

기댓값

- 확률 분포의 기댓값.

\rightarrow 만약 확률 변수가 라즈고 있는 확률 모형, 정확히는 확률 일도함수를 알고 있는 경우에, 과정과 같은 수식을 사용하여 이론적인 평균을 구할 수 있다. 이러한 이론적 평균을 확률 변수의 기댓값(expectation)이라고 한다.

- 확률 모형이 존재한다는 것이 물리적으로 확실한 경우에는 간단히 평균(mean)이라고 한다.

- 확률 변수의 기댓값을 구하는 연산자는 $E[\cdot]$ 로 표기한다. 그리스문자 μ_X 로 표기된다.

- 이산확률 변수 경우에는 확률질량함수 $P(x)$ 를 가중치로 x 를 곱하여 기댓값을 구한다.

$$\mu_X = E[X] = \sum x P(x)$$

- 연속확률 변수 경우에는 확률밀도함수 $f(x)$ 를 가중치로 x 를 곱하여 기댓값을 구한다.

$$\mu_X = E[X] = \int_{-\infty}^{\infty} x f(x) dx$$

- 확률 변수가 X 밖에 없을 때는 μ 라고 쓸 수 있다.

$$\mu_X = \mu.$$

확률밀도함수의 모양과 기댓값

- 기댓값은 여러가지 가능한 x 의 값들을 확률밀도값에 따라 가중합을 한 것이다.
가장 확률밀도가 높은 x 값 근처의 값이 된다.

\rightarrow 즉, 확률밀도가 높아 있는 근처의 위치를 표시낸다.

확률 변수의 변환

- 어떤 확률 변수 X 와 Y 를 가정하자.

- cX 는 확률 변수 X 에서 나온 값을 그 배하는 값.

- $X+Y$ 는 두 확률 변수 X 와 Y 를 더한 값.

- 이러한 기본 확률변수를 이용하여 새로운 확률 변수를 만드는 것이 "확률변수의 변환"이라고 한다.

기댓값의 성질

- 기댓값은 확률모형이라는 수식을 사용한 것이고 과정과 같은 성질을 가진다는 것을 수학적으로 증명 가능.

- 변환된 확률 변수의 기댓값을 계산을 할 때는 기댓값의 성질을 이용한다.

· 간접 변수가 아닌 고정된 것 c 에 대해

$$\Rightarrow E[c] = c$$

· 선형성

$$\Rightarrow E[cX] = cE[X]$$

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

생풀평균의 확률 분포.

- 확률변수부터 시기의 표본을 만들어 생풀 평균을 구하면 이 생풀평균값도 예측이 불가능한 확률변수라는 것을 알 수 있다. 생풀평균의 확률변수는 원래의 확률변수 이중에 bar 를 추가하여 X와 같이 표기한다. 예를 들어 확률변수 X에서 나온 표본으로 만들어진 생풀평균의 확률변수는 \bar{X} 로 표기).

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

- 위식에서는 X_i 는 i번재로 실현된 생풀을 의미한다.

연습문제 1

1. Numpy를 사용하여 20개의 숫자를 무작위로 생성한 후 히스토그램을 그리고 생풀평균을 구해라.
2. 1번과 같이 20개의 숫자 생성 및 생풀 평균 계산을 10번 반복하여, 10개의 Sample평균을 구해라.
3. 2번에서 구한 Sample평균의 히스토그램을 그리고, 생풀평균을 구해라.

기댓값과 생풀평균의 관계

- 생풀평균은 확률변수이고 기댓값이 존재한다. 생풀변수의 기댓값은 원래의 확률변수의 기댓값과 일치함을 수학적으로 증명할 수 있다.

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

(증명)

$$\begin{aligned} E[\bar{X}] &= E\left[\frac{1}{n} \sum_{i=1}^n X_i\right] \\ &= \frac{1}{n} \sum_{i=1}^n E[X_i] \\ &= \frac{1}{n} \times n \times E[X] \\ &= E[X] \end{aligned}$$

중앙값

- 확률분포부터 아론적 중앙값은 그 값보다 큰 값이 사�数 확률과 작은 값이 사�数 확률이 중립인 0.5이어야 하므로 그 값과 같이 누적 확률분포 $F(x)$ 에서 계산할 수 있다.

