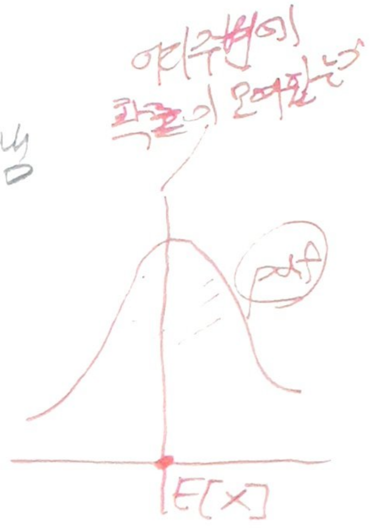
7.3 분산과 표준편차

→ 확률분포에서 확률이 얼마나 모여있는지 (퍼져있는지) 나타냄

기댓값	분산
확률변수에서 어떤 값이 나올지 예측	그 예측의 정확도 혹은 신뢰성을 표현



확률분포의 분산

확률밀도함수 $p(x)$의 수식을 알고 있다면, 이론적인 분산을 구할 수 있다. 분산을 구하는 연산은 영어 Variance 앞글자를 따서 Var[·]로 표기하고, 이 연산으로 계산된 분산값은 σ^2 으로 표기한다.

$$\sigma^2 = Var[X] = E[(X-\mu)^2]$$

이산확률변수의 분산은 평균으로부터 표본 데이터까지 거리의 제곱을 확률질량함수 $p(x)$로 가중하여 더한 값이다.

$$\sigma^2 = \sum_{x \in \Omega} (x_i - \mu)^2 p(x_i)$$

연속확률변수의 분산 ⇒ 평균으로부터 표본 데이터까지 거리의 제곱을 확률밀도함수 $p(x)$로 가중하여 적분한 값.

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 p(x) dx$$

분산의 성질

분산은 다음과 같은 성질을 만족한다.

- 분산은 항상 0 또는 양수이다.

$$Var[X] \geqq 0$$

- 확률변수가 아닌 상수값 c에 대해 다음 식이 성립한다.

$$Var[c] = 0$$

$$Var[cX] = c^2 Var[X]$$

또한 기댓값의 성질을 이용하여 다음 성질을 증명할 수 있다.

$$Var[X] = E[X^2] - (E[X])^2 = E[X^2] - \mu^2$$

또는

$$E[X^2] = \mu^2 + Var[X]$$

(증명)

$$\begin{aligned} Var[X] &= E[(X-\mu)^2] \\ &= E[(X^2 - 2\mu X + \mu^2)] \\ &= E[X^2] - 2\mu E[X] + \mu^2 \\ &= E[X^2] - 2\mu^2 + \mu^2 \\ &= E[X^2] - \mu^2 \\ &= E[X^2] - (E[X])^2 \end{aligned}$$

두 확률변수의 합의 분산

두 확률변수 X, Y의 합의 분산은 각 확률변수의 분산의 합과 다음과 같은 관계가 있다.

$$Var[X+Y] = \underbrace{Var[X]}_{\geq 0} + \underbrace{Var[Y]}_{\geq 0} + \underbrace{2E[(X-\mu_X)(Y-\mu_Y)]}_{\text{양수 OR 음수}}$$

(증명)

우선 확률변수 X+Y의 기댓값은 기댓값의 성질로부터 각 확률변수의 기댓값의 합과 같다.

$$\boxed{E[X+Y] = \mu_X + \mu_Y}$$

분산의 정의와 기댓값의 성질로부터 다음이 성립한다.

$$\begin{aligned} Var[X+Y] &= E[(X+Y-(\mu_X+\mu_Y))^2] \\ &= E[((X-\mu_X)+(Y-\mu_Y))^2] \\ &= E[(X-\mu_X)^2+(Y-\mu_Y)^2+2(X-\mu_X)(Y-\mu_Y)] \\ &= Var[X]+Var[Y]+2E[(X-\mu_X)(Y-\mu_Y)] \end{aligned}$$

확률변수의 독립

두 확률변수가 서로 독립 (independent) ... 두 확률변수가 가질수 있는 모든 사건 조합에 대해, 결합사건의 확률이 각 사건 확률과 동일하다

= 두 확률변수가 서로에게 영향을 미치지 않는다.

두 확률변수에서 하나의 확률변수 값이 특정한 값이면 다른 확률변수의 확률분포가 영향을 받아 변하게 되면 종속 (dependent)이라고 한다.

주사위를 두번 던져 나오는 합의 분포 ... 주사위 값 2개는 서로 종속.

