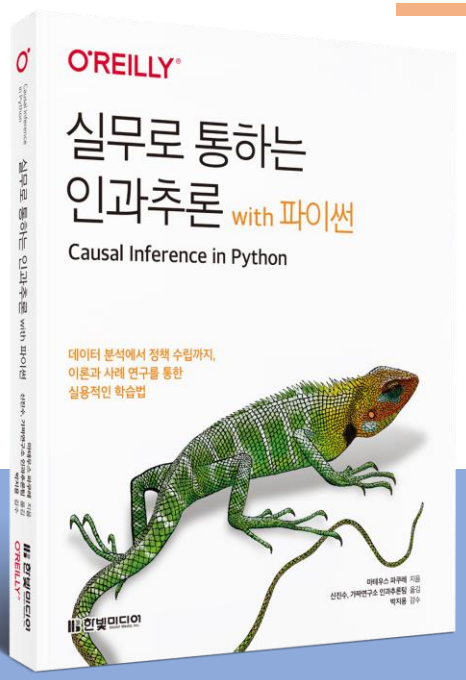
가짜연구소

Causal-Lab

『실무로 통하는 인과추론 with 파이썬』 특강



유저 이탈 관점에서 그래프 인과모형 소개

박시온

발표자 소개

Causal-Lab

박시온

- 게임회사 **데이터 분석가**
  - 이상탐지 업무
- 통계학 전공
  - 학부, 석사
  - 연구주제: 선형 구조 방정식 추론
- 관심 분야 : 운동, 요즘 일본어 공부, 행복하게 살기

Contents

Causal-Lab

0. 왜 그래프 인과모형인가?

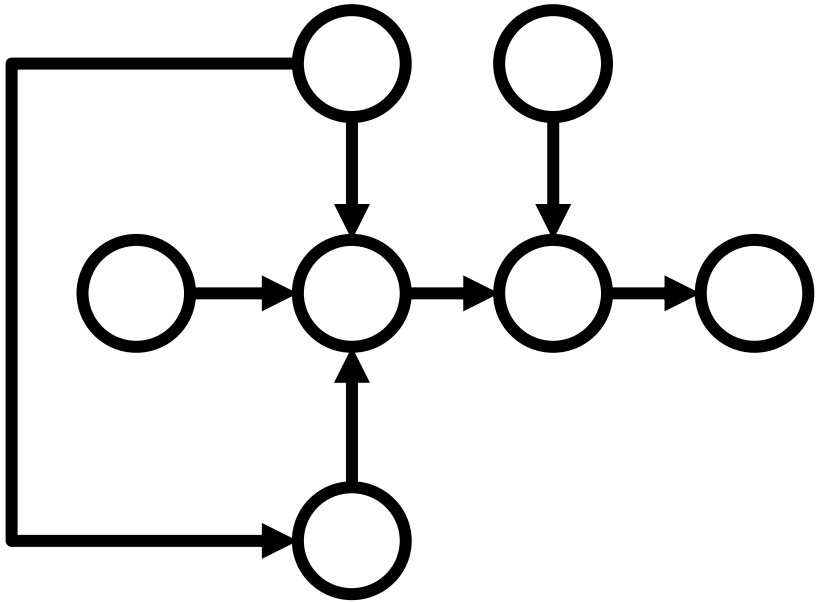
Causal-Lab

그래프 인과모형은 왜 배우는 걸까?

0. 왜 그래프 인과모형인가?

Causal-Lab

그래프 인과모형은 왜 배우는 걸까?

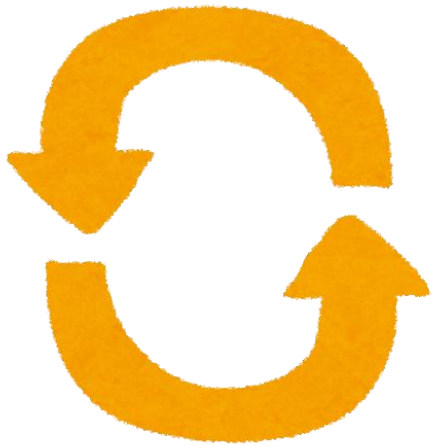


1. 그래프 인과모형 기초

Causal-Lab

(0) 그래프 인과모형 소개

DAG (Directed Acyclic Graph, 유향 비순환 그래프)



- 인과 그래프를 나타내는 방법
- 순환이 없음을 가정함

용어

- 노드(Node) : 변수, 보통 그래프에서 도형으로 나타냄
- 엣지(Edge) : 변수 간의 인과관계의 방향을 나타냄, 보통 그래프에서 화살표로 나타냄

1. 그래프 인과모형 기초

(0) 그래프 인과모형 소개

Tip

개인적으로는 가족관계를 이용해서 이해하는 방법이 쉽다.

모든 그래프 인과모형을 가족관계로 이해하는 것은 다소 무리가 있지만,
3개의 노드 만을 가진 그래프에서는 적절한 예시이다.

1. 그래프 인과모형 기초

Causal-Lab

(0) -1 조건부 독립이란?

어떠한 조건이 주어졌을 때 두 변수가 독립이다.

수식으로 표현하면

$P(X, Y \mid Z) = P(X|Z)P(Y|Z)$ 일 때 $X \perp Y \mid Z$ ➔ Z 가 주어졌을 때 X, Y 는 서로 독립이다.

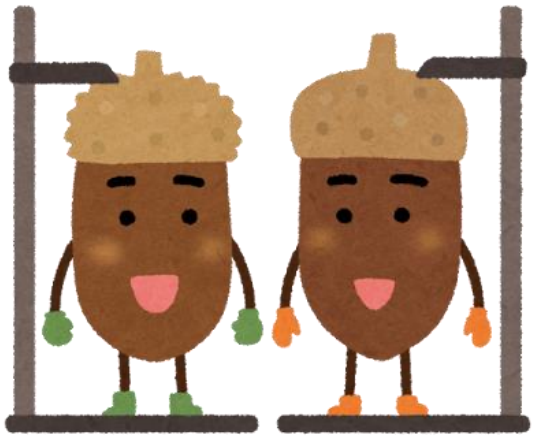
주어진 조건에 대한 정보를 알고 있는 상태인 것

예) 자식의 키 = 부모의 키 평균 $+ \varepsilon$ 이라고 하자. (여기서 ε 은 독립적인 변수)

일반적으로 첫째와 둘째의 키는 상관관계가 있다.

그럼 이때 자식의 키 – 부모의 키 평균 값은 첫째와 둘째가 서로 상관이 있을까?

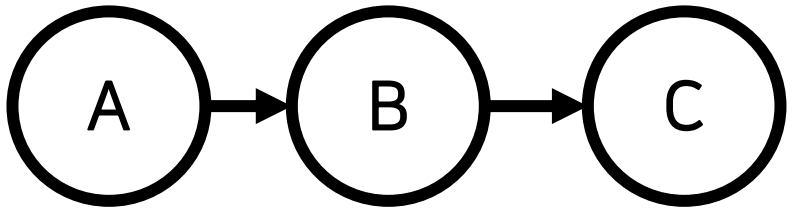
➔ 이미 알고 있는 정보(조건)는 빼고 비교해보자.



