1. 조건부 확률

조건부 확률 P(A|B)는 사건 B가 일어난 상황에서 사건 A가 발생할 확률을 의미한다.

$$P(A \cap B) = P(B)P(A|B)$$

베이즈 정리는 조건부확률을 이용해 정보를 갱신하는 방법을 알려준다. 사건 A가 일어났을 때 B가 일어날 확률을 아래와 같이 계산할 수 있다.

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

먼저 용어정리를 해보자. D 데이터는 새로 관찰하는 데이터, 세타는 모수라고 생각할 수 있다. 사후 확률이란 데이터가 주어졌을 때 이 파라미터가 성립할 확률을 의미한다. 사후확률인 이유는 데이터를 이미 관찰한 이후에 측정을 하는 확률이기 때문이다. 사전 확률은 데이터를 분석하기 전에 미리 가정해 놓는 것이다.

$$P(\theta|\mathcal{D}) = P(\theta)\frac{P(\mathcal{D}|\theta)}{P(\mathcal{D})}$$

$P(\mathcal{D}|\theta)$ → 가능도 (likelihood)

$P(\theta|\mathcal{D})$ → 사후확률 (posterior)

$P(\theta)$ → 사전확률 (prior)

$P(\mathcal{D})$ → Evidence

예제: COVID-99의 발병률이 10%로 알려져 있다. COVID-99에 실제로 걸렸을 때 검진될 확률은 99%, 실제로 걸리지 않았을 때 오검진될 확률이 1%라고 할 때, 어떤 사람이 질병에 걸렸다고 검진결과가 나왔을 때 정말로 COVID-99에 감염되었을 확률은?

$$P(\mathcal{D}|\theta) = 0.99$$
$$P(\theta) = 0.1 \quad P(\mathcal{D}|\neg\theta) = 0.01$$

$$P(\mathcal{D}) = \sum_{\theta} P(\mathcal{D}|\theta)P(\theta) = 0.99 \times 0.1 + 0.01 \times 0.9 = 0.108$$
$$P(\theta|\mathcal{D}) = 0.1 \times \frac{0.99}{0.108} \approx 0.916$$

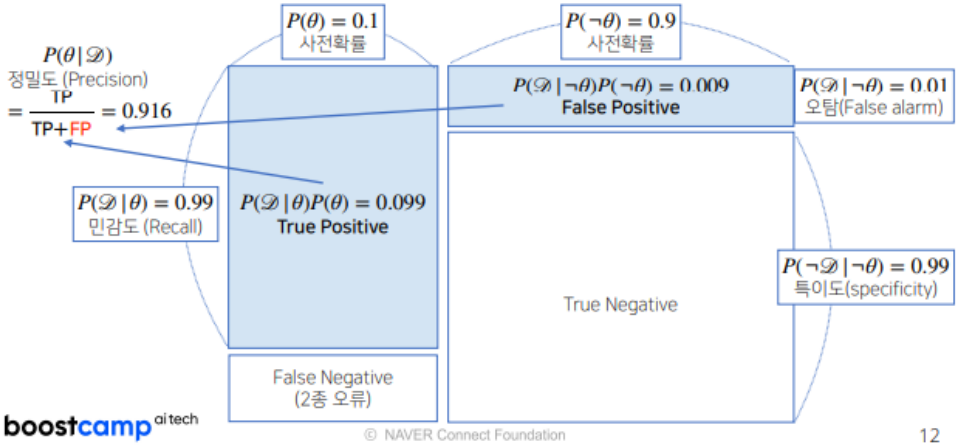
만약 오검진될 확률이 10%라고 할 때 확률은 어떻게 될까?

$$P(\mathcal{D}) = \sum_{\theta} P(\mathcal{D}|\theta)P(\theta) = 0.99 \times 0.1 + 0.1 \times 0.9 = 0.189$$
$$P(\theta|\mathcal{D}) = 0.1 \times \frac{0.99}{0.189} \approx 0.524$$

오답률(False alarm)이 오르면 테스트의 정밀도(Precision)가 떨어진다.

