I. 원자료를 이용한 공분산 및 상관계수 계산

II. 결합확률분포표를 이용한 공분산 및 상관계수 계산

1. 공식을 이용한 계산

- 확률변수 X의 분산: $Var(X) = \sigma_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n-1}(\sum_{i=1}^{n}X_i^2 - \overline{X}\sum_{i=1}^{n}X_i)$
- 확률변수 Y의 분산: $Var(Y) = \sigma_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \frac{1}{n-1}(\sum_{i=1}^{n}Y_i^2 - \overline{Y}\sum_{i=1}^{n}Y_i)$
- 확률변수 X의 표준편차: $\sigma_X = \sqrt{(Var(X)}$
- 확률변수 Y의 표준편차: $\sigma_Y = \sqrt{(Var(Y)}$
- 두 확률변수 X, Y의 공분산: $Cov(X,Y) = \sigma_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) = \frac{1}{n-1}(\sum_{i=1}^{n}X_iY_i - \overline{X}\sum_{i=1}^{n}Y_i)$
- 두 확률변수 X, Y의 상관계수: $Corr(X,Y) = \rho_{XY} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}$

(예)

다음 두 확률변수 X,Y의 자료를 이용하여 계산하하

X	2	3	4	5	6
Y	4	4	6	6	10

$$\sum X_i = 20,\ \sum Y_i = 30,\ \sum X_i Y_i = 134,\ \overline{X} = 4,\ \overline{Y} = 6,\ \sum X_i^2 = 90,\ \sum Y_i^2 = 204$$

$$\sum (X_i - \overline{X})^2 = \sum X_i^2 - \overline{X}\sum X_i = 90 - (4)(20) = 10$$

$$\sum (Y_i - \overline{Y})^2 = \sum Y_i^2 - \overline{Y}\sum Y_i = 204 - (6)(30) = 204 - 180 = 24$$

$$\sum (X_i - \overline{X})(Y_i - \overline{Y}) = \sum X_i Y_i - \overline{X}\sum Y_i = 134 - (4)(30) = 14$$

$$\sigma_X^2 = \frac{10}{4} = 2.5$$

$$\sigma_Y^2 = \frac{24}{4} = 6$$

$$\sigma_X = \sqrt{(2.5)} = 1.581139$$

$$\sigma_Y = \sqrt{(6)} = 2.44949$$

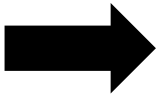
$$\sigma_{XY} = \frac{14}{4} = 3.5$$

$$\rho_{XY} = \frac{3.5}{(1.581139)*(2.44949)} = 0.9036961$$

b1-ch3-cor.py

```
import numpy as np
xx = [2,3,4,5,6]
yy = [4,4,6,6,10]
print("Data of x is :", xx)
print("Data of y is :", yy)
x = np.array(xx)
y = np.array(yy)


n = len(x)
sum_x = np.sum(x)
sum_y = np.sum(y)
m_x = np.mean(x)
m_y = np.mean(y)
sum_xy = np.sum(np.multiply(x, y))
sum_xsq = np.sum(x**2)
sum_ysq = np.sum(y**2)
print("Number of Sample is :", n)
print("Sum of x is :", sum_x)
print("Sum of y is :", sum_y)
print("Mean of x is :", m_x)
print("Mean of y is :", m_y)
print("Sum of y is :", sum_y)
print("Sum of x*y is :", sum_xy)
print("Sum of x-square is :", sum_xsq)
print("Sum of y-square is :", sum_ysq)
```



```
Data of x is : [2, 3, 4, 5, 6]
Data of y is : [4, 4, 6, 6, 10]
Number of Sample is : 5
Sum of x is : 20
Sum of y is : 30
Mean of x is : 4.0
Mean of y is : 6.0
Sum of x*y is : 134
Sum of x-square is : 90
Sum of y-square is : 204
```

