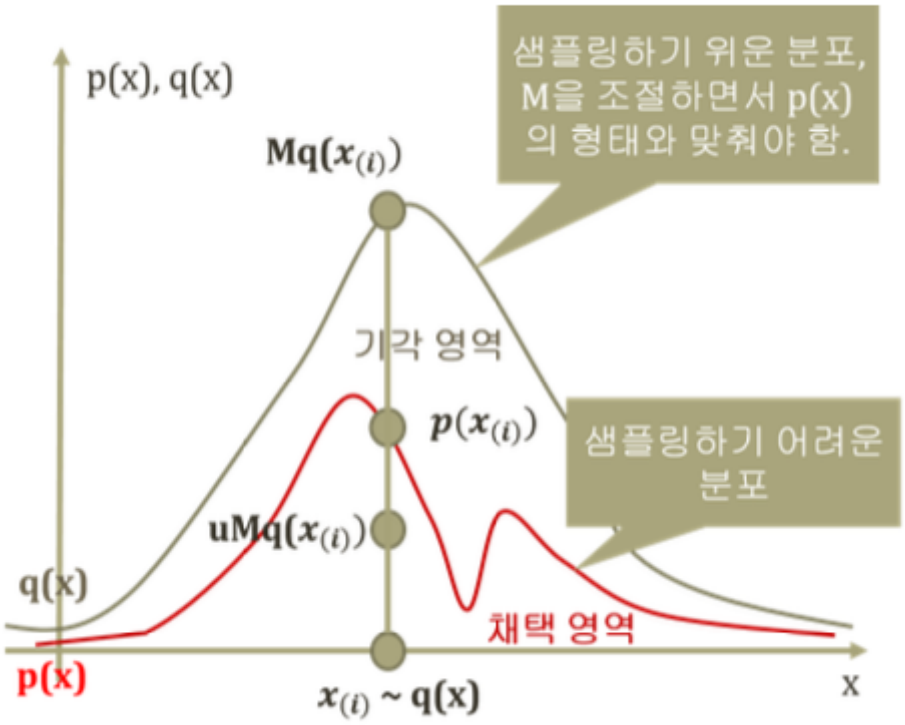
왜 샘플링이 필요한가?



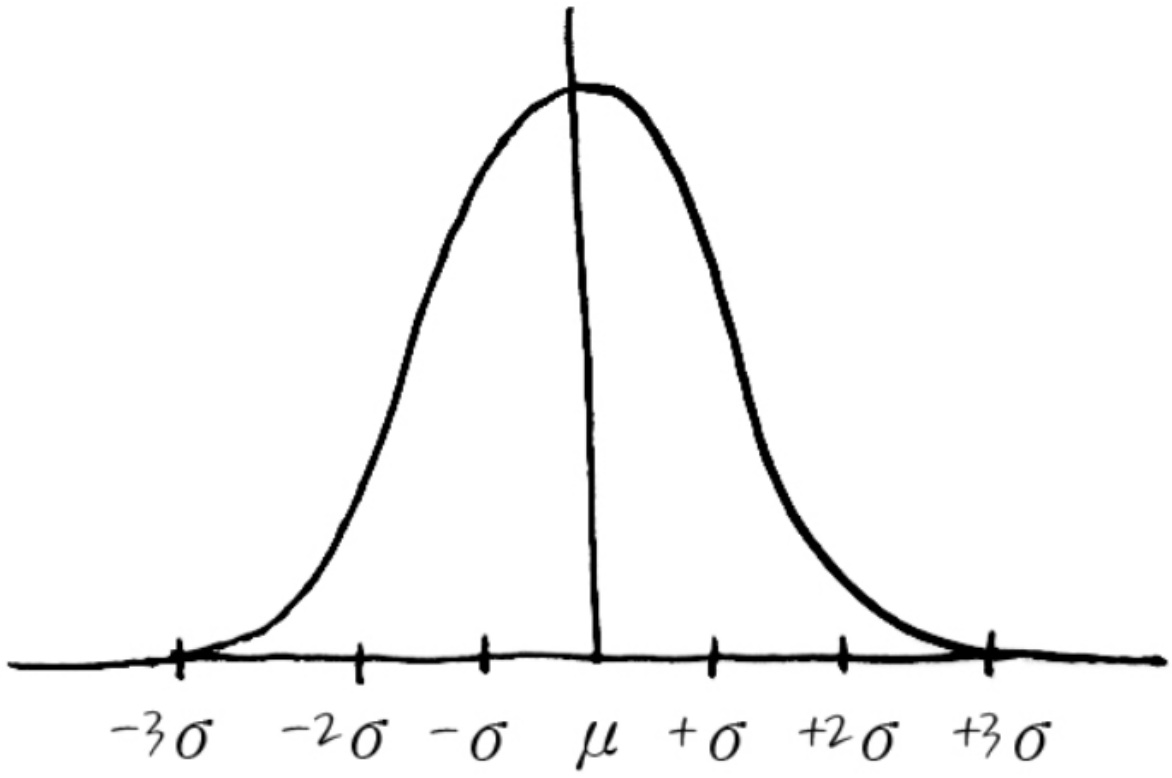
Rejection sampling 강의내용에서

- 우리는 확률분포 혹은 확률분포로 표현될 수 있는 p(x)알고 있음
- p(x)를 evoloped 하는 함수 q(x)를 정의하여 sampling을 수행

궁금증: p(x)를 아는데 굳이 따로 Sampling을 해야하나?

> ***샘플링(sampling) 이란 무엇일까?***
> 간단하게 말해, 샘플링은 어떤 자료에서 일부 값을 추출하는 것을 의미한다. 통계에서의 샘플링은 다른 말로 "표집"이라고 불린다.
>
> ***즉, 샘플링은 모집단에서 표본을 추출하는 일인 것이다.***

예시)
10000개 원소를 가진 모집단 가우시안 정규분포를 따르고 있을때 100개의 샘플로 표본집단을 만들어라.



아래와 같은 사고 흐름이 생김

- x in [-inf, inf] ~= [-10sigma, 10sigma]
- [-10sigma, 10sigma]에서 100개를 uniform하게 뽑음
- 이때 해당 분포가 모집단을 잘 표현할 수 있을까?