* 확률변수 X, Y가 서로 독립이면 다음 식이 성립한다.

$$E[(X-\mu_X)(Y-\mu_Y)] = 0$$

(증명은 추후 학습)

이 등식을 이용하면, 서로 독립인 두 확률변수의 합의 분산은 각 확률변수의 분산의 합과 같다는 것을 보일 수 있다.

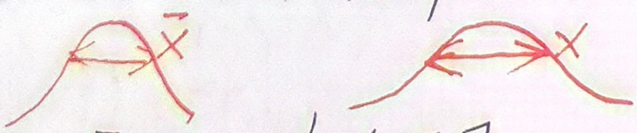
$$Var[X+Y] = Var[X] + Var[Y]$$

표본평균의 분산

확률변수 X의 표본평균 $\bar{X}$도 확률변수이고, 그 기댓값 $E[\bar{X}]$은 원래 확률변수 X의 기댓값 $E[X]$와 같다는 것을 증명한 적이 있다.

$$E[\bar{X}] = E[X]$$

표본평균 $\bar{X}$의 분산 $Var[\bar{X}]$은 원래 확률변수 X의 분산 $Var[X]$와 다음 관계를 갖는다.



$$Var[\bar{X}] = \frac{1}{N} Var[X]$$

따라서 표본평균을 계산한 표본 갯수가 커지면, 표본평균의 값의 변동은 작아진다. 표본의 갯수가 무한대가 되면, 표본평균의 값은 항상 일정한 값이 나온다. 즉 확률적인 값이 아니라 결정론적인 값이 된다.

(증명)

$$Var[\bar{X}] = E\left[(\bar{X} - E[\bar{X}])^2\right]$$

$$= E\left[(\bar{X} - \mu)^2\right]$$

$$= E\left[\left(\frac{1}{N}\sum_{i=1}^{N} X_i - \mu\right)^2\right]$$

$$= E\left[\left(\frac{1}{N}\sum_{i=1}^{N} X_i - \frac{1}{N}N\mu\right)^2\right]$$

$$= E\left[\left(\frac{1}{N}\left(\sum_{i=1}^{N} X_i - N\mu\right)\right)^2\right]$$

" summation의 제곱 "

$$= E\left[\frac{1}{N^2}\left(\sum_{i=1}^{N}(X_i - \mu)\right)^2\right]$$

$$= E\left[\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(X_i - \mu)(X_j - \mu)\right]$$

$$= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} E\left[(X_i - \mu)(X_j - \mu)\right]$$

i번째 표본값은 j번째 표본값에 영향을 미치지 않으므로, X_i와 X_j $(i \neq j)$는 독립이다. 따라서,

$$E[(X_i - \mu)(X_j - \mu)] = 0 \quad (i \neq j)$$

라는 사실을 이용하여 $i = j$인 항, 즉 제곱항만 남는다.

(확인)

$$Var[\bar{X}] = \frac{1}{N^2}\sum_{i=1}^{N} E[(X_i - \mu)^2]$$

$$= \frac{1}{N^2}\boxed{\sum_{i=1}^{N}} E[(X - \mu)^2]$$

$E(X_i) = E(X)$

모두 동일한 값.

$$= \frac{1}{N^2}\boxed{N} E[(X-\mu)^2]$$

$$= \frac{1}{N} \underline{E[(X-\mu)^2]}$$

$$= \frac{1}{N} \underline{Var[X]}$$

위 식의 의미

original distribution.

- 데이터를 생성하는 확률변수 X의 기댓값을 구하려면, 확률밀도함수(PDF) $p(x)$의 수식을 알아야 한다.
- 그런데 우리는 데이터를 생성하는 확률변수 X의 확률밀도함수 $p(x)$의 수식을 정확히 알지 못한다.
- 하지만 표본평균이라는 새 확률변수 $\bar{X}$의 기댓값 $E[\bar{X}]$는 원래 확률변수 X의 기댓값 $E[X]$와 같으므로, 표본평균 $\bar{x}$는 원래 확률변수 X의 기댓값 $E[X]$와 비슷한 값이 나올 것이다. (하지만 정확한 값은 아니다)
- 만약 표본 개수 N이 크면, 표본평균 $\bar{x}$의 분산이 아주 작아지므로 ($\frac{1}{N}$), 표본평균의 값 $\bar{x}$는 항상 표본평균의 기댓값 $E[\bar{X}] = E[X]$ 근처의 거의 일정한 값이 나올 것이다.
- 따라서, 표본 개수 N이 크면 표본평균 $\bar{x}$는 원래 확률변수 X의 기댓값 $E[X]$의 근삿값이라고 할 수 있다.

