

공분산과 상관계수

- 📖 공분산과 상관계수의 사용
- 📖 계산 방법 및 성질
- 📖 R-언어의 함수

Loren ipsum dolor sit amet, ius an molestie facilisi erroribus, mutat natorum delectus ei vis. Has ornatus conclusionemque id, an vide molestatis sit. In etqui praesent sit. An vel agan porro comprehensan, ad ludus constituto nea, et ius utroque scaevola assuaverit.


Vis cu nodus nulla feugait, oratio facilisi ex usu, eili vitae sea te. Ea fabulas accusamus dissentias sea, facete tacinates definitiones et per. Nihil dicant mediocrem pro eu, no mei nostro sensibus platonem. Qui id sunno perpetus neglegantur. Vel ipsum novum copiosae ut. Quo et liber detracto probatus. Nam augue scribentur an. Sea oporteat percipitur incidereit at. Qui viris nemore an.





공분산과 상관계수의 사용



 키와 몸무게가 어느 정도 상관성이 있는지, 소득과 저축의 관계는 어떠한지와 같이 두 속성의 연관을 수치로 나타내기 위한 방법으로 공분산이나 상관계수를 사용할 수 있음

 공분산은 같은 자료라도 측정단위에 따라 달라지므로 주로 상관계수가 사용됨





계산 방법 및 성질



공분산의 계산방법

- + 두 속성 x 와 y 에 대한 자료가 $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ 로 n 개의 짝으로 얻어진 경우 두 자료의 공분산은 다음과 같이 정의함

$$S_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n - 1} = \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{n - 1}$$

- + x 가 증가(감소)할 때 y 가 증가(감소)하면 공분산은 양의 값을 가짐
- + x 가 증가(감소)할 때 y 가 감소(증가)하면 공분산은 음의 값을 가짐





계산 방법 및 성질



상관계수의 계산방법

- + 상관계수라고 하면 보통 피어슨(Pearson)의 선형상관계수를 말하며 표본에서 얻은 상관계수는 보통 r 로 표시함
- + 상관계수 r 은 다음과 같이 정의함

$$r = \frac{Cov(x, y)}{S_x S_y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

- + 여기서 S_x 와 S_y 는 각각 x 와 y 의 표준편차

S_y





계산 방법 및 성질



상관계수의 계산방법



상관계수 r

- $-1 \leq r \leq 1$ 의 값을 가짐
- x 가 증가(감소)할 때 y 가 증가(감소)하면 상관계수는 양의 값을 가지며, x 가 증가(감소)할 때 y 가 감소(증가)하면 상관계수는 음의 값을 가짐
- 상관이 높을수록 상관계수의 절댓값이 1에 가까워지며, $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ 의 모든 점이 한 직선 위에 존재하면 상관계수는 1 또는 -1 의 값을 가짐
- 두 변수 독립이면 표본에서 얻은 상관계수 r 은 0에 가까운 값을 가지며, 역은 성립하지 않음
 - 즉 상관계수가 0(또는 0에 가까운 값)이라고 하더라도 두 변수는 독립이 아닐 수 있음



