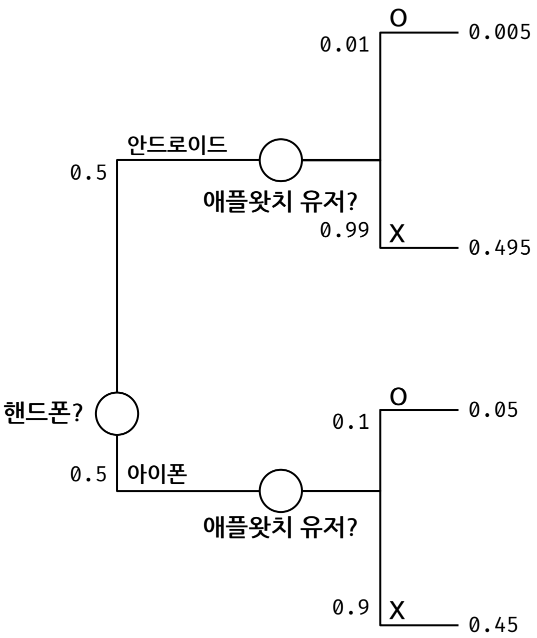
1. 관심변수에 대한 믿음(prior)으로 시작
2. 데이터의 관측
3. 데이터 관측의 가능성(likelihood) 고려
4. **관심변수에 대한 믿음 업데이트 (→ posterior)**

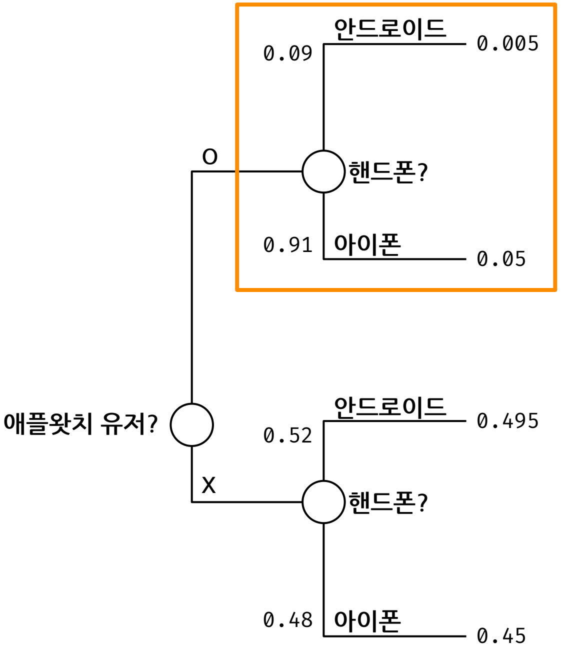


안드로이드나 아이폰 유저일 확률이 같다고 믿음.
그런데, 애플왓치를 차고 있는 것을 목격.
안드로이드 유저일 확률은?

1. 관심변수에 대한 믿음(prior)으로 시작
2. 데이터의 관측
3. 데이터 관측의 가능성(likelihood) 고려
4. **관심변수에 대한 믿음 업데이트 (→ posterior)**

베이지안 확률모형: 손계산의 예





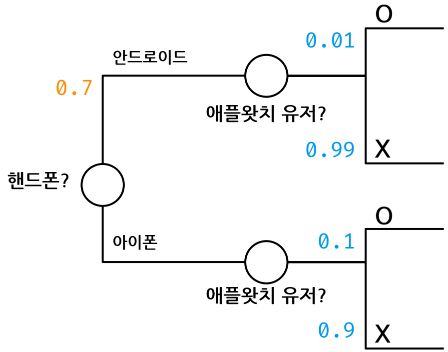
안드로이드나 아이폰 유저일 확률이 같다고 믿음.
그런데, 애플왓치를 차고 있는 것을 목격.
안드로이드 유저일 확률은?

1. 관심변수에 대한 믿음(prior)으로 시작
2. 데이터의 관측
3. 데이터 관측의 가능성(likelihood) 고려
4. 관심변수에 대한 믿음 업데이트 (→ posterior)

베이지안 확률모형: 손계산의 예

연습문제 1

만약 사전확률(prior)이, "안드로이드 70%"였고,
데이터에 대한 가능성(likelihood)는 그대로라면?



연습문제 2: 몬티 홀 문제의 확률 계산

- 세 개의 문 중에 하나를 선택하여 문 뒤에 있는 선물을 가질 수 있는 게임쇼에 참가
- 한 문 뒤에는 자동차가 있고, 나머지 두 문 뒤에는 꽝
- 세 개 문 중 처음 하나를 선택
- 게임쇼 진행자가 나머지 두 개의 문 중 하나를 열어 꽝임을 보여준 후 “처음 선택을 바꿀 기회를 주겠다”고 함

선택하지도, 진행자가 열지도 않은
<u>나머지 문 뒤에 자동차가 있을 확률</u>은?

연습문제 3

- 회사에서 A, B, C 세 명이 승진 후보자
- 세 명 중 한 명만 다음주 승진이 확정됨
- A(=여러분)는
  - 셋 다 승진할 확률이 똑같다고 생각 (각 ⅓)
  - 매니저와 각별한 친분이 있어,
    “B나 C 중 누가 떨어졌나요?” 물었더니,
    매니저 왈: “B는 확실히 승진 못해”
- 그 말을 들은 후 A가 생각하는
  - 본인이 승진할 확률은?
  - C가 승진할 확률은?

다음: 연습문제 풀이

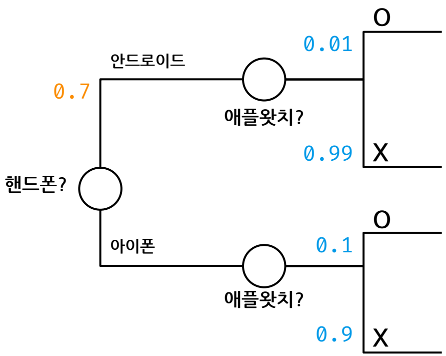
1부: 불확실성과 데이터

연습문제 풀이

데이터와 의사결정 | 정종빈

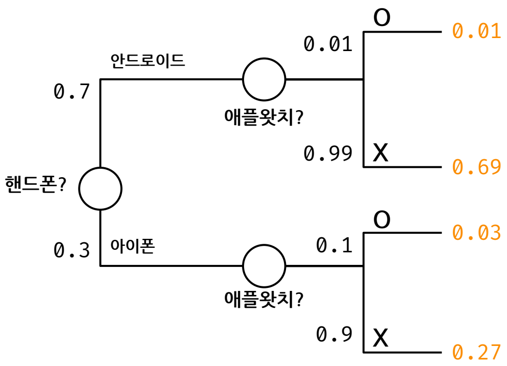
연습문제 1

만약 사전확률(prior)이, “안드로이드 70%”였고,
데이터에 대한 가능성(likelihood)는 그대로라면?



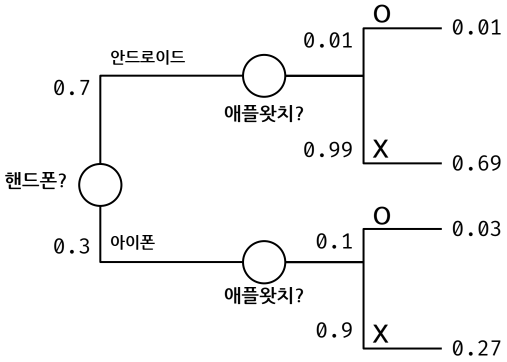
연습문제 1

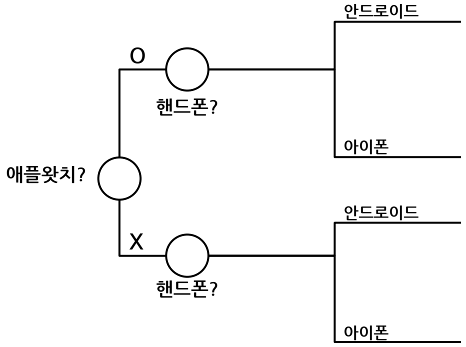
만약 사전확률(prior)이, "안드로이드 70%"였고,
데이터에 대한 가능성(likelihood)는 그대로라면?



연습문제 1

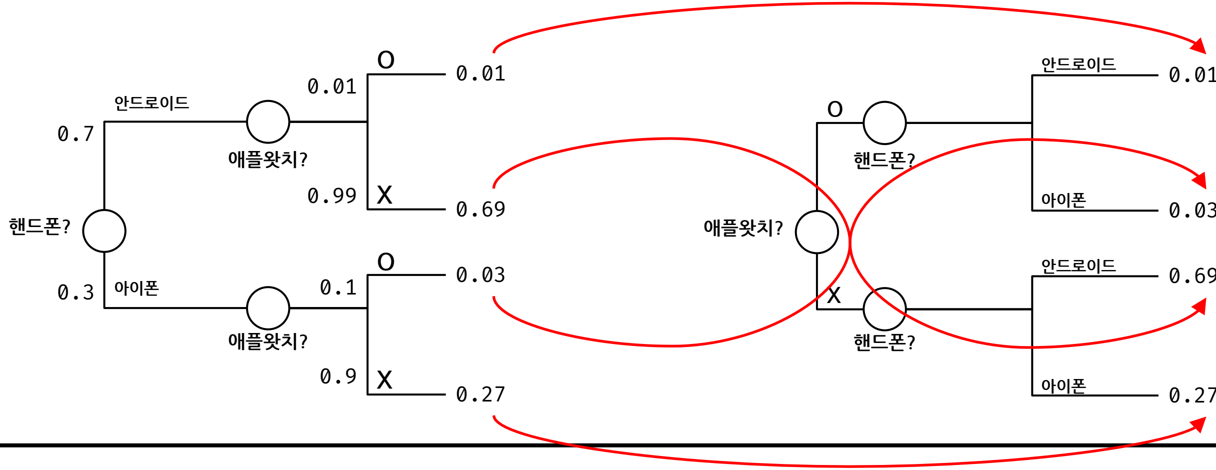
만약 사전확률(prior)이, “안드로이드 70%”였고,
데이터에 대한 가능성(likelihood)는 그대로라면?





