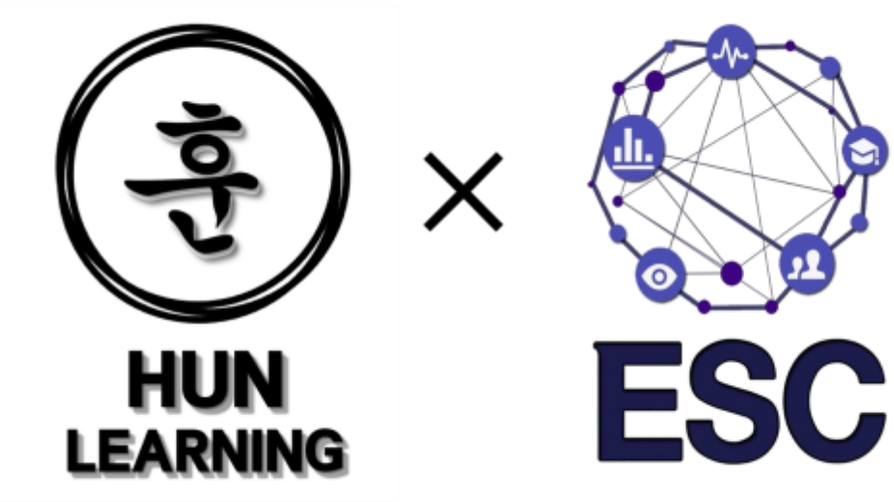
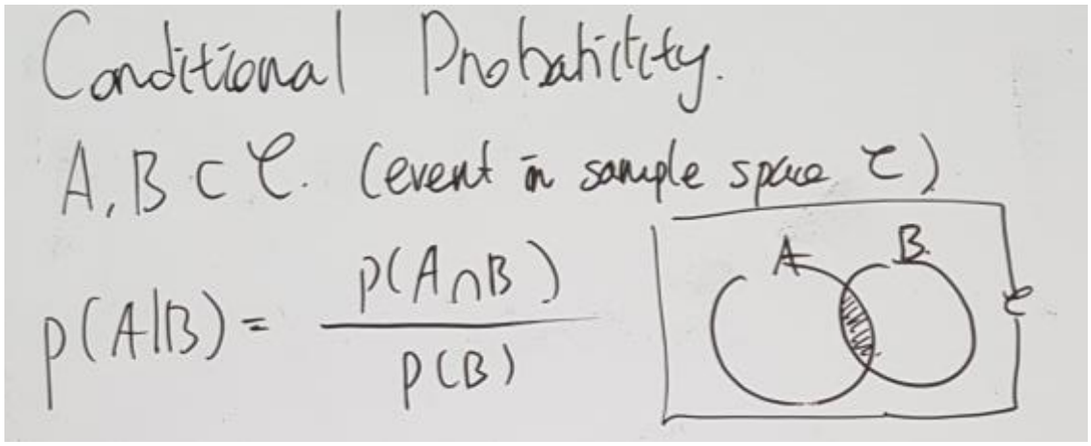
PROBABILITY DENSITY, LIKELIHOOD AND BAYES RULE

왜 베이즈 추론은 지극히 자연스러울 수 밖에 없는가?

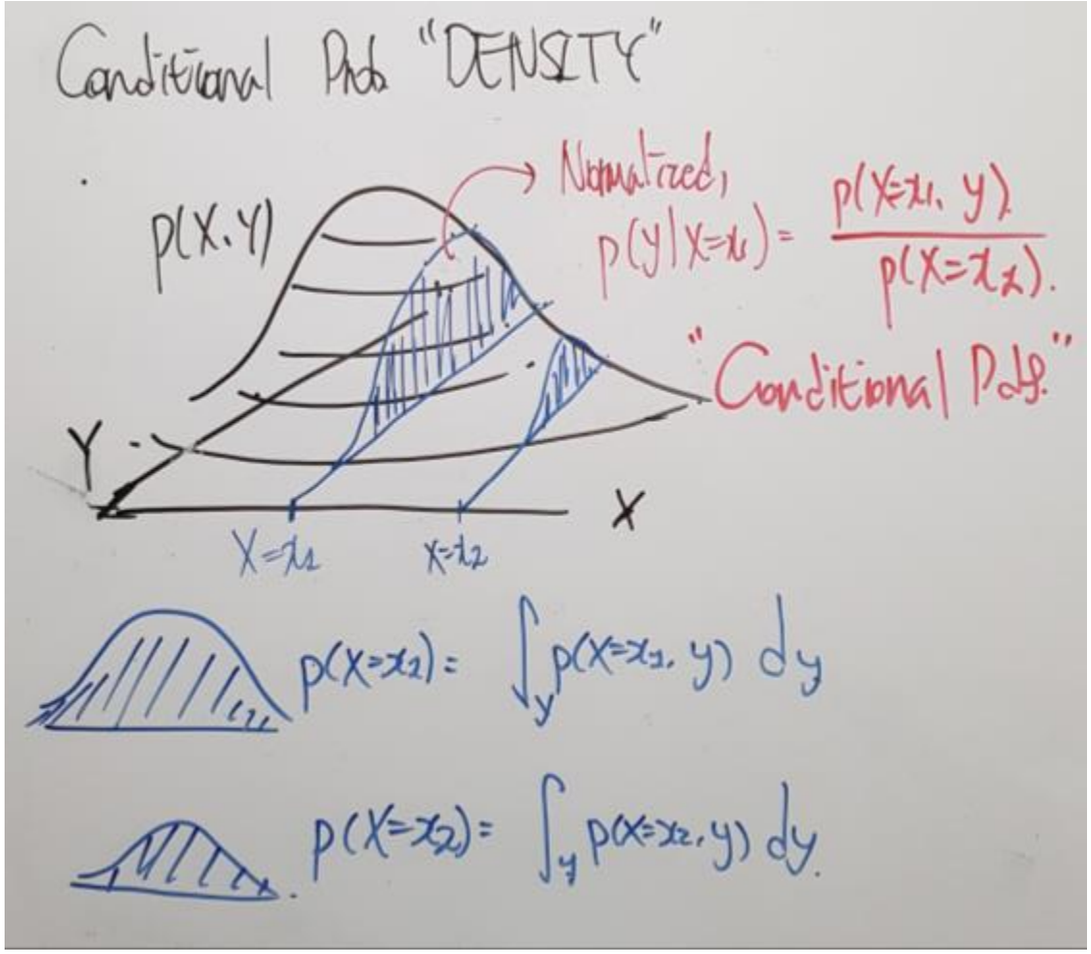


베이즈를 공부하기 이전에…

@hun_learning



일단 조건부확률과
조건부분포의 차이를 짚고
넘어가자. 조건부확률은 말
그대로 한 사건이 이미
일어났을 때 (주어졌을 때)
다른 사건이 발생할
확률이다. $p(B)$는 확률.



