R 공분산 및 상관분석

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

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

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

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

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

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

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

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

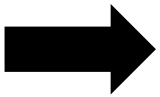
$$\sigma_Y = \sqrt{(6)} = 2.44949$$

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

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

b1-ch3-cor.R

```
x<-c(2,3,4,5,6)
y<-c(4,4,6,6,10)

(sumx<-sum(x))
(sumy<-sum(y))
(mx=mean(x))
(my=mean(y))
(xy<-x*y)
(sumxy<-sum(xy))
(sumxsq<-sum(x^2))
(sumysq<-sum(y^2))

(varx<-(sumxsq-mx*sumx)/4)
(vary<-(sumysq-my*sumy)/4)

(sdx<-sqrt(varx))
(sdy<-sqrt(vary))

(covxy<-(sumxy-mx*sumy)/4)
(corxy<-(covxy/(sdx*sdy)))
```



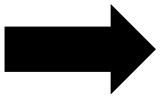
```
> x<-c(2,3,4,5,6)
> y<-c(4,4,6,6,10)
>
> (sumx<-sum(x))
[1] 20
> (sumy<-sum(y))
[1] 30
> (mx=mean(x))
[1] 4
> (my=mean(y))
[1] 6
> (xy<-x*y)
[1]  8 12 24 30 60
> (sumxy<-sum(xy))
[1] 134
> (sumxsq<-sum(x^2))
[1] 90
> (sumysq<-sum(y^2))
[1] 204
>
> (varx<-(sumxsq-mx*sumx)/4)
[1] 2.5
> (vary<-(sumysq-my*sumy)/4)
[1] 6
>
> (sdx<-sqrt(varx))
[1] 1.581139
> (sdy<-sqrt(vary))
[1] 2.44949
>
> (covxy<-(sumxy-mx*sumy)/4)
[1] 3.5
> (corxy<-(covxy/(sdx*sdy)))
[1] 0.9036961
```

2. 함수를 이용한 계산

b1-ch3-cor.R

앞에서 계속

```
(vx<-var(x))
(vy<-var(y))

(sx<-sd(x))
(sy<-sd(y))

(cov(x,y))
(cor(x,y))

nx<-10*x
ny<-10*y

(vnx<-var(nx))
(vny<-var(ny))

(snx<-sd(nx))
(sny<-sd(ny))

(cov(nx,ny))
(cor(nx,ny))
```



```
> (vx<-var(x))
[1] 2.5
> (vy<-var(y))
[1] 6
>
> (sx<-sd(x))
[1] 1.581139
> (sy<-sd(y))
[1] 2.44949
>
> (cov(x,y))
[1] 3.5
> (cor(x,y))
[1] 0.9036961
> nx<-10*x
> ny<-10*y
>
> (vnx<-var(nx))
[1] 250
> (vny<-var(ny))
[1] 600
>
> (snx<-sd(nx))
[1] 15.81139
> (sny<-sd(ny))
[1] 24.4949
>
> (cov(nx,ny))
[1] 350
> (cor(nx,ny))
[1] 0.9036961
```

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

X \ Y	-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-new-loop.R

```
data<-c(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)
mat<-matrix(data, nrow=5, byrow=T)
rownames(mat)<-c("5", "7", "-4","15","f(y)")
colnames(mat)<-c("-1", "6", "2", "20","f(x)")
mat
mu_x<-5*mat[1,5]+7*mat[2,5]+(-4)*mat[3,5]+15*mat[4,5]
mu_y<-(-1)*mat[5,1]+6*mat[5,2]+2*mat[5,3]+20*mat[5,4]
mu_x
mu_y
var_x<-5^2*mat[1,5]+7^2*mat[2,5]+(-4)^2*mat[3,5]+15^2*mat[4,5]-mu_x^2
var_y<-(-1)^2*mat[5,1]+6^2*mat[5,2]+2^2*mat[5,3]+20^2*mat[5,4]-mu_y^2
var_x
var_y
p_xy<-data[c(1:4,6:9,11:14,16:19)] ; p_xy
xy<-c(-5,30,10,100,-7,42,14,140,4,-24,-8,-80,-15,90,30,300)
x<-c(5,7,-4,15)
y<-c(-1,6,2,20)
z<-matrix(data=NA, nrow=4, ncol=4, byrow=T)
for(i in 1:4) {
  for(j in 1:4) {
     z[i,j]<-x[i]*y[j]    }
   }
xy_z<-c(z[1,],z[2,],z[3,],z[4,]) ; xy_z
mu_xy<-sum(p_xy*xy) ; mu_xy
cov_xy<-sum(p_xy*xy)-(mu_x*mu_y) ; cov_xy
corr_xy<-cov_xy/sqrt(var_x*var_y) ; corr_xy
```

```
> mat
      -1   6   2  20 f(x)
5    0.1 0.0 0.0 0.0  0.1
7    0.0 0.4 0.0 0.0  0.4
-4   0.0 0.0 0.3 0.0  0.3
15   0.0 0.0 0.0 0.2  0.2
f(y) 0.1 0.4 0.3 0.2  1.0
```

```
> mu_x
[1] 5.1
> mu_y
[1] 6.9
```

```
> var_x
[1] 45.89
> var_y
[1] 48.09
```

```
> mu_xy
[1] 73.9
```

```
> cov_xy
[1] 38.71
```

```
> corr_xy
[1] 0.8240181
```

```
> p_xy
 [1] 0.1 0.0 0.0 0.0 0.0 0.4 0.0 0.0 0.0 0.0 0.3 0.0 0.0 0.0 0.0 0.2
```

