Fast campus Copyright FASTCAMPUS Corp.All Rights Reserved

선수 지식 - 통계
공분산과 상관계수

공분산과 상관계수 | 딥러닝의 기초가 되는 확률 개념 알아보기

강사 나동빈

fast campus Copyright FASTCAMPUS Corp.All Rights Reserved

선수 지식 - 통계

공분산과 상관계수

fast campus Copyright FASTCAMPUS Corp.All Rights Reserved

공분산(Covariance)

- 확률 변수가 하나일 때에 대해서 **분산(variance)**을 계산할 수 있었다.
- 만약 **변수가 여러 개일 때**(다변수 확률분포)의 분산은 어떻게 계산할 수 있을까?

fast campus Copyright FASTCAMPUS Corp.All Rights Reserved

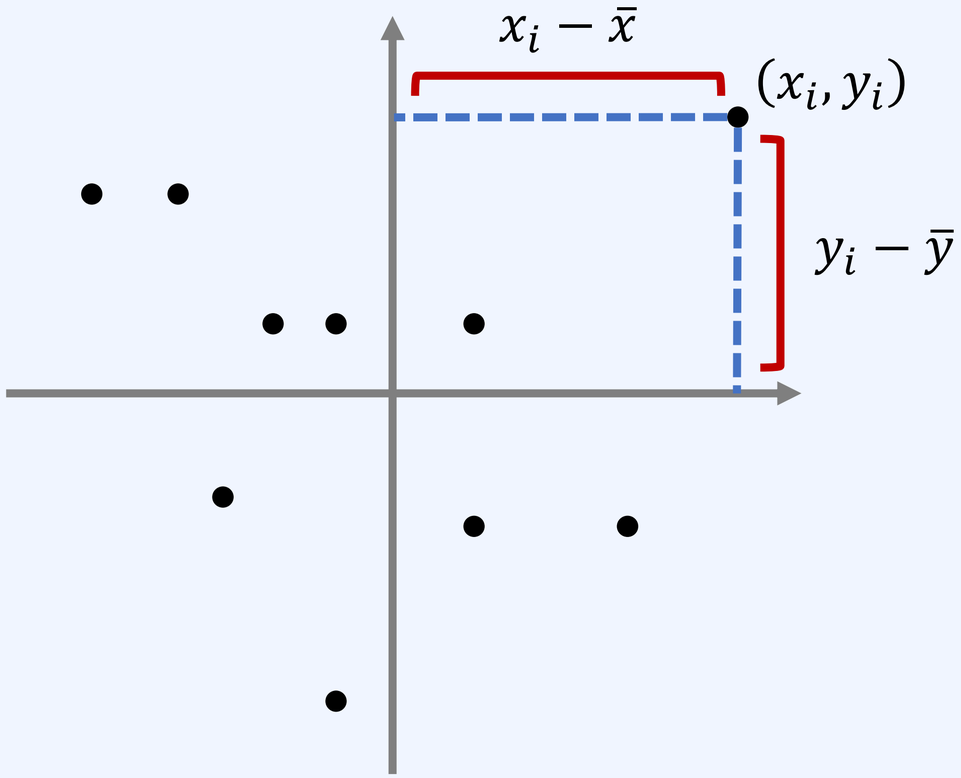
공분산(Covariance)

- 공분산(covariance)의 공식은 다음과 같다.
- $s_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)$
- 분산과 마찬가지로, 데이터가 평균으로부터 얼마나 멀리 떨어져 있는지를 나타낸다.
- 간단히 이해하자면, 평균값의 위치와 표본 위치 사이의 **사각형 면적**을 사용한다.

fast campus Copyright FASTCAMPUS Corp.All Rights Reserved

공분산(Covariance)

- 공분산에서는 평균값의 위치와 표본 위치 사이의 **사각형 면적**을 사용한다.



fast campus Copyright FASTCAMPUS Corp.All Rights Reserved

공분산(Covariance)