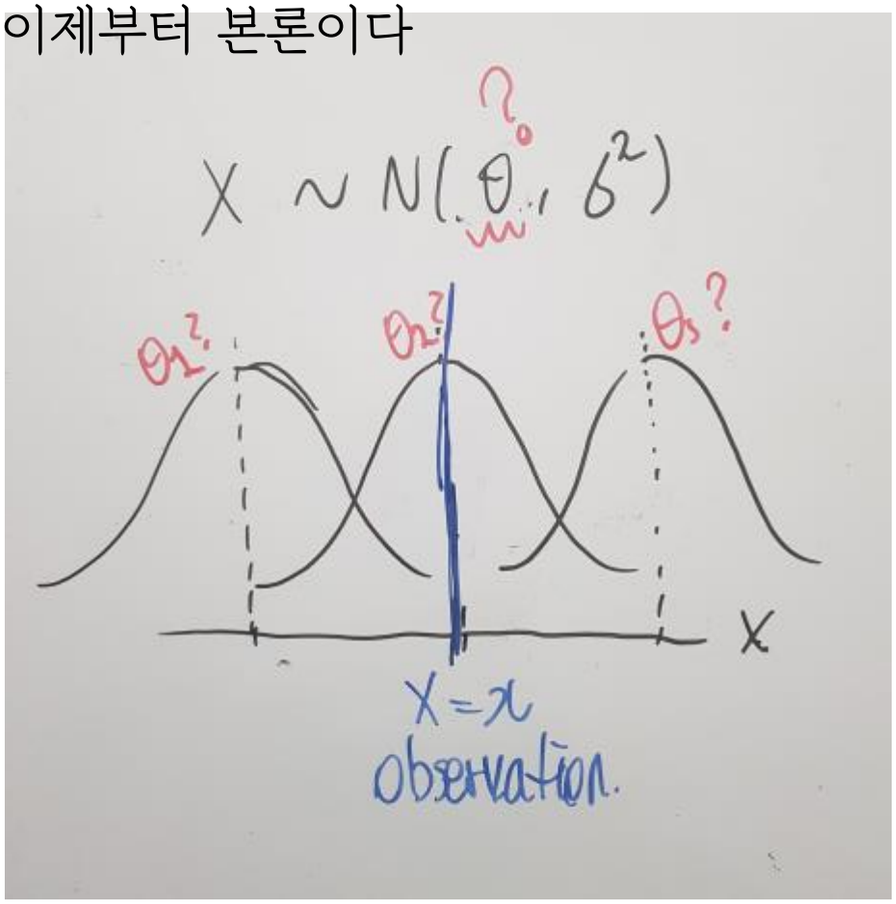
두 x, y 연속형 확률변수에서
조건부**분포**는

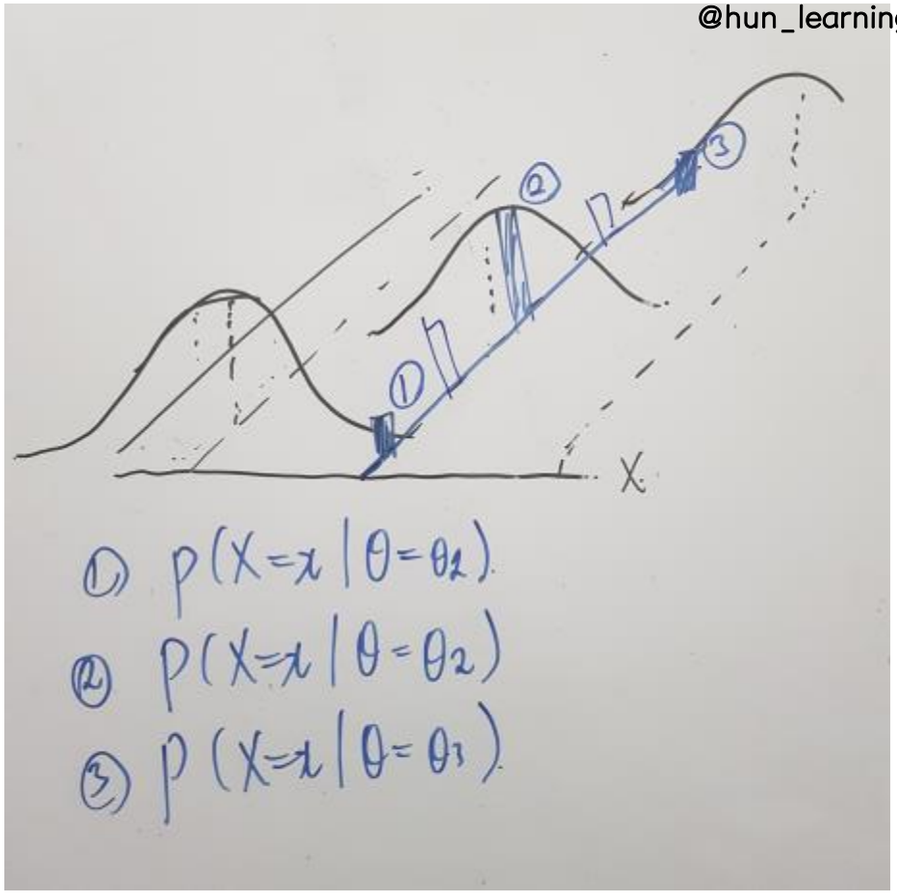
$$f(Y \mid X = x) = \frac{f(X = x, Y)}{f(X = x)}$$

인데, **여기서 $f(X = x)$는
확률이 아니다!**

그냥 적분값인데, 이 값은
**$f(X, Y)$라는 둔덕을, $X = x$라는
지점에서 칼로 잘랐을 때
나오는 단면의 넓이이다.**

잘라낸 단면 자체는 적분해서
1이 안된다. 때문에 잘라낸
단면의 분포를 보고 싶을때
넓이만큼 나눠서 분포로
만들어주는거고 (Normalize!!),
이게 바로 조건부 분포의
의미다!