$$\text{median} = F^{-1}(0.5)$$

$$0.5 = F(\text{median})$$

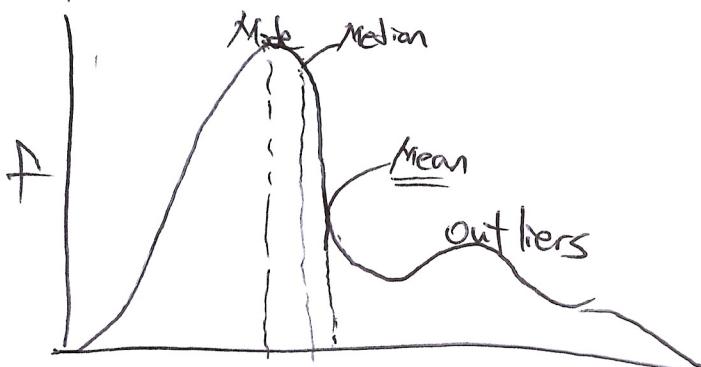
최빈값

→ 이산형 확률 분포에서는 가장 높은 값이 큰 수를 최빈값이라고 한다. 하지만 연속 확률 분포의 경우 어느 값에 대해서나 특정한 값이 사용 확률은 0 (zero)으로 고용과 같이 확률밀도함수의 값이 가장 큰 확률 변수의 값으로 정의한다. 즉, 확률 밀도 함수의 최대값의 위치

$$\text{mode} = \arg \max_x f(x)$$

기댓값, 중앙값, 최빈값 비교

- 확률 분포 즉, 확률밀도함수가 대칭인 경우에는 기댓값, 중앙값, 최빈값 모두 같다.
- 그러나 분포가 어느 한쪽으로 치그러진 (skewed) 경우에는 고용 그림과 같이 다를 수 있다.
- 예상량으로 비교하면 기댓값 < 중앙값 < 최빈값. 최빈값은 최종학 과정을 통하여 구할 수 있으나 예상량이 가장 많으며 오차가 크다.
- 기댓값은 이상값이나 한쪽으로 치그러진 상황에서는 큰 영향을 받지만, 중앙값이나 최빈값은 이에 대한 영향이 적다.



부산과 풋볼전차

- 흐름이나 기랫값이 본토의 위치를 재포하는 것이면, 본선은 본토의 폭(width)을 재포한 값.

$$\text{표준편차} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$$

쪽글봉트의 부산

- 각률변포는 각률변환 $T(x)$ 로 한도전체의 고양을 정확하게 정의할 수 있으므로
다음과 같이 이를 주인보석은 구할 수 있다.

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

- 이산확률변수의 경우에는 총합질량함수 $P(X)$ 를 사용하여 보간을 구한다.

$$\sigma^2 = \text{Var}[X] = E[(X - \mu)^2] = \sum (x - \mu)^2 p(x)$$

- 연속변수의 경우에는 차를 일으킬 수 있는지를 사용하여 보상을 구한다.

$$\sigma^2 = \text{Var}[X] = E[(X-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx.$$

- 주, 본인은 평균으로부터 Data까지의 거리제곱합~~을~~ $P(x)$ 또는 쪽을 빼고 $f(x)$ 를
가중치로하여 확장한 것으로 볼 수 있다.

부산의 성질

나는 분수를 자유롭고 같은 성질을 만족한지

1) 0 또는 양수 $\text{Var}[X] \geq 0$

ii) 랜덤 변수가 아닌 상수값에 대해서 $\text{Var}[c] = 0$

$$\text{Var}[cX] = c^2 \text{Var}[X]$$

- 또한 기대값 성질을 이용하여, 다음 성질을 증명할 수 있다.

$$\text{Var}[X] = E[X^2] - (E[X])^2 = \Sigma [x^2] - \mu^2$$

- ५८

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

<Pop>

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

$$= \int x^2 - 2mx + m^2$$

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

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

주 확률 변수의 합의 분산

$$- \text{Var}[X+Y] = \text{Var}[X] + \text{Var}[Y] + 2E[(X-\mu_X)(Y-\mu_Y)]$$

〈증명〉

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

- However, X 와 Y 가 독립이면 (μ_X)

$$\begin{aligned} \text{Var}[X+Y] &= \text{Var}[X] + \text{Var}[Y] \\ E[(X-\mu_X)(Y-\mu_Y)] &= 0. \end{aligned}$$

샘플 평균의 분산

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

〈증명〉

$$\begin{aligned} \text{Var}[\bar{X}] &= \text{Var}\left[\frac{1}{N} \sum_{i=1}^N X_i\right] \\ &= E\left[\left(\frac{1}{N} \sum_{i=1}^N X_i - \mu\right)^2\right] \\ &= E\left[\left(\frac{1}{N} \sum_{i=1}^N X_i - \frac{1}{N} N\mu\right)^2\right] \\ &= E\left[\left(\frac{1}{N} \left(\sum_{i=1}^N X_i - N\mu\right)\right)^2\right] \\ &= E\left[\left(\frac{1}{N} \sum_{i=1}^N X_i - N\mu\right)^2\right] \\ &= E\left[\frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N (X_i - \mu)(X_j - \mu)\right] \quad \Rightarrow \text{이차항식과 생간계비.} \end{aligned}$$

- If 서로 독립적 (X_i, X_j) $\Rightarrow E[(X_i - \mu)(X_j - \mu)] = 0$ 라는 사실을 이용하면,
 $i=j$ 일 때, 즉 제곱항만 남는다. $\text{Var}[\bar{X}] = \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N E[(X_i - \mu)^2] = \frac{1}{N^2} \sum_{i=1}^N E[(X_i - \mu)^2] = \frac{1}{N^2} N \text{Var}[X] = \frac{1}{N} \text{Var}[X] = \boxed{\frac{1}{N} \text{Var}[X]}$