아래는 표는 조건부 확률을 시각화한 표이다. 양성이 나왔을 때 실제 병이 관찰될 확률을 True Positive, 음성이 나왔을 때 실제로 질병이 걸리지 않은 경우는 True Negative라고 하자. 반대로 양성이 나왔을 때 병에 걸리지 않은 경우는 False Positive라고 부르고, 1종 오류라고 부른다. 그리고 음성이 나왔는데 병에 걸린 경우는 False Negative라고하고 2종 오류라고 부른다.

데이터 분석의 성격에 따라 1종 오류를 줄이는게 중요한지 2종 오류를 줄이는게 중요한지 결정한다.



2. 베이즈 정리를 통한 정보 갱신

베이즈 정리를 통해 새로운 데이터가 들어왔을 때 **앞서 계산한 사후확률을 사전확률로 사용해 갱신된 사후확률을 계산할 수 있다.**

$$P(\theta|\mathcal{D}) = P(\theta)\frac{P(\mathcal{D}|\theta)}{P(\mathcal{D})}$$

사후확률 (posterior)

$$P(\theta|\mathcal{D}) = P(\theta)\frac{P(\mathcal{D}|\theta)}{P(\mathcal{D})}$$

갱신된 사후확률 (posterior)

앞서 COVID-99 판정을 받은 사람이 두 번째 검진을 받았을 때도 양성이 나왔을 때 진짜 COVID-99에 걸렸을 확률을 구해보자. (오검진 확율이 10%인 사례)

$$P(\theta|\mathcal{D}) = 0.1 \times \frac{0.99}{0.189} \approx 0.524$$

$$P(\mathcal{D}|\theta) = 0.99$$
$$P(\mathcal{D}|\neg\theta) = 0.1$$

$$P(\mathcal{D}^*) = 0.99 \times 0.524 + 0.1 \times 0.476 \approx 0.566$$

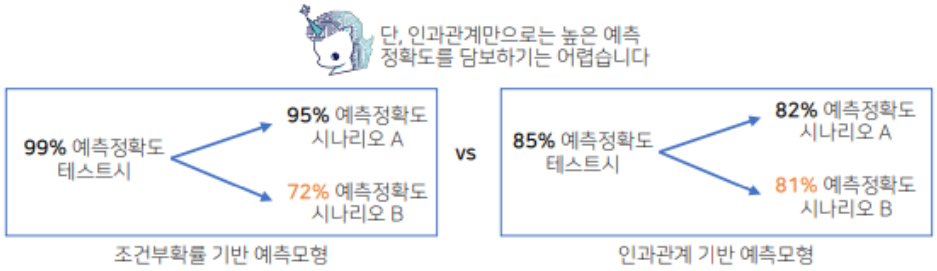
갱신된 사후확률 (posterior)

$$P(\theta|\mathcal{D}^*) = 0.524 \times \frac{0.99}{0.566} \approx 0.917$$

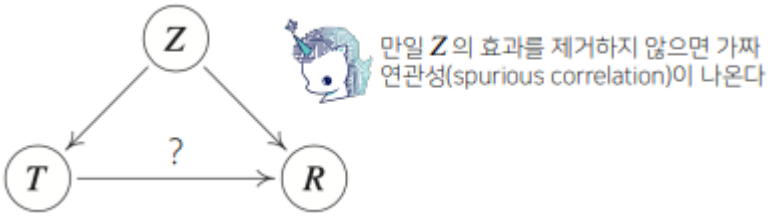
3. 조건부 확률 -> 인과관계

조건부 확률은 유용한 통계적 해석을 제공하지만 인과관계(causality)를 추론할 때 함부로 사용해서는 안 된다. 예를 들어, A가 B의 원인이라는 해석은 하면 안 된다. 데이터가 많아져도 조건부 확률만 가지고 인과관계 추론은 불가능하다.