- 공분산의 경우 데이터의 위치에 따라서 부호가 다르게 반영된다.
- **양수 부호**: 1사분면, 3사분면
- **음수 부호**: 2사분면, 4사분면

fast campus Copyright FASTCAMPUS Corp.All Rights Reserved

공분산(Covariance)

- 공분산은 데이터가 어떻게 분포되어 있는지에 대한 크기와 방향성을 같이 보여준다.
- **크기**: 원점에서 얼마나 멀리 떨어져 있는지 알 수 있다.
- **방향**: 양수/음수에 따라 어느 방향을 가지는지 알 수 있다.

fast campus Copyright FASTCAMPUS Corp.All Rights Reserved

공분산(Covariance)

- **양의 상관관계**: 공분산이 양수의 값을 가지는 경우
- **음의 상관관계**: 공분산이 음수의 값을 가지는 경우

fast campus Copyright FASTCAMPUS Corp.All Rights Reserved

상관계수(Correlation Coefficient)

- 공분산은 **크기**와 **방향성** 정보를 같이 가지고 있다.
- 우리는 일반적으로 공분산에서 크기 그 자체보다는 **상관성(방향성)**만을 보고자 한다.
- 따라서, 아래의 공식을 이용해 정규화를 진행할 수 있다.
- s_x^2: x의 분산, s_y^2: y의 분산, s_{xy}: x와 y의 공분산

$$r_{xy} = \frac{s_{xy}}{\sqrt{s_x^2 \cdot s_y^2}}$$

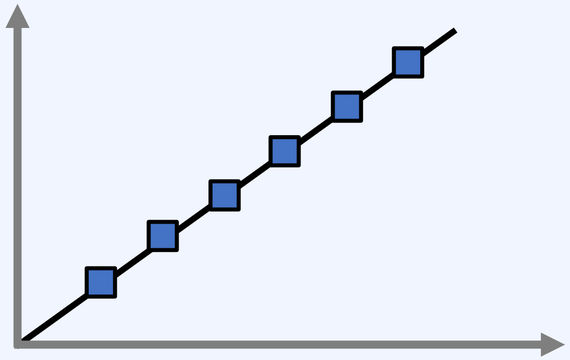
→ 이를 **피어슨(Pearson) 상관계수**라고도 한다.

fast campus Copyright FASTCAMPUS Corp.All Rights Reserved

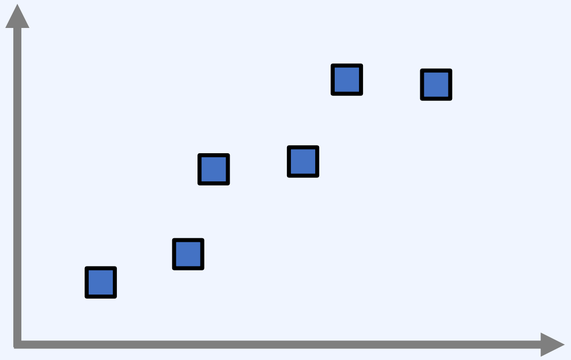
상관계수(Correlation Coefficient)

- **피어슨(Pearson) 상관계수**를 그림으로 이해할 수 있다.