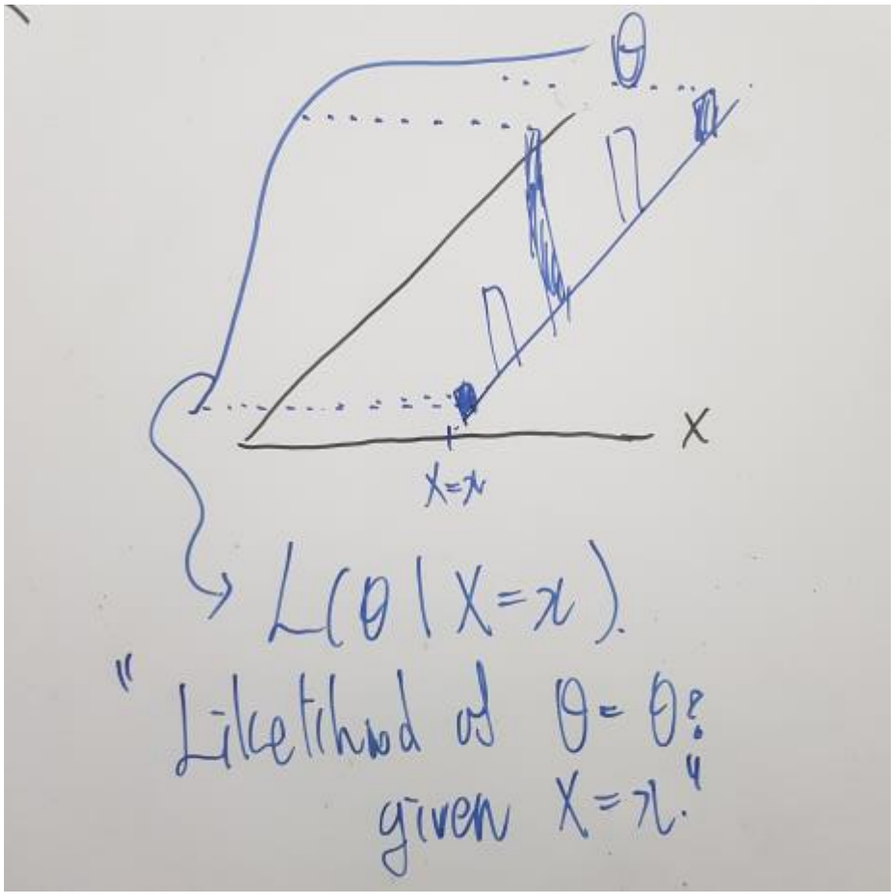


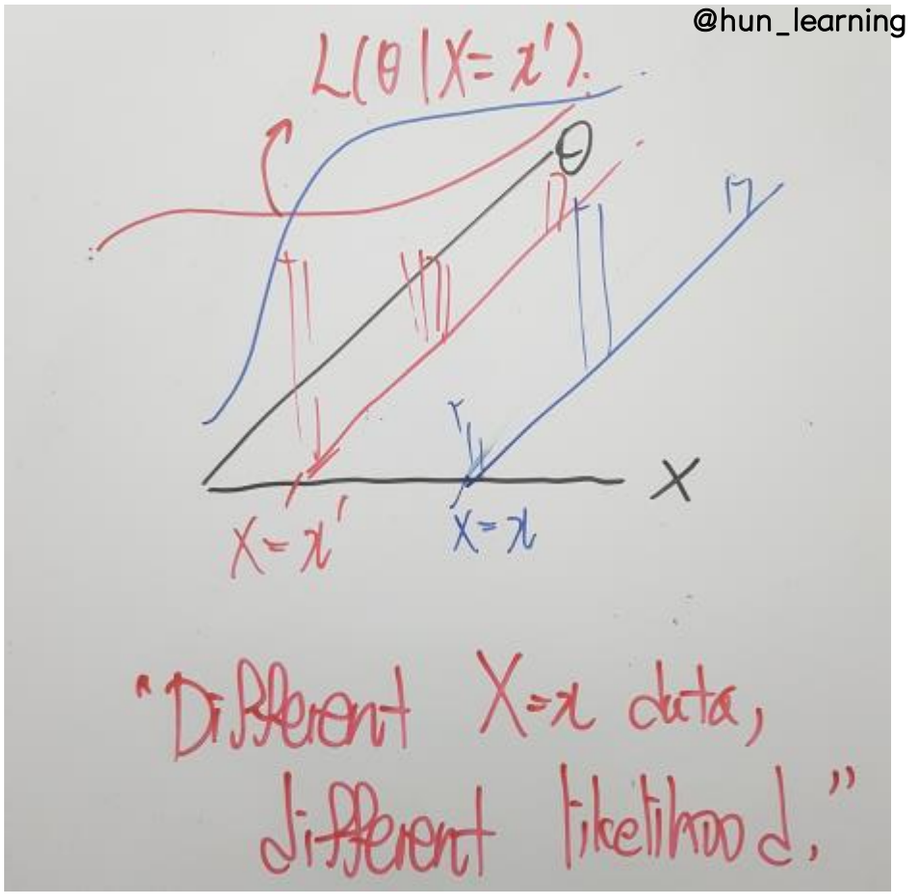
우리가 관측하는 데이터 x의 분포가, 분산을 아는 정규분포라고 하자. 우리의 관심은 모평균 θ이다. 정규분포의 pdf는 잘 안다. 그러나, 모수를 모른다면 pdf는 아무 쓰잘데기가 없다! 어떻게 그릴건데!! $\theta_1?\ \theta_2?\ \theta_3?$ 뭘까?

일단 각 모수별로 pdf $p(X \mid \theta = ?)$를 다 그려보면 왼쪽 그림이 나온다. 우리의 관측치는 $X = x$로 표시되어있다. 왼쪽 그림처럼 다 겹쳐서보면 보기 힘드니까, 오른쪽 그림처럼 입체적으로 생각해보자.

오른쪽 그림을 보면, 우리의 데이터 $X = x$를 기준으로, 각각의 모수에서의 이 데이터에 해당하는 density $p(X = x \mid \theta = ?)$ 가 보인다. *(파란색 막대, 연속형 확률변수니까 함수값 자체가 확률은 아니다.)* 직관적으로 생각해보면 θ가 데이터에 가까울 때 데이터가 나올 확률이 높으니, θ_2에 해당하는 pdf에서 데이터의 density가 가장 높다.

이처럼 각 모수에서의 데이터의 density를 한 눈에 보는 방법은 없을까?





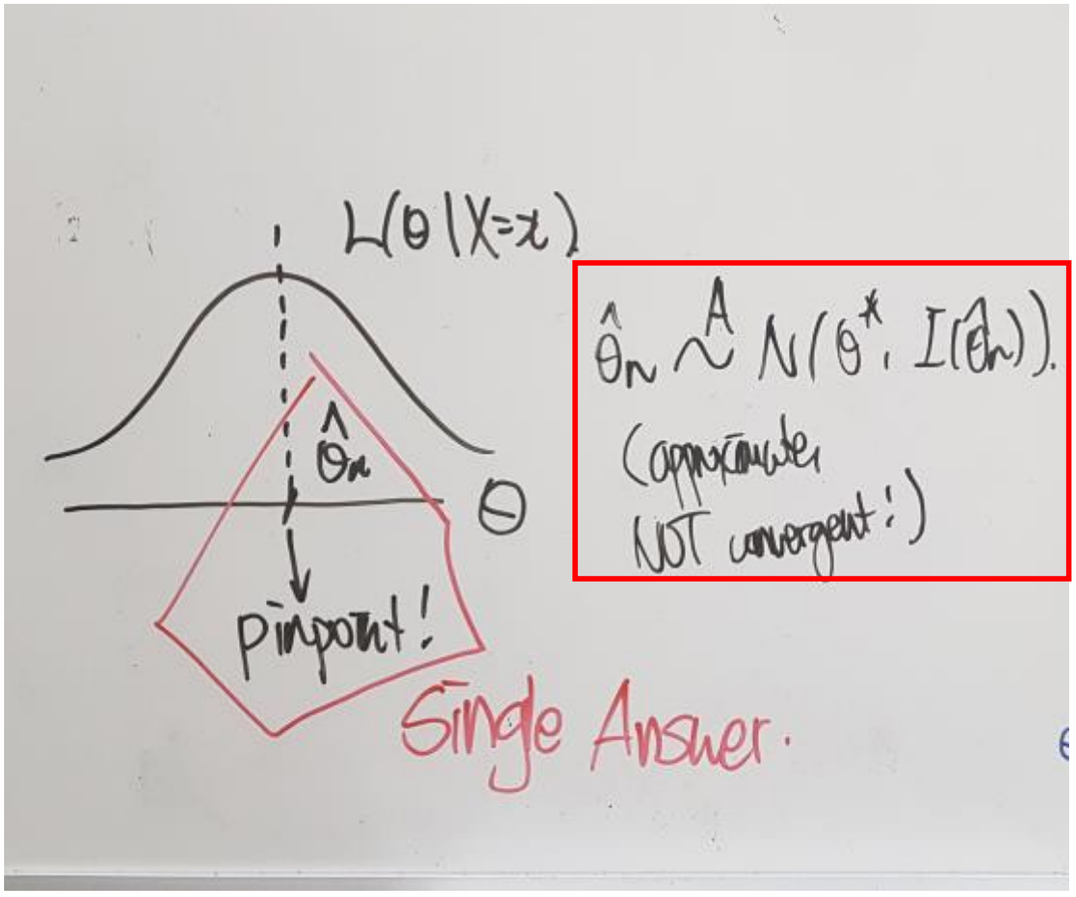
이처럼 각 모수에서의 데이터의 density를 한 눈에 보는 방법은 없을까?

