1. 인과 추론의 정의 및 인과 추론이란?

생성일	@2022년 2월 3일 오후 2:25
태그	

Causal Inference

Causality: An Introduction

원인파악 / 효과측정을 위해 제일 좋은 방법은?

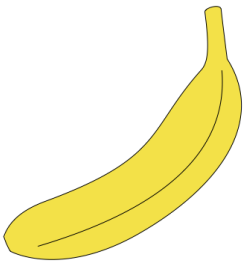
측정하고자 하는 변수 이외에는 모든 것들을 고정시키고,
확인하고 싶은 항목만 변경해서 테스트해본다

Randomized Control Test

- 흔히들 말하는 A/B 테스트
- 모든 변수를 완벽하게 고정하는 대신, 무작위 배정으로 평균적인 효과를 측정할 수 있다

그런데 A/B 테스트를 할 수 없는 상황이라면 어떻게 해야 할까?

- 이것은 인과 관계에 대한 3개의 시리즈 중 첫 번째 게시물입니다.
- 인과관계를 **우리가 생각하는 방식과 연결**하고, **전통적인 통계의 문제를 강조**하고, **인과관계를 수학적으로 표현**하는 과정을 진행할 것 입니다.



The Banana

위의 사진을 보면 무엇이 떠오르나요 ? 잠시 시간을 드리겠습니다.

가장 많이 떠올릴 질문은 "바나나와 인과 관계는 무슨 상관이 있을까?"입니다. 물론 바나나와 인과-관계는 아무런 관계가 없습니다. 하지만 의도적이든 아니든 당신은 아마도 커넥션을 찾으려고 시도했을 것입니다.

우리는 언제나 왜 ? 라는 근본적인 질문을 던집니다.
”바나나는 왜 여기있나요?”, “내가 이 블로그 게시물을 클릭한 이유는 뭐였죠?”, “나는 왜 아

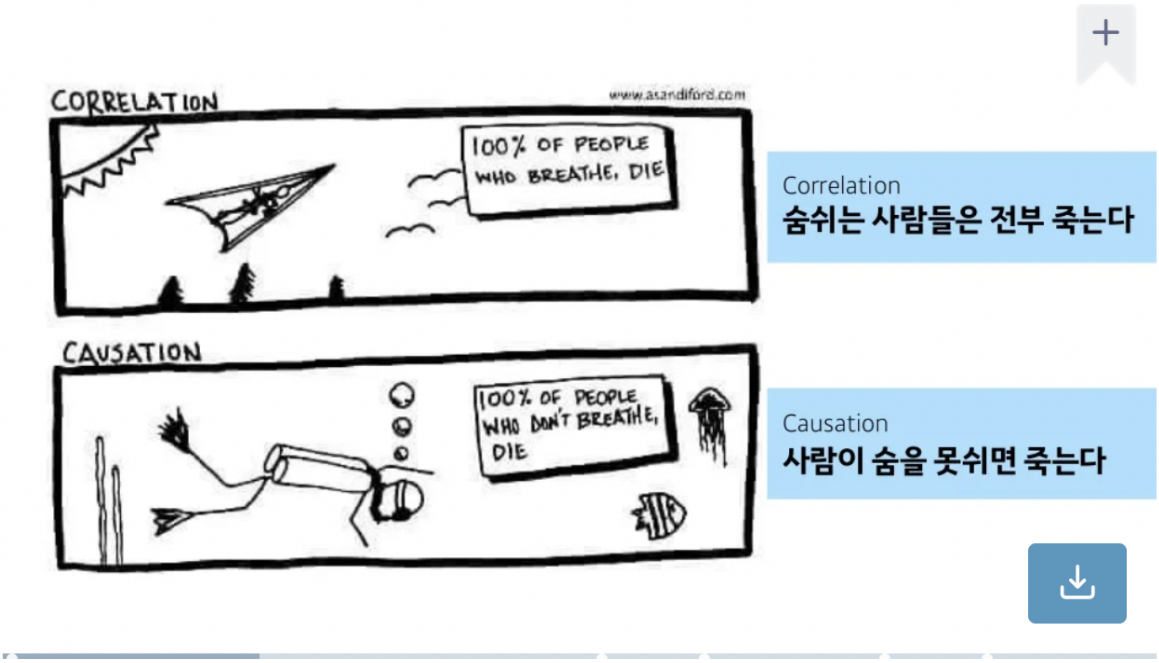
직도 이것을 읽고 있나요...?”와 같은 질문 말입니다.