표본분산의 기댓값

앞에서 표본평균의 기댓값을 구하면 이론적인 평균, 즉 기댓값과 같아진다는 것을 증명했다. 그런데 표본분산 S^2의 기댓값을 구하면 이론적인 분산 σ^2과 같아지는 것이 아니라 이론적인 분산값의 $\frac{N-1}{N}$배가 된다.

$$E[S^2] = \frac{N-1}{N}\sigma^2$$

즉, 표본분산값이 이론적인 분산값보다 더 작아진다.

증명은 다음과 같다.

$$E[S^2] = E\left[\frac{1}{N}\sum_{i=1}^{N}(X_i-\bar{X})^2\right] = E\left[\frac{1}{N}\sum_{i=1}^{N}\{(X_i-\mu)-(\bar{X}-\mu)\}^2\right]$$

$$= E\left[\frac{1}{N}\sum_{i=1}^{N}\{(X_i-\mu)^2-2(X_i-\mu)(\bar{X}-\mu)+(\bar{X}-\mu)^2\}\right]$$

$$= E\left[\frac{1}{N}\sum_{i=1}^{N}(X_i-\mu)^2\right] - 2E\left[\frac{1}{N}\sum_{i=1}^{N}(X_i-\mu)(\bar{X}-\mu)\right] + E\left[\frac{1}{N}\sum_{i=1}^{N}(\bar{X}-\mu)^2\right]$$

이때 첫 번째 항은,

$$E\left[\frac{1}{N}\sum_{i=1}^{N}(X_i-\mu)^2\right] = E\left[\frac{1}{N}\sum_{i=1}^{N}(X-\mu)^2\right]$$

$$= E\left[\frac{1}{N}N(X-\mu)^2\right]$$

$$= E\left[(X-\mu)^2\right]$$

$$= \mathrm{Var}[X]$$

$$= \sigma^2$$

두 번째 항은,

$$E\left[\frac{1}{N}\sum_{i=1}^{N}(X_i-\mu)(\bar{X}-\mu)\right] = E\left[\frac{1}{N}\sum_{i=1}^{N}(X_i-\mu)\left(\frac{1}{N}\sum_{i=1}^{N}X_i-\mu\right)\right]$$

$$= E\left[\frac{1}{N}\sum_{i=1}^{N}(X_i-\mu)\left(\frac{1}{N}\sum(X_j-\mu)\right)\right]$$

$$= E\left[\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(X_i-\mu)(X_j-\mu)\right]$$

X_i와 X_j $(i \neq j)$는 서로 독립일 때,

$$E\left[(X_i-\mu)(X_j-\mu)\right] = 0 \quad (i \neq j)$$

라는 성질을 이용하면

$$E\left[\frac{1}{N}\sum_{i=1}^{N}(X_i-\mu)(\bar{X}-\mu)\right] = E\left[\frac{1}{N^2}\sum_{i=1}^{N}(X_i-\mu)^2\right]$$

$$= \frac{1}{N}E\left[\frac{1}{N}\sum_{i=1}^{N}(X_i-\mu)^2\right]$$

$$= \frac{1}{N}E\left[\frac{1}{N}\sum_{i=1}^{N}(X-\mu)^2\right]$$

$$= \frac{1}{N}E\left[\frac{1}{N}N(X-\mu)^2\right]$$

$$= \frac{1}{N}E\left[(X-\mu)^2\right]$$

$$= \frac{1}{N}Var[X]$$

$$= \frac{\sigma^2}{N}$$

세 번째 항은 다음과 같아진다.

$$E\left[\frac{1}{N}\sum_{i=1}^{N}(\bar{x}-\mu)^2\right] = E\left[\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{N}\sum_{j=1}^{N}X_j-\mu\right)^2\right]$$

$$= E\left[\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{N}\sum_{j=1}^{N}(X_j-\mu)\right)^2\right]$$

$$= E\left[\frac{1}{N^3}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{N}(X_j-\mu)(X_k-\mu)\right]$$

X_j와 X_k $(j\neq k)$가 독립일 때,

$$E[(X_j-\mu)(X_k-\mu)] = 0 \quad (j\neq k)$$

라는 성질을 이용하면

$$E\left[\frac{1}{N}\sum_{i=1}^{N}(\bar{x}-\mu)^2\right] = E\left[\frac{1}{N^3}\sum_{i=1}^{N}\sum_{j=1}^{N}(X_j-\mu)^2\right]$$

$$= E\left[\frac{1}{N^3}N\sum_{j=1}^{N}(X_j-\mu)^2\right]$$

$$= E\left[\frac{1}{N^2}\sum_{j=1}^{N}(X_j-\mu)^2\right]$$

$$= \frac{1}{N}E\left[\frac{1}{N}\sum_{j=1}^{N}(X_j-\mu)^2\right]$$

$$= \frac{1}{N}Var[X]$$

$$= \frac{\sigma^2}{N}$$

따라서 세 항의 합은 다음과 같다.