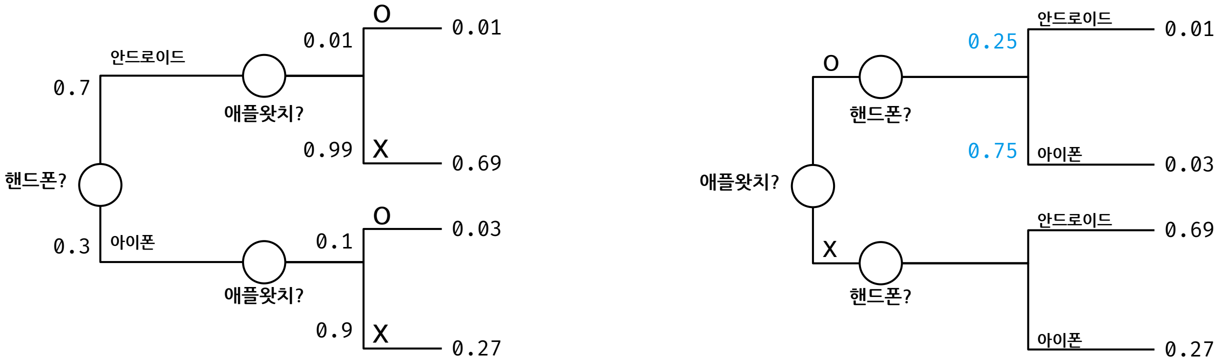
연습문제 1

만약 사전확률(prior)이, “안드로이드 70%”였고,
데이터에 대한 가능성(likelihood)는 그대로라면?



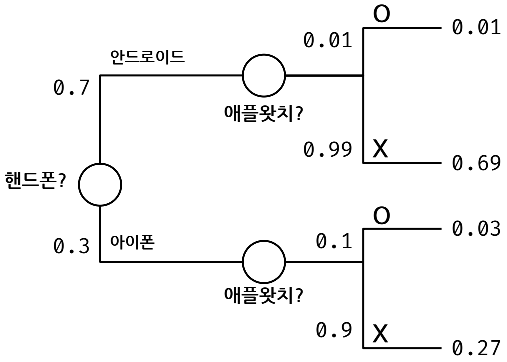
연습문제 1

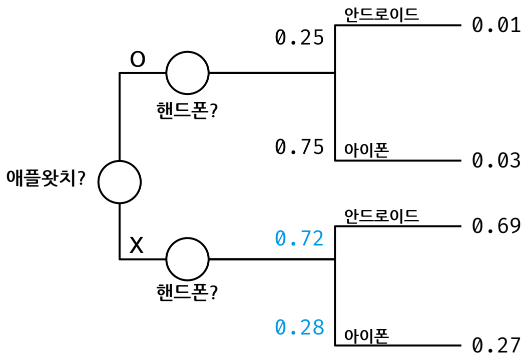
만약 사전확률(prior)이, "안드로이드 70%"였고,
데이터에 대한 가능성(likelihood)는 그대로라면?



연습문제 1

만약 사전확률(prior)이, “안드로이드 70%”였고,
데이터에 대한 가능성(likelihood)는 그대로라면?





연습문제 2: 몬티 홀 문제의 확률 계산

- 세 개의 문 중에 하나를 선택하여 문 뒤에 있는 선물을 가질 수 있는 게임쇼에 참가
- 한 문 뒤에는 자동차가 있고, 나머지 두 문 뒤에는 꽝
- 세 개 문 중 처음 하나를 선택
- 게임쇼 진행자가 나머지 두 개의 문 중 하나를 열어 꽝임을 보여준 후 “처음 선택을 바꿀 기회를 주겠다”고 함

선택하지도, 진행자가 열지도 않은
나머지 문 뒤에 자동차가 있을 확률은?

연습문제 2: 몬티 홀 문제의 확률 계산

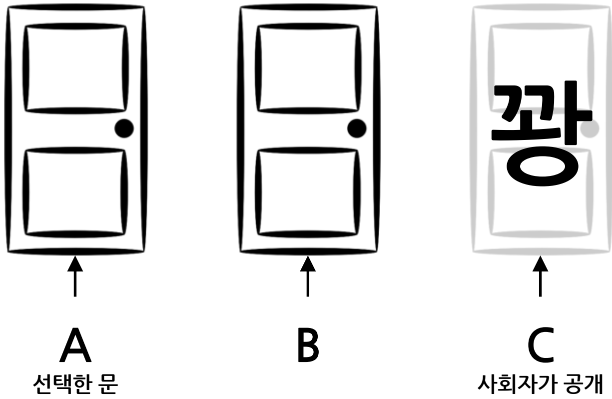
- 세 개의 문 중에 하나를 선택하여 문 뒤에 있는 선물을 가질 수 있는 게임쇼에 참가
- 한 문 뒤에는 자동차가 있고, 나머지 두 문 뒤에는 꽝
- 세 개 문 중 처음 하나를 선택
- 게임쇼 진행자가 나머지 두 개의 문 중 하나를 열어 꽝임을 보여준 후 “처음 선택을 바꿀 기회를 주겠다”고 함

선택하지도, 진행자가 열지도 않은

<u>**나머지 문 뒤에 자동차가 있을 확률**</u>은?

= <u>**선택한 문**</u> 뒤에 자동차가 있을 확률?

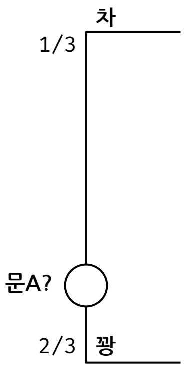
연습문제 2: 몬티 홀 문제의 확률 계산



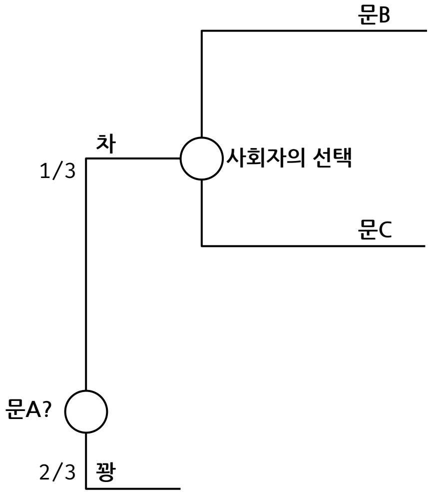
알고 싶은 것(불확실성): 문A 뒤에는 차일까 꽝일까?

