8.4 정규분포와 중심극한정리

정규분포(normal distribution) 또는 가우스 정규분포(Gaussian normal distribution)라는 분포는 자연현상에서 나타나는 숫자를 확률모형으로 모형화할 때 많이 사용한다.

* 정규분포의 확률밀도함수 (pdf)

$$N(x;\mu,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

분산의 역수를 정밀도(precision) β라고 부르기도 한다.

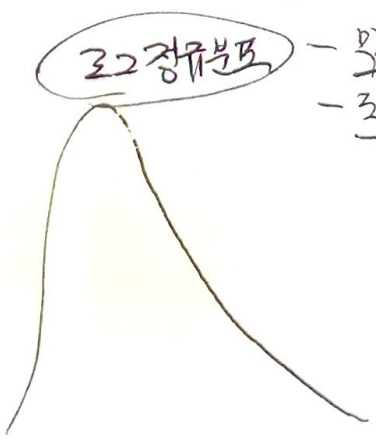
$$\beta=\frac{1}{\sigma^2}$$

↳ 역수를 취하는 이유 = 미분이 편함

정규분포 중에서도 평균이 0이고 분산이 1인 ($\mu=0$, $\sigma^2=1$) 정규분포를 표준정규분포 (standard normal distribution)라고 한다.

정규분포의 확률밀도함수는 아래와 같은 성질을 가진다.

- $x=\mu$ 일 때 확률밀도가 최대가 된다.
- $x=\infty$ 이거나 $x=-\infty$ 이 값으로 확률밀도는 0으로 가까워진다.



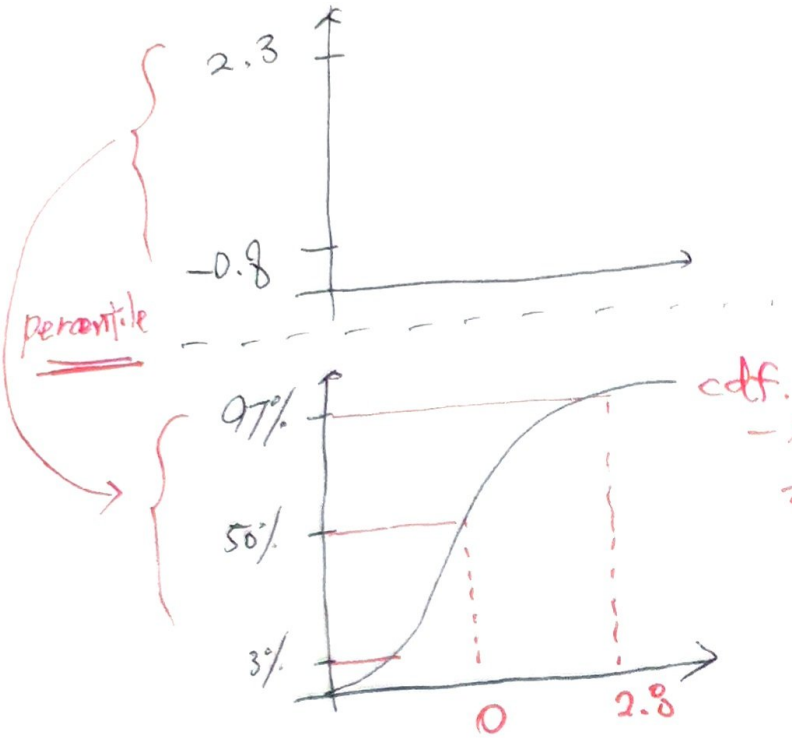
- 모든 양수 (ex. 거래량)
- 로그 취하면 정규분포 형태

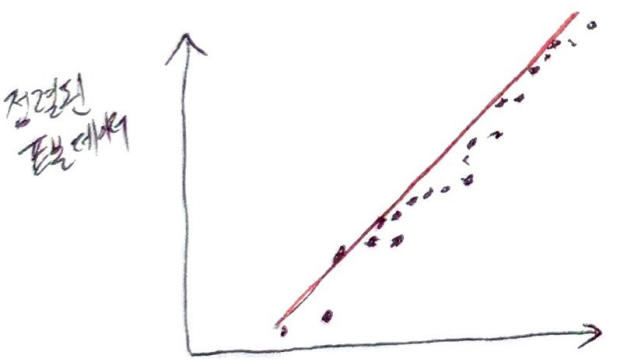
Q-Q 플롯

* 정규분포는 여러 연속확률분포 중에서도 가장 널리 사용되는 확률분포.
→ 어떤 확률변수의 분포가 정규분포인지 아닌지 확인하는 것은 중요한 통계적분석 중 하나