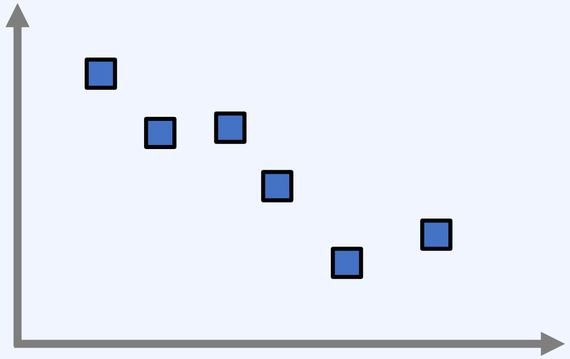
$r = +1$: 완벽한 양의 상관관계



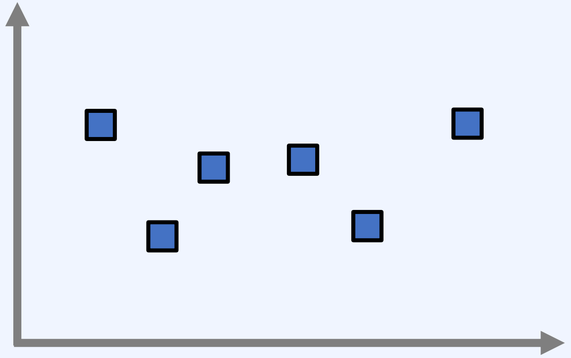
r가 $+1$에 가까울 때: 강한 양수(+) 연관성



r가 -1에 가까울 때: 강한 음수(-) 연관성



r가 0에 가까울 때: 관련이 적음



fast campus Copyright FASTCAMPUS Corp.All Rights Reserved

확률 변수의 공분산과 상관 계수

- 두 확률변수 X와 Y의 공분산은 다음과 같이 정의된다.

$$Cov[X, Y] = E[(X - E[X])(Y - E[Y])]$$

- 두 확률변수 X와 Y의 상관 계수는 다음과 같이 정의된다.

$$\rho[X, Y] = \frac{Cov[X, Y]}{\sqrt{Var[X] \cdot Var[Y]}}$$

- 상관계수(ρ)는 항상 -1 이상, 1 이하의 값을 가진다.
- $\rho = 1$: 완전 선형 양수(+) 상관관계
- $\rho = -1$: 완전 선형 음수(-) 상관관계

fast campus Copyright FASTCAMPUS Corp.All Rights Reserved

공분산 행렬(Covariance Matrix)

- 기계학습 분야에서는 **다변수 확률변수**(벡터 형태의 표본 값)를 가정하는 경우가 많다.
- 예를 들어 하나의 얼굴을 세 개의 특징으로 표현한다고 해보자.

→ 이때, 하나의 데이터는 3개($d = 3$)의 원소를 가지는 벡터이다.

→ 얼굴 데이터 $x = [\text{얼굴 길이}, \text{코 길이}, \text{입술 두께}]$

- 이러한 데이터가 N개 있다고 해보자.
- N개의 얼굴($d = 3$) 데이터를 하나의 행렬로 표현하면 $N \times 3$ 행렬이다.

$$X = \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ & \vdots & \\ x_{N1} & x_{N2} & x_{N3} \end{bmatrix}$$

fast campus Copyright FASTCAMPUS Corp.All Rights Reserved

공분산 행렬(Covariance Matrix)

- 얼굴의 길이와 코의 길이는 양의 상관관계가 있을 것으로 예상할 수 있다.
- **공분산 행렬**을 이용해 그러한 상관 정도가 얼마나 큰지 표현할 수 있다.
- 결과적으로 3개의 서로 다른 확률 변수의 모든 조합에 대하여, 공분산을 한꺼번에 표기할 수 있다.
- 이를 **공분산 행렬(covariance matrix)**라고 한다.

$$\Sigma = \begin{pmatrix} \sigma_{x_1}^2 & \sigma_{x_1x_2} & \sigma_{x_1x_3} \\ \sigma_{x_1x_2} & \sigma_{x_2}^2 & \sigma_{x_2x_3} \\ \sigma_{x_1x_3} & \sigma_{x_2x_3} & \sigma_{x_3}^2 \end{pmatrix}$$

fast campus Copyright FASTCAMPUS Corp.All Rights Reserved

공분산 행렬(Covariance Matrix)

- 공분산 행렬 Σ는 다음과 같이 정의된다. (데이터 개수: N개, 특징의 수: d개일 때)
- **대각 성분(diagonal)**은 각 확률변수의 **분산**이다.
- **비 대각 성분(off-diagonal)**은 두 확률변수의 **공분산**이다.