그 분포로 histogram(a.k.a. PMF)를 그리면 uniform distribution이 나올 것임
따라서 확률분포(빈도)를 고려하여 sampling을 수행했다고 볼 수 없음.

그럼 확률분포로 sampling을 수행한다는 것은 무슨 의미일까?

$$y = f(x)$$

위와 같은 분포를 알고 있을때 우리는 y 즉, $f(x)$를 알고 있음.
따라서 x를 알고 싶다는 것은 $f^{-1}(y)$를 알고 있다는 것과 동치
하지만 모든 분포 아니, 대부분의 분포가 역함수가 존재하지 않음

Rejection sampling

Rejection Sampling은 우리가 Target function의 PDF는 알고 있지만, 그 함수에서 직접 샘플링 하는것이 매우 어렵거나 불가능할때(항상 어려움) 효율적으로 샘플링 하기 위해 사용되는 방법이다.

예시)
아래와 같은 함수를 알고 있을때,

$$f(x) = 0.3\exp(-0.2x^2) + 0.7\exp(-0.2(x-10)^2)$$

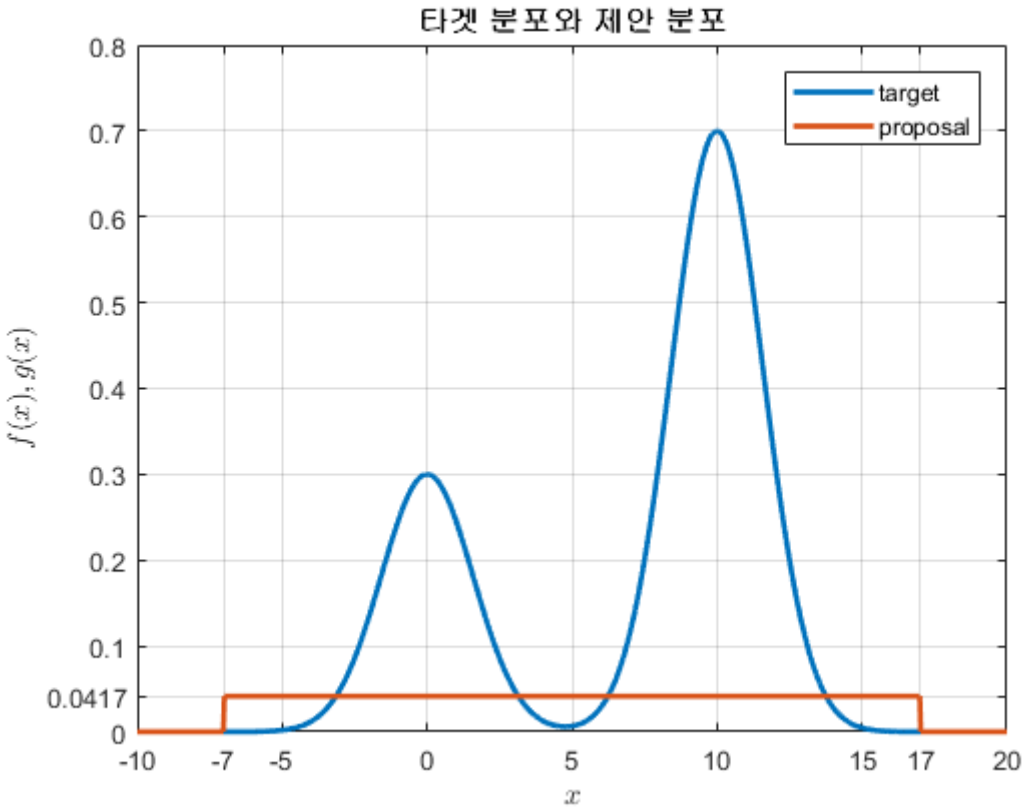
해당 분포를 target distrubtion이라고 부름

Step 1. 제안분포

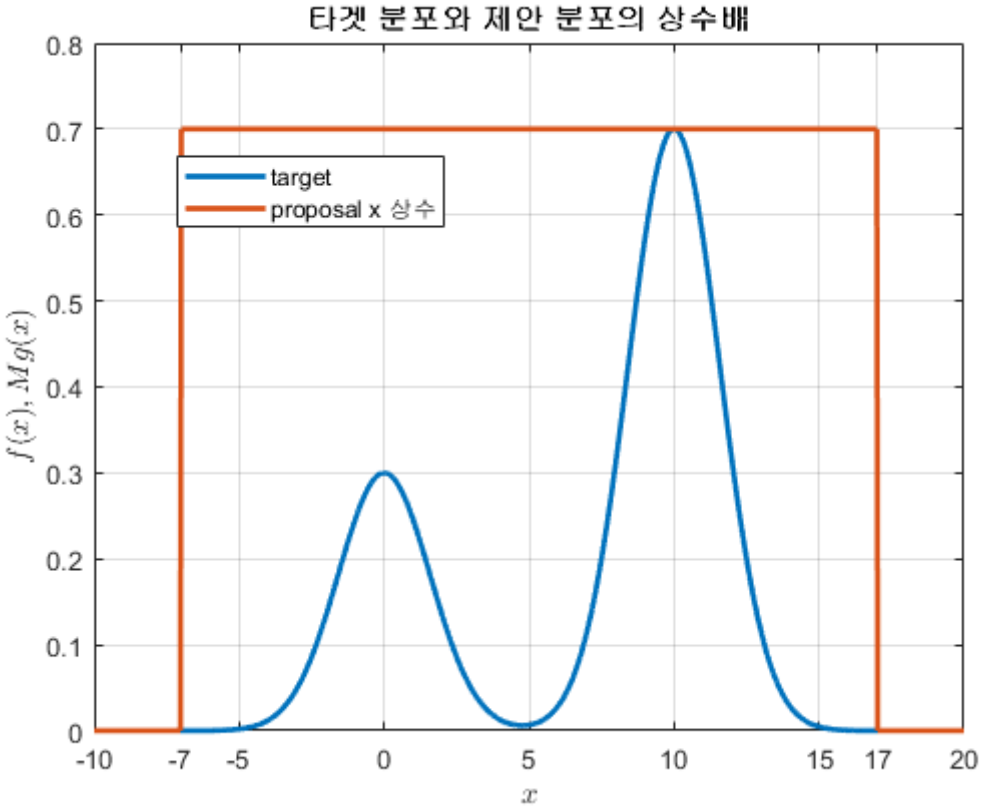
처음 우리가 쉽게 분포를 생성할 수 있는 분포를 가정하여 $g(x)$라고 생각해보자.
보통은 uniform distribution이 가장 분포를 만들기 편하다.
또한, 일반적으로 타겟 분포는 무한대의 정의역을 가지니, 유의미한 범위에서의 정의역을 새로 정의해야 한다.

$$x = \{x | -7 \leq x < 17\}$$

그러면 제안분포는 아래와 같이 생긴다.