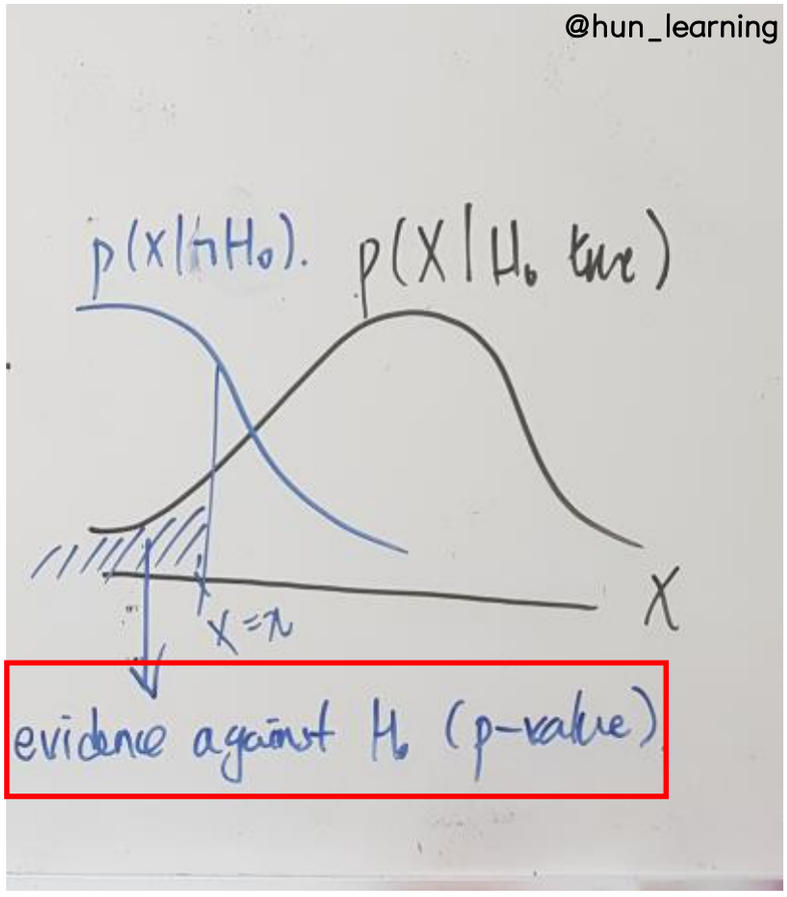
주어진 데이터 $X = x$에 대해 각 모수별로 그려진 파란색 막대를 하나의 축에 모두 그려보자. 그려면 우리는 파란색 선으로 그려진 Likelihood를 얻는다.

$$PDF \qquad p(X = ? \mid \theta\ known)$$
$$Likelihood \quad p(X = x\ given \mid \theta = ?)$$

모수 θ에 의해 결정되는 똑같은 함수 $p(X;\theta)$에 대해서, 만일 모수를 알고 데이터를 모르면 이게 probability density function이 되는거고, 아까 보았듯이 X축에 대하여 그려지는 것이다. 반면에 데이터를 알고 모수를 모른다면, 위 왼쪽 그림에서처럼 파란색으로 θ에 대해서 그려지는 Likelihood 함수가 되는 것이다.

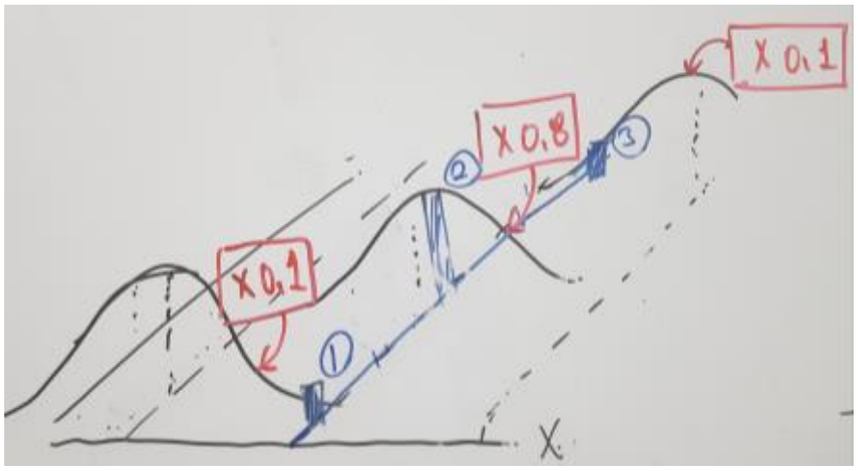
이때 어떤 데이터를 관측하냐에 따라 그려지는 Likelihood 함수는 다르다!
그러니까 Likelihood는 관측치에 따라 다르게 그려지는 거다!!

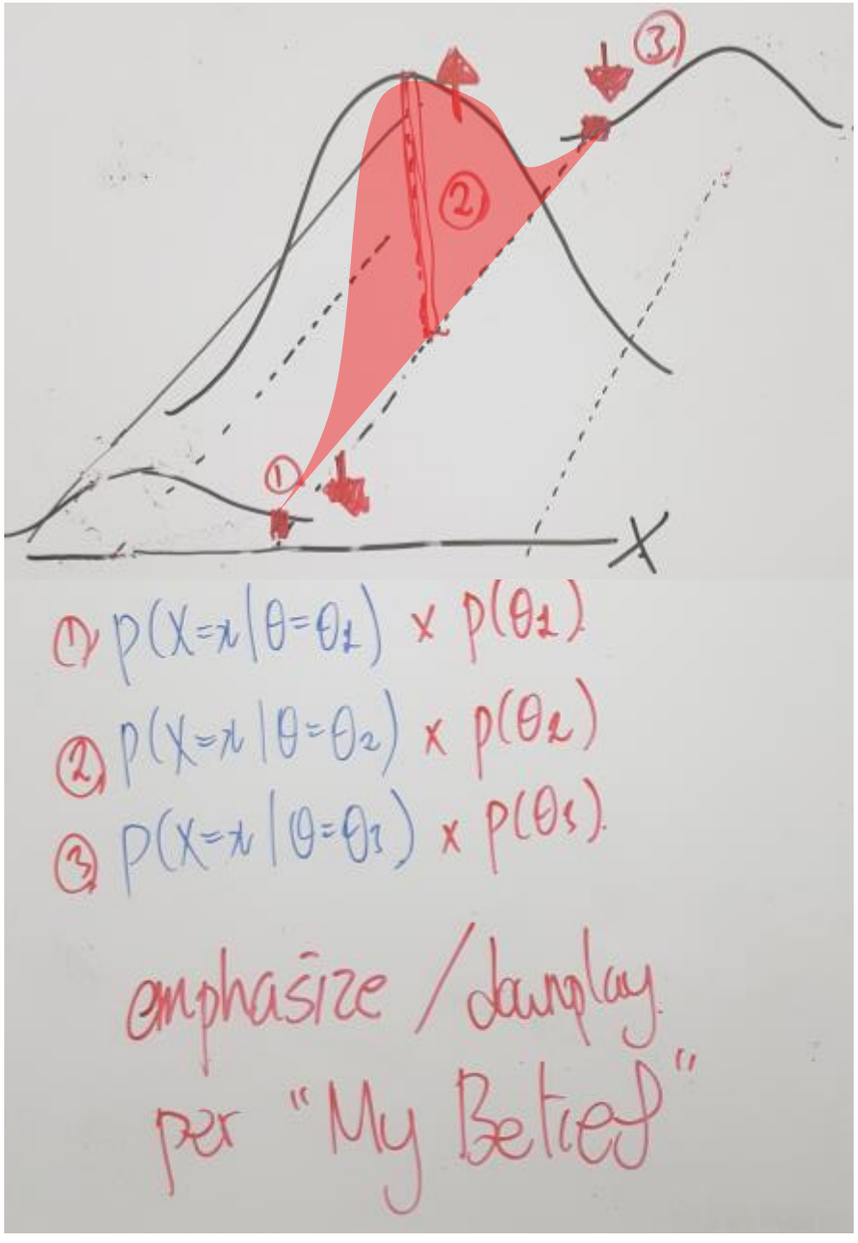




우리가 익숙한 Frequentist 관점에서의 통계적 추론은 바로 이 Likelihood 함수를 이용하여 이뤄진다.