- **(예 2) 두 확률변수의 공분산 및 상관계수를 계산하라**

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

b1-ch3-11-rev.R

```
data<-c(0.167, 0.083, 0.083, 0.333, 0.083, 0.5, 0.083, 0.667, 0.25, 0.583, 0.167, 1)
mat<-matrix(data, nrow=3, byrow=T)
rownames(mat)<-c("1", "3", "합계")
colnames(mat)<-c("0", "1", "2", "합계")
mat
mu_x<-1*mat[1,4]+3*mat[2,4]
mu_y<-0*mat[3,1]+1*mat[3,2]+2*mat[3,3]
mu_x
mu_y
var_x<-1^2*mat[1,4]+3^2*mat[2,4]- mu_x^2
var_y<-0^2*mat[3,1]+1^2*mat[3,2]+2^2*mat[3,3]-mu_y^2
var_x
var_y
p_xy<-c(0.167, 0.083, 0.083, 0.083, 0.5, 0.083)
xy<-c(0,1,2,0,3,6)
cov_xy<-sum(p_xy*xy)-(mu_x*mu_y)
cov_xy
corr_xy<-cov_xy/sqrt(var_x*var_y)
corr_xy
```

```
> mat
          0     1     2  합계
1     0.167 0.083 0.083 0.333
3     0.083 0.500 0.083 0.667
합계 0.250 0.583 0.167 1.000
```

```
> mu_x
[1] 2.334
> mu_y
[1] 0.917
```

```
> var_x
[1] 0.888444
> var_y
[1] 0.410111
```

```
> cov_xy
[1] 0.106722
```

```
> corr_xy
[1] 0.1768024
```

b1-ch3-11-rev.R

(계속)

```
frac<-c("1/6","1/12","1/12","1/3","1/12","1/2","1/12","2/3","1/4","7/12","1/6","1")
dd<-sapply(frac, function(x) eval(parse(text=x)))
d<-as.numeric(dd)
matd<-matrix(d, nrow=3, byrow=T)
rownames(matd)<-c("1", "3", "합계")
colnames(matd)<-c("0", "1", "2", "합계")
matd
mu_x_d<-1*matd[1,4]+3*matd[2,4]
mu_y_d<-0*matd[3,1]+1*matd[3,2]+2*matd[3,3]
mu_x_d
mu_y_d
var_x_d<-1^2*matd[1,4]+3^2*matd[2,4]- mu_x_d^2
var_y_d<-0^2*matd[3,1]+1^2*matd[3,2]+2^2*matd[3,3]-mu_y_d^2
var_x_d
var_y_d
frac_d <- c("1/6","1/12","1/12","1/12","1/2","1/12")
ddd<-sapply(frac_d, function(x) eval(parse(text=x)))
d_d<-as.numeric(ddd)
p_xy_d<-d_d
xy<-c(0,1,2,0,3,6)
cov_xy_d<-sum(p_xy_d*xy)-(mu_x_d*mu_y_d)
cov_xy_d
corr_xy_d<-cov_xy_d/sqrt(var_x_d*var_y_d)
corr_xy_d
```

```
> matd
                0         1         2      합계
1    0.16666667 0.08333333 0.08333333 0.3333333
3    0.08333333 0.50000000 0.08333333 0.6666667
합계 0.25000000 0.58333333 0.16666667 1.0000000
```

```
> mu_x_d
[1] 2.333333
> mu_y_d
[1] 0.9166667
```

```
> var_x_d
[1] 0.8888889
> var_y_d
[1] 0.4097222
```

```
> cov_xy_d
[1] 0.1111111
```

```
> corr_xy_d
[1] 0.1841149
```