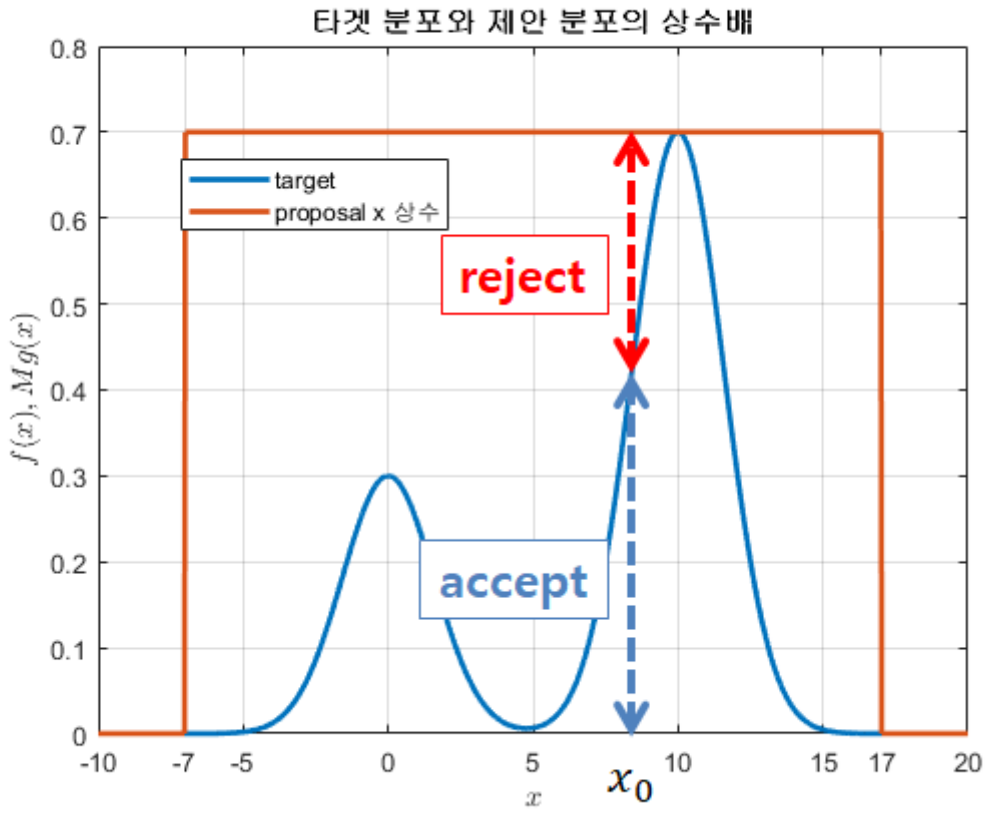


또한 임의의 적당한 상수배를 취하여 타겟 분포를 모두 포함할 수 있도록 해주자



Step 2. 샘플링 과정

$$\frac{f(x_0)}{(Mg(x_0))}$$



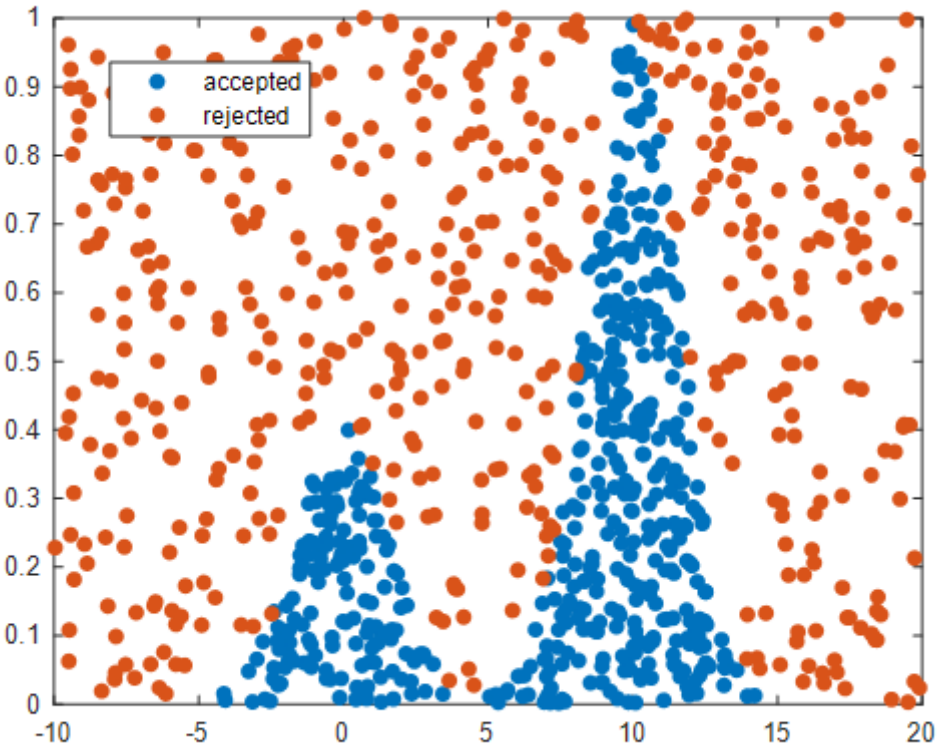
샘플링 하나(initial sample, x_0)를 추출함.

$$r = \frac{f(x_0)}{(Mg(x_0))}$$

Step 3. 비교

위의 값을 랜덤하게 형성되는 $u = \{u|0 \leq u < 1\}$의 값과 비교함

$r > u$ 이면 accept, 그렇지 않으면 reject



Vola!