관찰한 것: 사회자가 문C를 열었다

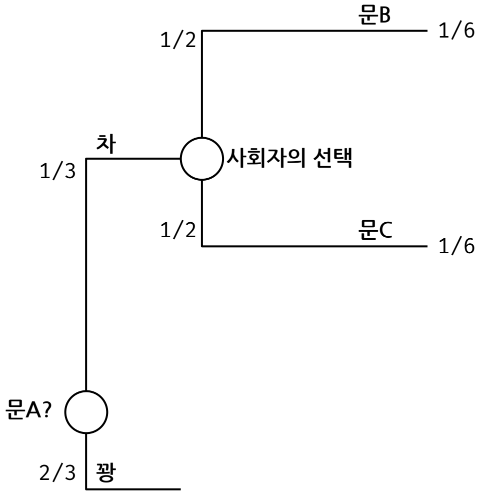
연습문제 2: 몬티 홀 문제의 확률 계산



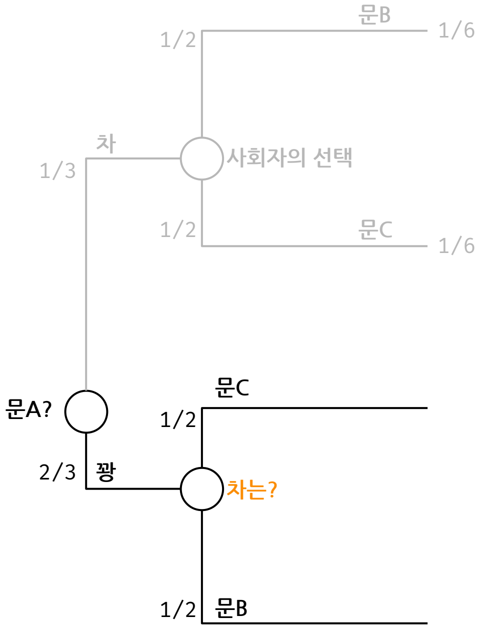
연습문제 2: 몬티 홀 문제의 확률 계산



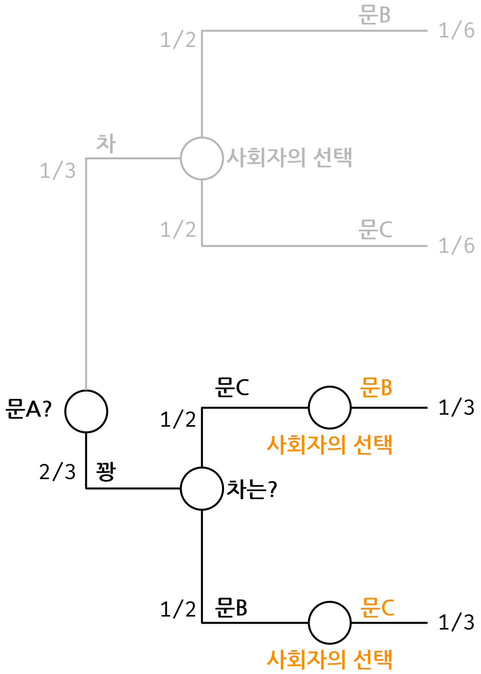
연습문제 2: 몬티 홀 문제의 확률 계산



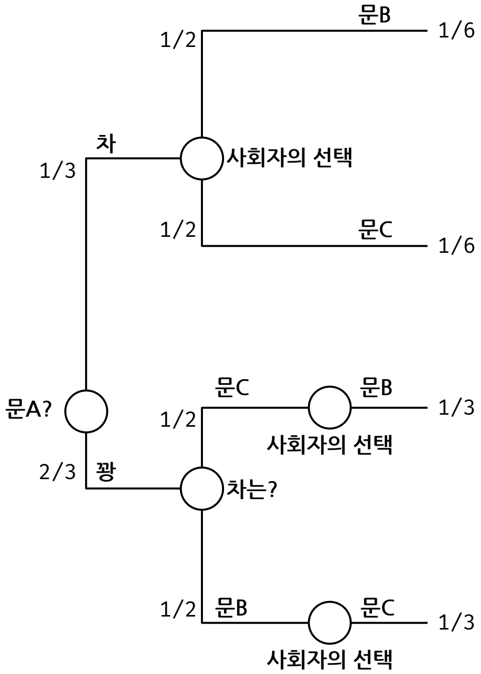
연습문제 2: 몬티 홀 문제의 확률 계산



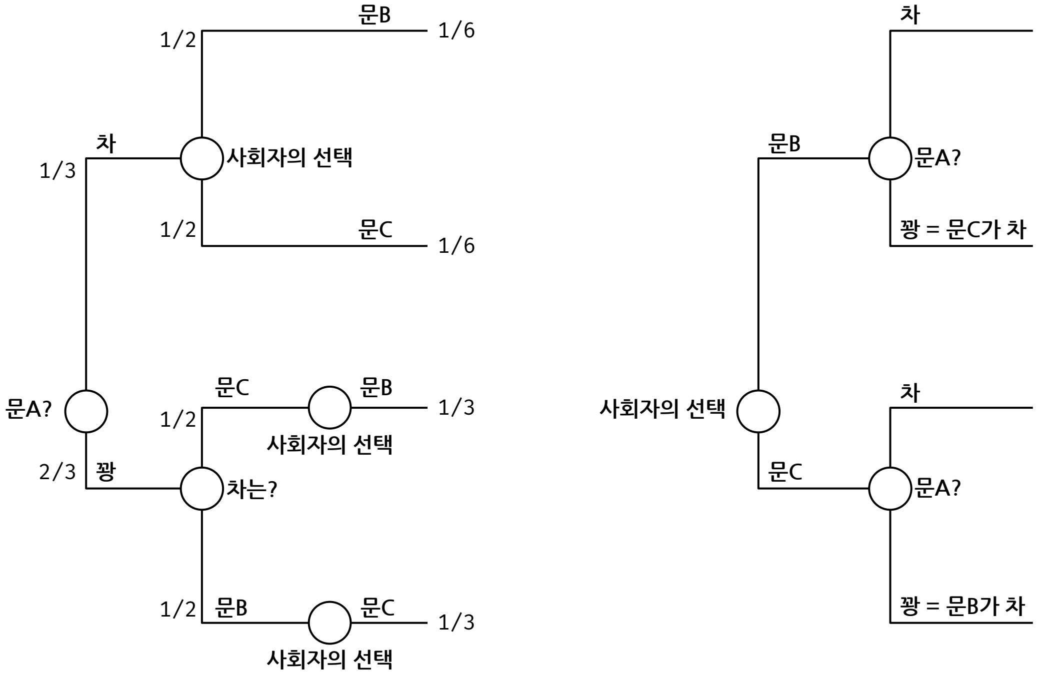
연습문제 2: 몬티 홀 문제의 확률 계산



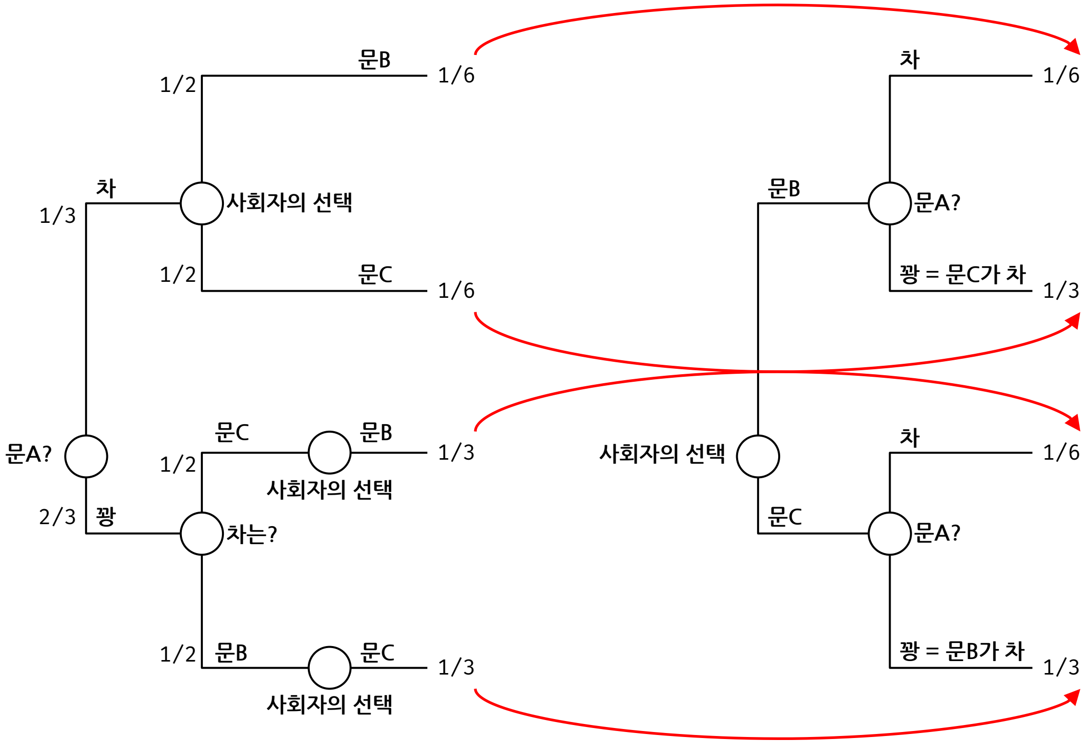
연습문제 2: 몬티 홀 문제의 확률 계산



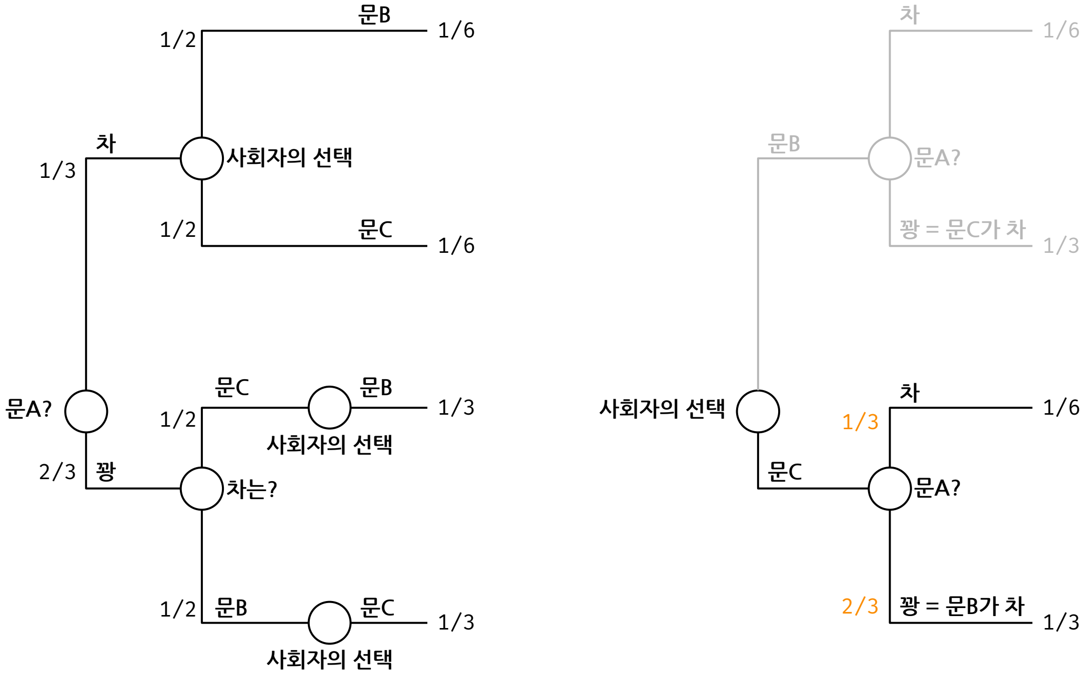
연습문제 2: 몬티 홀 문제의 확률 계산



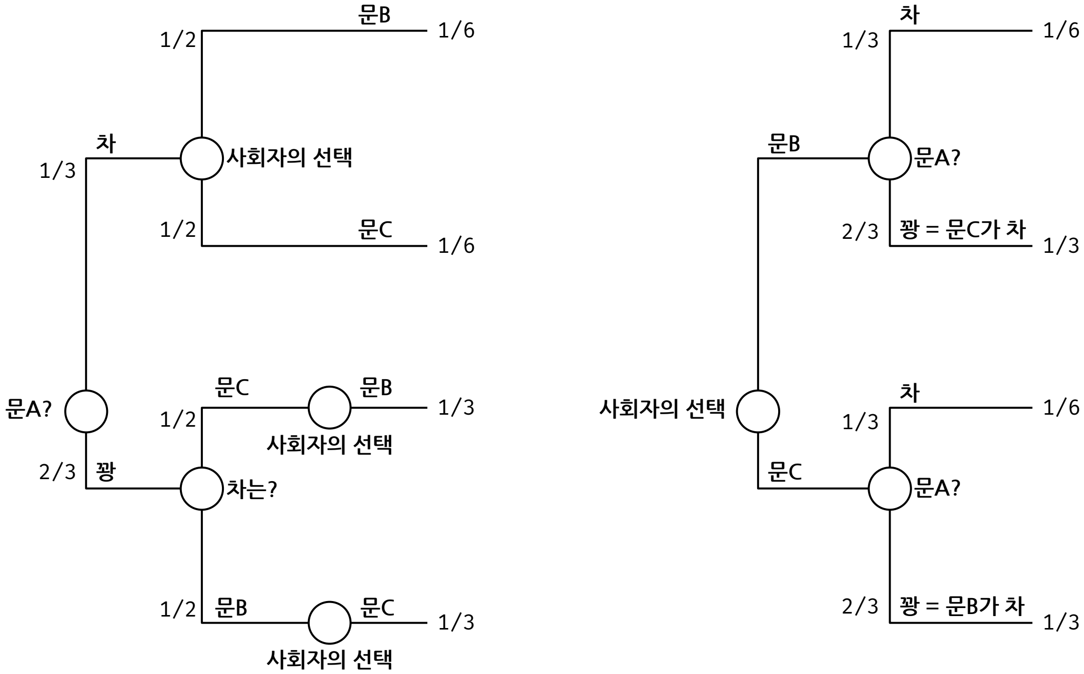
연습문제 2: 몬티 홀 문제의 확률 계산



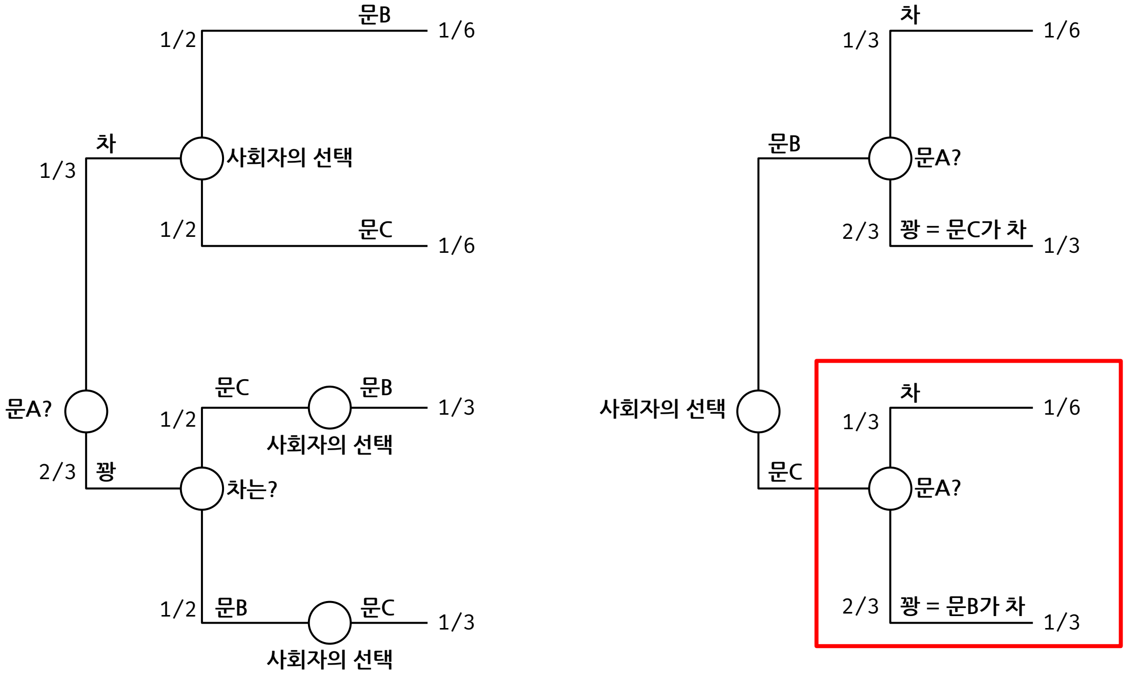
연습문제 2: 몬티 홀 문제의 확률 계산



연습문제 2: 몬티 홀 문제의 확률 계산



연습문제 2: 몬티 홀 문제의 확률 계산



연습문제 3

- 회사에서 A, B, C 세 명이 승진 후보자
- 세 명 중 한 명만 다음주 승진이 확정됨
- A(=여러분)는
  - 셋 다 승진할 확률이 똑같다고 생각 (각 ⅓)
  - 매니저와 각별한 친분이 있어,
    "B나 C 중 누가 떨어졌나요?" 물었더니,
    매니저 왈: "B는 확실히 승진 못해"
- 그 말을 들은 후 A가 생각하는
  - 본인이 승진할 확률은?
  - C가 승진할 확률은?

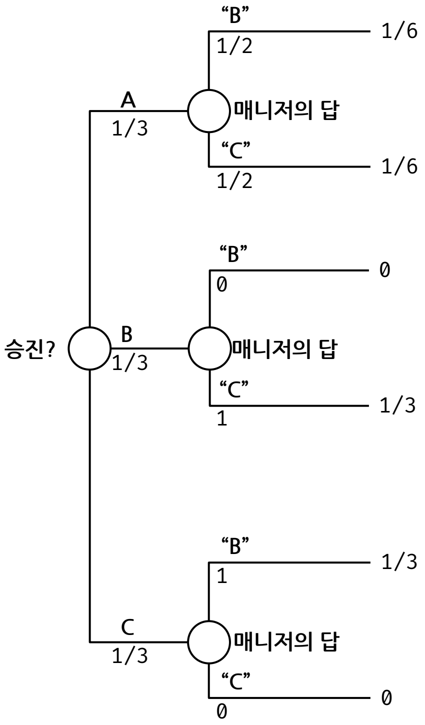
연습문제 3: 흔한 오류