인과관계는 데이터 분포의 변화에 강건한 예측모형을 만들 때 필요하다. 인과 관계를 고려하지 않고 조건부 확률 기반의 예측 모형을 만들면 테스트에는 높은 예측을 가지는 모델을 만들 수 있다. 하지만 데이터의 유입이 바뀌거나 새로운 치료법을 개발했을 때 예측 확률이 크게 변할 수 있다. 하지만 인과 관계를 유입하면 높은 예측 정확도를 조건부확률 기반만큼 담보하기는 어렵다. 하지만 데이터 분포 변화에는 강건한 모습을 보인다.

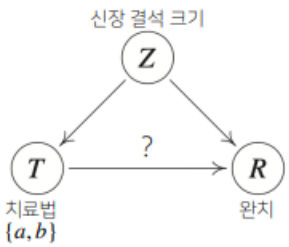


인과관계를 알아내기 위해서는 중첩요인(confounding factor)의 효과를 제거하고 원인에 해당하는 변수만의 인과관계를 계산해야 한다.



아래는 인과관계 추론 예제이다.

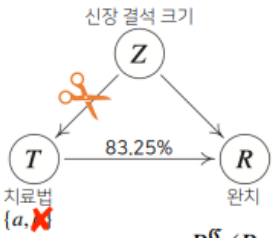
아래를 보면 치료법 a가 치료법 b보다 전체적으로 치료될 확률이 적지만 small stones, large stones 각각의 경우에 치료 확률이 높다. 이를 해결하기 위해서는 신장 결석 크기에 따른 중첩효과를 제거해야 실제 정확한 완치율 계산이 가능하다.



		$Z=0$	$Z=1$
	Overall	Patients with small stones	Patients with large stones
Treatment a: Open surgery	78% (273/350)	**93%** (81/87)	**73%** (192/263)
Treatment b: Percutaneous nephrolithotomy	**83%** (289/350)	87% (234/270)	69% (55/80)

출처: Elements of Causal Inference, Peters et al.

아래와 같이 조건부 확률로 계산한 치료효과와 달리 정반대의 결과가 나온다.



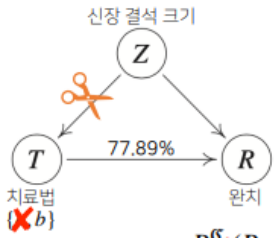
		$Z=0$	$Z=1$
	Overall	Patients with small stones	Patients with large stones
Treatment a: Open surgery	78% (273/350)	**93%** (81/87)	**73%** (192/263)
Treatment b: Percutaneous nephrolithotomy	**83%** (289/350)	87% (234/270)	69% (55/80)

출처: Elements of Causal Inference, Peters et al.

$$P^{\mathfrak{G}_a}(R=1) = \sum_{z\in\{0,1\}} P^{\mathfrak{G}}(R=1 \mid T=a, Z=z)P^{\mathfrak{G}}(Z=z)$$

do$(T=a)$ 라는 조정(intervention) 효과를 통해 Z의 개입을 제거한다

$$= \frac{81}{87} \times \frac{(87+270)}{700} + \frac{192}{263} \times \frac{(263+80)}{700} \approx 0.8325$$



		$Z=0$	$Z=1$
	Overall	Patients with small stones	Patients with large stones
Treatment a: Open surgery	78% (273/350)	**93%** (81/87)	**73%** (192/263)
Treatment b: Percutaneous nephrolithotomy	**83%** (289/350)	87% (234/270)	69% (55/80)

출처: Elements of Causal Inference, Peters et al.

$$P^{\mathfrak{G}_b}(R=1) = \sum_{z\in\{0,1\}} P^{\mathfrak{G}}(R=1 \mid T=b, Z=z)P^{\mathfrak{G}}(Z=z)$$

조건부확률로 계산한 치료효과와 정반대의 결과가 나오게 된다

$$= \frac{234}{270} \times \frac{(87+270)}{700} + \frac{55}{80} \times \frac{(263+80)}{700} \approx 0.7789$$