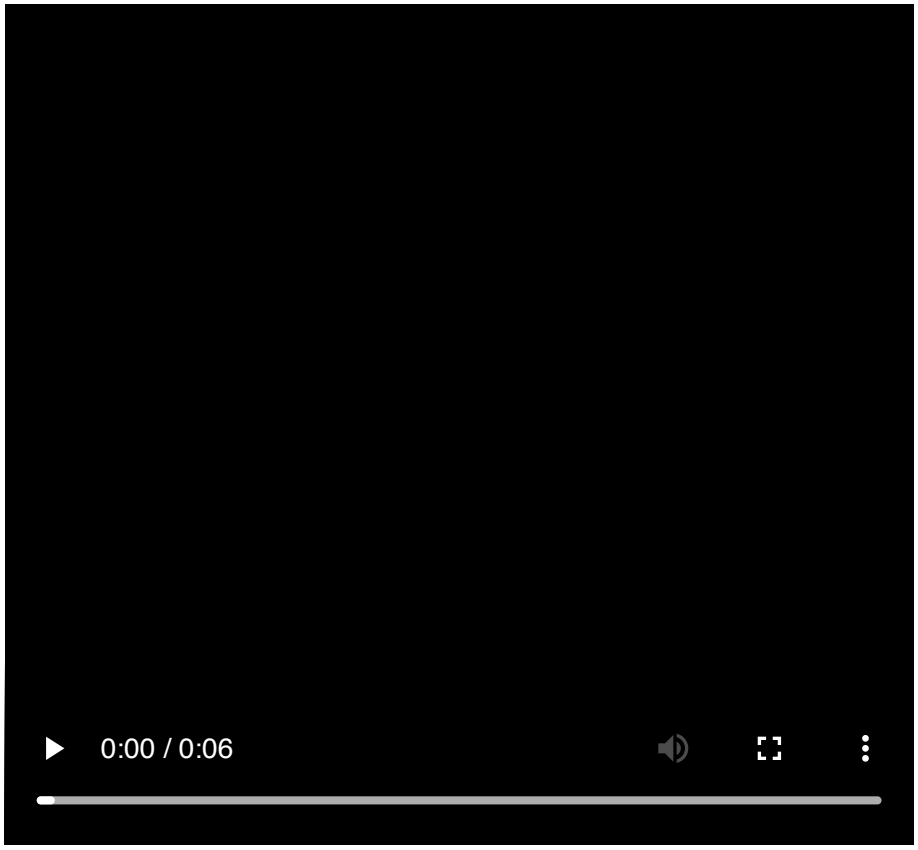
위와 같은 방법은 다음과 같은 문제점들을 가지고 있음

- 적절한 제안분포와 상수배(M) 찾기
- 이전에 관측치와 새로운 관측치가 서로 무관하여 불필요한 과정을 반복함

Markov chain Monte Carlo (MCMC)

Monte Carlo

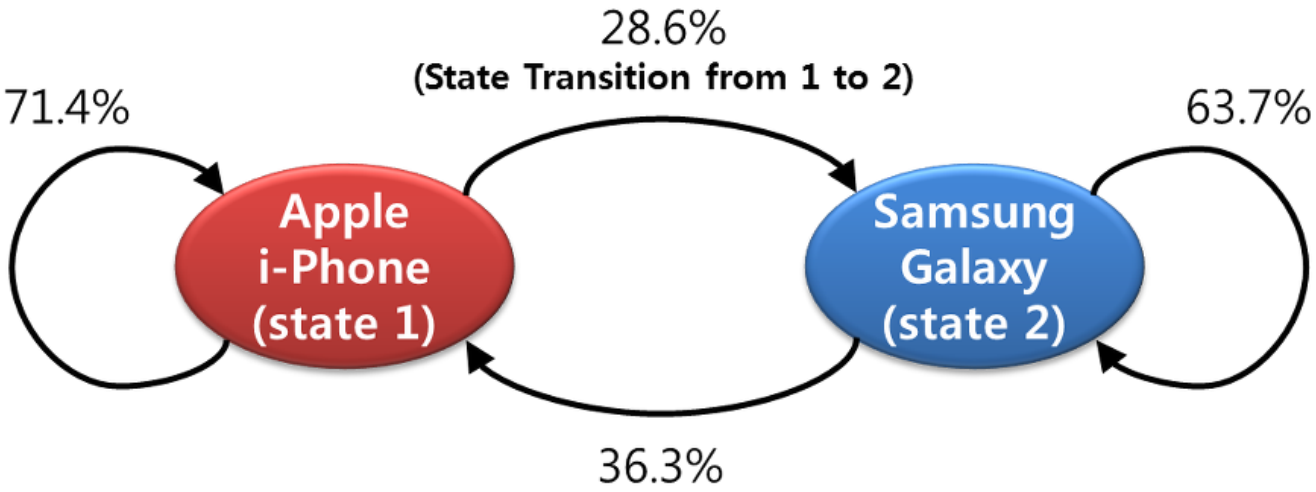
- 통계적인 특성을 이용해 무수히 많이 시도하여 원하는 수치적 모델링의 결과를 얻어내는 방식



Markov chain

현재 상태가 다음 상태에만 영향을 받는 확률 과정
예시)

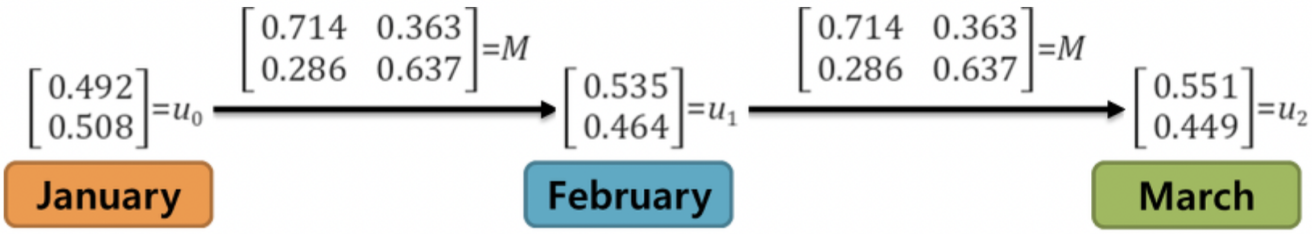
- 애플을 사용하던 사람이 그대로 애플을 사용할 확률은 71.4%, 그렇지 않은 경우는 모두 갤럭시로 이동
- 갤럭시를 사용하던 사람이 그대로 갤럭시를 사용할 확률을 63.7% 그렇지 않은 경우는 모두 애플로 이동
- 초기 시장 점유율은 애플과 삼성이 각각 49.2%, 50.8%