$$\Sigma = \begin{pmatrix} \sigma_{x_1}^2 & \sigma_{x_1x_2} & \sigma_{x_1x_3} & \cdots & \sigma_{x_1x_d} \\ \sigma_{x_1x_2} & \sigma_{x_2}^2 & \sigma_{x_2x_3} & \cdots & \sigma_{x_2x_d} \\ \sigma_{x_1x_3} & \sigma_{x_2x_3} & \sigma_{x_3}^2 & \cdots & \sigma_{x_3x_d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{x_1x_d} & \sigma_{x_2x_d} & \sigma_{x_3x_d} & \cdots & \sigma_{x_d}^2 \end{pmatrix}$$

fast campus Copyright FASTCAMPUS Corp.All Rights Reserved

공분산과 독립

- 독립(Independent)에 대하여 다시 생각해 보자.
- $P_{XY}(x, y) = P_X(x)P_Y(y)$일 때 독립이었다.
- 결과적으로 X와 Y가 독립이라면 다음의 공식이 성립한다.

$\rightarrow E(XY) = E(X)E(Y)$

- 이때 공분산 $Cov(X, Y) = E(XY) - E(X)E(Y) = 0$이다.

→ 두 확률 변수가 독립이라면, 공분산은 0이다.

[참고] 역은 성립하지 않는다.

→ 공분산이 0이라고 해서 두 확률 변수가 독립이라는 보장은 없다.

fast campus Copyright FASTCAMPUS Corp.All Rights Reserved

공분산(Covariance) 계산 예시 – 수학/영어 성적

[두 데이터가 양의 상관관계를 가지는 경우]

- 확률 변수 X의 값이 크면, Y의 값도 큰 경우를 의미한다.
- 현실에서는 높은 수학 성적을 받은 사람은, 영어 성적도 높은 경향이 있다.
- 성적이 낮은 학생은 수학/영어 모두에서 낮은 성적을 보이는 경향이 있다.
- 반대로 공분산이 음수 값을 가지면, “음의 상관관계”를 가진다고 표현한다.

fast campus Copyright FASTCAMPUS Corp.All Rights Reserved

공분산(Covariance) 계산 예시 – 수학/영어 성적

- 총 10명의 학생에 대하여, **수학과 영어 성적 예시**를 확인해 보자.

학생 번호	수학	영어	학생 번호	수학	영어
1	97	100	6	15	28
2	85	92	7	33	57
3	26	31	8	83	45
4	54	61	9	88	92
5	76	83	10	91	93

fast campus Copyright FASTCAMPUS Corp.All Rights Reserved

공분산(Covariance) 계산 예시 – 수학/영어 성적

- 수학 성적과 영어 성적에 대하여 **평균**, **분산**, **공분산**을 계산할 수 있다.