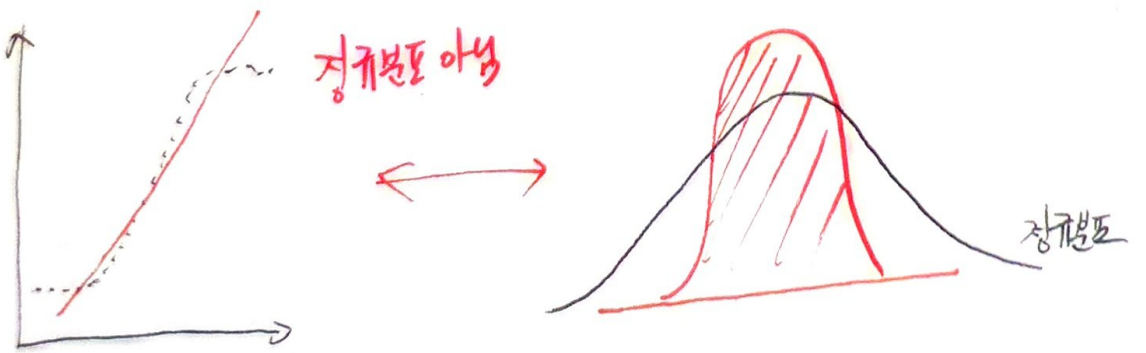
Q-Q (Quantile-Quantile) 플롯
→ 분석할 표본 데이터의 분포와 정규분포의 분포형태 비교
→ 표본 데이터가 정규분포를 따르는지 검사하는 간단한 시각적 도구
→ 동일 분위수에 해당하는 정상 분포의 값과 주어진 데이터값을 pairing 하여 그린 scatter plot

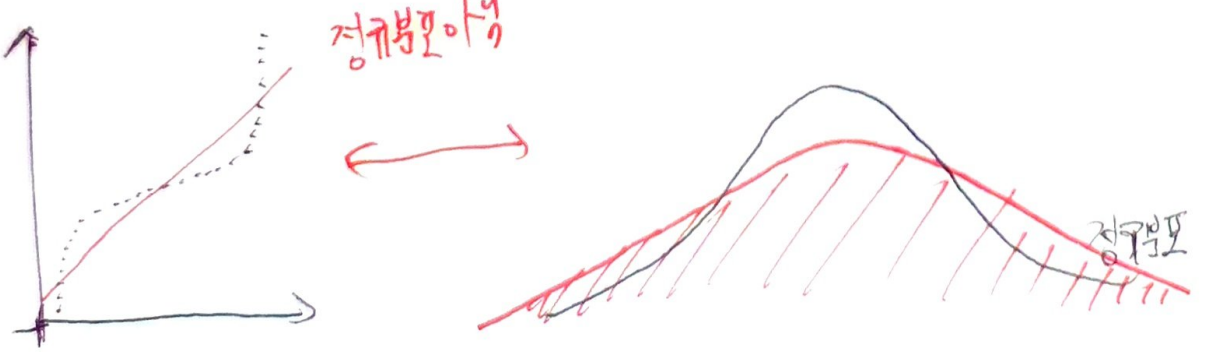
Q-Q process (대략적 방법론)
① 표본 데이터 정렬
② 해당하나의 표본데이터가 전체 데이터 중의 몇 %에 해당하는지 위치값 계산
• 위치값으로는 특정 순위 (order) 값이 나타날 가능성이 높은 값을 뜻하는 순서통계량 (order statistics) 이라는 값 사용 (절반씩 데이터를 가능성이 가장 높은 값)
③ 각 표본 데이터의 위치값에 정규분포의 누적확률함수 (cdf) 값이 되는 표준정규분포의 표본값 계산
• 즉, 확률값에 대한 누적확률함수의 역함수 값 구하기
= 표본정규분포의 분위함수 (quantile function)
(ex) 표본정규분포의 1% 분위함수값은 $F^{-1}(0.01)$, 약 -2.326
④ 정렬된 표본데이터 (ordered values) 와 그에 대응하는 분위수 (theoretical quantiles)를 하나의 쌍으로 간주하여 2차원 공간에 하나의 점 (point) 으로 그린다.
⑤ 모든 표본에 대해 ② - ④ 반복하여 Q-Q 플롯 완성