"B를 빼고나면 나(A)랑 C 사이에 한 명이 승진하는거니까,
이제 각각 50%의 확률!"

알고 싶은 것: A, B, C 각각이 승진할 확률

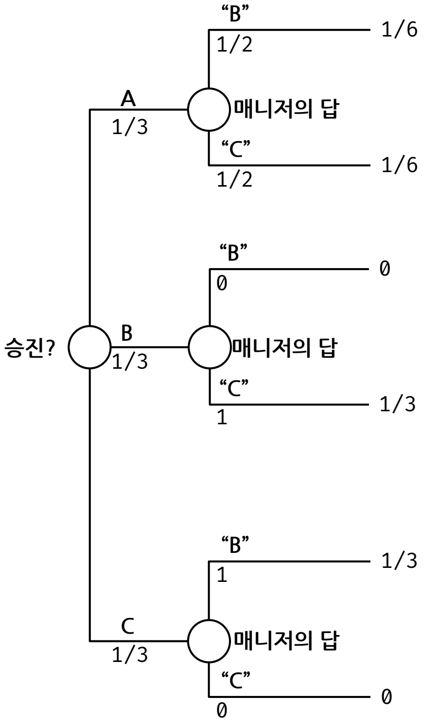
관찰한 것: 매니저는 "B, C 중 B는 승진 못한다"고 말함

연습문제 3

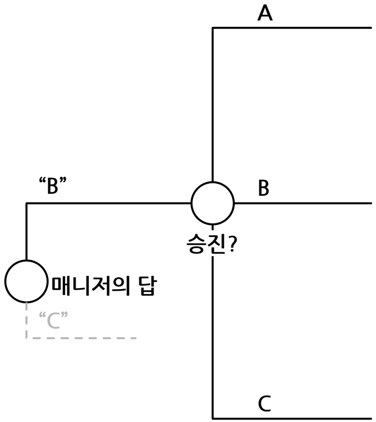


승진?
A
1/3
B
1/3
C
1/3
매니저의 답
“B”
1/2
1/6
“C”
1/2
1/6
매니저의 답
“B”
0
0
“C”
1
1/3
매니저의 답
“B”
1
1/3
“C”
0
0

연습문제 3

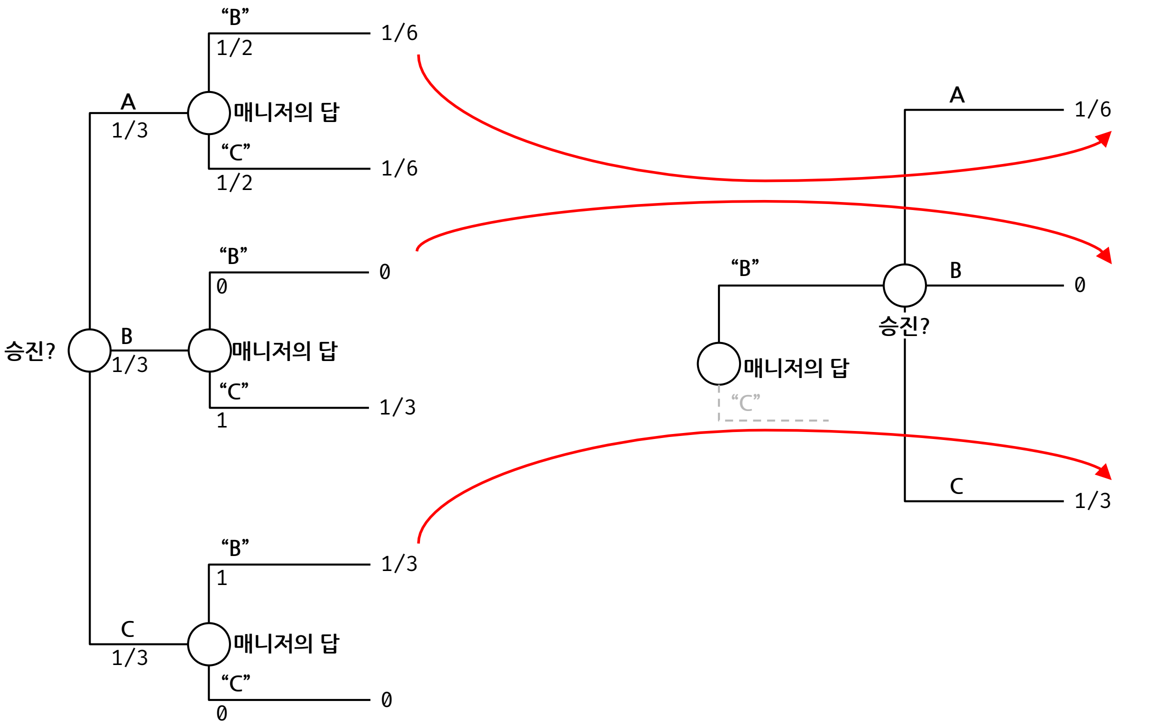


승진?
A
1/3
B
1/3
C
1/3
매니저의 답
"B"
1/2
1/6
"C"
1/2
1/6
매니저의 답
"B"
0
0
"C"
1
1/3
매니저의 답
"B"
1
1/3
"C"
0
0



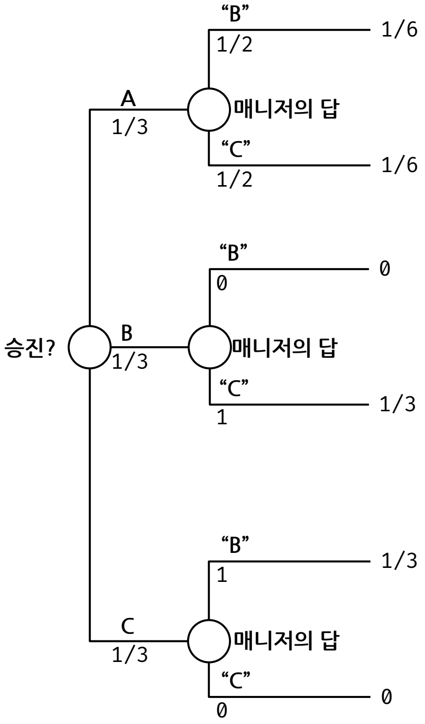
매니저의 답
"B"
"C"
승진?
A
B
C

연습문제 3

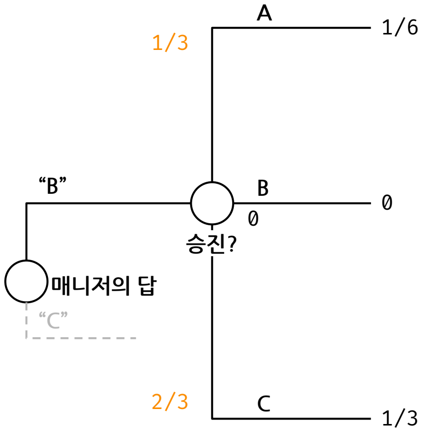


승진?
A
1/3
매니저의 답
"B"
1/2
1/6
"C"
1/2
1/6
B
1/3
매니저의 답
"B"
0
0
"C"
1
1/3
C
1/3
매니저의 답
"B"
1
1/3
"C"
0
0
매니저의 답
"B"
"C"
승진?
A
1/6
B
0
C
1/3

연습문제 3



승진?
A
1/3
B
1/3
C
1/3
매니저의 답
"B"
1/2
1/6
"C"
1/2
1/6
매니저의 답
"B"
0
0
"C"
1
1/3
매니저의 답
"B"
1
1/3
"C"
0
0



매니저의 답
"B"
"C"
승진?
A
1/3
1/6
B
0
0
C
2/3
1/3

연습문제 3: 오해의 근원

"B를 빼고나면 나(A)랑 C 사이에 한 명이 승진하는거니까,
이제 각각 50%의 확률!"