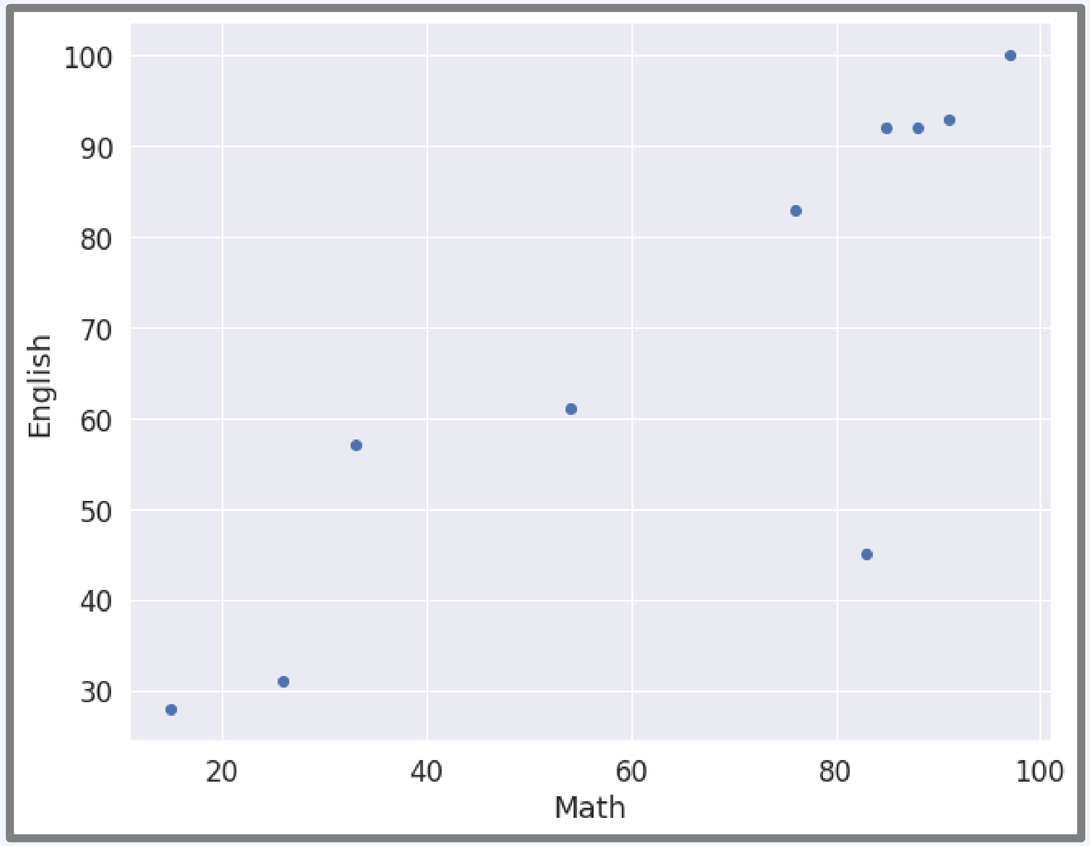
```
import matplotlib.pyplot as plt

X = [97, 85, 26, 54, 76, 15, 33, 83, 88, 91]
Y = [100, 92, 31, 61, 83, 28, 57, 45, 92, 93]
plt.plot(X, Y, 'o')
plt.xlabel("Math")
plt.ylabel("English")
plt.show()
```

fast campus Copyright FASTCAMPUS Corp.All Rights Reserved

공분산(Covariance) 계산 예시 – 수학/영어 성적

- 수학 성적과 영어 성적에 대하여 **평균**, **분산**, **공분산**을 계산할 수 있다.



fast campus Copyright FASTCAMPUS Corp.All Rights Reserved

공분산(Covariance) 계산 예시 – 수학/영어 성적

- 수학 성적과 영어 성적에 대하여 **평균**, **분산**, **공분산**을 계산할 수 있다.

```
x_mean = 0 # 평균(mean)
for x in X:
    x_mean += x / len(X)
x_var = 0 # 분산(variance)
for x in X:
    x_var += ((x - x_mean) ** 2) / (len(X) - 1)

print(f"x_mean = {x_mean:.3f}, x_var = {x_var:.3f}")

y_mean = 0 # 평균(mean)
for y in Y:
    y_mean += y / len(Y)
y_var = 0 # 분산(variance)
for y in Y:
    y_var += ((y - y_mean) ** 2) / (len(Y) - 1)

print(f"y_mean = {y_mean:.3f}, y_var = {y_var:.3f}")
```

[실행 결과]

```
x_mean = 64.800, x_var = 915.511
y_mean = 68.200, y_var = 743.733
```

fast campus Copyright FASTCAMPUS Corp.All Rights Reserved

공분산(Covariance) 계산 예시 – 수학/영어 성적

- 수학 성적과 영어 성적에 대하여 **평균**, **분산**, **공분산**을 계산할 수 있다.

```
import numpy as np
np.set_printoptions(precision=3)
import math

# 공분산(covariance)
covar = 0
for x, y in zip(X, Y):
    covar += ((x - x_mean) * (y - y_mean)) / (len(X) - 1)
print(f"Sample covariance: {covar:.3f}")
print("[Sample covariance (NumPy)]")
print(np.cov(X, Y))

# 상관 계수(correlation coefficient)
correlation_coefficient = covar / math.sqrt(x_var * y_var)
print(f"Correlation coefficient: {correlation_coefficient:.3f}")
print("[Correlation coefficient (NumPy)]")
print(np.corrcoef(X, Y))
```