1. Likelihood 함수가 최대가 되는 지점 $\hat{\theta}$ 딱 하나(MLE)를 콕 집는다!
   주어진 데이터에 견주어 봤을 때 가장 그럴듯한 모수(pdf)니까!!
2. 추정량 $\hat{\theta}$의 분포를 구한다. (극한분포 혹은 Bootstrap 노가다)
   가장 만만하고 좋은 MLE의 극한분포는 이미 알려져있다.
   (**주의**: MLE도 결국 CLT인데, 중심극한정리는 데이터 크기가 늘어나면 표본평균이 정규분포로 분포"수렴"한다는 게 아니다!!!! 아니라고!!! 표본평균들의 분포는 그냥 모평균이라는 하나의 점으로 수렴한다. 다만 그 수렴하는 모양이 정규분포이기 때문에, 충분히 큰 데이터에 대해서는 대충 표본평균의 분포를 정규분포로 퉁치겠다는 거다. MLE도 마찬가지)
3. 추정량의 분포를 바탕으로 신뢰구간, p-value, 가설검정 등등 하고 싶은 이상한 것들 다 한다.





근데 그냥 저러케 끝내버리면 노잼 확률론을 제대로 한번 이용해보자.

각각의 모수를 일종의 "시나리오"라고 생각해보고, 각각의 시나리오가 얼마나 신빙성이 있는지, 확실한지에 대한 믿음에 따라서 가중치를 주어보자. 왼쪽 그림에서 0.1, 0.8, 0.1 이렇게.

그럼 내가 그렇게 부여한 가중치에 따라서 각각의 시나리오의 비중이 달라질거다. 내가 힘을 많이 실어준 시나리오 $p(X = x \mid \theta = \theta_2)$는 높은 $p(\theta_2)$ 덕분에 상대적으로 강조될 것이고, 나머지 시나리오는 비교적 비중이 줄어들 것이다.