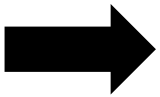
b1-ch3-cor.py

(앞에서 계속)

```
var_x = (sum_xsq-m_x*sum_x)/(n-1)
var_y = (sum_ysq-m_y*sum_y)/(n-1)
std_x = np.sqrt(var_x)
std_y = np.sqrt(var_y)

sum_xy = np.sum(np.multiply(x, y))
cov_xy = (sum_xy-m_x*sum_y)/(n-1)
corr_xy = cov_xy/(std_x*std_y)

print("Variance of x :", var_x)
print("Variance of y :", var_y)
print("Standard deviation of x :", round(std_x, 6))
print("Standard deviation of y :", round(std_y, 6))
print("Covariance of x and y :", cov_xy)
print("Correlation of x and y :", round(corr_xy, 6))
```



```
Variance of x : 2.5
Variance of y : 6.0
Standard deviation of x : 1.581139
Standard deviation of y : 2.44949
Covariance of x and y : 3.5
Correlation of x and y : 0.903696
```

Ⅱ. 결합확률분표를 이용한 공분산 및 상관계수의 계산

- 두 이산형 확률변수의 확률분포표가 주어지면 이를 이용하여 다음을 구할 수 있음
  - 개별 확률변수의 분산 및 표준편차
  - 두 확률변수의 공분산 및 상관계수
- (예 1) 두 확률변수의 공분산 및 상관계수를 계산하라

	-1	6	2	20	f(x)
5	0.1	0	0	0	0.1
7	0	0.4	0	0	0.4
-4	0	0	0.3	0	0.3
15	0	0	0	0.2	0.2
f(y)	0.1	0.4	0.3	0.2	1.0

$$E(X) = \sum xf(x) = 5(0.1) + 7(0.4) + (-4)(0.3) + 15(0.2) = 5.1$$

$$E(Y) = \sum yf(y) = (-1)(0.1) + 6(0.4) + 2(0.3) + 20(0.2) = 6.9$$

$$Var(X) = E(X^2) - [E(X)]^2 = 45.89$$
$$Var(Y) = E(Y^2) - [E(Y)]^2 = 48.09$$

$$E(XY) = \sum \sum xyf(x,y) = 73.9$$

$$Cov(X, Y) = E(XY) - E(X)E(Y) = 73.9 - (5.1)(6.9) = 38.71$$

$$\rho_{X,Y} = \frac{Cov(X, Y)}{\sigma_X \sigma_Y} = \frac{38.71}{\sqrt{45.89}\sqrt{48.09}} = 0.824$$

b1-ch3-11-1.py

```
import numpy as np
from numpy import sqrt
x = np.matrix([[0.1,0,0,0,0.1],[0,0.4,0,0,0.4],[0,0,0.3,0,0.3],[0,0,0,0.2,0.2],[0.1,0.4,0.3,0.2,1]])
x
print("Matrix is : \n", x)
print(type(x))
x_11=x[0,0];x_11
x_12=x[0,1];x_12
x_13=x[0,2];x_13
x_14=x[0,3];x_14
x_15=x[0,4];x_15
x_21=x[1,0];x_21
x_22=x[1,1];x_22
x_23=x[1,2];x_23
x_24=x[1,3];x_24
x_25=x[1,4];x_25
x_31=x[2,0];x_31
x_32=x[2,1];x_32
x_33=x[2,2];x_33
x_34=x[2,3];x_34
x_35=x[2,4];x_35
x_41=x[3,0];x_41
x_42=x[3,1];x_42
x_43=x[3,2];x_43
x_44=x[3,3];x_44
x_45=x[3,4];x_45
x_51=x[4,0];x_51
x_52=x[4,1];x_52
x_53=x[4,2];x_53
x_54=x[4,3];x_54
x_55=x[4,4];x_55
```

```
Matrix is :
 [[0.1 0.  0.  0.  0.1]
 [0.  0.4 0.  0.  0.4]
 [0.  0.  0.3 0.  0.3]
 [0.  0.  0.  0.2 0.2]
 [0.1 0.4 0.3 0.2 1. ]]
```