$P(X_1 = 애플) = P(X_0 = 애플) * P(X_1 = 애플)|X_0 = 애플) + P(X_0 = 애플) * P(X_1 = 애플)|X_0 = 삼성)$
이런식으로 기존 스테이트에만 영향을 받게 됨
여기서

$$\begin{matrix} & Apple & Sam \\ Apple \\ Sam \end{matrix} \begin{bmatrix} 0.714 & 0.363 \\ 0.286 & 0.637 \end{bmatrix} = M \quad , \quad \begin{bmatrix} 0.492 \\ 0.508 \end{bmatrix} = u \quad \begin{matrix} Current\ state \end{matrix}$$

위와 같이 수식을 전개 할 수 있고,
연속적인 시간의 흐름으로 생각하면,

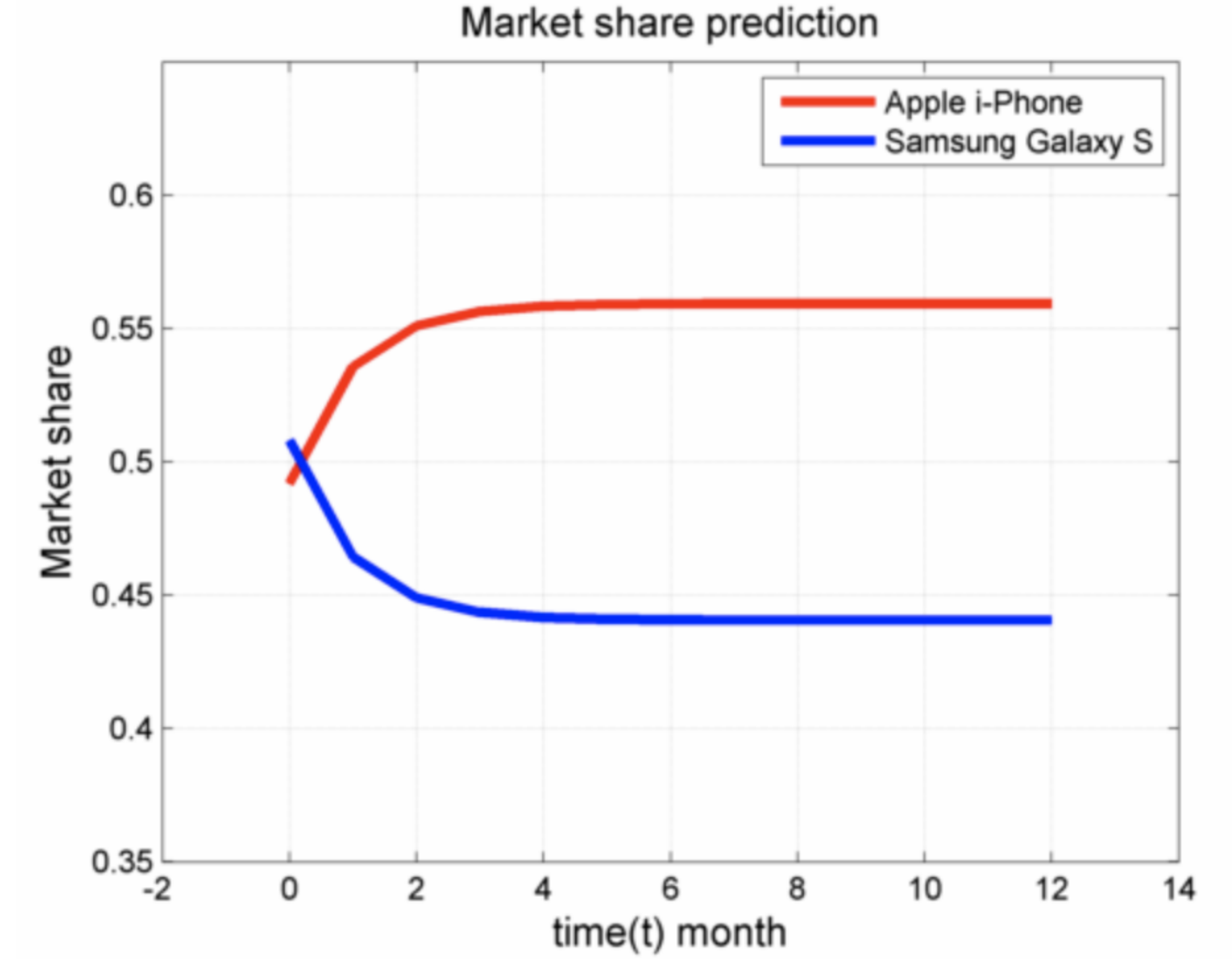


위의 그림처럼 표현이 가능함.

일반화하면

$$S_i = M^i * S_0$$

확률이기 때문에 하다보면 언젠가 수렴함



Markov Chain Monte Carlo

무작위 하게 샘플링을 하여 원하는 테스크를 수행하지만, 샘플링하는 과정에서 전 단계의 샘플링이 뒷 단계의 샘플링에 영향을 준다는 의미.