1. 그래프 인과모형 기초

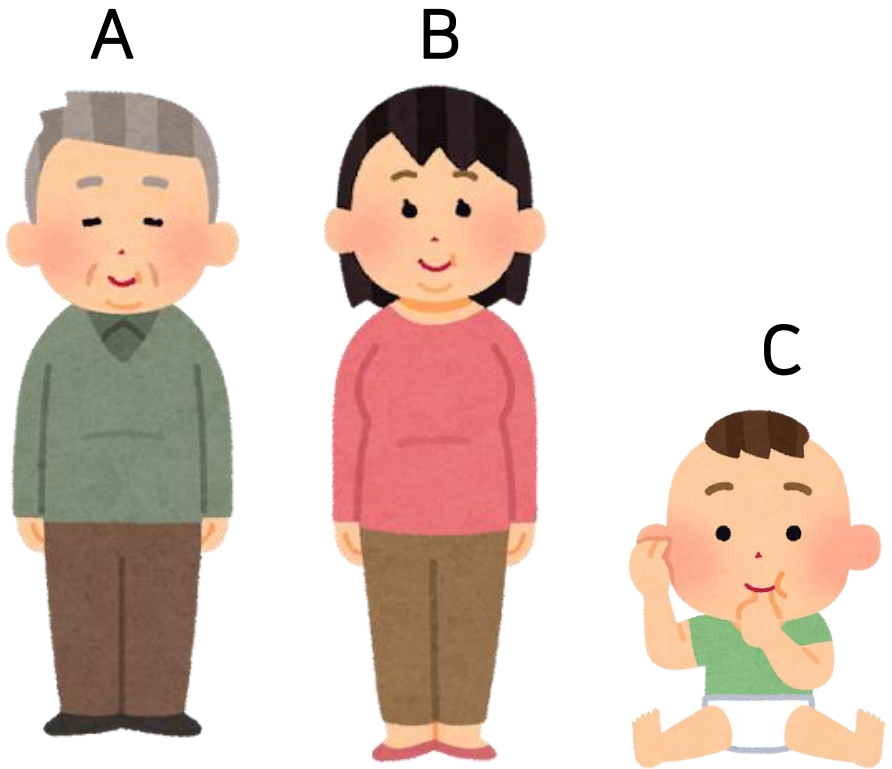
Causal-Lab

(1) 사슬



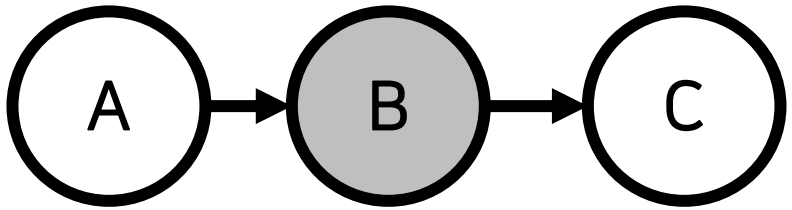
예) 조부모와 부모와 자식의 관계

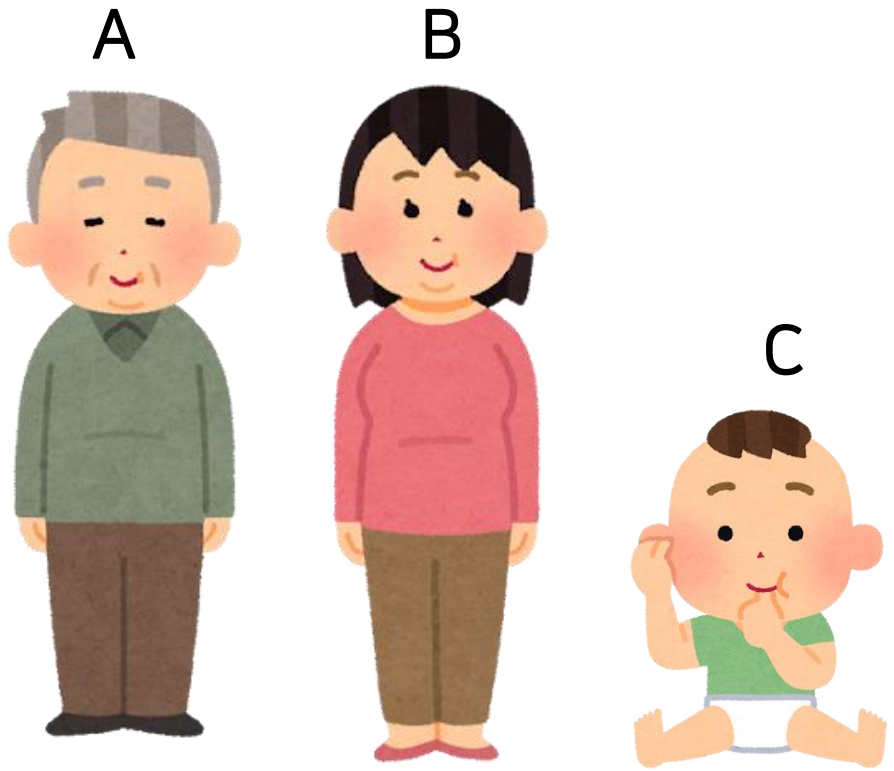
조부모(A) 와 자식(C)의 키는 서로 **종속**이다.



1. 그래프 인과모형 기초

(1) 사슬





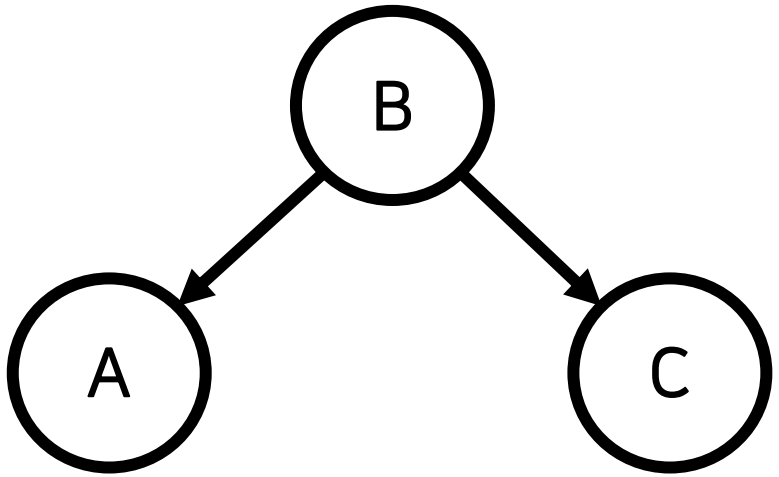
예) 조부모와 부모와 자식의 관계

만약 키가 180cm인 어머니들만 모은 집단이 있다면?

이 집단에서 할아버지들과 아이들의 키는 서로 **독립**이다. $(A \perp C \mid B)$

1. 그래프 인과모형 기초

(2) 분기

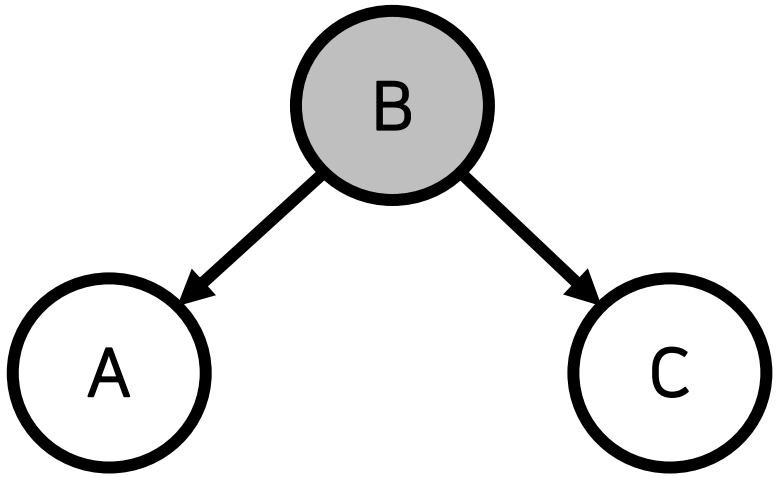


예) 부모와 형제들의 관계

첫째(A) 와 둘째(C)의 키는 서로 **종속**이다.