생물 분산의 기댓값

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

〈증명〉

$$\begin{aligned} E[S^2] &= E\left[\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2\right] = E\left[\frac{1}{N} \sum_{i=1}^N \{(x_i - \mu) - (\bar{x} - \mu)\}^2\right] \\ &= E\left[\frac{1}{N} \sum_{i=1}^N \{(x_i - \mu)^2 - 2(x_i - \mu)(\bar{x} - \mu) + (\bar{x} - \mu)^2\}\right] \\ &= E\left[\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2\right] - 2E\left[\frac{1}{N} \sum_{i=1}^N (x_i - \mu)(\bar{x} - \mu)\right] + E\left[\frac{1}{N} \sum_{i=1}^N (\bar{x} - \mu)^2\right] \end{aligned}$$

(1) (2) (3)

(1) 항

$$\begin{aligned} 4E\left[\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2\right] &= E\left[\frac{1}{N} \sum_{i=1}^N (x - \mu)^2\right] = E\left[\frac{1}{N} \times N (x - \mu)^2\right] \\ &= E[(x - \mu)^2] = \text{Var}[x] = \underline{\underline{6^2}}. \end{aligned}$$

(2) 항

$$\begin{aligned} 4E\left[\frac{1}{N} \sum_{i=1}^N (x_i - \mu)(\bar{x} - \mu)\right] &= E\left[\frac{1}{N} \sum_{i=1}^N \left(\left(\frac{1}{N} \sum_{j=1}^N x_j - \mu\right) - \mu\right)(\bar{x} - \mu)\right] \\ &= E\left[\frac{1}{N} \sum_{i=1}^N (x_i - \mu) \left(\frac{1}{N} \sum_{j=1}^N (x_j - \mu)\right)\right] \\ &= E\left[\frac{1}{N} \sum_{i=1}^N (x_i - \mu) \left(\frac{1}{N} \sum_{j=1}^{N-1} (x_j - \mu)\right)\right] \end{aligned}$$

생물 분산 = $E\left[\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^{N-1} (x_i - \mu)(x_j - \mu)\right]$

$$= \frac{1}{N} \text{Var}[x] = \underline{\underline{\frac{6^2}{N}}}$$

(3) 항

$$\begin{aligned} 4E\left[\frac{1}{N} \sum_{i=1}^N (\bar{x} - \mu)^2\right] &= E\left[\frac{1}{N} \sum_{i=1}^N \left(\frac{1}{N} \sum_{j=1}^N (x_j - \mu)\right)^2\right] \\ &= E\left[\frac{1}{N} \sum_{i=1}^N \left(\frac{1}{N} \sum_{j=1}^N (x_j - \mu)^2\right)\right] \\ &= E\left[\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^N (x_i - \mu)(x_j - \mu)\right] = \frac{1}{N} \text{Var}[x] = \underline{\underline{\frac{6^2}{N}}} \end{aligned}$$

$\therefore E[S^2] = 6^2 - \frac{26}{N} + \frac{6^2}{N} = \frac{N-1}{N} 6^2$

따라서 평균의 기대값이 정확하게 이산분포의 평균을 구할 때,
 흔히 시이 아니라 시-1이 되어야 한다.

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

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

오멘트

- 앞서 구한 기댓값, 분산등의 특징값을 확률분포의 Moment라고 한다.
- 오멘트는 차수를 갖는데 기댓값은 1차오멘트, 분산은 2차오멘트, ...

- 1차 Moment = $E[X]$: 기댓값(Expectation)
- 2차 Moment = $E[(X-\mu)^2]$: 분산(Variance)
- 3차 Moment = $E[(X-\mu)^3]$: 스케니스(Skewness)
- 4차 Moment = $E[(X-\mu)^4]$: 커터시스(Kurtosis)

- 1) 기대값은 "확률변수의 값이 어느정도 균형에 수렴하는지를 나타내는 특징값."
- 2) 분산은 "그 값 균형에 어느정도 조밀하게 수렴하는지를 나타내는 특징값."
- 3) Skewness는 "전체 분포의 모양이 대칭(Symmetric)인지 아닌지를 표현하는 특징값."
 \hookrightarrow Skewness = 0 이면, 대칭인 확률분포, 양수면 오른쪽으로 치우침.
 4) Kurtosis는 "확률분포가 중앙으로 얼마나 수렴하는지 혹은 반대로 끝쪽이 얼마나 있는지를 나타내는 특징값."

오멘트와 확률분포

오멘트는 두 가지 확률분포에서 찾아낼 수 있는 특징값의 집합이다. 만약 두 개의 확률분포가 있고 1차부터 무한대수까지 모든 오멘트값이 같다면, 같은 확률분포이다.

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

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

$$E[(X-\mu_X)^4] = E[(Y-\mu_Y)^4]$$

$$X = Y$$