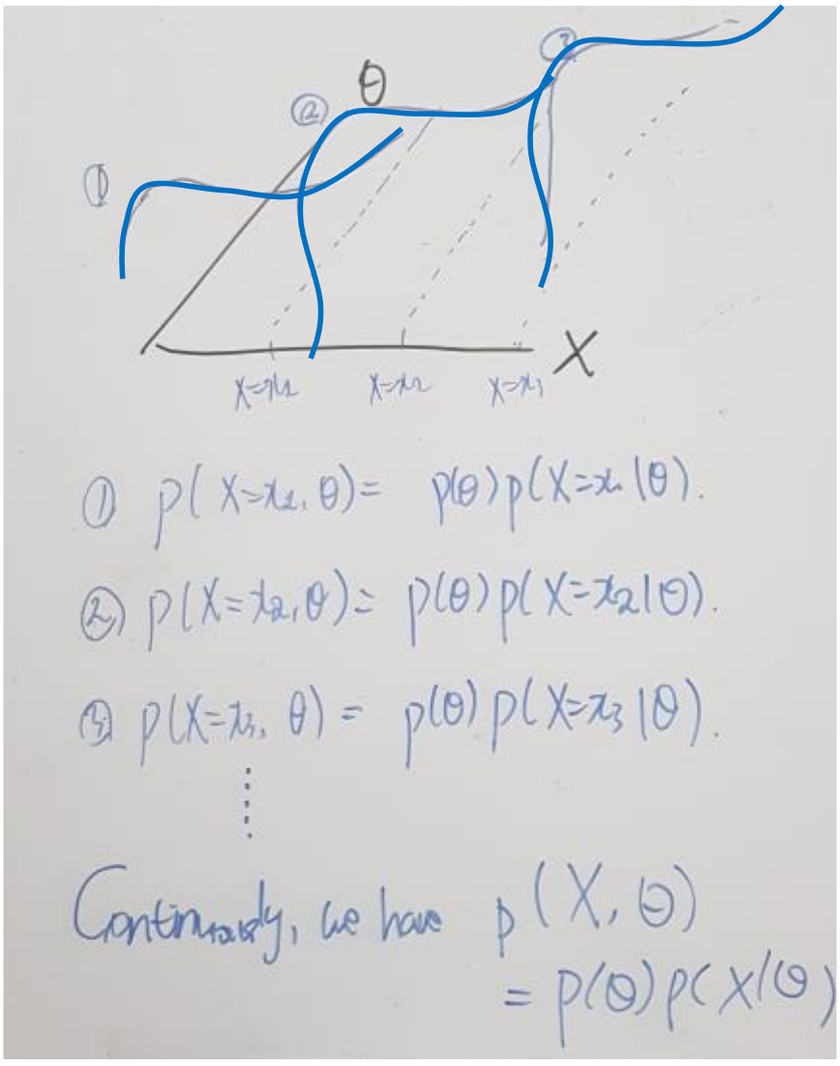
이렇게 주어진 데이터 $X = x$에 대해 가능한 각각의 시나리오들을

$$p(X = x \mid \theta)$$

라고 하고, 각 시나리오 별로 내가 주는 가중치를

$$p(\theta)$$

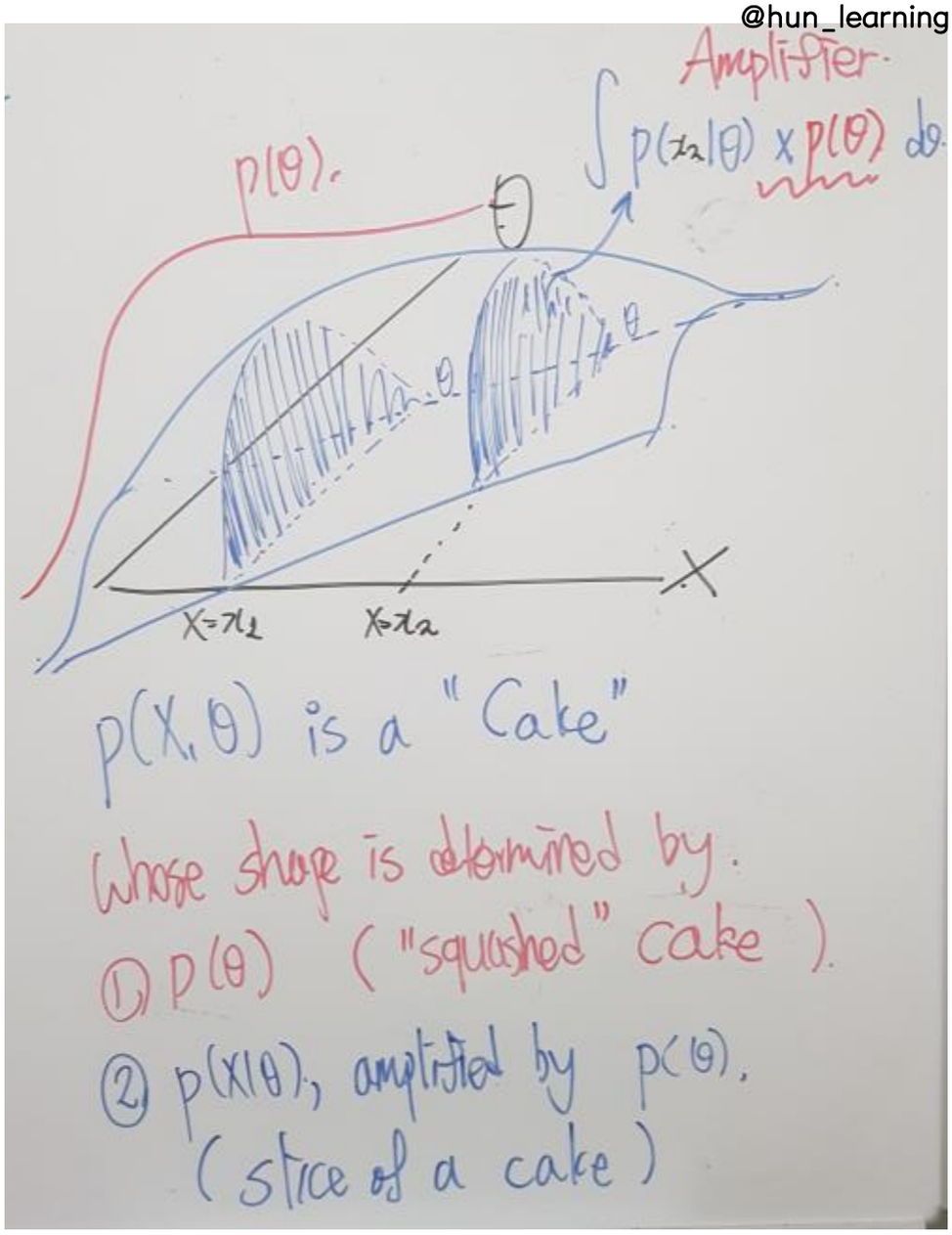
나의 믿음이라고 해보자.
이때 빨간색으로 표시된 선의 의미는 무엇일까?



우리는 믿음 $p(\theta)$와 시나리오 $p(X \mid \theta)$를 모두 정의했으니, 사실상

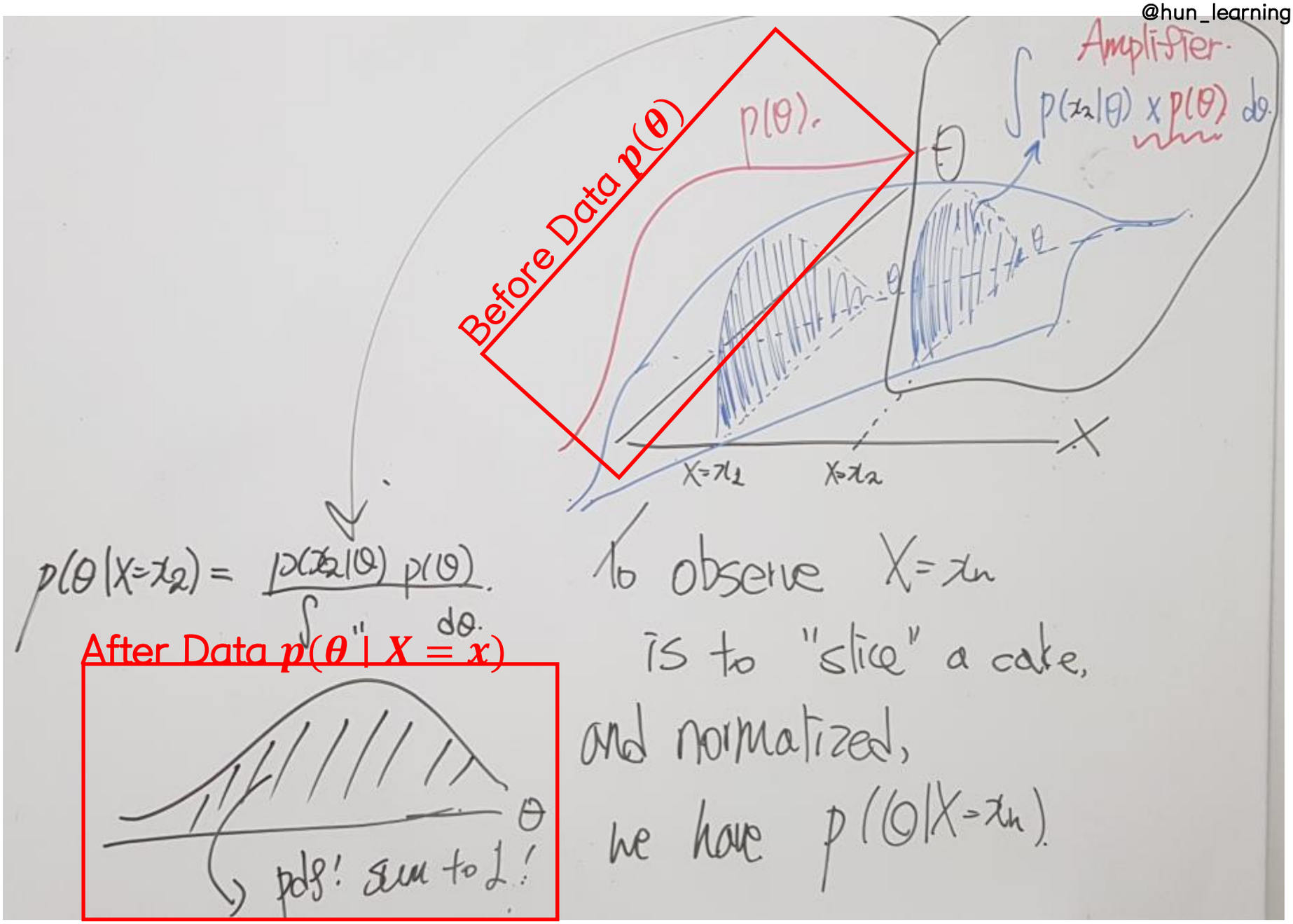
$$p(X, \theta) = p(\theta)p(X \mid \theta)$$

라는 Full Probability Model을 정의한 것과 마찬가지이다.
이 $p(X, \theta)$라는 둔덕에서도 또한, 처음 슬라이드에서 본 것처럼 조건부 분포를 구할 수 있다.
먼저 이 둔덕을 $X = x$라는 지점들을 따라 자르면 $p(X = x, \theta)$라는 θ의 식이 나오는데, 이게 파란 선들이다.



이 파란 선의 의미는, 주어진 데이터 $X = x$에 대해, 각 시나리오들 $p(X = x \mid \theta = ?)$이 나의 믿음 $p(\theta = ?)$에 따라 "확대/축소"된 것이다.

이 파란 선들도 그 자체로는 확률분포가 아니므로, Normalize를 해줘야 한다.
파란 색 영역은, 모든 시나리오에 걸쳐 주어진 데이터의 density $p(X = x)$이다.