1. 그래프 인과모형 기초

(2) 분기



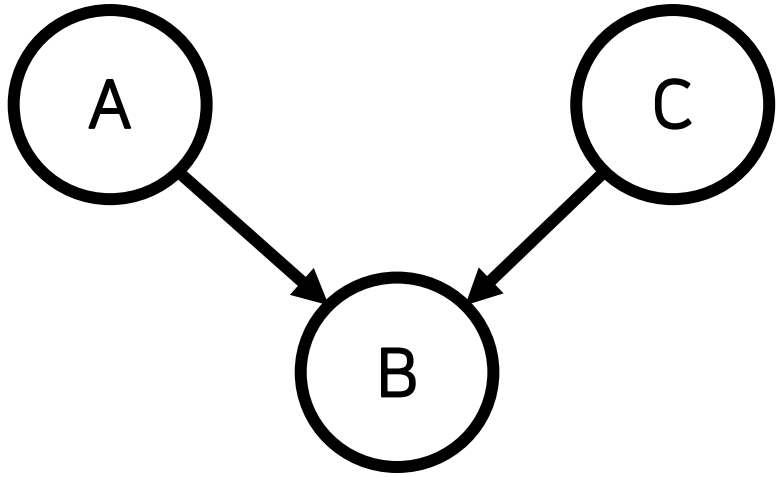
예) 부모와 형제들의 관계

만약 2명의 자식을 가지고 키가 160cm인 아버지들만 모아 놓으면?

이때 첫째들과 둘째들의 키는 서로 **독립**이다. $(A \perp C \mid B)$

1. 그래프 인과모형 기초

(3) 충돌부

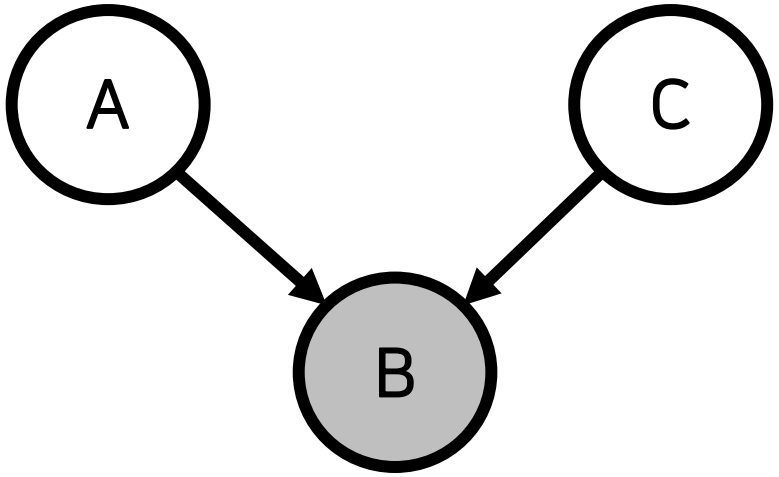


예) 부모와 자식의 관계

남편과 아내의 키는 독립이다.

1. 그래프 인과모형 기초

(3) 충돌부



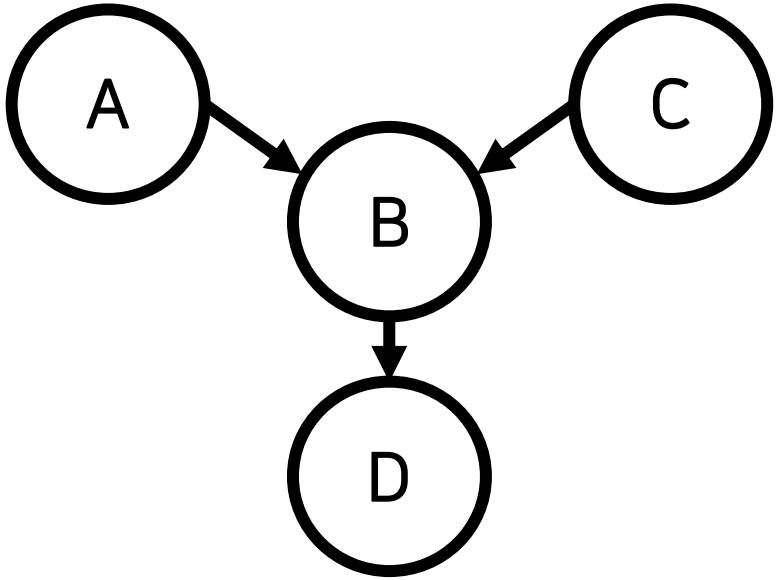
예) 부모와 자식의 관계

하지만 키가 190cm인 자식들만 모아 놓으면?

남편과 아내의 키는 서로 **종속**이다. $(A \not\perp C \mid B)$

1. 그래프 인과모형 기초

(3)-1 충돌부 심화

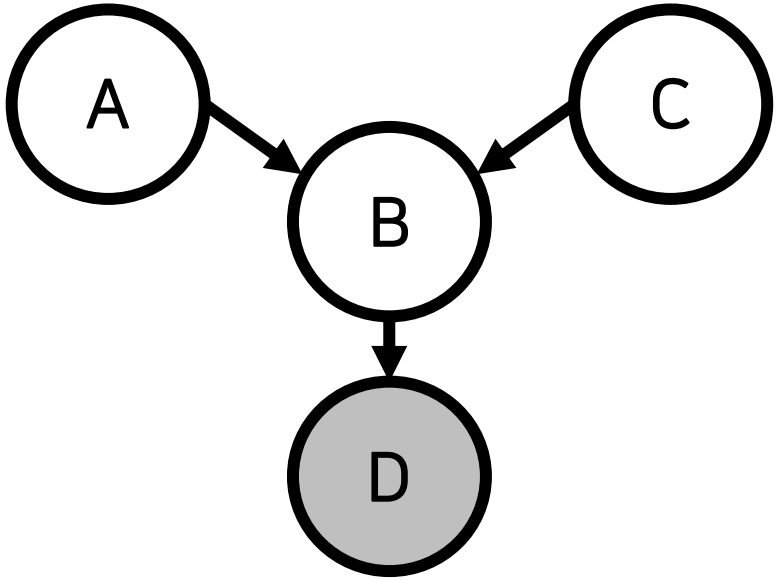


예) 부모와 자식의 관계

만약 키가 190cm인 손자들만 모아 놓으면?