$$E[S^2] = \sigma^2 - \frac{2\sigma^2}{N} + \frac{\sigma^2}{N} = \frac{N-1}{N}\sigma^2$$

그러므로 표본분산의 기대값이 정확하게 σ^2이 되려면 평균과의 거리의 제곱의 평균을 구할 때, 분모가 N이 아니라 $N-1$으로 써야 한다.

$$\sigma^2 = \frac{N}{N-1}E[S^2]$$

$$= \frac{\cancel{N}}{N-1}E\left[\frac{1}{\cancel{N}}\sum(X_i-\bar{x})^2\right]$$

$$= E\left[\frac{1}{N-1}\sum(X_i-\bar{x})^2\right]$$

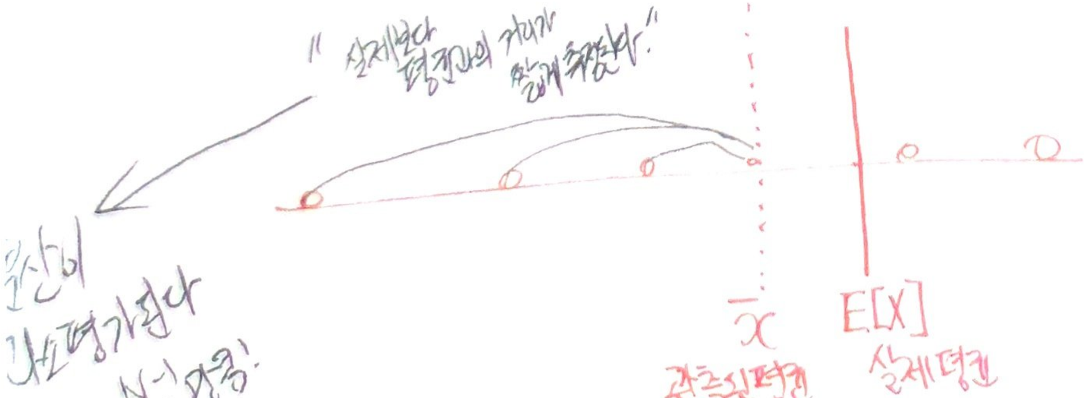
따라서 기댓값이 정확한 분산값과 일치하는 비편향 표본분산은 다음처럼 정의한다.

$$S^2_{unbiased} = \frac{1}{N-1}\sum(X_i-\bar{x})^2$$

* $\frac{1}{N}\sum(X_i-\bar{x})^2$가 실제보다 작으므로 $\frac{N}{N-1}$을 곱하여 크게 만들어주는 것.

이렇게 표본분산이 실제 분산보다 작아지는 이유는 다음과 같다.

1. 표본분산을 계산할 때 사용하는 표본평균의 값이 데이터가 많이 몰려있는 쪽으로 편향되게 나온다.
2. 이렇게 데이터가 몰려있는 위치에 있는 표본평균을 기준으로 각 데이터까지의 거리를 계산하면 원래의 기댓값으로부터의 거리보다 작게 나올 수 있다.



비대칭도와 첨도

비대칭도 (skew)

- 3차 모멘트 값에서 계산하고, 확률밀도함수의 비대칭 정도를 가리킨다.
- 비대칭도가 0이면 확률분포가 대칭이다.

$$E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\mu_3}{\sigma^3}$$

첨도 (kurtosis)

- 4차 모멘트 값에서 계산하며, 확률이 정규분포와 대비하여 중심에 모여있는지, 바깥으로 퍼져있는지를 나타낸다.

$$E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = \frac{\mu_4}{\sigma^4}$$

모멘트

앞서 구한 기댓값이나 분산은 확률분포의 모멘트(moment)의 하나다.

$$\mu_n = E[(X-\mu)^n] = \int (x-\mu)^n p(x)\,dx$$

모멘트는 확률분포에서 계산한 특징값이다. 만약 두 확률분포 X, Y가 있고 1차부터 무한 대차수에 이르기까지 두 확률분포의 모든 모멘트값이 서로 같다면 두 확률분포는 같은 확률분포다.

$$E[X] = E[Y]$$

$$E[(X-\mu_X)^2] = E[(Y-\mu_Y)^2]$$

$$E[(X-\mu_X)^3] = E[(Y-\mu_Y)^3]$$

$$\vdots$$

$$\Downarrow$$

$$X \stackrel{d}{=} Y$$ ($\stackrel{d}{=}$: 같은 확률분포)