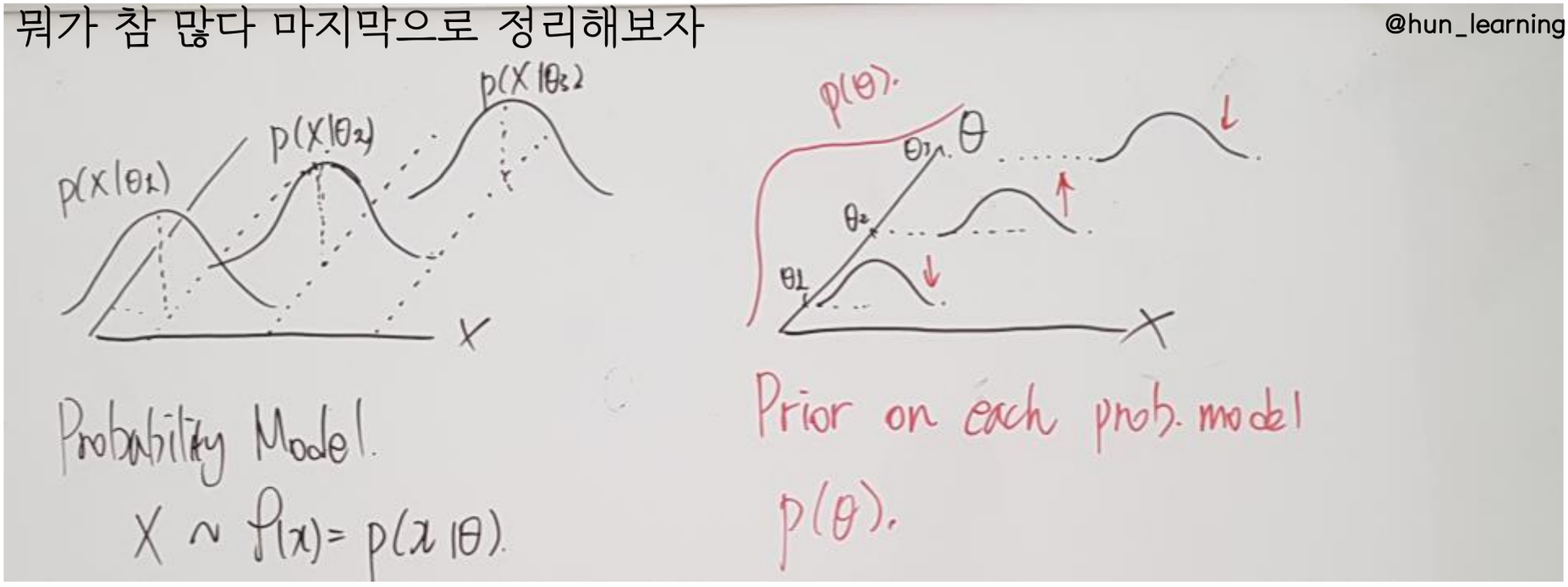


파란 선으로 그려진 선은 분포가 아니므로, Normalize를 해줘서 조건부 분포를 구해야 한다. 그 과정이 바로 Bayes Rule이다!

$$p(\theta \mid X=x) = \frac{p(X=x \mid \theta)p(\theta)}{p(X=x)} = \frac{p(X=x \mid \theta)p(\theta)}{\int p(X=x \mid \theta)p(\theta)\, d\theta}$$

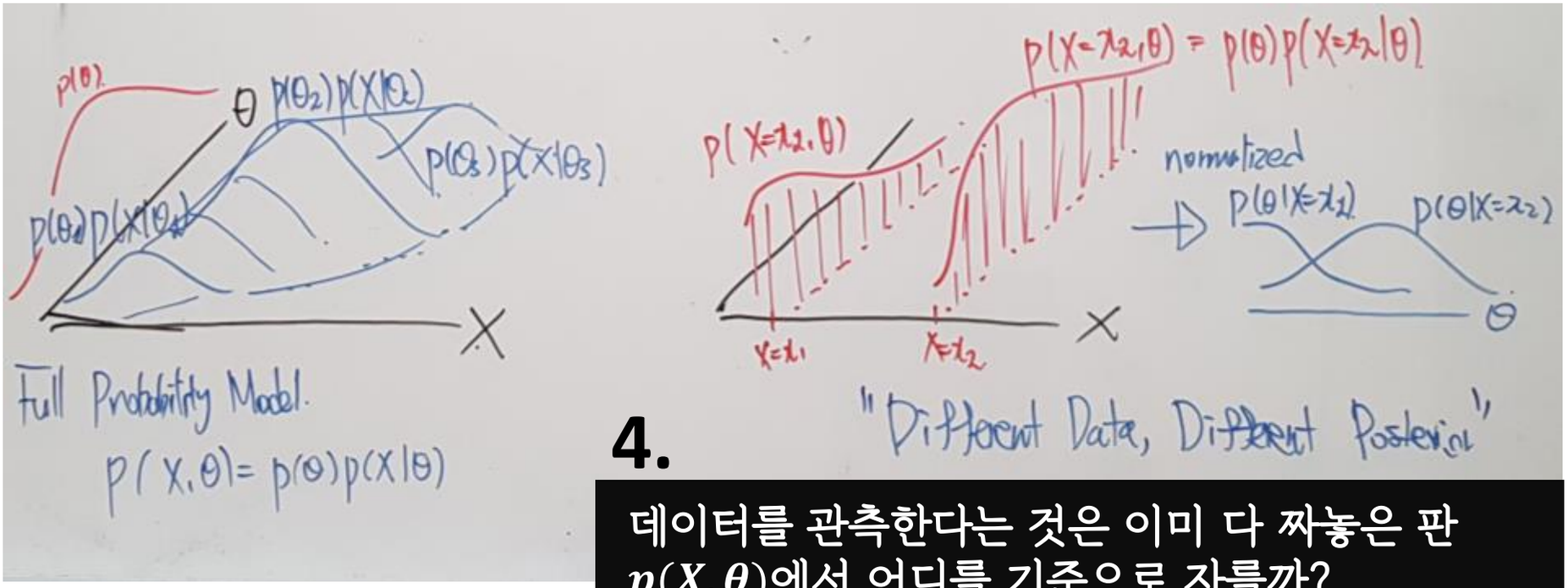
$p(\theta \mid X=x)$가 의미하는 바는? 니 믿음이 $p(\theta)$인데, 데이터가 $X=x$이네? 그럼 이걸 보고 수정된 너의 믿음이고 요놈이고, 베이즈 룰은 수정하는 방법!