1. 그래프 인과모형 기초

(3)-1 충돌부 심화



예) 부모와 자식의 관계

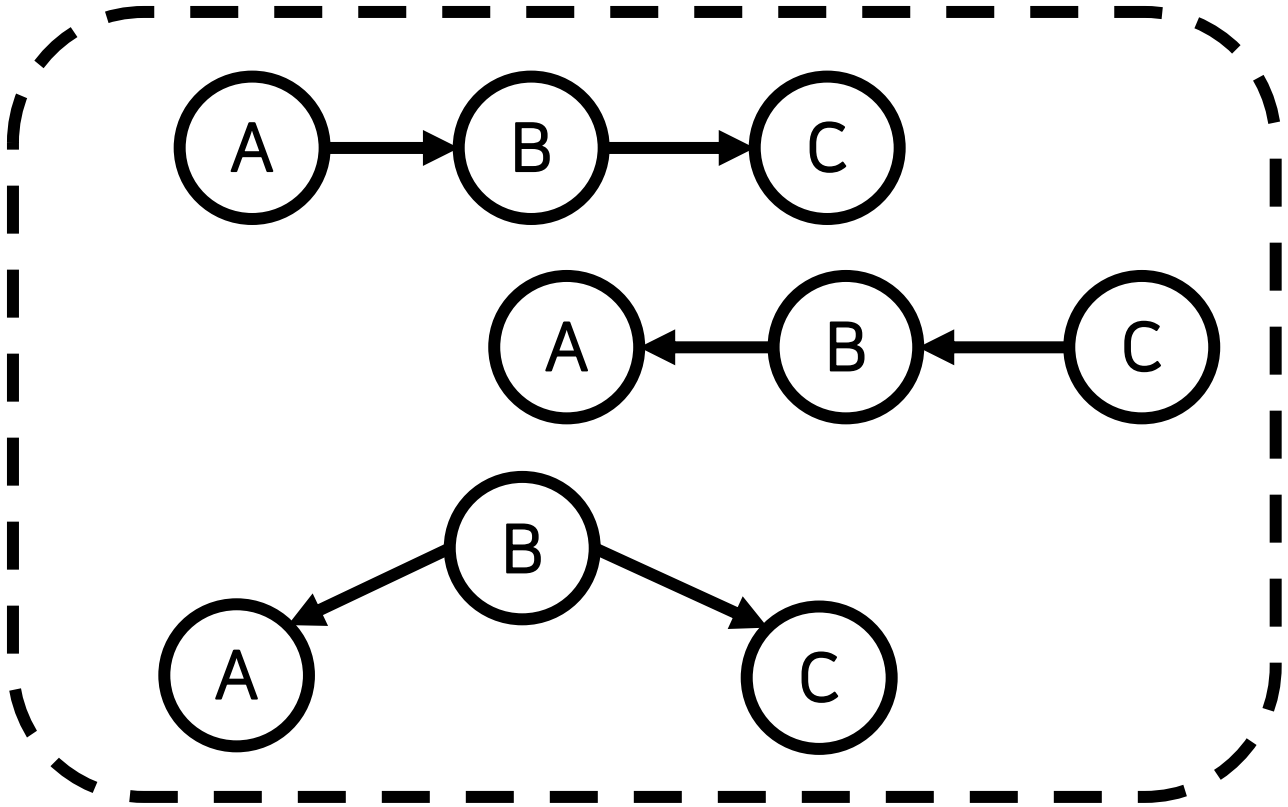
만약 키가 190cm인 손자들만 모아 놓으면?

남편과 아내의 키는 서로 **종속**이다. $(A \not\perp C \mid D)$ ➔ 손자가 자식의 키에 대한 정보를 일부 가지고 있기 때문

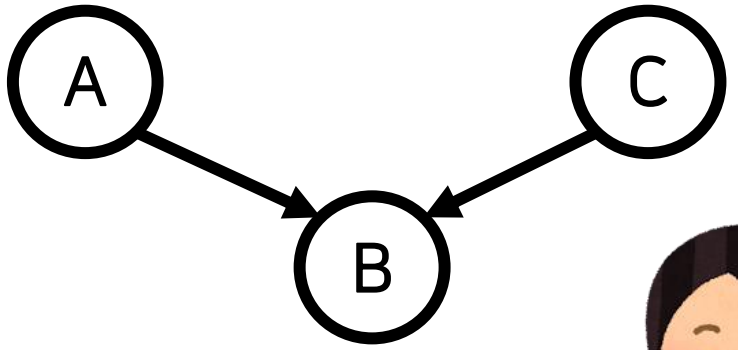
1. 그래프 인과모형 기초

Causal-Lab

(3)-2 충돌부로 알아보는 인과관계



세 그래프는 상관관계 구조가
모두 같아 **식별이 불가능**하다.

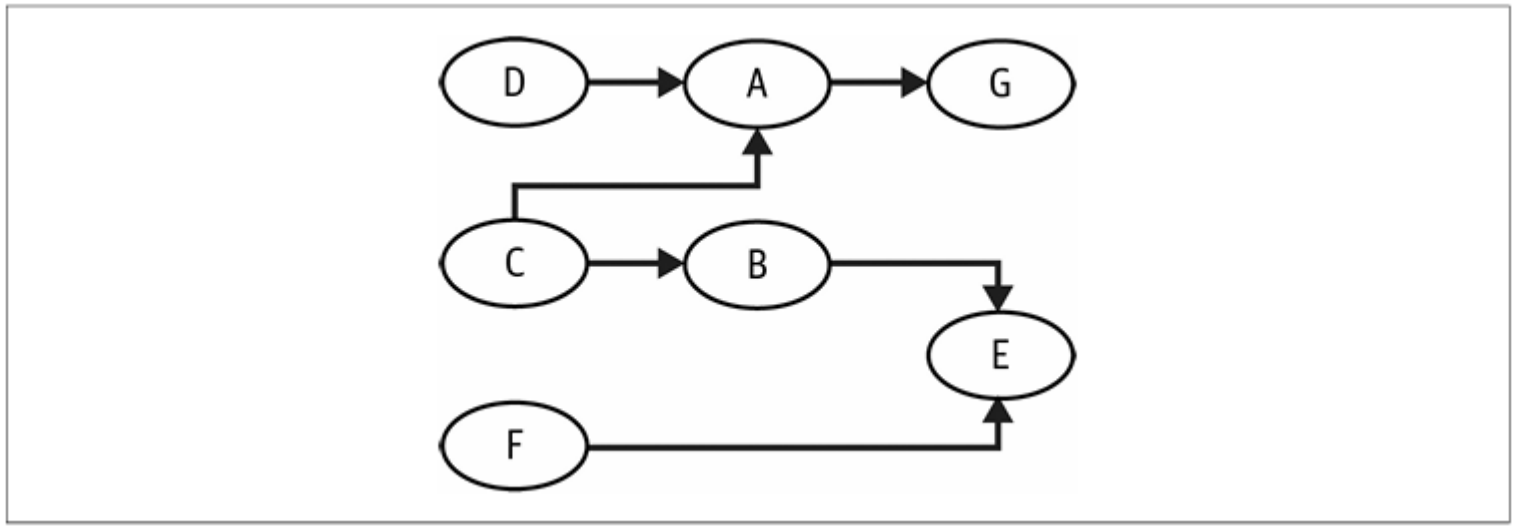


충돌부는 유일한 상관관계 구조를
가져 **식별할 수 있다**.

1. 그래프 인과모형 기초

Causal-Lab

(4) 실전 그래프 예시



(p 114)

D 와 C는 종속일까요?

A와 B 는 종속일까요?

G와 F는 종속일까요?

E가 주어진 경우, G와 F는 종속일까요?