다양한 알고리즘이 사용되지만 (*Metropolis*)이 주로 사용 됌

Metropolis 알고리즘

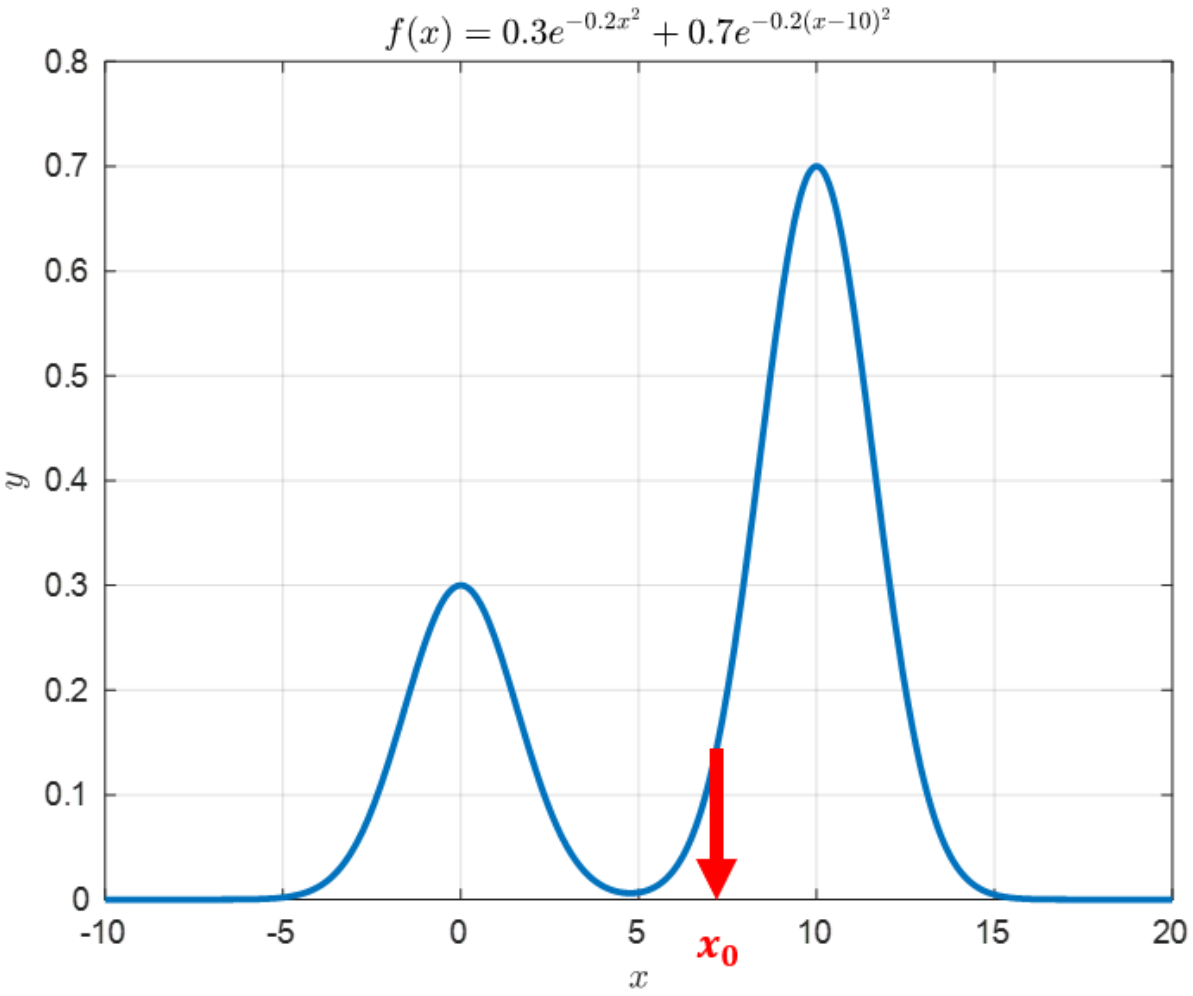
강의내용 뒤에서 배울 내용임

$$f(x) = 0.3\exp(-0.2x^2) + 0.7\exp(-0.2(x-10)^2)$$

위와 같은 함수를 알고 있다고 가정

1. 초기화 과정

   변수를 아무거나 하나 먼저 고름

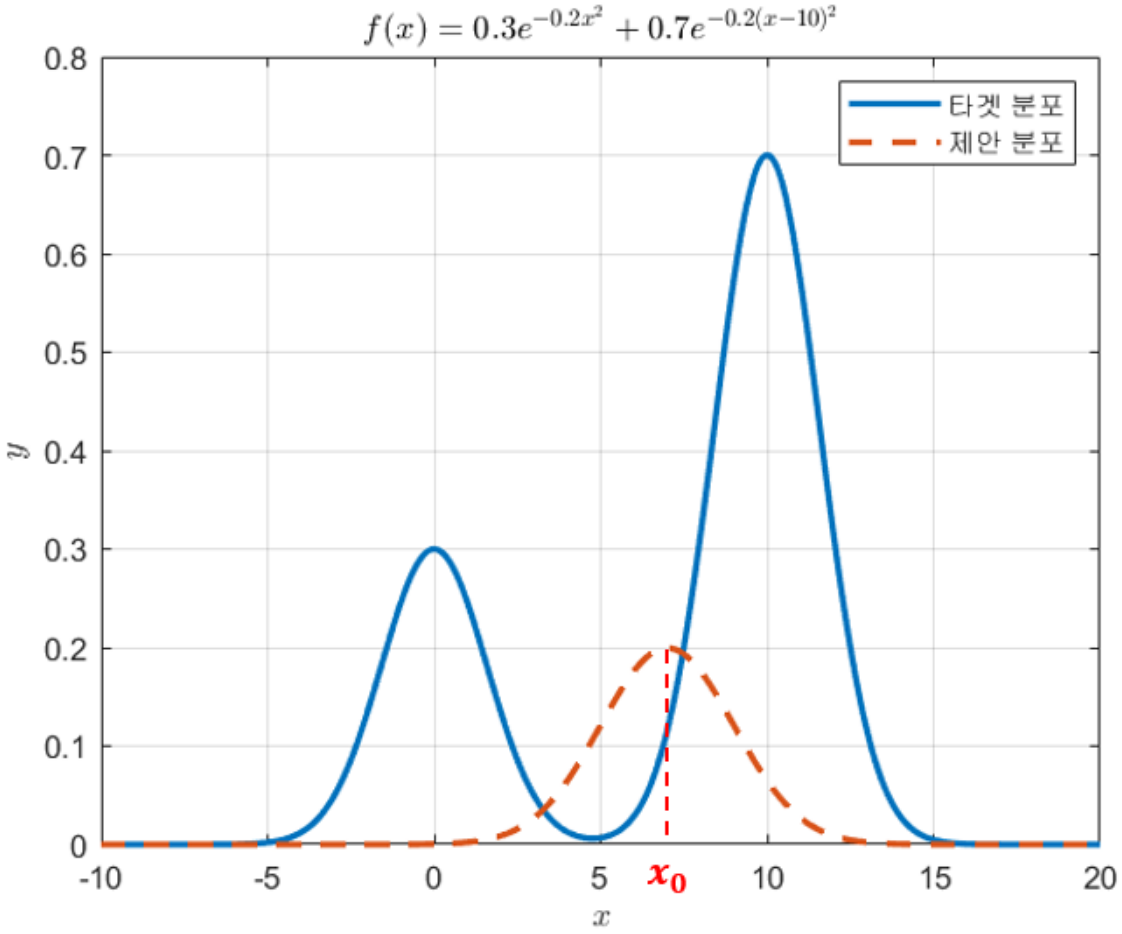


2. 제안분포로부터 다음 포인트를 추천받기

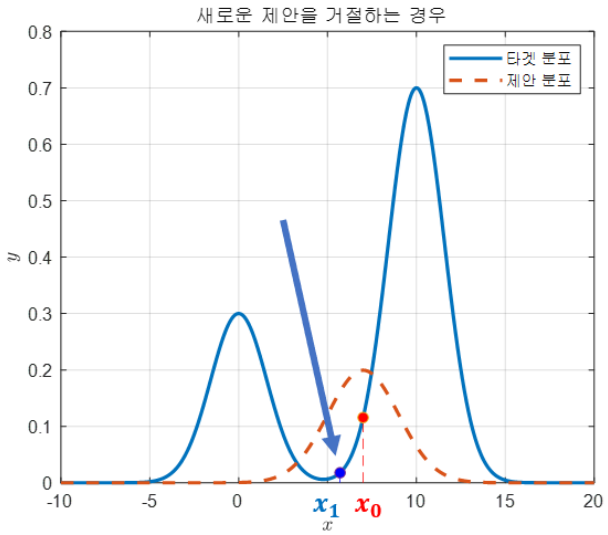