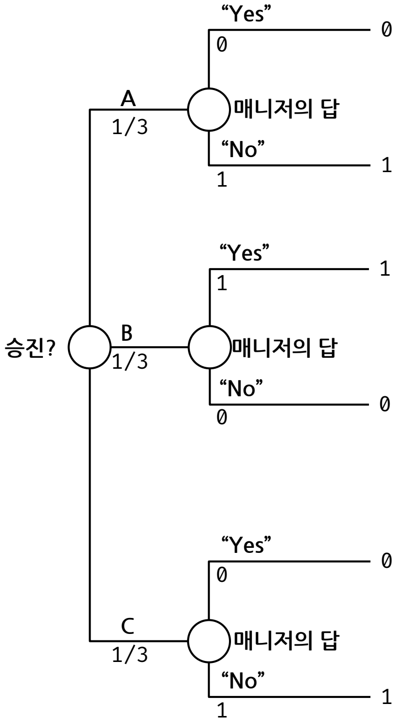
알고 싶은 것: A, B, C 각각이 승진할 확률

관찰한 것: 매니저는 "B, C 중 B는 승진 못한다"고 말함

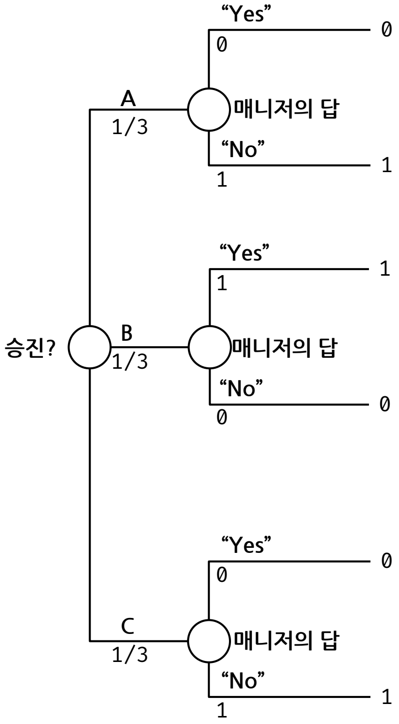
→ "B, C 중 누가 떨어지나요?"의 답에는 정보가 없음!

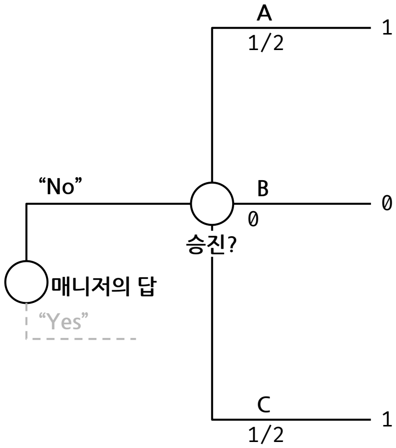
→ "B가 승진하나요?"

연습문제 3 변형: "B가 승진하나요?"



연습문제 3 변형: “B가 승진하나요?”





다음: Computational Bayes

1부: 불확실성과 데이터

Computational Bayes

데이터와 의사결정 | 정종빈

1부 강의 목적 및 개요

체계적이고 일반적인 <u>의사결정 모형</u> 소개

불확실성을 계량화하는 두 가지 접근법:

- **빈도주의(frequentist) 통계**
  Estimand, estimator, estimate, 신뢰구간, bootstrap
- **베이지안(Bayesian) 확률**
  Prior, likelihood, posterior (손계산 → 복잡한 모델링)

수식 경고!

표기법 정리

$$P(X)$$

X라는 **확률변수**가 가질수 있는 모든 값에 대한 확률 **분포**

$$P(X = x)$$

X라는 **확률변수**가 x라는 **구체적인 값**을 가질 확률

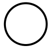
$$P(Y = y \mid X = x)$$

조건부확률

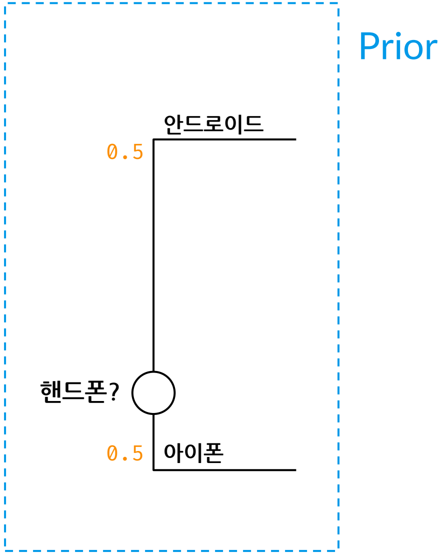
X라는 확률변수가 x라는 구체적인 값을 **가졌을 때**

Y라는 **확률변수**가 y라는 **구체적인 값**을 가질 확률

$p \in \{0, 1\}$: 핸드폰 OS (안드로이드면 1)

핸드폰? 

베이지안 확률모형의 수학적 정의



$p \in \{0, 1\}$: 핸드폰 OS (안드로이드면 1)

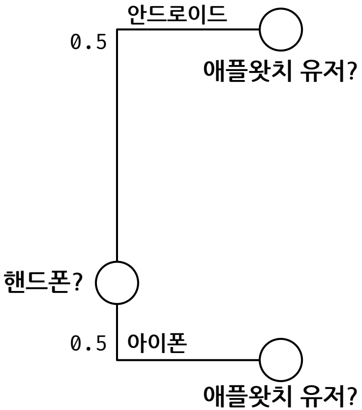
$P(p)$: 핸드폰이 안드로이드 일 확률 (prior)

베이지안 확률모형의 수학적 정의

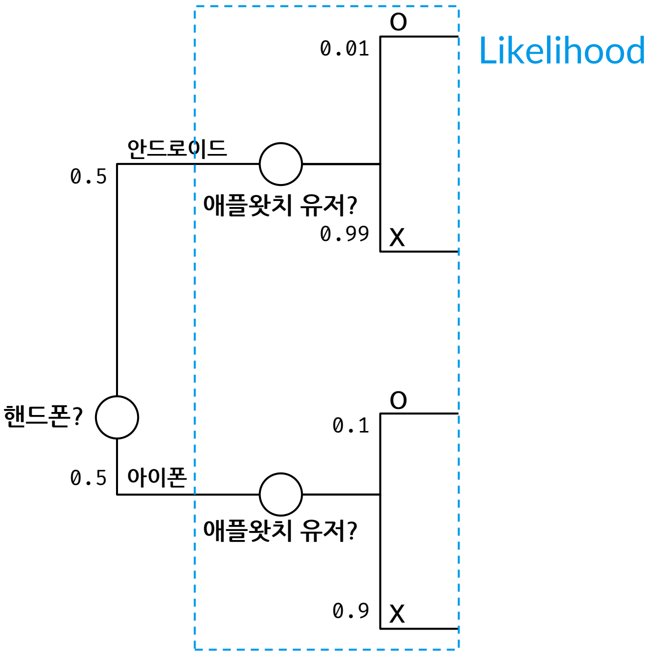
$p \in \{0, 1\}$: 핸드폰 OS (안드로이드면 1)

$P(p)$: 핸드폰이 안드로이드 일 확률 (prior)

$w \in \{0, 1\}$: 관찰한 데이터 (왓치유저면 1)



베이지안 확률모형의 수학적 정의



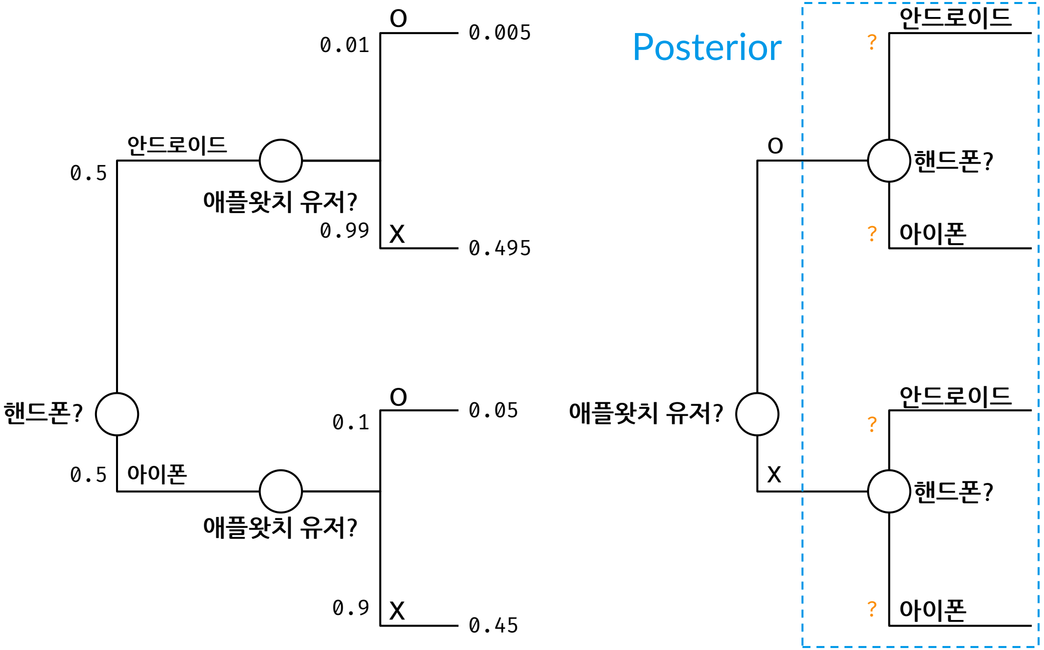
$p \in \{0, 1\}$: 핸드폰 OS (안드로이드면 1)

$P(p)$: 핸드폰이 안드로이드 일 확률 (prior)

$w \in \{0, 1\}$: 관찰한 데이터 (왓치유저면 1)

$P(w \mid p)$: 사전 확률분포를 가정 할 때, 데이터를 관찰할 확률

베이지안 확률모형의 수학적 정의



$p \in \{0, 1\}$: 핸드폰 OS (안드로이드면 1)