1. 그래프 인과모형 기초

Causal-Lab

(5) 뒷문 경로

$$E[Y \mid T=1] - E[Y \mid T=0] = \underbrace{E\left[Y_1 - Y_0 \mid T=1\right]}_{ATT} + \underbrace{\left\{\left[Y_0 \mid T=1\right] - E\left[Y_0 \mid T=0\right]\right\}}_{\text{BIAS}}$$

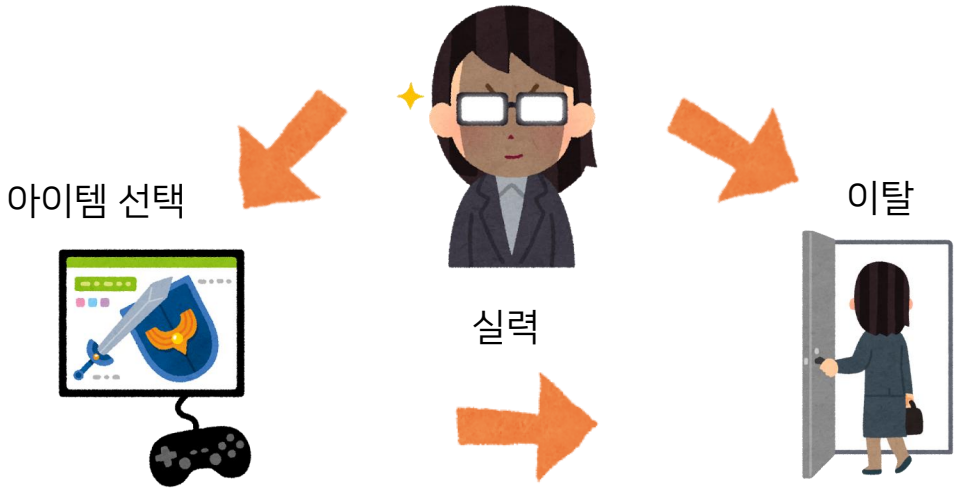
이 BIAS 는 무엇이며 왜 생길까?

예) 아이템 선택에 따른 이탈 분석

아이템 선택 박스에서 좋은 아이템을 선택하는 경우와 그렇지 않은 경우의 이탈률 차이를 비교하고 싶다.

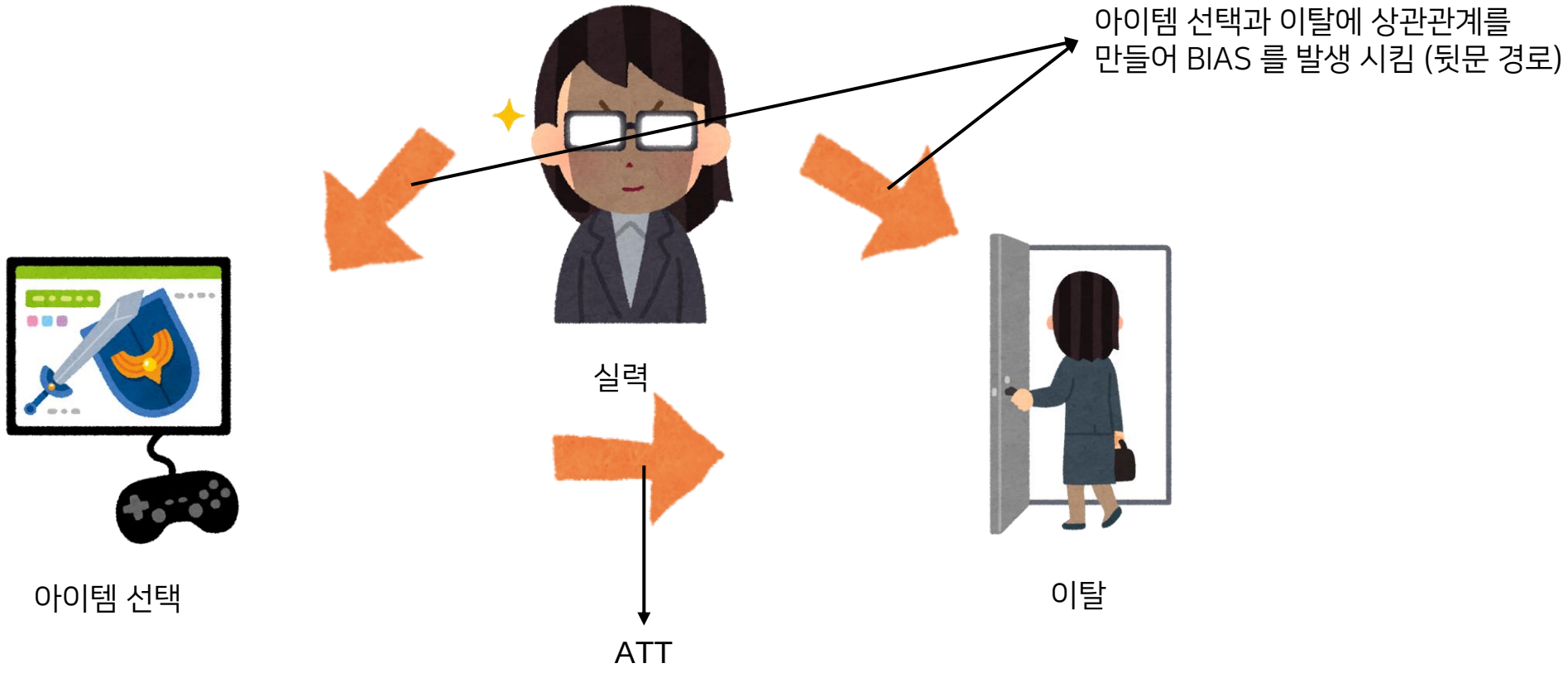
아이템을 선택한 경우와 그렇지 않은 경우를 직접 비교하면?

BIAS 가 발생한다.



1. 그래프 인과모형 기초

(5) 뒷문 경로



뒷문 경로(backdoor path): 직접적인 인과 경로와 공통 원인 때문에 교란 받는 비인과 경로

2. 인과추론의 가정

Causal-Lab

(1) 조건부 독립성 가정

$$(Y_0, Y_1) \perp T \mid X$$

공변량 수준이 동일한 경우에 동일한 대상을 비교하면 잠재적 결과는 평균적으로 같음을 말한다. (p.120)

공변량 수준이 동일하면 처치가 마치 무작위로 배정된 것처럼 보인다.

앞의 아이템 선택에 따른 이탈 예제

실력이 비슷한 사람들을 모집하여 아이템 선택과 이탈을 비교해보자!