즉, 우리는 끊임없이 스스로에게 그 이유를 묻고 있습니다. 우리는 무의미한 세상을 이해하는데 도움이 되는 인과 관계의 이야기를 계속해서 엮으려고 합니다.

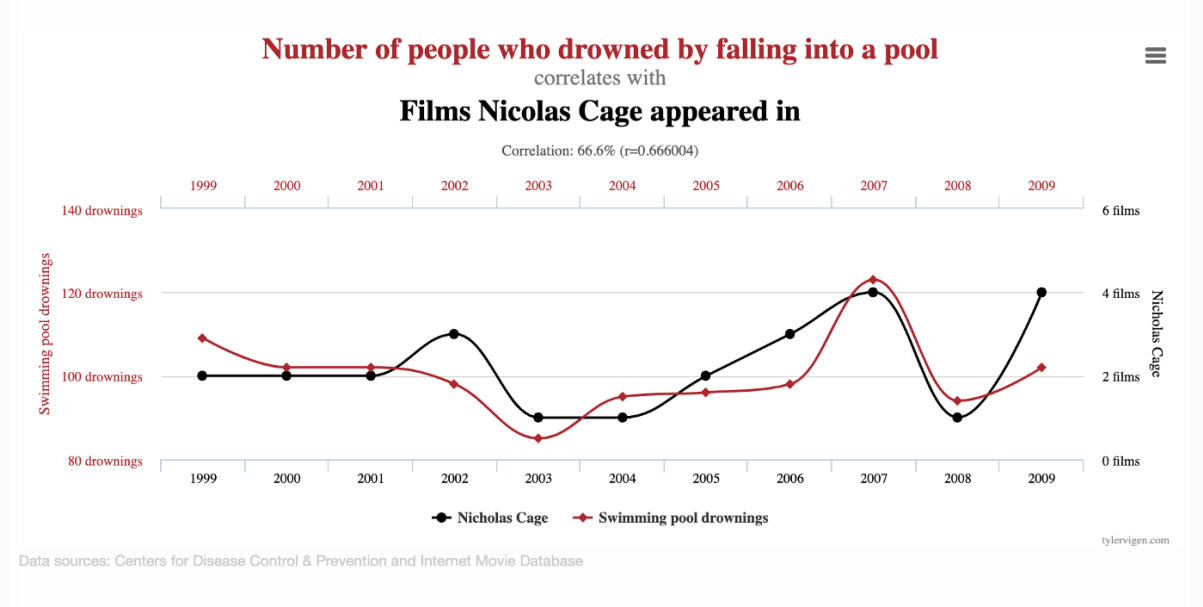
이러한 인과적 사고 방식은 우리에게 자연스러운 것으로 느껴지지만, 우리가 생각하는 전통적인 통계는 인과 관계를 쉽게 처리할 수 없습니다. 세 가지 주요 포인트에서 전통적인 통계의 부적절함을 설명해보겠습니다. 이 포인트들을 ***3 traps of statistics*** 라고 부르겠습니다.

3 Traps of Statistics

Trap 1: Spurious Correlation : 비 논리적인 상관관계



우리는 <상관관계 = 인과관계>가 아니라는 오래된 사실을 알고 있습니다. 이것은 Spurious Correlation를 잘 설명해줍니다. 아래 그림을 보고 생각해보십시오. (이번에는 바나나를 사용하지 않겠다고 약속합니다).수영장에서 익사한 사람의 연간 수와 니콜라스 케이지 배우의 영화의 수 사이에는 강한 상관관계(66%)가 있습니다. 분명히, 그리고 아주 재미있게, 이 두 변수는 상관 관계에 있음에도 불구하고 어떤 식으로든 인과 관계가 없습니다.



Correlation between yearly number of people who drowned by pool and number of Nick Cage movies. Source: https://tylervigen.com/spurious-correlations

웃긴 예시임에도 불구하고 이 예시는 우리가 상관 관계를 다룰 때 주의해야 함을 상기시켜줍니다. 통계적 유의성에 대해서 포인트를 잡으면, 두가지의 장애물을 볼 수 있습니다.