   Metropolis는 synmmetric한 분포를 사용하여 제안분포를 만드는 것을 고안함.

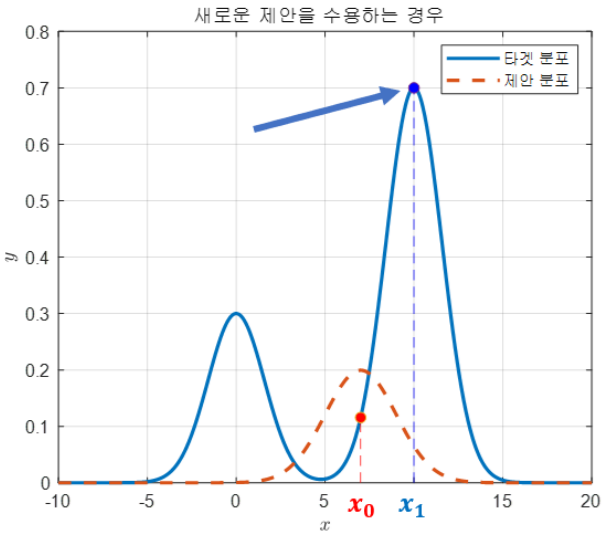
   $g(x)$를 제안분포로 정의. 제안분포는 시작점을 평균으로 갖고 너비는 임의의 설정된 값을 갖는 정규분포

를 따름



제안분포에서 임의의 값을 설정하여 그 값을 수용할지 승낙할지 결정한다.





왼쪽 그림에서는 제안 분포의 값이 더 신뢰가 높기 때문에 거절하고,
오른쪽 그림에서는 기존 분포의 값이 더 신뢰가 높기 때문에 승낙함
즉, 수용 기준은 아래와 같다