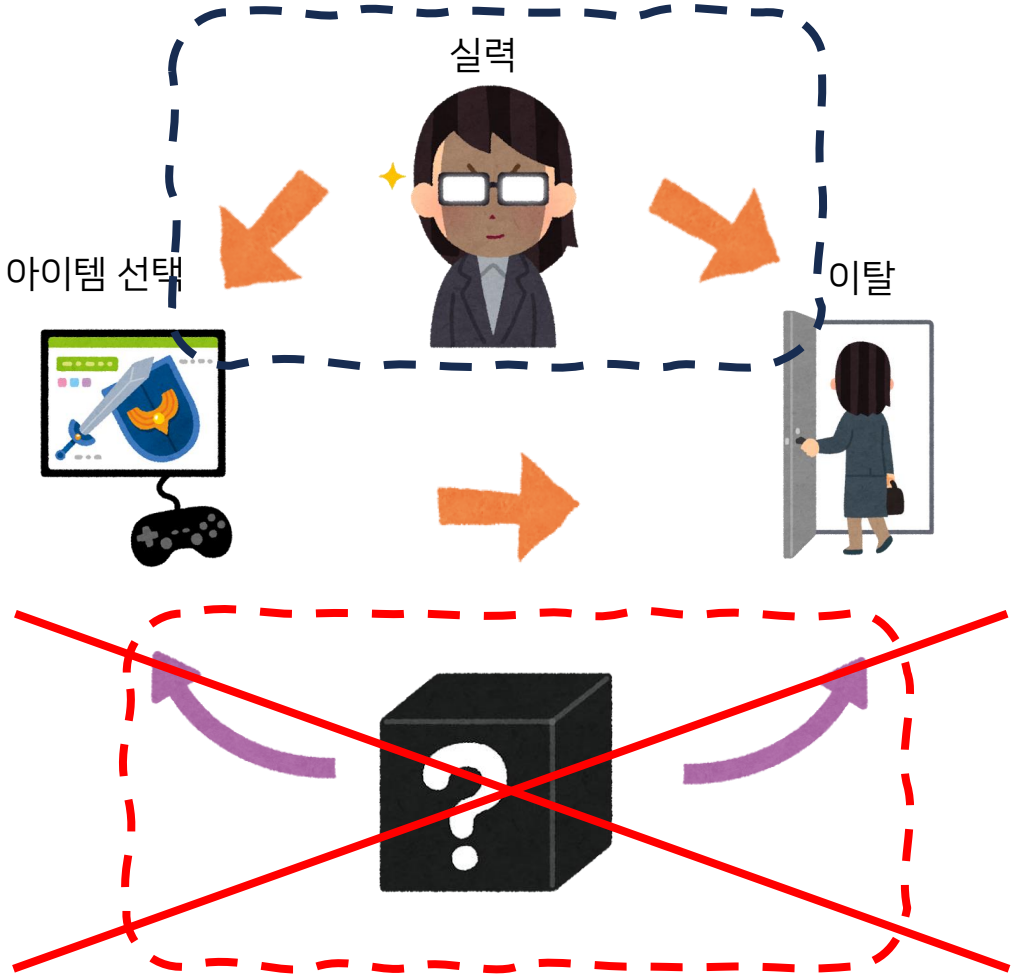
2. 인과추론의 가정

(1) 조건부 독립성 가정

더 이상 아이템 선택과 이탈에 비인과적 상관 관계가 발생하지 않음.

그래프의 측면에서 보면 조건부 독립성 가정은 이 공변량(X)가
뒷문 경로를 완전히 차단함을 의미.

예제에서 실력 외에 다른 뒷문 경로가 없음을 의미한다.



2. 인과추론의 가정

Causal-Lab

(2) 보정 공식

조건부 독립성 가정이 잘 만족한다면 보정공식을 이용해 인과 효과의 평균을 얻을 수 있다.

$$
\begin{aligned}
ATE &= E_X[E[Y \mid T=1] - E[Y \mid T=0]] \\
ATE &= \sum_x \{(E[Y \mid T=1, X=x] - E[Y \mid T=0, X=x])P(X=x)\} \\
&= \sum_x \{E[Y \mid T=1, X=x]P(X=x) - E[Y \mid T=0, X=x]P(X=x)\}
\end{aligned}
\quad \text{(p.120)}
$$

뒷문 경로로 흐르는 비인과 효과를 차단하기 때문에 **뒷문 보정**이라고 불림.
이 보정 공식에서는 아래와 같은 양수성 가정을 만족하는 것이 중요하다.

양수성 가정

$$0 < P(T|X) < 1$$

X의 모든 그룹에 실험군과 대조군이 반드시 존재하여야 함

2. 인과추론의 가정

Causal-Lab

(2)-1 보정 공식 적용해보기

실력	아이템 선택	이탈 여부
상	1	0
상	1	0
상	0	1
하	1	1
하	0	0
하	0	1

$P(\text{실력} = \text{상}) = 0.5$

$P(\text{실력} = \text{하}) = 0.5$

$E[\text{이탈} \mid \text{아이템선택} = 1, \text{실력} = \text{하}] = 1$

$E[\text{이탈} \mid \text{아이템선택} = 0, \text{실력} = \text{하}] = 0.5$

$E[\text{이탈} \mid \text{아이템선택} = 1, \text{실력} = \text{상}] = 0$

$E[\text{이탈} \mid \text{아이템선택} = 0, \text{실력} = \text{상}] = 1$

$$ATE = (E[\text{이탈} \mid \text{아이템선택} = 1, \text{실력} = \text{상}] - E[\text{이탈} \mid \text{아이템선택} = 0, \text{실력} = \text{상}]) \cdot P(\text{실력} = \text{상})$$
$$+(E[\text{이탈} \mid \text{아이템선택} = 1, \text{실력} = \text{하}] - E[\text{이탈} \mid \text{아이템선택} = 0, \text{실력} = \text{하}]) \cdot P(\text{실력} = \text{하})$$
$$= (0 - 1) \cdot 0.5 + (1 - 0.5) \cdot 0.5 = -0.25$$

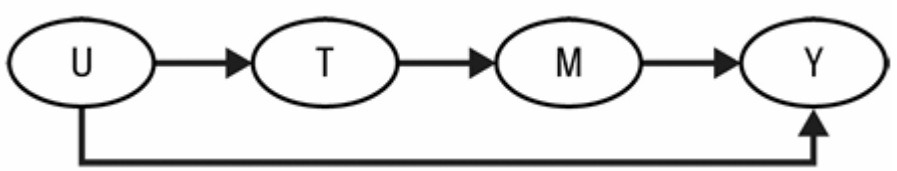
아이템 선택을 하면 이탈율이 25% 감소함 (뒷문 조정을 이용해 합리적인 인과 효과를 추정할 수 있다.)

2. 인과추론의 가정

Causal-Lab

(3) 앞문 조정

앞문 조정도 당연히 있습니다.



(p.123)

이 사슬 구조에서도 **U -> Y의 직접적인 인과 효과**를 추론하기 위해 T 로 통하는 상관관계를 차단할 수 있습니다.

하지만 이러한 사슬구조에서 상관 관계 흐름은 인과 관계 구조상 **자연스러운 흐름**입니다.
그렇기 때문에 이 흐름 또한 인과 효과라고 볼 수 있습니다.

즉, 굳이 **U -> Y로의 인과 효과를 추정하기 위해서는 굳이 앞문 조정을 할 필요가 없다**고 생각됩니다.

3. 편향

(1) 교란 편향

비인과적으로 연관성이 흐르는 열린 뒷문 경로가 있을 때 발생

앞의 아이템 선택과 이탈률 예제에서 설명한 것들이 모두 교란 편향

인과 효과를 추정할 때에 교란 편향은 아주 빈번하게 나타나는 문제이지만, 이를 바로 알기는 어렵습니다.
➔ 도메인 지식을 통해 처치와 결과 사이에 나타날 수 있는 인과 관계를 그래프로 잘 정리하기