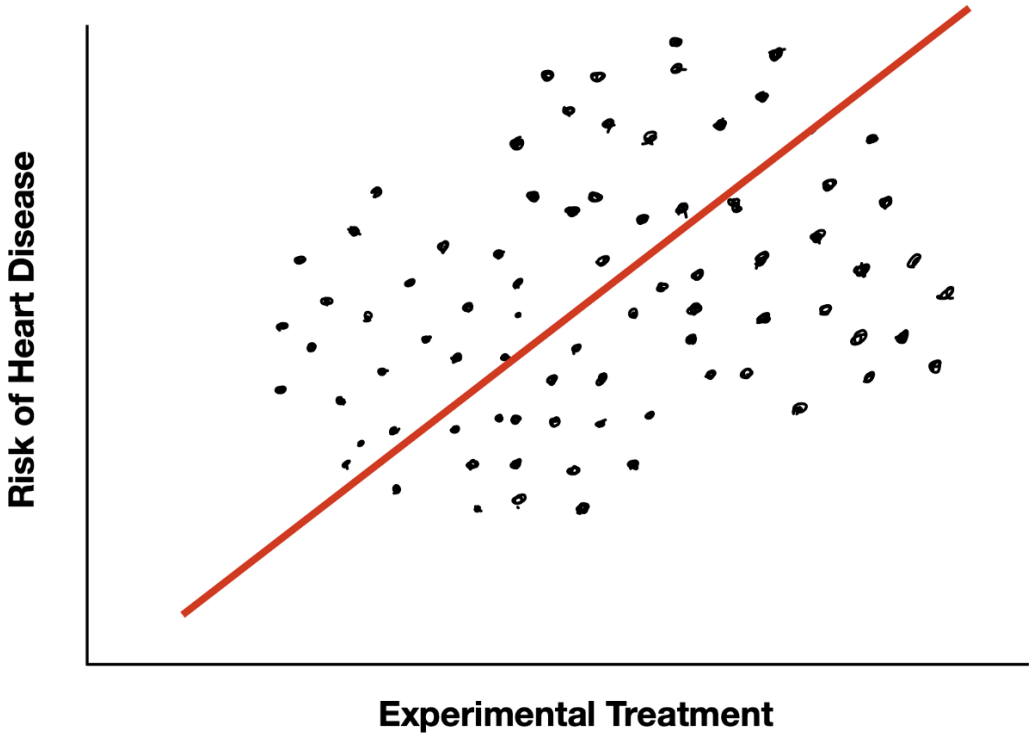
첫째, 이것은 실제로(Real world) 일어나는 것보다 훨씬 쉽게 이야기 되고
(좋은 p-값이란 무엇이고?, 얼마나 많은 데이터가 필요합니까?)

둘째, 통계적 유의성 조차도 인과 관계를 결론짓기에 충분한 것은 아닙니다.