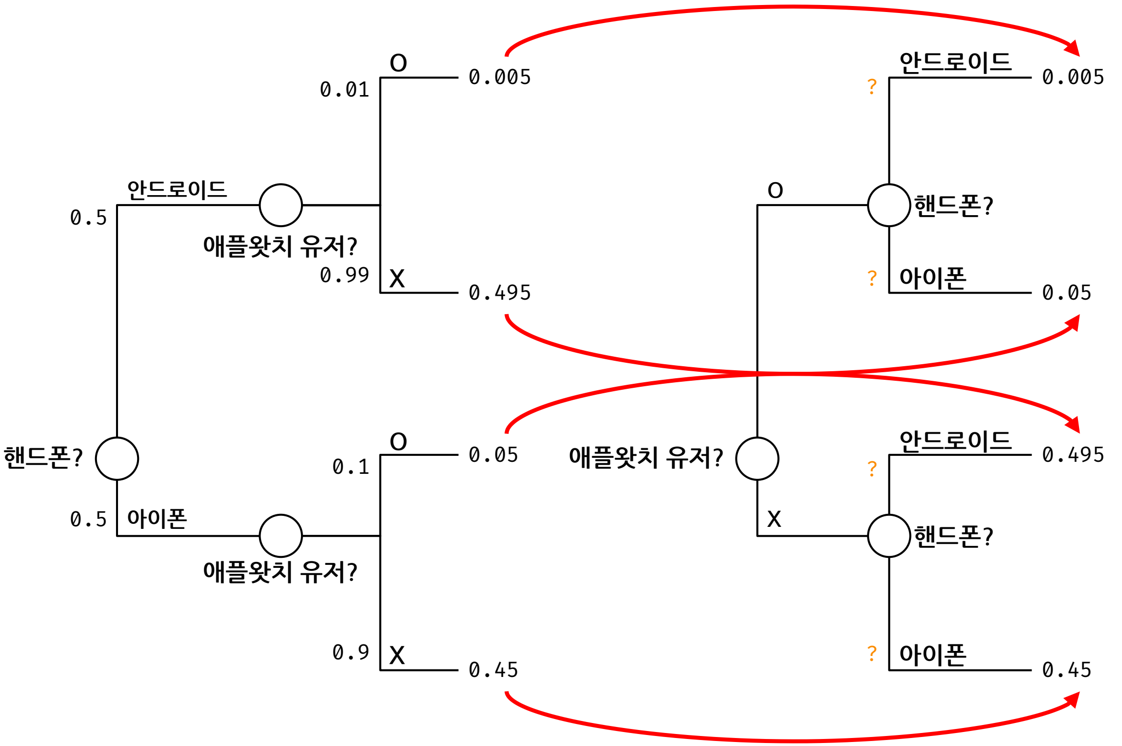
$P(p)$: 핸드폰이 안드로이드 일 확률 (prior)

$w \in \{0, 1\}$: 관찰한 데이터 (왓치유저면 1)

$P(w \mid p)$: 사전 확률분포를 가정 할 때, 데이터를 관찰할 확률

$P(p \mid w)$: 데이터 관찰 후의 p에 대한 확률분포

베이지안 확률모형의 수학적 정의



$p \in \{0, 1\}$: 핸드폰 OS (안드로이드면 1)

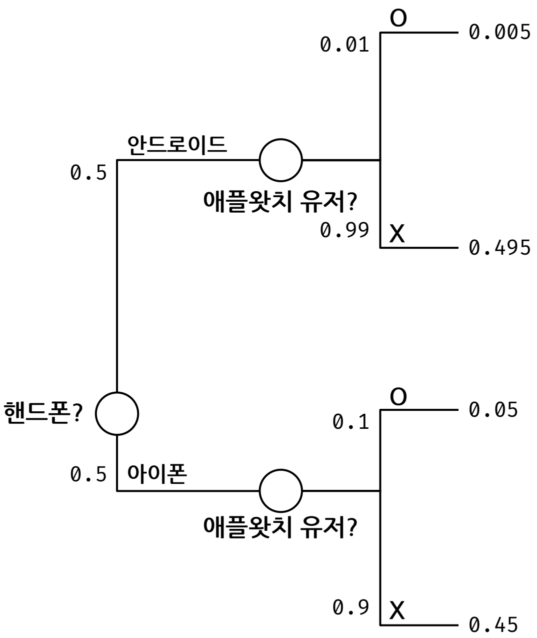
$P(p)$: 핸드폰이 안드로이드 일 확률 (prior)

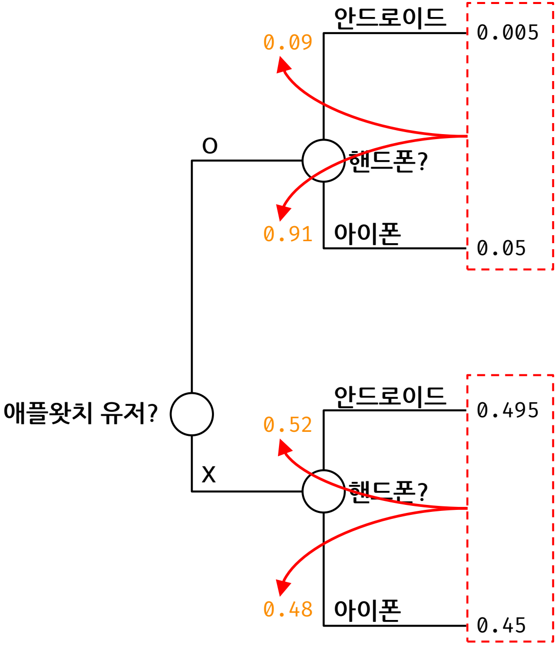
$w \in \{0, 1\}$: 관찰한 데이터 (왓치유저면 1)

$P(w \mid p)$: 사전 확률분포를 가정 할 때, 데이터를 관찰할 확률

$$P(p \mid w) \qquad P(p)P(w \mid p)$$

베이지안 확률모형의 수학적 정의





$p \in \{0, 1\}$: 핸드폰 OS (안드로이드면 1)

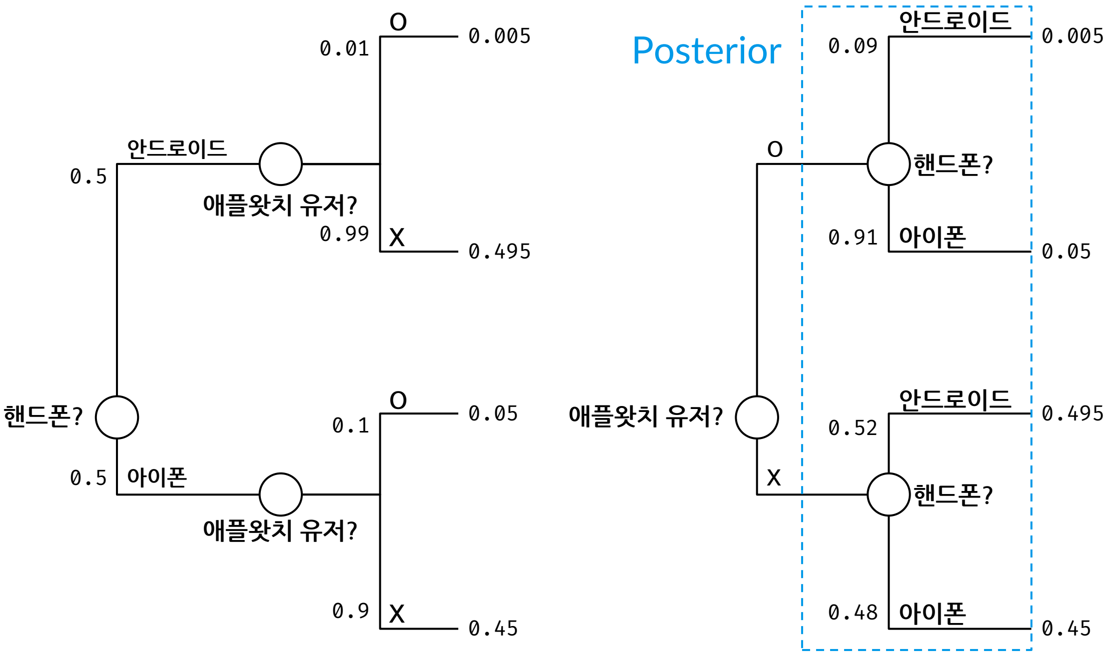
$P(p)$: 핸드폰이 안드로이드 일 확률 (prior)

$w \in \{0, 1\}$: 관찰한 데이터 (왓치유저면 1)

$P(w \mid p)$: 사전 확률분포를 가정 할 때, 데이터를 관찰할 확률

$$P(p \mid w) = \frac{P(p)P(w \mid p)}{C}$$

베이지안 확률모형의 수학적 정의



$p \in \{0, 1\}$: 핸드폰 OS (안드로이드면 1)

$P(p)$: 핸드폰이 안드로이드 일 확률 (prior)

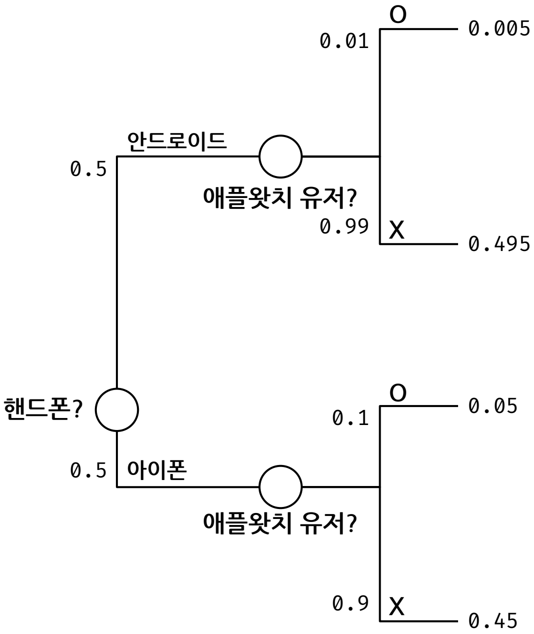
$w \in \{0, 1\}$: 관찰한 데이터 (왓치유저면 1)

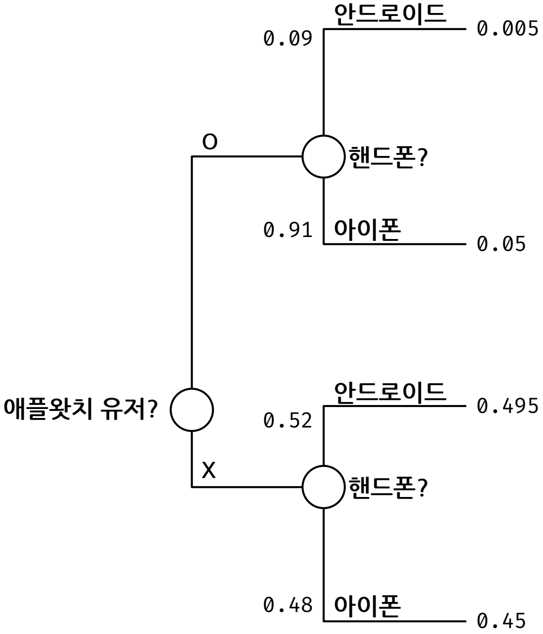
$P(w \mid p)$: 사전 확률분포를 가정 할 때, 데이터를 관찰할 확률

$$P(p \mid w) = \frac{P(p)P(w \mid p)}{C}$$

$$P(p \mid w) \propto P(p)P(w \mid p)$$

베이지안 확률모형의 수학적 정의





$p \in \{0, 1\}$: 핸드폰 OS (안드로이드면 1)