교란 편향

교란 편향

3. 편향

(1)-1 교란 편향 - 대리 교란 요인

교란 요인을 측정할 수 없을 때, 교란요인을 대신할 수 있는 요인

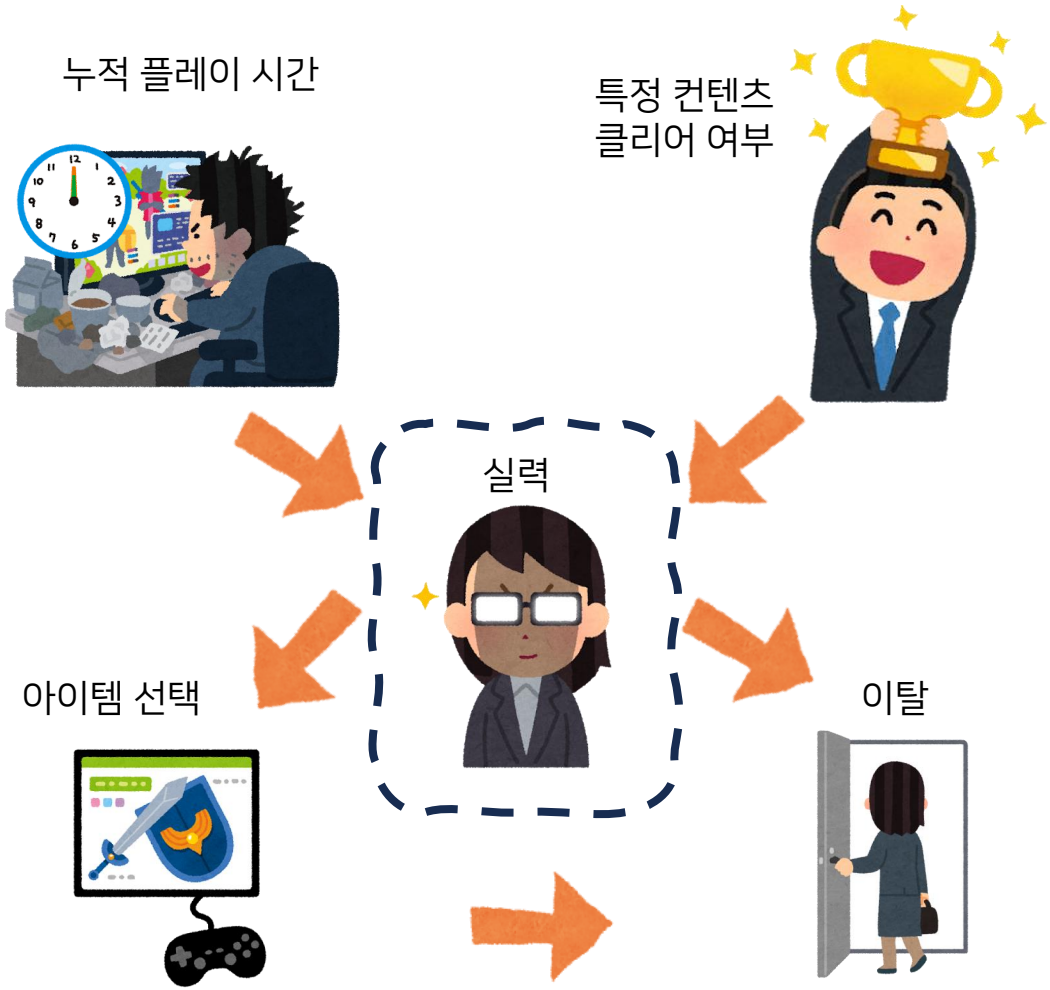
대리 교란 요인을 이용해서 교란 편향을 줄일 수 있다.
(완전히 제거할 수 없다.)

예) 앞의 아이템 선택에 대한 이탈률 예제

실력이란 변수는 실제로 측정하기 어려움.

측정가능한 누적 플레이 시간, 특정 컨텐츠 클리어 여부를 조건으로 두어 **아이템 선택에 다른 이탈률의 인과효과를 추정할 수 있다**.

왜?
누적 플레이 시간, 특정 컨텐츠 클리어 여부를 통해 실력에 대한 정보를 간접적으로 알 수 있기 때문임.



3. 편향

(1)-2 교란 편향 - 랜덤화

처치를 랜덤하게 줌으로 써 교란요인의 영향을 받지 않도록 함.

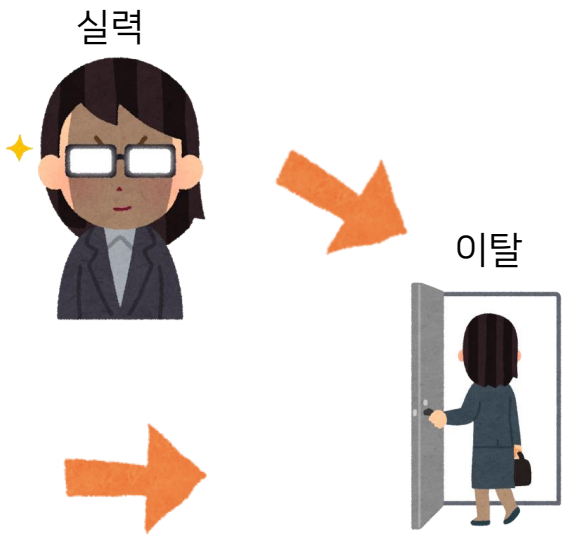
교란요인의 영향을 받지 않으므로 뒷문 경로가 사라져 비인과적 상관관계가 흐르지 않음.
교란 편향 값이 0이 됨.

예) 앞의 아이템 선택에 대한 이탈률 예제

유저별로 랜덤하게 아이템 부여

그러면 아이템 선택에 따른 이탈 효과를 바로 측정할 수 있다.

$$E[\text{이탈}|\text{아이템선택}=1] - E[\text{이탈}|\text{아이템선택}=0]$$



너무 좋은 방법이지만 대부분의 상황에서 불가능 하다. ➔ 우리가 인과 추론을 배우는 이유

3. 편향

(2) 선택 편향

인과 효과 추정을 위한 표본 선택으로 나타나는 편향

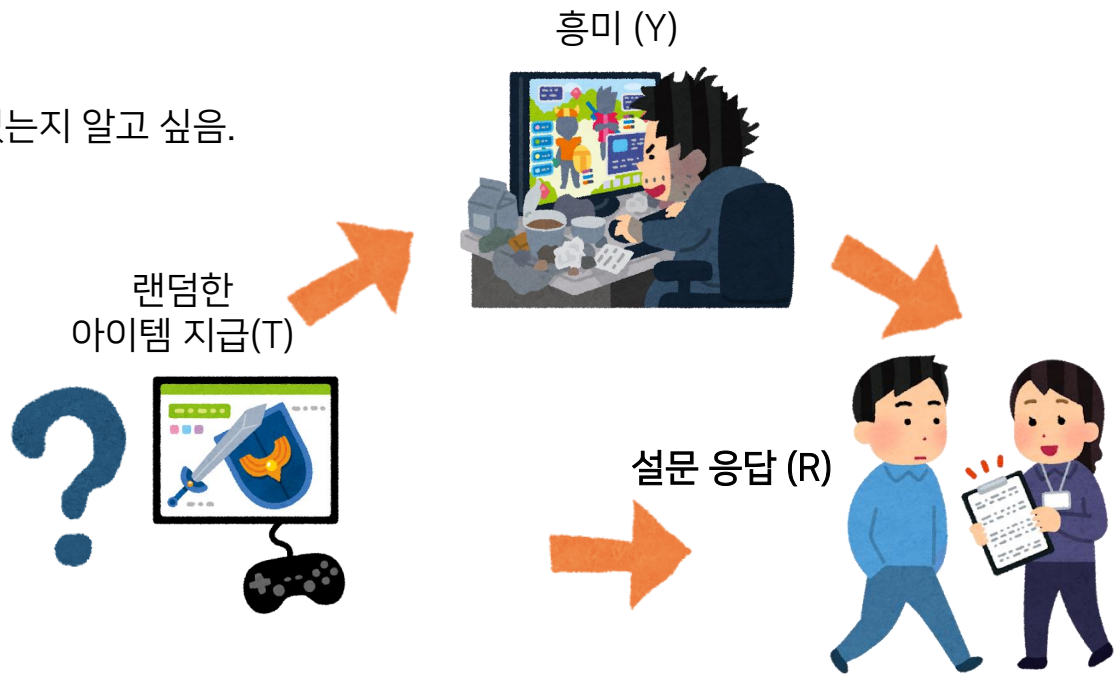
예) 랜덤 아이템 지급에 따른 흥미 수준 비교

랜덤하게 아이템을 지급하고 좋은 아이템을 지급 받는 경우에 흥미가 있는지 알고 싶음.