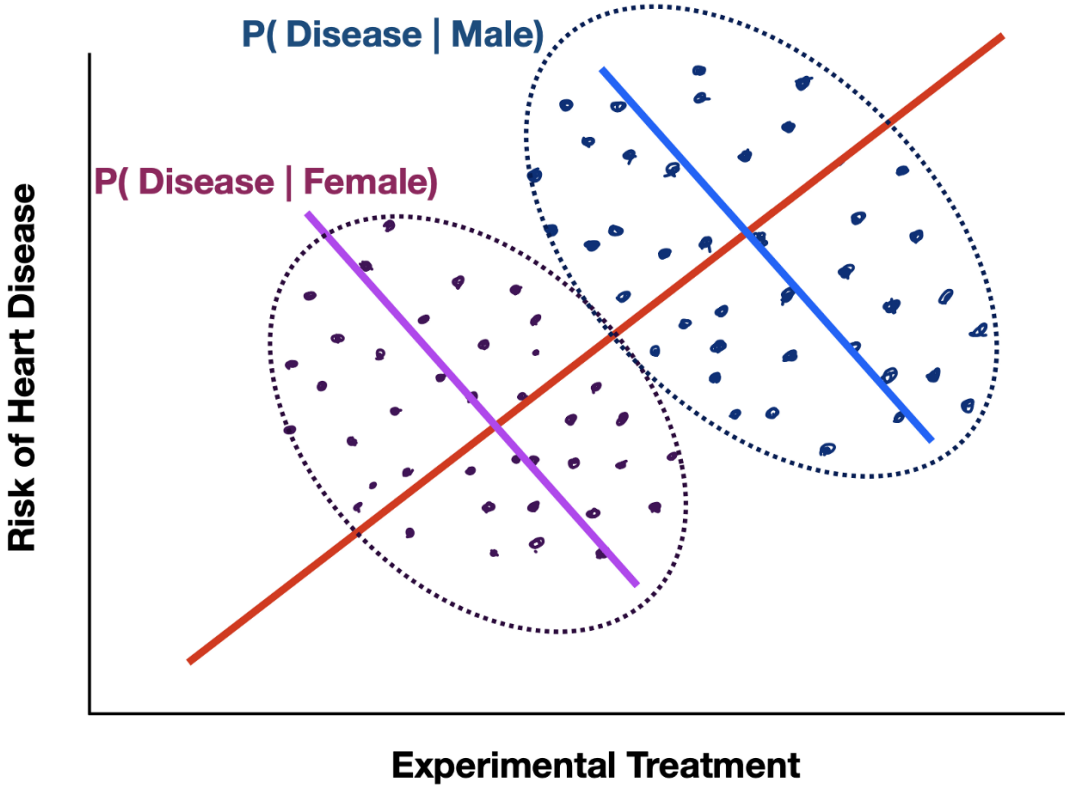
Trap 2: Simpson’s Paradox 심슨의 역설

Spurious Correlation는 잘 알려진 통계 Trap 이므로 찾기 쉽습니다. 덜 알려진 Trap은 심슨의 역설(Simpson's Paradox)인데, 이는 동일한 데이터에서 보는이의 시선에 따라 모순되는 결론이 내려지는 것을 말합니다.

심장병에 대한 실험할만한 치료법이 있다고 가정합니다. 우리는 일련의 참가자로부터 데이터를 수집하고 치료와 결과 사이의 관계를 플로팅합니다. 그런 다음 아래 그림과 같이 데이터를 플로팅합니다.



언뜻 보기에 이것은 끔찍한 치료법이라고 결론지을 수 있습니다. 누군가가 약을 더 많이 복용하거나 정해진 행동을 하면 할수록 심장병의 위험이 더 커집니다. 그러나 이제 아래 그림과 같이 연구 참가자의 두 하위 집단을 본다고 가정합니다.



각 하위 모집단을 별도로 고려하면 이전과 반대의 결과를 얻습니다. 각 그룹, 즉 남성과 여성에서 치료 증가는 심장 질환 위험 감소와 관련이 있습니다. 즉, "여성에게는 좋지만 사람에게는 나쁜 치료법이 있습니다." 라고 요약할 수 있습니다.

현실에서 다른 예를 보겠습니다. 아래 그림에는 1995년과 1996년 데릭 지터와 데이비드 저스티스의 타율이 나와 있습니다. 두 해를 따로 떼어 놓고 보면 데이비드 저스티스가 더 나은 성과를 보였습니다. 그러나 2년을 합치면 Derek Jeter가 더 나은 성과를 보였습니다. 혼란스럽죠?

Batter \ Year	1995		1996		Combined	
Derek Jeter	12/48	.250	183/582	.314	195/630	**.310**
David Justice	104/411	**.253**	45/140	**.321**	149/551	.270

Simpson의 역설은 데이터를 보는 방식이 중요하다는 점을 강조합니다.

따라서 문제는 데이터를 어떻게 나누어 보냐는 것입니다. 통계에는 이에 대한 표준 방법이 없지만 인과 추론은 이 문제를 처리하는 formulism을 제공합니다. 그것은 모두 적절한 교란 요인(틀리게 만드는 요인)을 조정한 후 변수가 다른 변수에 미치는 영향을 정량화하는 인과 관계로 요약됩니다.

Trap 3: Symmetry 대칭, 균형

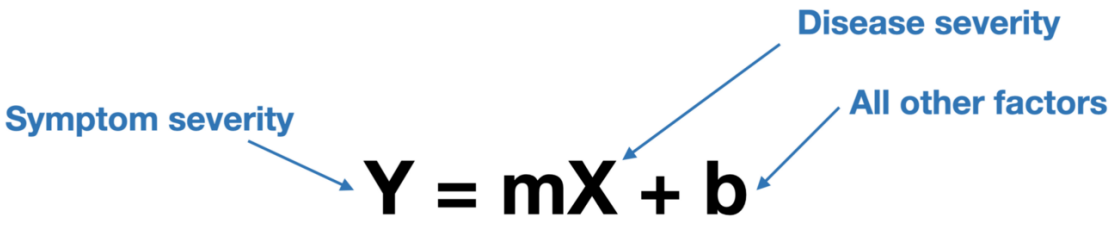
인과 관계에 대해 생각할 때 전통적인 통계의 문제는 대수학의 기본 속성인 대칭에서 비롯됩니다

방정식의 좌변은 우변과 같습니다(algebra point). 등호는 대칭을 의미합니다. 그러나 인과 관계는 근본적으로 비대칭입니다. 즉, 원인이 결과를 낳고 그 반대는 아닙니다.