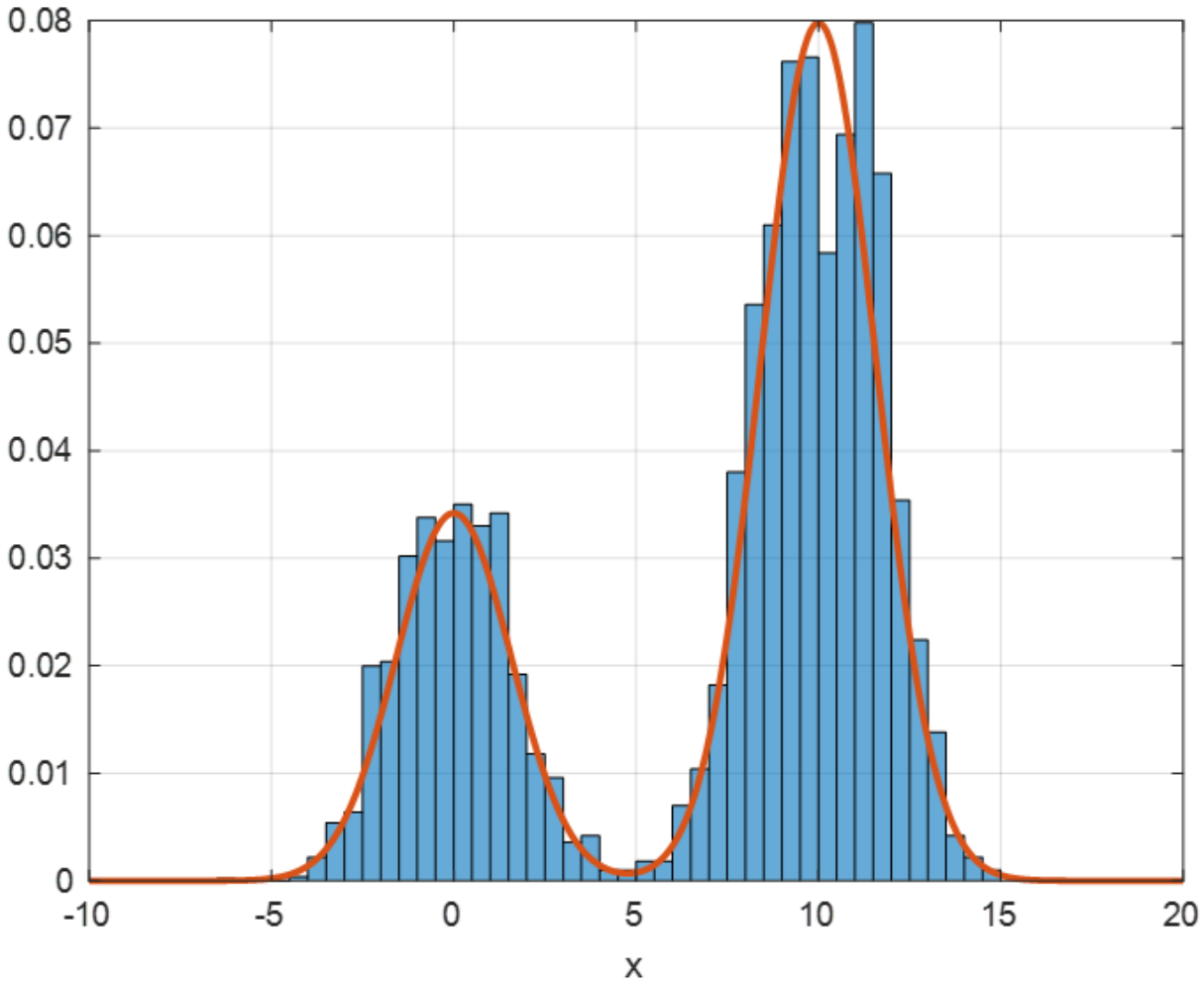
$$\frac{f(x_{i+1})}{f(x_i)} > 1$$

해당 조건으로만 수용조건을 만들면, 제안분포가 타겟분포를 envoloped 하게 만들어지면 모두 수용하지 못함.
따라서 제안분포에 대한 불확실성 문제를 해소하기 위해서 다음 단계에서 수용할 수 있는 방법을 마련함
3. 패자부활전

[0,1] 사이에 값에서 uniform distribution을 따르는 u에서 아래 조건을 만족하면 수용

$$\frac{f(x_{i+1})}{f(x_i)} > u$$

2~3 과정을 무한히 반복하면 결국 수렴함.

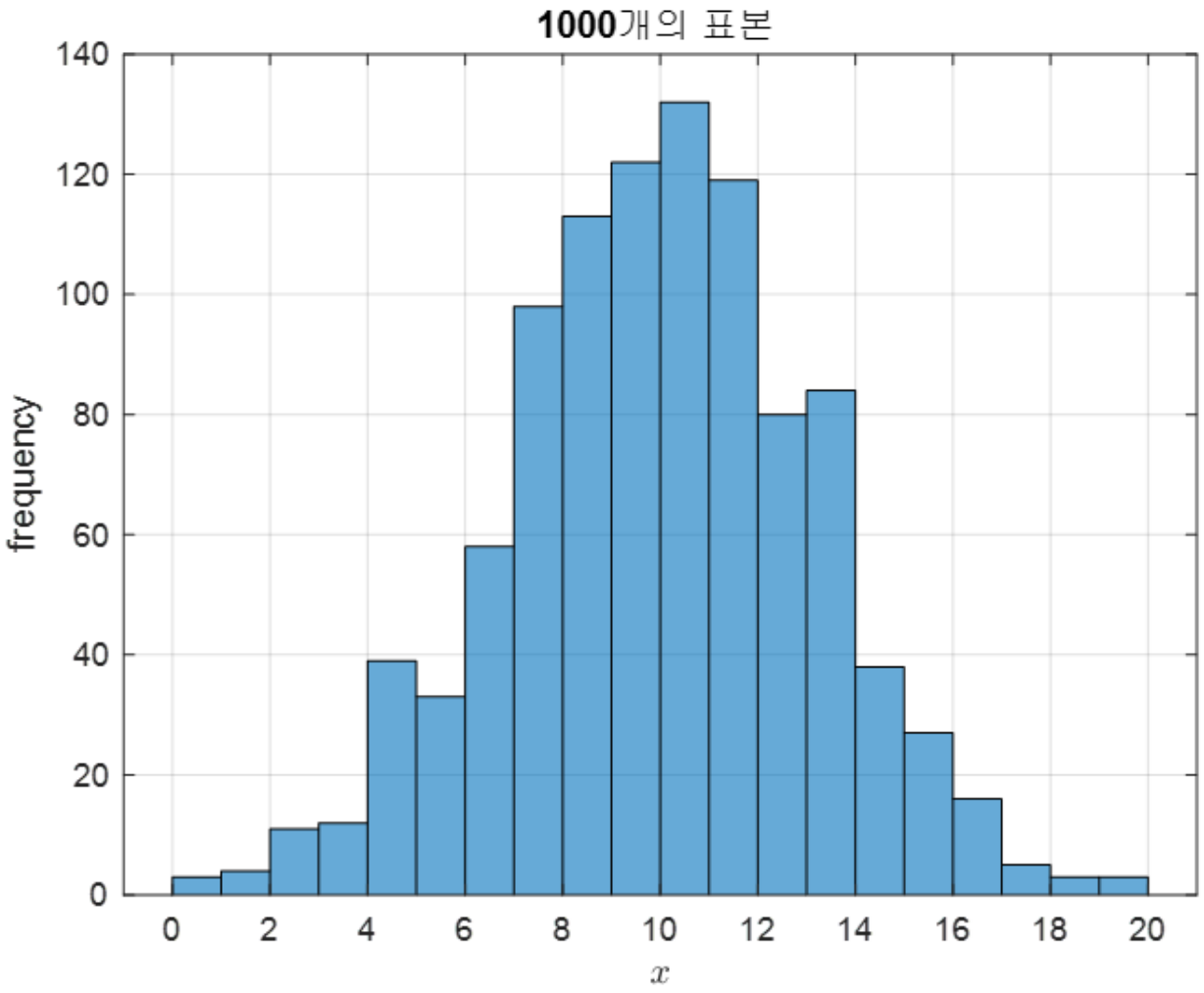


MCMC를 활용한 Bayesian estimation

표본집단에서 모집단의 특성(평균, 분산 등)을 설명하는 파라미터를 추정하는데에도 MCMC가 사용 될 수 있음

예시)

모집단은 30000개의 원소로 이루어진 평균 10, 표준편차가 3인 정규분포를 따름
표본집단은 1000개의 표본만 추출함



자세한 설명은 아래 링크 참조....

---

https://angeloyeo.github.io/2020/09/16/rejection_sampling.html
https://www.secmem.org/blog/2019/01/11/mcmc/
https://untitledtblog.tistory.com/134
https://m.blog.naver.com/jinis_stat/221648406160