아이템을 랜덤으로 지급하고 이에 따라 이탈에 따른 설문 응답을 한다

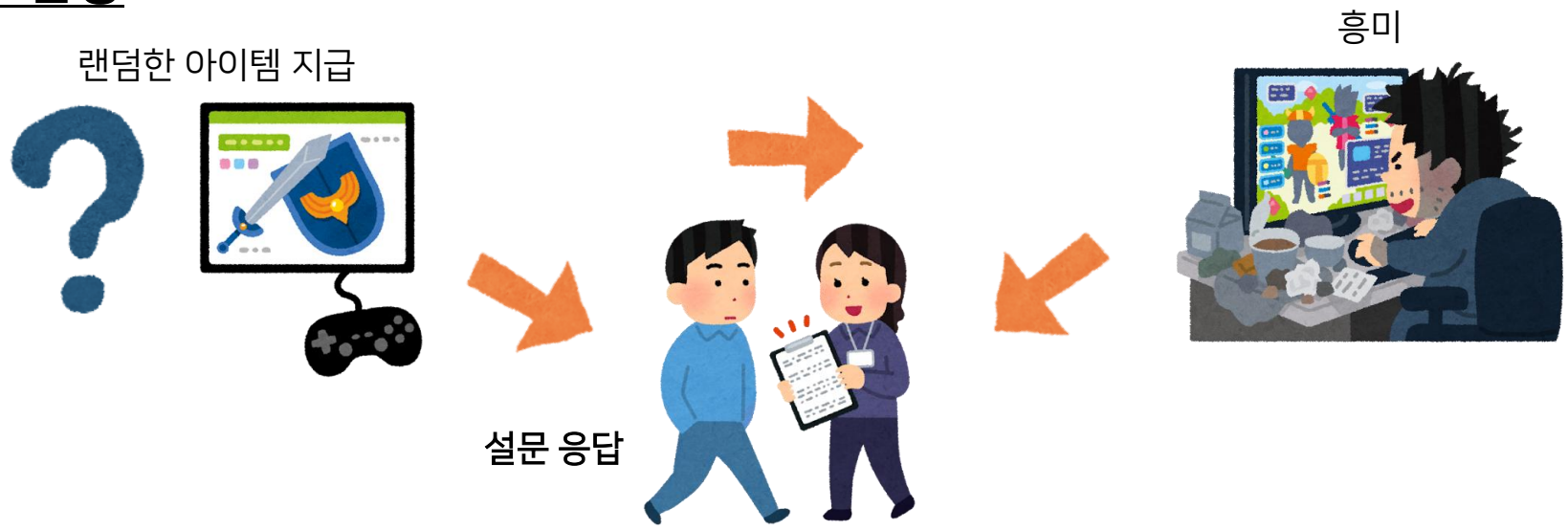
$$E[Y \mid T=1,R=1]-E[Y \mid T=1,R=1] = \underbrace{E\left[Y_1-Y_0 \mid R=1\right]}_{ATE} + \underbrace{E\left[Y_0 \mid T=0,R=1\right]-E\left[Y_0 \mid T=1,R=1\right]}_{\text{선택편향}}$$

(p.130)



3. 편향

(2) 선택 편향



인과 그래프를 이용한 해석: **설문 응답(R)을 조건**으로 하여 랜덤한 아이템 지급 (T) 와 흥미 (Y) 에 **비인과적 상관관계**가 흐름

$$\underbrace{E\left[Y_0 \mid T=0, R=1\right] - E\left[Y_0 \mid T=1, R=1\right]}_{\text{(1)} \quad \text{선택편향} \quad \text{(2)}}$$

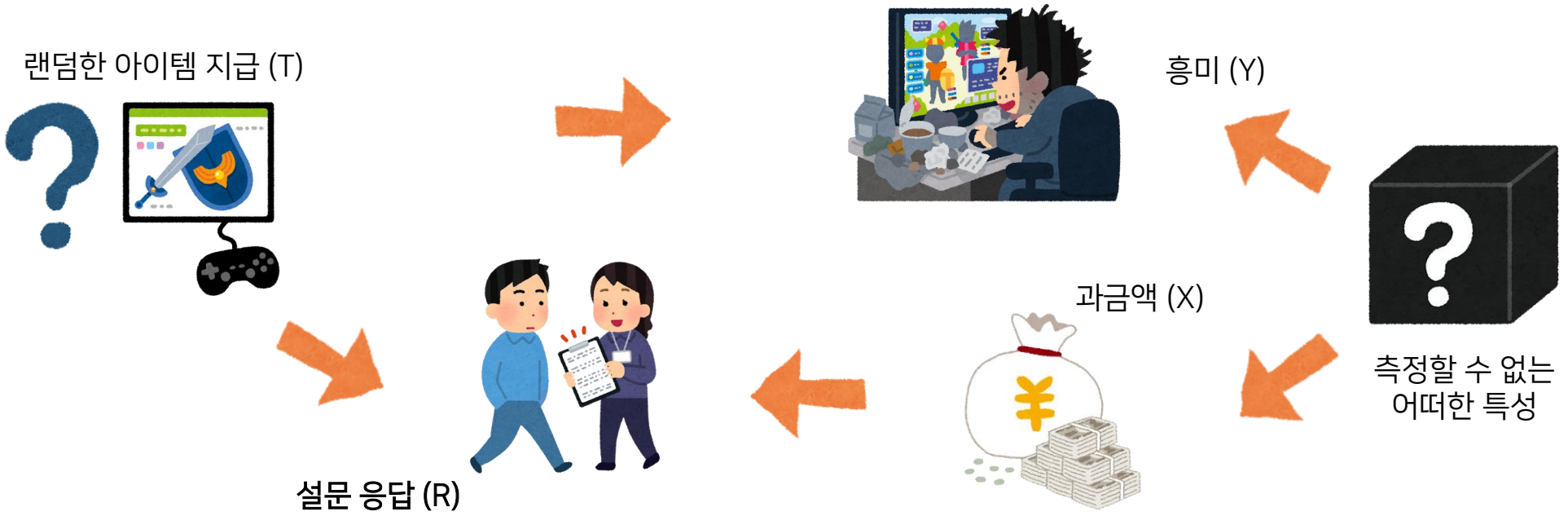
(1) 기존에 흥미가 적은 사람도 좋은 아이템을 지급을 받아 설문에 우호적일 수도 있음
➔ 좋은 아이템을 지급 받지 않은 경우에 흥미가 낮았을 것임, 물론 기존에 흥미가 있던 유저도 다수 포함

(2) 좋은 아이템을 지급받지 않더라도 흥미가 있는 사람은 설문에 응답할 확률이 높음
➔ 기존에 흥미가 없는 사람들은 응답률이 낮음

3. 편향

(2)-1 선택 편향 보정

그럼 어떻게 선택편향을 제거할 수 있을까?



흥미 (Y) 에 대한 설문 응답 (R)으로의 상관관계를 끊어 줌으로 써 (과금액 (X)을 조건으로 둠) 비인과적 상관관계를 제거할 수 있다.

$$ATE = \sum_{x} \{(E[Y \mid T=1, R=1, X] - E[Y \mid T=0, R=1, X])P(X \mid R=1)\}$$

4. 요약

Causal-Lab

인과 그래프를 잘 알고 있다면, 앞으로 배울 여러 인과 추론 방법에 대해 쉽게 이해할 수 있습니다.

그래프 인과모형의 구조를 이용해 비인과적 상관관계가 흐르는 경로를 알 수 있다.
➔ 비인과적 상관관계가 흐르지 않도록 설계할 수 있음.

편향은 어디에나 숨어 있고 이를 명시적으로 알기에는 어렵습니다.
➔ 각 분야별 도메인 지식이 중요합니다.

편향