간단한 예를 살펴보겠습니다.

아래 표현을 사용하여 질병과 질병이 유발하는 증상 간의 관계를 모델링한다고 가정합니다.

Y는 증상의 중증도, X는 질병의 중증도, m은 둘 사이의 연관성, b는 기타 모든 요인을 나타냅니다.



대수학 규칙을 사용하여 위의 방정식을 반전하여 다음 표현식을 얻을 수 있습니다.

$$\Rightarrow X = (Y - b)/m$$

여기 문제가 있습니다.

첫 번째 방정식을 질병이 증상을 유발한다고 해석하면
두 번째 방정식은 증상이 질병을 일으키는 것으로 해석해야 합니다! ⇒ 물론 사실이 아닙니다.

데이터를 처리할 때 X 와 Y 간의 상관 관계는 Y 와 X 간의 동일한 상관 관계를 의미합니다. 이는 통계적 종속성 보다 일반적인 개념으로 더 나아가서 즉, <X 와 Y 간의 종속성 = Y와 X 간의 종속성>을 의미합니다.

이러한 사실은 오로지 데이터(Ex, ML)에만 기반한 모델을 인과 모델로 해석할 가능성을 제거합니다. 그러나 적절한 가정이 있으면 어느정도 가능성은 있을 수 있습니다.

따라서 대칭이 대수학과 통계의 핵심 속성이라면 비대칭 관계를 처리하여 인과 관계를 나타낼 수 있는 다른 공식이 필요합니다. 이를 수행하는 한 가지 방법은 다음 섹션에 설명되어 있습니다.

Causality

인과관계를 수학적으로 표현하기 전에 인과관계란 무엇인가? 라는 더 깊은 질문에 답해야 합니다. 인과 관계는 시스템의 인과 관계를 설명하기 위한 상관 관계 또는 일반적 통계적 종속성(x→y)을 넘어선 것입니다.

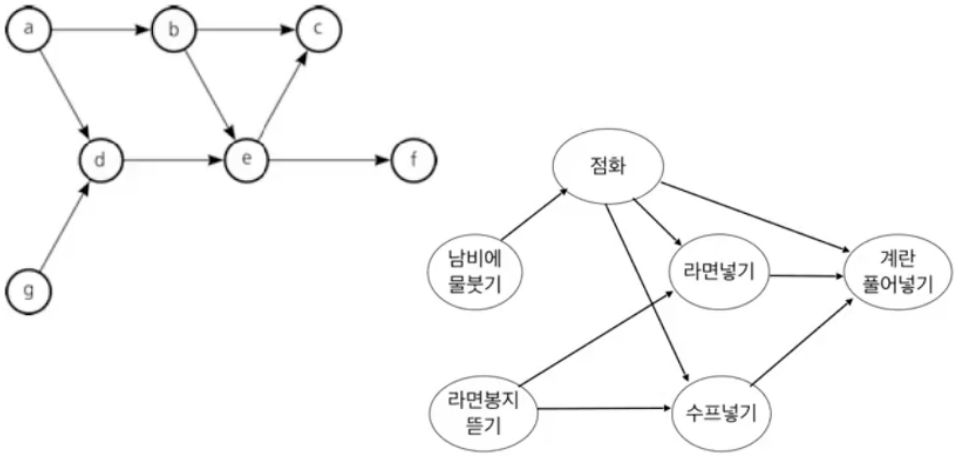
변수 X는 모든 교란 요인을 조정했을 때 다른 변수 Y를 유발한다고 할 수 있습니다. 즉, 교란 요인을 통제했을때 X에 대한 개입으로 Y가 변경되지만, 그렇다고해서 Y에 대한 개입이 반드시 X의 변경으로 이어지는 것은 아닙니다.

인과 관계는 SCM(구조적 인과 모델 : **Structural Causal Models**)을 통해 수학적으로 표현됩니다.