2.3
-0.8
percentile
97%
50%
3%
0
2.8
cdf.
- 정규분포라고
가정하고
cdf matching


정렬된
표본데이터
일직선이 형성되면
정규분포의 가깝다는 의미.
(정규분포)이론적 위치값.

qq plot


정규분포 아님
정규분포


정규분포아님
정규분포

중심극한정리 Central Limit Theorem

실세계에서 발생하는 현상 중 많은 것들을 정규분포로 모형화가능.
→ 그 이유 중 하나: 중심극한정리 (Central Limit Theorem)

→ "독립확률변수의 합이 정규분포과 비슷한 분포를 이루는 현상"

분포 $X_1, X_2, \cdots, X_N$이 기댓값이 μ이고 분산이 σ^2로 동일한 분포 (기댓값과 분산의 값이 동일할 뿐이며, 분포의 모양은 달라도 됨)이며 서로 독립인 확률변수들이라 하자.

$X_1, X_2, \cdots, X_N$에서 뽑은 각각의 표본데이터 $x_1, x_2, \cdots, x_N$의 표본평균

$$\bar{x}_N = \frac{1}{N}(x_1 + \cdots + x_N)$$

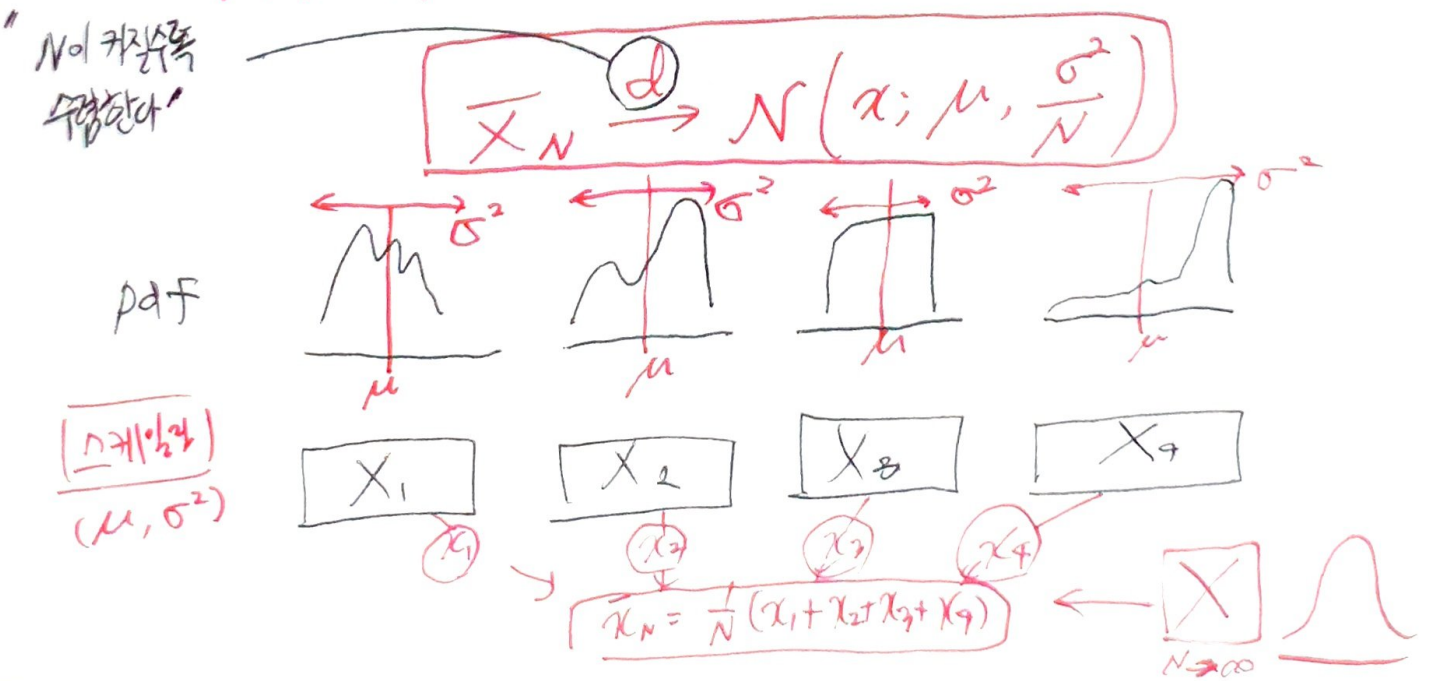
도 마찬가지로 예측할 수 없는 확률변수다. 이 확률변수를 $\bar{X}_N$이라고 하자.