[실행 결과]

```
Sample covariance: 703.267
[Sample covariance (NumPy)]
[[915.511 703.267]
 [703.267 743.733]]
Correlation coefficient: 0.852
[Correlation coefficient (NumPy)]
[[1.    0.852]
 [0.852 1.   ]]
```

fast campus Copyright FASTCAMPUS Corp.All Rights Reserved

공분산(Covariance) 계산 예시 – 성적과 수면 시간

[두 데이터가 음의 상관관계를 가지는 경우]

- 확률 변수 X의 값이 크면, Y의 값은 작은 경우를 의미한다.
- 중간고사 평균 성적이 높은 경우, 잠을 자는 시간이 적은 경우가 많다.
- 중간고사 평균 성적을 확률 변수 X, 평균 수면 시간을 확률 변수 Y라고 하자.

fast campus Copyright FASTCAMPUS Corp.All Rights Reserved

공분산(Covariance) 계산 예시 – 성적과 수면 시간

- 총 10명의 학생에 대하여, **성적과 수면 시간 예시**를 확인해 보자.

학생 번호	성적	수면 시간	학생 번호	성적	수면 시간
1	97	5.5	6	19	8
2	100	7	7	41	10
3	25	8	8	97	7.5
4	42	9	9	95	6
5	55	8.5	10	91	6

fast campus Copyright FASTCAMPUS Corp.All Rights Reserved

공분산(Covariance) 계산 예시 – 성적과 수면 시간

- 성적과 수면 시간에 대하여 **평균**, **분산**, **공분산**을 계산할 수 있다.

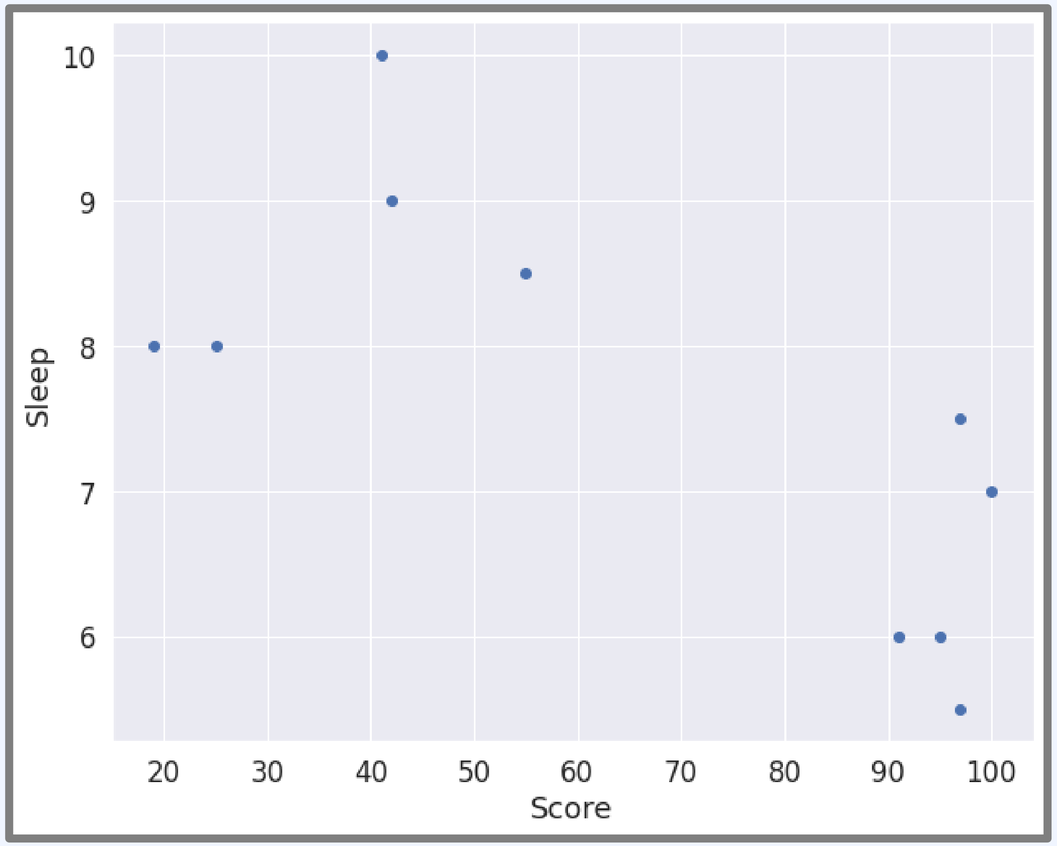
```
import matplotlib.pyplot as plt

X = [97, 100, 25, 42, 55, 19, 41, 97, 95, 91]
Y = [5.5, 7, 8, 9, 8.5, 8, 10, 7.5, 6, 6]
plt.plot(X, Y, 'o')
plt.xlabel("Score")
plt.ylabel("Sleep")
plt.show()
```