SCM의 두 가지 핵심 요소는 그래프와 방정식입니다.

보다 구체적으로, 그래프는 방향성이 있는 비순환 그래프 **Directed Acyclic Graphs** (DAG)이고 방정식 세트는 SEM **Structural Equation Model (**구조화 방정식 모델)입니다.

Directed Acyclic Graphs

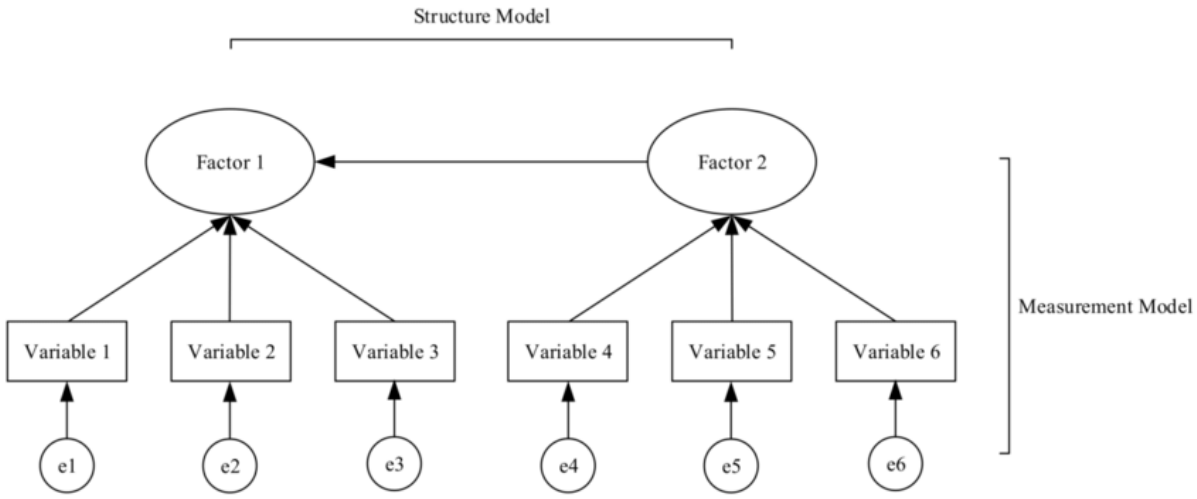


DAG는 누가 누구의 말을 듣는지, 또는 더 중요하게는 누가 누구의 말을 듣지 않는지 보여줌으로써 인과 구조를 나타냅니다.

DAG는 모든 edges가 방향을 띄고((information flow is in one direction) 순환이 존재하지 않는(vertex_꼭짓점을 떠나는 정보가 꼭짓점으로 돌아갈 수 없습니다.) 특수한 종류의 그래프입니다.

인과 관계 DAG의 꼭짓점(원)은 변수를 나타내고 가장자리(화살표)는 인과 관계를 나타냅니다. 여기서 변수는 부모에 의해 직접 발생합니다.

Structural Equation Model



SEM은 변수 간의 관계를 나타냅니다. 이 방정식에는 두 가지 특징이 있습니다.

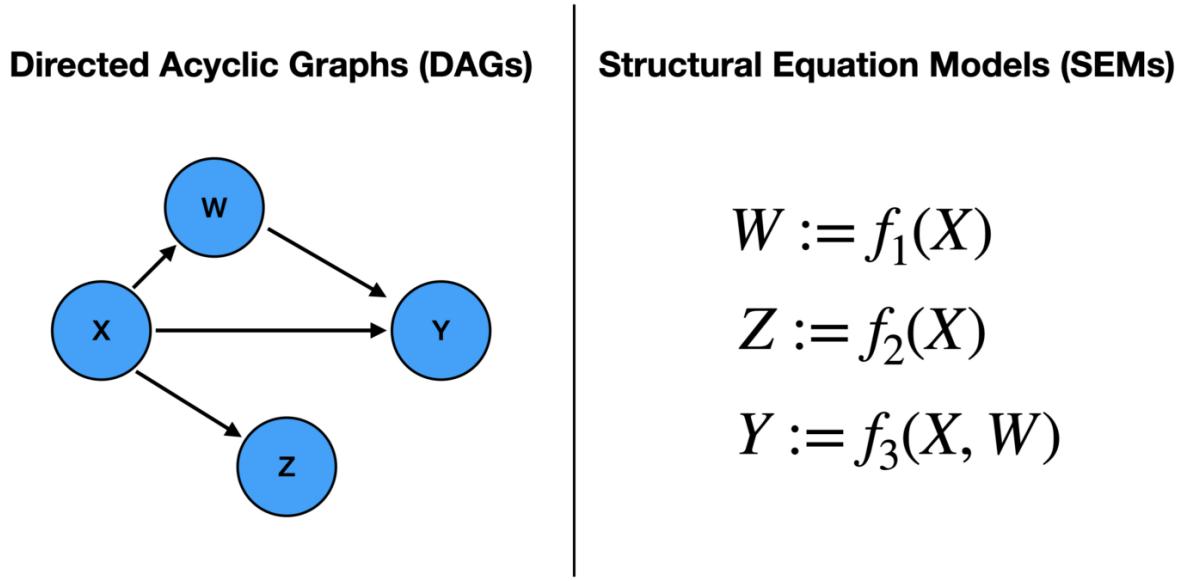
첫째, 방정식은 비대칭(성립 x) 이므로 방정식이 한 방향으로만 작동합니다. 이것은 SEM이 SEM 방정식을 유도하기 위해 반전될 수 없음을 의미합니다.