b1-ch3-11-1.py

(앞에서 계속)

```
mu_x=5*x_15+7*x_25+(-4)*x_35+15*x_45
mu_x
mu_y=(-1)*x_51+6*x_52+2*x_53+20*x_54
mu_y
print("Mean of x is : \n", round(mu_x,6))
print("Mean of y is : \n", round(mu_y,6))

var_x=5**2*x_15+7**2*x_25+(-4)**2*x_35+15**2*x_45-mu_x**2
var_x
var_y=(-1)**2*x_51+6**2*x_52+2**2*x_53+20**2*x_54-mu_y**2
var_y
print("Variance of x is : \n", round(var_x,6))
print("Variance of y is : \n", round(var_y,2))

"""
sum_p_xy = ((-5)*x_11+30*x_12+10*x_13+100*x_14
        +(-7)*x_21+42*x_22+14*x_23+140*x_24
        +(4)*x_31+(-24)*x_32+(-8)*x_33+(-80)*x_34
        +(-15)*x_41+90*x_42+30*x_43+300*x_44)
sum_p_xy
print("Sum of multiplication of p_xy and xy is : \n", sum_p_xy)
"""
```

```
Mean of x is :
 5.1
Mean of y is :
 6.9
```

```
Variance of x is :
 45.89
Variance of y is :
 48.09
```

b1-ch3-11-1.py

(앞에서 계속)

```
xx = np.matrix([[0.1,0,0,0],[0,0.4,0,0],[0,0,0.3,0],[0,0,0,0.2]])
p_xy = np.array(xx)

yy= np.matrix([[-5,30,10,100],[-7,42,14,140],[4,-24,-8,-80],[-15,90,30,
300]])
xy = np.array(yy)

cov_xy = np.sum(p_xy*xy) - mu_x*mu_y
cov_xy
print("Covariance of x and y is :", round(cov_xy,2))

#cov_xy = sum_p_xy - mu_x*mu_y
#cov_xy
#print("Covariance of x and y is :", round(cov_xy,6))

corr_xy = (cov_xy)/sqrt(var_x*var_y)
corr_xy
print("Correlation of x and y is :", round(corr_xy,3))
```

```
Sum of multiplication of p_xy and xy is :
 73.9
```

```
Covariance of x and y is : 38.71
Correlation of x and y is : 0.824
```

- (예 2) 두 확률변수의 공분산 및 상관계수를 계산하라

b1-ch3-11-2.py

```
import numpy as np
from numpy import sqrt

x = np.matrix([[1/6, 1/12, 1/12, 1/3], [1/12, 1/2, 1/12, 2/3], [1/4, 7/12, 1/6, 1]])
x
print("Matrix is : \n", x)
print(type(x))

x_11=x[0,0];x_11
x_12=x[0,1];x_12
x_13=x[0,2];x_13
x_14=x[0,3];x_14
x_21=x[1,0];x_21
x_22=x[1,1];x_22
x_23=x[1,2];x_23
x_24=x[1,3];x_24
x_31=x[2,0];x_31
x_32=x[2,1];x_32
x_33=x[2,2];x_33
x_34=x[2,3];x_34
```

X \ Y	0	1	2	합
1	1/6	1/12	1/12	1/3
3	1/12	1/2	1/12	2/3
합	1/4	7/12	1/6	1

```
Matrix is :
 [[0.16666667 0.08333333 0.08333333 0.33333333]
 [0.08333333 0.5        0.08333333 0.66666667]
 [0.25       0.58333333 0.16666667 1.        ]]
```

b1-ch3-11-2.py

(앞에서 계속)

```
mu_x=1*x_14+3*x_24
mu_x
mu_y=0*x_31+1*x_32+2*x_33
mu_y
print("Mean of x is : \n", round(mu_x,6))
print("Mean of y is : \n", round(mu_y,6))

var_x=1**2*x_14+3**2*x_24-mu_x**2
var_x
var_y=0**2*x_31+1**2*x_32+2**2*x_33-mu_y**2
var_y
print("Variance of x is : \n", round(var_x,6))
print("Variance of y is : \n", round(var_y,6))

sum_p_xy = 0*x_11+1*x_12+2*x_13+0*x_21+3*x_22+6*x_23
sum_p_xy
print("Sum of multiplication of p_xy and xy is : \n", sum_p_xy)

cov_xy = sum_p_xy - mu_x*mu_y
cov_xy
print("Covariance of x and y is :", round(cov_xy,6))

corr_xy = (cov_xy)/sqrt(var_x*var_y)
corr_xy
print("Correlation of x and y is :", round(corr_xy,6))
```

```
Mean of x is :
 2.333333
Mean of y is :
 0.916667
```

```
Variance of x is :
 0.888889
Variance of y is :
 0.409722
```

```
Covariance of x and y is : 0.111111
Correlation of x and y is : 0.184115
```