$P(p)$: 핸드폰이 안드로이드 일 확률 (prior)

$w \in \{0, 1\}$: 관찰한 데이터 (왓치유저면 1)

$P(w \mid p)$: 사전 확률분포를 가정 할 때, 데이터를 관찰할 확률

$$P(p \mid w) = \frac{P(p)P(w \mid p)}{C}$$

$$P(p \mid w) \propto P(p)P(w \mid p)$$

시뮬레이션을 통해 복잡한 사후확률분포(posterior)도 추정 가능
→ (Approximate) Computational Bayes

베이지안 확률모형의 수학적 정의

MCMC 샘플링 데모

https://chi-feng.github.io/mcmc-demo/

다음: (조금) 복잡한 확률모형

다시 지하철로 돌아가서 …

1부: 불확실성과 데이터

(조금) 복잡한 확률모형

데이터와 의사결정 | 정종빈

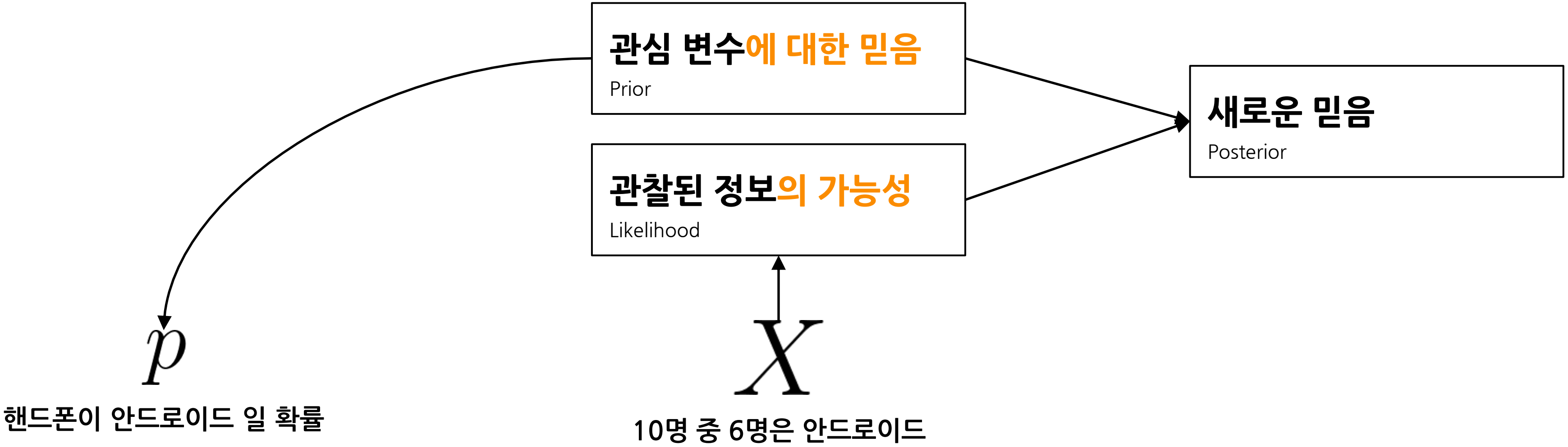
1부 강의 목적 및 개요

체계적이고 일반적인 <u>의사결정 모형</u> 소개

불확실성을 계량화하는 두 가지 접근법:

- **빈도주의(frequentist) 통계**
  Estimand, estimator, estimate, 신뢰구간, bootstrap
- **베이지안(Bayesian) 확률**
  Prior, likelihood, posterior (손계산 → 복잡한 모델링)

지하철로 돌아가서 …



지하철 상황의 실제 확률모형

관심 변수에 대한 믿음
Prior

관찰된 정보의 가능성
Likelihood

새로운 믿음
Posterior

p

핸드폰이 안드로이드 일 확률

0과 1 사이의 값 → Beta, 정규분포 + 변형 (e.g., logistic)

지하철 상황의 실제 확률모형

관심 변수에 대한 믿음
Prior

관찰된 정보의 가능성
Likelihood

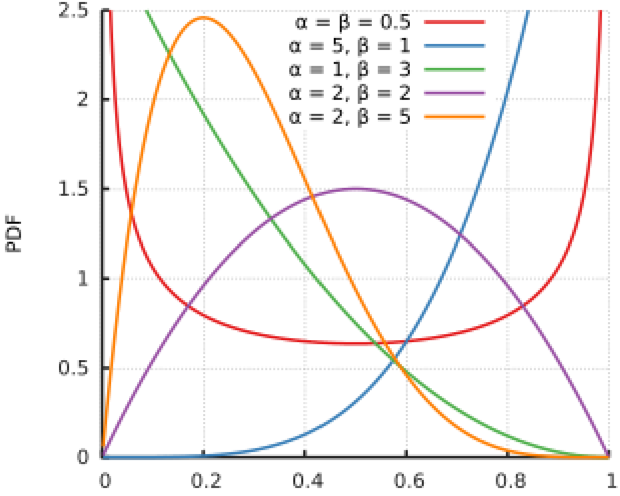
새로운 믿음
Posterior

p

핸드폰이 안드로이드 일 확률

0과 1 사이의 값 → **Beta**

- 두 개의 파라미터 a, b; 평균은 a / (a + b)



https://en.wikipedia.org/wiki/Beta_distribution

지하철 상황의 실제 확률모형

지하철 상황의 실제 확률모형

관심 변수에 대한 믿음
Prior

관찰된 정보의 가능성
Likelihood

새로운 믿음
Posterior

p

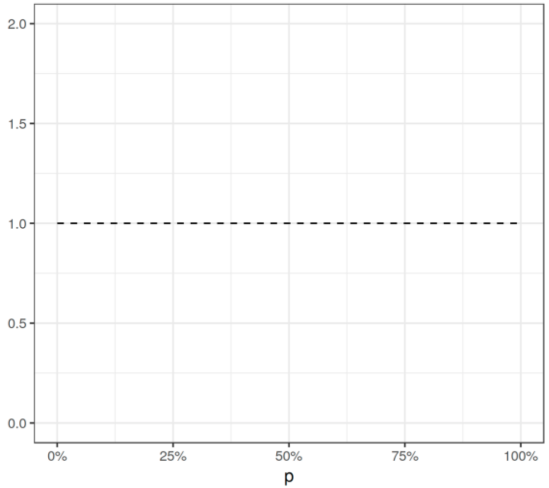
핸드폰이 안드로이드 일 확률

0과 1 사이의 값 → **Beta**

- 두 개의 파라미터 a, b; 평균은 a / (a + b)

예:

- 아무것도 모르겠다 → Beta(1, 1)



관심 변수에 대한 믿음
Prior

관찰된 정보의 가능성
Likelihood

새로운 믿음
Posterior

p

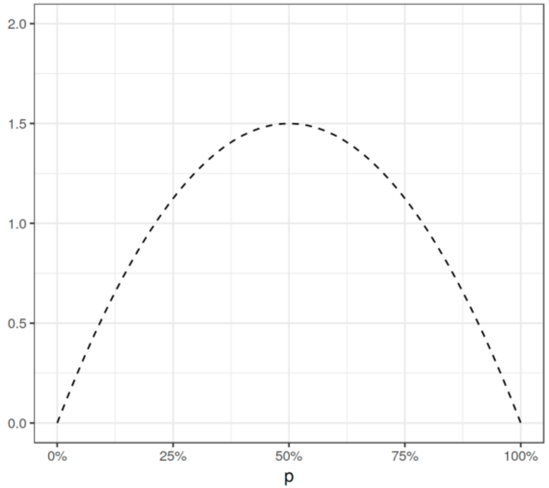
핸드폰이 안드로이드 일 확률

0과 1 사이의 값 → **Beta**

- 두 개의 파라미터 a, b; 평균은 a / (a + b)

예:

- 아무것도 모르겠다 → Beta(1, 1)
- 대충 반반쯤 되지 않나? → Beta(2, 2)



지하철 상황의 실제 확률모형

관심 변수에 대한 믿음
Prior

관찰된 정보의 가능성
Likelihood

새로운 믿음
Posterior

p

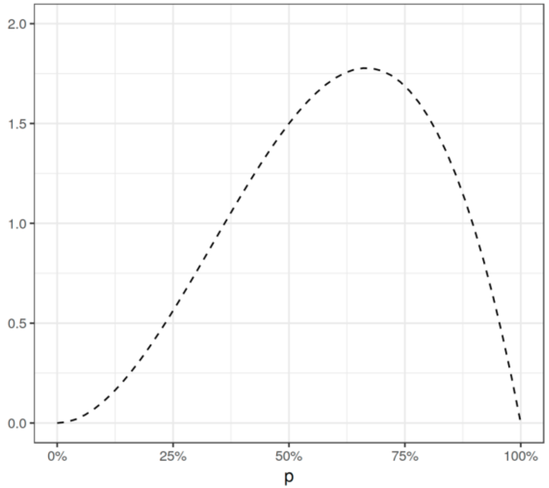
핸드폰이 안드로이드 일 확률

0과 1 사이의 값 → **Beta**

- 두 개의 파라미터 a, b; 평균은 a / (a + b)

예:

- 아무것도 모르겠다 → Beta(1, 1)
- 대충 반반쯤 되지 않나? → Beta(2, 2)
- 0.5 보다는 좀 큰가? → Beta(3, 2)



지하철 상황의 실제 확률모형

관심 변수에 대한 믿음
Prior

새로운 믿음
Posterior

관찰된 정보의 가능성
Likelihood

X

10명 중 6명은 안드로이드

p

핸드폰이 안드로이드 일 확률

0과 1 사이의 값 → **Beta**

- 두 개의 파라미터 a, b; 평균은 a / (a + b)

예:

- 아무것도 모르겠다 → Beta(1, 1)
- 대충 반반쯤 되지 않나? → Beta(2, 2)
- 0.5 보다는 좀 큰가? → Beta(3, 2)

“성공”확률이 있는 어떤 이벤트를
N = 10 번 반복해서 6번 “성공”
→ **이항분포(Binomial)**

- 두 개의 파라미터
  - N: 시행 횟수
  - p: 성공확률 (우리의 prior)

지하철 상황의 실제 확률모형

지하철 상황의 실제 확률모형

관심 변수에 대한 믿음
Prior

관찰된 정보의 가능성
Likelihood

새로운 믿음
Posterior

$$P(p \mid X) \propto P(p)P(X \mid p)$$

p ~ Beta(a, b) = P(p)
X ~ Binomial(N, p) = P(X | p)

$$p$$

핸드폰이 안드로이드 일 확률

0과 1 사이의 값 → **Beta**

- 두 개의 파라미터 a, b; 평균은 a / (a + b)

예:

- 아무것도 모르겠다 → Beta(1, 1)
- 대충 반반쯤 되지 않나? → Beta(2, 2)
- 0.5 보다는 좀 큰가? → Beta(3, 2)

$$X$$

10명 중 6명은 안드로이드

“성공”확률이 있는 어떤 이벤트를
N = 10 번 반복해서 6번 “성공”
→ **이항분포(Binomial)**

- 두 개의 파라미터
  - N: 시행 횟수
  - p: 성공확률 (우리의 prior)

Computational Bayesian Inference

Stan https://mc-stan.org/

+ 특화된 소프트웨어, documentation이 우수, 연구를 바탕으로한 우수한 개발팀 및 커뮤니티
- 실무 적용 인프라, 전용 언어/표기법

PyMC https://github.com/pymc-devs/pymc

+ Aesara (과거 Theano) 기반, 빠르게 발전/개발
- 빠르게 발전/개발 → 아직 안정화 X

Edward http://edwardlib.org/

+ Tensorflow/Keras를 이용한 확장성 및 실무 활용 가능성
- 커뮤니티의 부재, 개발 지연 (2021년 12월 현재, 2018년이 마지막 업데이트)

다음: 코딩실습 (PyMC)

1부: 불확실성과 데이터

PyMC 코딩 실습

데이터와 의사결정 | 정종빈

Google Colab에서 진행

PyMC, numpy, seaborn 이 설치된 환경 어디든 실습 가능

다음: Analytical solution

1부: 불확실성과 데이터

Analytical solution

데이터와 의사결정 | 정종빈

1부 강의 목적 및 개요

체계적이고 일반적인 <u>의사결정 모형</u> 소개

불확실성을 계량화하는 두 가지 접근법:

- **빈도주의(frequentist) 통계**
  Estimand, estimator, estimate, 신뢰구간, bootstrap
- **베이지안(Bayesian) 확률**
  Prior, likelihood, posterior (손계산 → 복잡한 모델링)

수식 경고!

반드시 알아야 할 중요한 내용은 **아님.**
내용의 “완성도”를 위해 다루니, 참고만 하세요.

관심 변수에 대한 믿음
Prior

관찰된 정보의 가능성
Likelihood

새로운 믿음
Posterior

$$P(p \mid w) \propto P(p)P(w \mid p)$$

p

핸드폰이 안드로이드 일 확률

0과 1 사이의 값 → **Beta**

- 두 개의 파라미터 a, b; 평균은 a / (a + b)

예:

- 아무것도 모르겠다 → Beta(1, 1)
- 대충 반반쯤 되지 않나? → Beta(2, 2)
- 0.5 보다는 좀 큰가? → Beta(3, 2)

X

10명 중 6명은 안드로이드

“성공”확률이 있는 어떤 이벤트를
N = 10 번 반복해서 s = 6번 “성공”
→ **이항분포(Binomial)**

- 두 개의 파라미터
  - N: 시행 횟수
  - p: 성공확률 (우리의 prior)

p ~ Beta(a, b)
X ~ Binomial(N, p)

특이 케이스: Conjugate priors

$$P(p) = \frac{p^{(a-1)}(1-p)^{(b-1)}}{B(a,b)}$$

관심 변수에 대한 믿음
Prior

관찰된 정보의 가능성
Likelihood

새로운 믿음
Posterior

$$P(p \mid w) \propto P(p)P(w \mid p)$$

$$p$$

핸드폰이 안드로이드 일 확률

0과 1 사이의 값 → **Beta**

- 두 개의 파라미터 a, b; 평균은 a / (a + b)

예:

- 아무것도 모르겠다 → Beta(1, 1)
- 대충 반반쯤 되지 않나? → Beta(2, 2)
- 0.5 보다는 좀 큰가? → Beta(3, 2)

$$X$$

10명 중 6명은 안드로이드

"성공"확률이 있는 어떤 이벤트를
N = 10 번 반복해서 s = 6번 "성공"
→ **이항분포(Binomial)**

- 두 개의 파라미터
  - N: 시행 횟수
  - p: 성공확률 (우리의 prior)

p ~ Beta(a, b)
X ~ Binomial(N, p)

특이 케이스: Conjugate priors

$$P(p) = \frac{p^{(a-1)}(1-p)^{(b-1)}}{B(a,b)}$$

$$P(w \mid p) = \frac{N!}{s!(N-s)!)} p^s (1-p)^{(N-s)}$$

$$p$$

핸드폰이 안드로이드 일 확률

0과 1 사이의 값 → **Beta**

- 두 개의 파라미터 a, b; 평균은 a / (a + b)

예:

- 아무것도 모르겠다 → Beta(1, 1)
- 대충 반반쯤 되지 않나? → Beta(2, 2)
- 0.5 보다는 좀 큰가? → Beta(3, 2)

관심 변수에 대한 믿음
Prior

관찰된 정보의 가능성
Likelihood

$$X$$

10명 중 6명은 안드로이드

"성공"확률이 있는 어떤 이벤트를

N = 10 번 반복해서 s = 6번 "성공"

→ **이항분포(Binomial)**

- 두 개의 파라미터
  - N: 시행 횟수
  - p: 성공확률 (우리의 prior)

새로운 믿음
Posterior

$$P(p \mid w) \propto P(p)P(w \mid p)$$

p ~ Beta(a, b)
X ~ Binomial(N, p)

특이 케이스: Conjugate priors

$$P(p) = \frac{p^{(a-1)}(1-p)^{(b-1)}}{B(a,b)} \qquad P(w \mid p) = \frac{N!}{s!(N-s)!)} p^{s}(1-p)^{(N-s)}$$

$$P(p \mid w) \propto P(p)P(w \mid p)$$

계산해 볼까요?

$$P(p) = \frac{\boxed{p^{(a-1)}(1-p)^{(b-1)}}}{B(a,b)} \qquad P(w \mid p) = \frac{N!}{s!(N-s)!}\boxed{p^s(1-p)^{(N-s)}}$$

$$\begin{aligned} P(p \mid w) &\propto P(p)P(w \mid p) \\ &\propto \boxed{p^{(a-1)}(1-p)^{(b-1)}}\boxed{p^s(1-p)^{(N-s)}} \end{aligned}$$

계산해 볼까요?

$$P(p) = \frac{p^{(a-1)}(1-p)^{(b-1)}}{B(a,b)}$$

$$P(w \mid p) = \frac{N!}{s!(N-s)!)} p^{s}(1-p)^{(N-s)}$$

$$P(p \mid w) \propto P(p)P(w \mid p)$$
$$\propto p^{(a-1)}(1-p)^{(b-1)} p^{s}(1-p)^{(N-s)}$$
$$\propto p^{(a-1+s)}(1-p)^{(b-1+N-s)}$$

계산해 볼까요?

$$P(p) = \frac{\boxed{p^{(a-1)}(1-p)^{(b-1)}}}{B(a,b)} \qquad P(w \mid p) = \frac{N!}{s!(N-s)!)} p^{s}(1-p)^{(N-s)}$$

$$\begin{aligned} P(p \mid w) &\propto P(p)P(w \mid p) \\ &\propto p^{(a-1)}(1-p)^{(b-1)} p^{s}(1-p)^{(N-s)} \\ &\propto \boxed{p^{(a-1+s)}(1-p)^{(b-1+N-s)}} \end{aligned}$$

계산해 볼까요?

$$P(p) = \frac{p^{(a-1)}(1-p)^{(b-1)}}{B(a,b)} \qquad P(w \mid p) = \frac{N!}{s!(N-s)!)} p^s (1-p)^{(N-s)}$$

$$\begin{aligned} P(p \mid w) &\propto P(p)P(w \mid p) \\ &\propto p^{(a-1)}(1-p)^{(b-1)} p^s (1-p)^{(N-s)} \\ &\propto p^{(a-1+s)}(1-p)^{(b-1+N-s)} \\ &\propto \mathrm{Beta}(a+s, b+(N-s))) \end{aligned}$$

계산해 볼까요?

정리

Prior: Beta(a, b)
Likelihood: Binomial(N, p)
“성공” 횟수가 s인 경우

정리

Prior: Beta(a, b)
Likelihood: Binomial(N, p)
“성공” 횟수가 s인 경우
Posterior: Beta(a + s, b + (N - s))

정리

Prior: Beta(a, b)
Likelihood: Binomial(N, p)
“성공” 횟수가 s인 경우
Posterior: Beta(a + s, b + (N - s))
지하철의 예: 아무것도 모르는 상태(a = 1, b = 1)에서
N = 10 명 중 s = 6 명이 안드로이드 유저인 걸 봤으면,
사후분포는 Beta(7, 4)을 따름. 평균은 7/(7+4) = 0.64
→ Beta-Binomial *conjugate* prior

Conjugate prior

Beta-Bernoulli
Beta-Binomial
Gamma-Poisson
Normal-Normal
Gamma-Normal
Gamma-Exponential
…

빠른 계산 vs 모델의 유연성

다음: 1부 마무리

1부: 불확실성과 데이터

마무리

데이터와 의사결정 | 정종빈

1부 강의 목적 및 개요

체계적이고 일반적인 의사결정 모형 소개

불확실성을 계량화하는 두 가지 접근법:

빈도주의(frequentist) 통계

Estimand, estimator, estimate, 신뢰구간, bootstrap

베이지안(Bayesian) 확률

Prior, likelihood, posterior (손계산 → 복잡한 모델링)

빈도주의? 베이지안?

사후분포 ∝ 데이터(likelihood) x 사전분포(prior)

1. 사전정보가 <u>전혀</u> 없다면?
    a. 정.말. <u>아무것도</u> 모르는가?
2. 데이터가 압도적으로 많다면?
    a. “극한”의 사전분포 → 데이터가 필요한가?
    b. 다른 사전분포 → 없는거나 마찬가지

베이지안의 강점

사전 정보가 중요할 때. 데이터가 적을 때*.

“앞/뒷면이 있는 동전을 두 번 던져서 앞면만 두 번 나옴”

- 빈도주의: “앞면이 나올 확률 1”
- 베이지안: “아마도 1보다는 0.5에 더 가깝지 않을까?”

* 소위 “빅데이터” 시대에도 데이터가 적은 경우 많음 (종로3가 80대 여성의 PS5 광고 CTR?)

베이지안의 약점

정밀하게(precise) 생각하지 못할 때

- 경험/지식으로 인한 잘못된 편견
- 문제에 대한 명확하지 못한 이해/분석 (승진의 예제)

분석가(의사결정자)의 능력/판단력이 중요

포괄적인 접근의 필요

빈도주의 기법의 적용
상식적으로(사전정보) 말이 되는가?

베이지안 기법의 적용
분석 과정/결과를 "신뢰"할 수 있는가 (평행우주의 관점에서)

공통적인 난제
"데이터는 어디로부터 왔는가?"

다음: 1부 실전 연습문제