둘째, 방정식은 비 모수적일 수 있으므로 functional form을 알 수 없습니다.

SEM은 액면 그대로 DAG보다 더 많은 정보를 포함할 수 있습니다. DAG는 인과 관계의 개요를 설명하는 반면 SEM은 인과 관계와 관계의 세부 정보를 표시할 수 있습니다.

이 말만 듣고보면 DAG가 불필요해 보일 수 있지만, SEM을 통해 인과관계가 명확하지 않은 것들을 DAG를 통해 명확하게 보는 것은 엄청난 가치가 있기때문에 DAG는 중요성을 가집니다.

DAG 묘사의 선택은궁극적으로 인과관계를 평가하기 위한 d-separation의 방법을 사용한 것입니다.



Example structural causal model (SCM)

Causal Inference

이제 공식을 가지고 수학적으로 인과관계를 분석하는 방향으로 나아가 보도록 하겠습니다. 인과 추론의 목표는 문제의 인과 구조를 기반으로 질문에 답하는 것입니다.

인과 추론의 출발점은 인과 모델입니다. 즉, 최소한 어떤 변수가 서로를 듣고 말하는지 알아야 합니다. 예를 들어, X→ Y를 유발한다는 것은 알 수 있지만 상호 작용의 세부 사항은 알 수 없다는 것이 있습니다.

과적 추론에는 인과적 효과 추정 (**causal effects)**, do-calculus 사용, 혼란스러움 깨부수기 (breaking down **confounding)**와 같은 많은 측면이 있습니다.

Causal Discovery

인과 추론에서는 문제의 인과 구조가 추정되는 경우가 많습니다.즉, 상황을 나타내는 DAG를 추정해봅니다.

하지만 실제로는 시스템의 인과 관계를 알 수 없는 경우가 많습니다. 인과 관계 발견은 관찰 데이터에서 인과 구조를 밝히는 것을 목표로 합니다.

본질적으로 인과관계 발견은 다른, 어쩌면 반대라고 볼 수 있는 문제입니다. 기본 전략은 가정을 통해 가능한 솔루션의 범위를 좁히는 것입니다.

Conclusion

많은 기반이 마련되었지만, 인과관계는 아직 몇 가지 미해결 문제가 있는 젊은 분야입니다.

머신 러닝과 마찬가지로 인과 관계 또한 지속적으로 발전하는 분야입니다. 당연히 단일 블로그 게시물 또는 일련의 블로그 게시물에서 빠르게 성장하는 이 분야를 포괄적으로 설명할 수 있는 방법은 없습니다.

그러나 제 목표는 도움이 되는 몇 가지 핵심 아이디어와 리소스를 공유하는 것입니다.

이를 위해서 다음 몇 개의 게시물에서 인과 관계의 두 가지 큰 아이디어, 즉 인과 관계 추론(**causal inference)**과 인과 관계 발견(**causal discovery**)에 대해 논의할 것입니다.

참고

Causal Inference
https://towardsdatascience.com/causal-inference-962ae97cefda

Causal Inference : Primer (2019-06-01 잔디콘)
https://www.slideshare.net/lumiamitie/causal-inference-primer-20190601

참고 유튜브
Causality: An Introduction | Shawhin Talebi