중심극한정리는 다음과 같다.

N개의 임의 분포로부터 얻은 표본의 평균은
N이 증가할수록 기댓값이 μ, 분산이 $\frac{\sigma^2}{N}$인 정규분포로 수렴한다.

"N이 커질수록 수렴한다"

$$\bar{X}_N \xrightarrow{d} N\left(x; \mu, \frac{\sigma^2}{N}\right)$$



표본평균이 0, 분산이 1이 되도록 다음과 같이 정규화를 하면
다음과 같이 쓸 수 있다.

{ N개의 임의의 분포로부터 얻은 표본의 평균을 정규화하면
N이 증가할수록 표준정규분포로 수렴한다. }

$$\frac{\bar{X}_N - \mu}{\frac{\sigma}{\sqrt{N}}} \xrightarrow{d} \mathcal{N}(x; 0, 1)$$

* 중심극한정리가 중요한 이유?

현실에 존재하는 매우 복잡한 데이터들은 각각은 서로 매우 다르지만
많이 합칠수록 정규분포의 특성을 나타낸다

→ 잘 모르겠으면 일단 많이 모아서 정규분포로 분석해보자 ☆

정규분포의 통계량 분포

임의의 분포가 아닌 특수한 정규분포로부터 얻은 표본 데이터로 구한 표본평균은
어떤 분포를 가지게 될까?

N개의 정규분포로부터 얻은 표본의 합은 N과 상관없이 기댓값이 $N\mu$,
분산이 $N\sigma^2$인 정규분포다.

$$x_i \sim \mathcal{N}(\mu, \sigma^2) \rightarrow \sum_{i=1}^{N} x_i \sim \mathcal{N}(N\mu, N\sigma^2)$$

정규분포의 표본에 상수를 빼거나 곱해도 정규분포다. 이 경우에도 위와 같이 기댓값이 0, 표준편차가 1이 되도록 정규화를 하면 다음과 같이쓸수있다.

$$x_i \sim N(\mu, \sigma^2) \rightarrow \boxed{z} = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{N}}} \sim N(x; 0, 1)$$

정규분포 표본의 평균을 정규화한 통계량 = Z통계량

* 중심극한정리 — 무작위 표본, 무한대의 데이터가 정규분포가된다
* Z통계량 — 정규분포 표본, N 갯수와 상관없이 항상 정확하게 표준정규분포.

선형회귀모형과 정규분포

* 정규분포는 선형회귀모형에서 잡음(disturbance)을 모형화하는 데 사용

{ 선형회귀모형은 입력변수 $x_1, \cdots, x_N$이 종속변수 y에 선형적으로 영향을 미치는 모형이다. }

$$\hat{y} = w_1 x_1 + \cdots + w_N x_N \approx y$$

이 모형은 다음과같이 표현될 수 있다.

$$y = w_1 x_1 + \cdots + w_N x_N + \epsilon$$

ϵ는 잡음(disturbance)이라고 하며, 우리가 값을 측정할 수 없는 양을 뜻한다. 예측값과 실제값의 차이를 뜻하는 잔차(residual)와는 다르다.

$$\epsilon = \text{disturbance}$$
$$\neq$$
$$\text{residual}$$

잡음은 선형회귀모형을 만들때 하나하나의 영향력이 작거나 일일이 측정하기 힘들어서 무시된 수많은 변수들의 영향을 하나로 합친 것이다.