fast campus Copyright FASTCAMPUS Corp.All Rights Reserved

공분산(Covariance) 계산 예시 – 성적과 수면 시간

- 성적과 수면 시간에 대하여 **평균**, **분산**, **공분산**을 계산할 수 있다.



fast campus Copyright FASTCAMPUS Corp.All Rights Reserved

공분산(Covariance) 계산 예시 – 성적과 수면 시간

- 성적과 수면 시간에 대하여 **평균**, **분산**, **공분산**을 계산할 수 있다.

```
x_mean = 0 # 평균(mean)
for x in X:
    x_mean += x / len(X)
x_var = 0 # 분산(variance)
for x in X:
    x_var += ((x - x_mean) ** 2) / (len(X) - 1)

print(f"x_mean = {x_mean:.3f}, x_var = {x_var:.3f}")

y_mean = 0 # 평균(mean)
for y in Y:
    y_mean += y / len(Y)
y_var = 0 # 분산(variance)
for y in Y:
    y_var += ((y - y_mean) ** 2) / (len(Y) - 1)

print(f"y_mean = {y_mean:.3f}, y_var = {y_var:.3f}")
```

[실행 결과]

```
x_mean = 66.200, x_var = 1083.956
y_mean = 7.550, y_var = 2.081
```

fast campus Copyright FASTCAMPUS Corp.All Rights Reserved

공분산(Covariance) 계산 예시 – 성적과 수면 시간

- 성적과 수면 시간에 대하여 **평균**, **분산**, **공분산**을 계산할 수 있다.

```
import numpy as np
np.set_printoptions(precision=3)
import math

# 공분산(covariance)
covar = 0
for x, y in zip(X, Y):
    covar += ((x - x_mean) * (y - y_mean)) / (len(X) - 1)
print(f"Sample covariance: {covar:.3f}")
print("[Sample covariance (NumPy)]")
print(np.cov(X, Y))

# 상관 계수(correlation coefficient)
correlation_coefficient = covar / math.sqrt(x_var * y_var)
print(f"Correlation coefficient: {correlation_coefficient:.3f}")
print("[Correlation coefficient (NumPy)]")
print(np.corrcoef(X, Y))
```

[실행 결과]

```
Sample covariance: -34.844
[Sample covariance (NumPy)]
[[1083.956  -34.844]
 [ -34.844    2.081]]
Correlation coefficient: -0.734
[Correlation coefficient (NumPy)]
[[ 1.   -0.734]
 [-0.734  1.  ]]
```