뭐가 참 많다 마지막으로 정리해보자



1. 그냥 pdf임 모수 모르니까 하나로 못 그림 그래서 일단 졸라 그려봄 $p(X \mid \theta)$

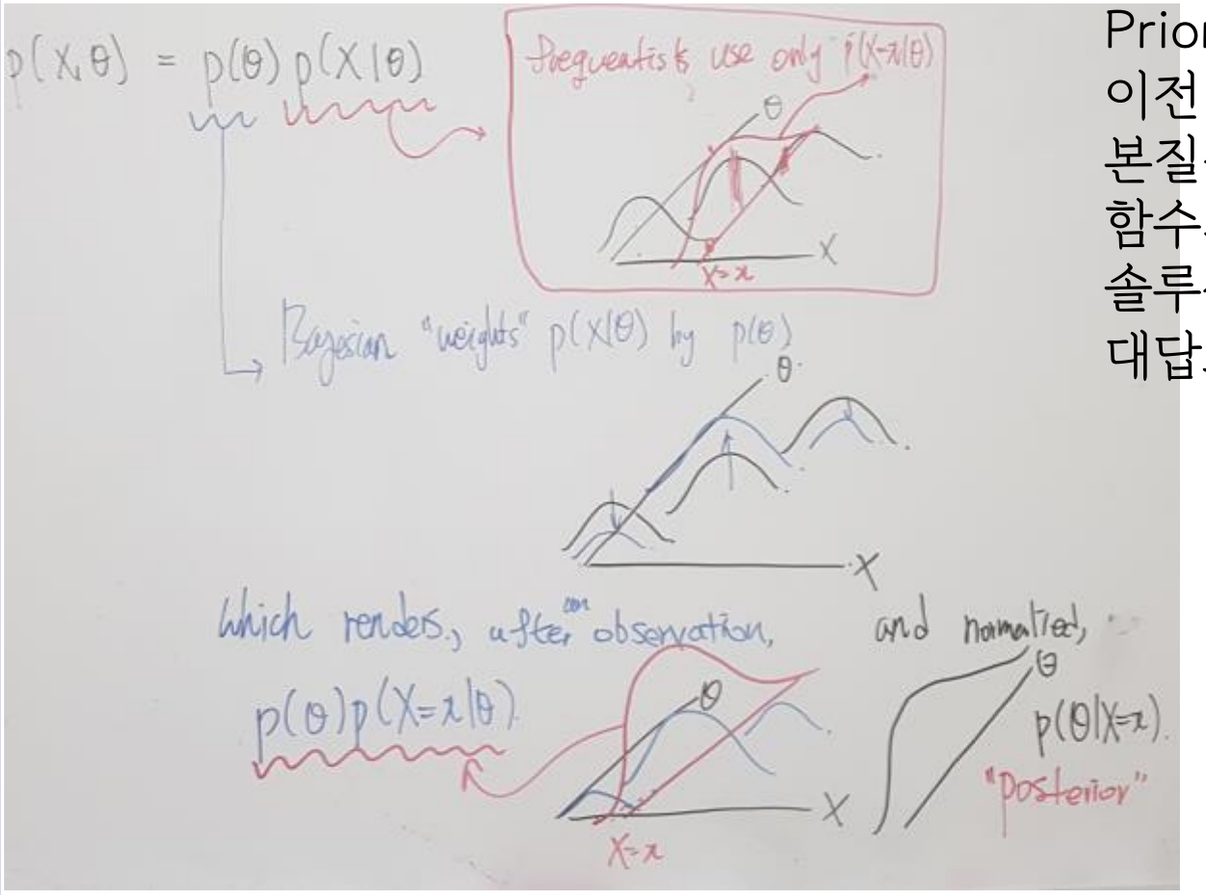
2. 각 pdf 별로 (시나리오 별로) 니 맘대로 가중치를 줘라 $p(\theta)$
이거 때매 어떤 놈은 강조되고 아니면 씹히고



3. 그럼 얼떨결에 $p(X, \theta)$ 만들어짐 아직 데이터는 관측 하지도 않았음 데이터 보기 전임

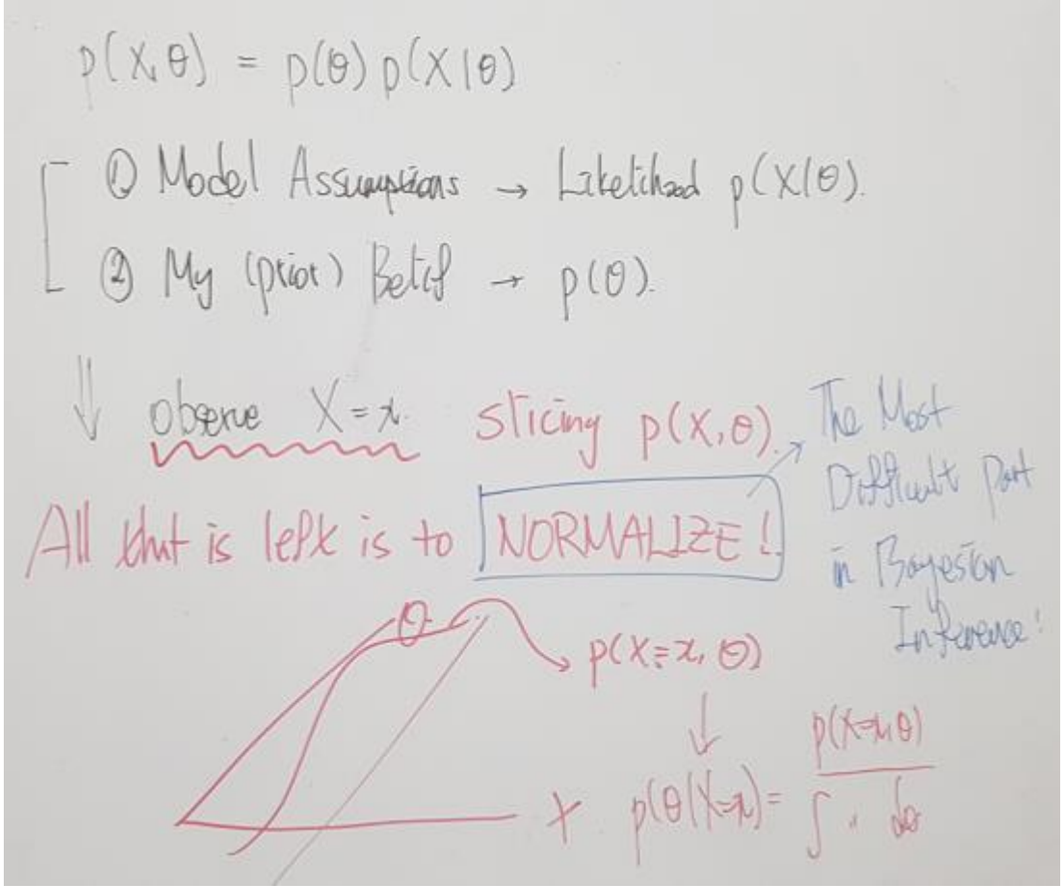
4.

데이터를 관측한다는 것은 이미 다 짜놓은 판 $p(X, \theta)$에서 어디를 기준으로 자를까?
를 보는 것에 불과함.
남은 단계는 그냥 잘라 놓은 단면을 pdf로 만들어주는 것 밖에 없음 (근데 이게 제일 어려움)



Prior $p(\theta)$를 정 쓰기 싫다면 그냥 이전 슬라이드 1. 에서 다 끝남 본질적으로 그냥 $L(\theta \mid X = x)$라는 함수의 최적화 푸는 거다. 최적화 솔루션은 딱 하나니까 모수에 대한 대답도 딱 하나의 숫자 $\hat{\theta}$로 나온다.

Prior $p(\theta)$를 쓴다는 것은 문제 해결에 있어서 확률론이라는 도구를 끝까지 사용한다는 것.
모수에 대한 대답도 분포 $p(\theta \mid X)$로 나온다.



데이터 형성 과정에 대한 가정 (iid, 노말 오차 등등…)을 수식으로 쓰면 그게 바로 Likelihood가 되는 거다.

그렇게 생긴 Likelihood에다가 냅다 최적화 풀지 말고, 모수에 대하여 Prior를 부여해서 조건부 분포를 구하는게 확률론을 이용하는 방식.

제일 빡센 부분이 바로 조건부 분포를 구하는 거다. Exactly 하게 구할 수 있는 경우는 흔하지 않아서 샘플링이나 아니면 형태를 단순화해서 근사한다. **끝! 파이팅!**