즉, 월래 y값은 $x_1, \dots, x_N, \dots$의 거의 무한한 갯수의 입력변수의 영향을 받는다.

$$y = w_1 x_1 + \dots + w_N x_N + w_{N+1} x_{N+1} + w_{N+2} x_{N+2} + \dots$$

하지만 이 중에서 입력변수 $x_1, \dots, x_N$만의 영향력이 크거나 측정이 쉽다면 다른 변수의 영향을 하나의 확률변수라고 합쳐서 표현할 수 있다.

$$\epsilon = w_{N+1} x_{N+1} + w_{N+2} x_{N+2} + \dots$$

(측정X, 구할수없는 어떤값.)

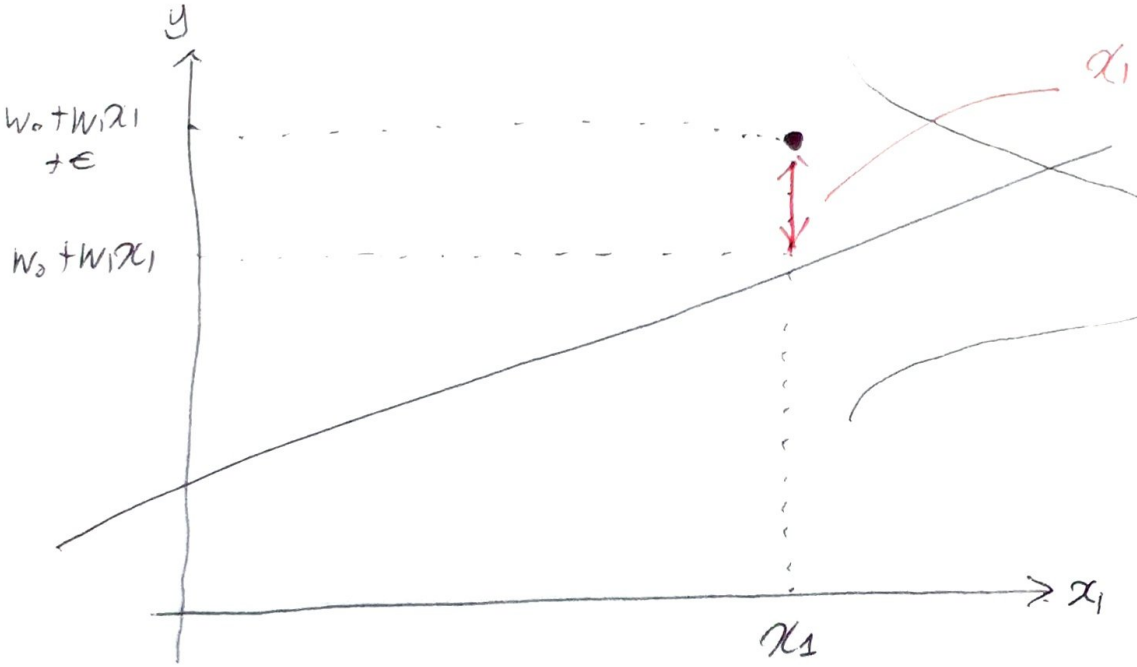
중심극한정리에 의해,
임의의 확률변수의 합은 정규분포와 비슷한 형태가 된다. 또한 ϵ의 기댓값이 0이 아니면 아래처럼 상수항 $w_0 = E[\epsilon]$을 추가하고 ϵ의 기댓값이 0이라고 할 수 있으므로,

$$y = w_0 + w_1 x_1 + \dots + w_N x_N + \epsilon$$

잡음 ϵ의 기댓값인 0인 정규분포라고 가정하는 것을 알아두어라.

$$\epsilon \sim N(0, \sigma^2)$$

$$y = w_0 + w_1 x_1 + \overbrace{w_2 x_2 + w_3 x_3 + \dots}^{\epsilon}$$



x_1 이외의 설명변수들이 설명하는 다른 영향들이 모여서,
우리가 분석하고 싶은 것 이외의 설명변수들이 누적되어 미친 영향의 크기
||
ϵ
(epsilon)

〈선형회귀모델의 2가지 정규〉

① ϵ을 배제

② ϵ가 정규분포에서 비롯된다고 가정하고, ϵ의 정보를 이용