계산 방법 및 성질

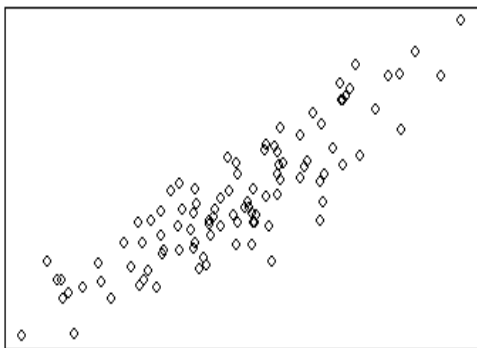


상관계수의 성질

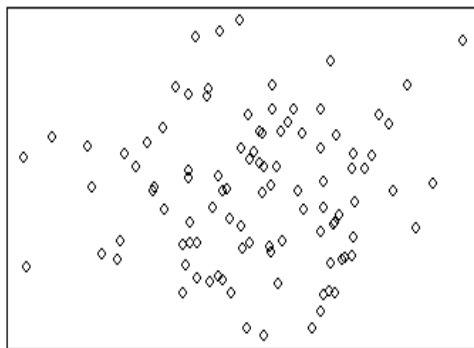


상관계수의 값에 따른 산점도

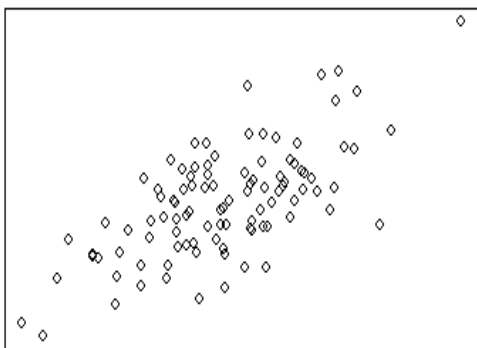
$\rho = 0.9$



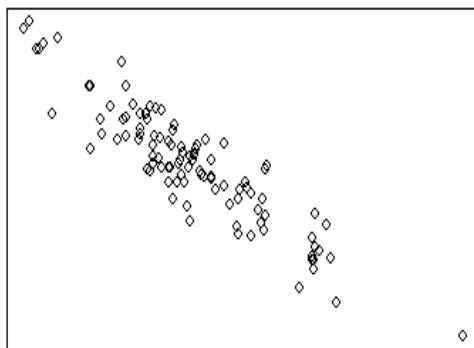
$\rho = 0$



$\rho = 0.6$



$\rho = -0.9$



- 상관계수가 양수이면 양의 상관을 보이며, 음수인 경우 음의 상관을 보임
- 상관계수의 절대값이 클수록 상관이 강함





R-언어의 함수



cor, cov 함수의 사용



상관계수 및 공분산은 각각 cor과 cov함수를 사용하여 얻을 수 있음



사용함수

```
cov(x, y = NULL, use=, ...)  
cor(x, y = NULL, use=, ...)
```



매개변수

- x : 상관계수 또는 공분산을 계산할 벡터, 행렬, 데이터 프레임 등을 설정함
- y : x가 벡터일 경우 y가 반드시 있어야 하며 x가 행렬이나 데이터 프레임과 같이 둘 이상의 열을 포함한 자료인 경우 x만 설정할 수 있음

- x, y가 모두 설정된 경우 x의 i번째 열과 y의 j번째 열의 상관계수 또는 공분산을 모든 가능한 조합에 대해서 계산하고 그 결과를 행렬로 반환함



R-언어의 함수



사용함수



매개변수

use= "everything", "all.obs", "complete.obs", "na.or.complete", "pairwise.complete.obs" 중의 하나로 설정하며 결측치를 어떻게 처리할지 설정함

- "everything" : 기본값이며 NA가 한 값이라도 있으면 모두 결과는 NA가 됨
- "all.obs" : 한 값이라도 NA가 있으면 결측치가 있다는 에러 메시지가 출력됨
- "complete.obs" : 분산의 계산에서는 각 변수에서 결측치를 제외하고 계산하고 공분산 계산은 x, y 둘 중의 한 값이라도 결측치이면 그 쌍을 제외하고 계산함
 - 모든 쌍의 자료가 하나 이상의 결측치를 가지면 에러 메시지를 출력함
- "na.or.complete" : "complete.obs"과 같은 방식으로 계산하나 모든 쌍의 자료가 하나 이상의 결측치를 가지면 NA를 출력함
- "pairwise.complete.obs" : x, y 둘 중의 한 값이라도 결측치이면 그 쌍을 제외하고 계산함
 - 모든 쌍의 자료가 하나 이상의 결측치를 가지면 에러 메시지를 출력함



R-언어의 함수



예시 1



단위에 따른 공분산과 상관계수

공분산은 단위에 따라 달라지나 상관계수는 측정단위에 무관함을 확인하기 위해 같은 자료 다른 단위인 5명의 학생의 키와 몸무게의 공분산과 상관계수 확인



우리나라 학생

```
htcm <- c(170, 179, 174, 167, 175)
wtkg <- c(62, 67, 70, 60, 72)
```



미국 학생

```
(1kg=2.2lb, 1cm=0.394inch)
htin <- c(67.0, 70.5, 68.6, 65.8, 69.0)
wtlb <- c(136.4, 147.4, 154.0, 132.0, 158.4)
```



R-언어의 함수



예시 1

- + 같은 자료에서 다른 단위를 사용한 앞의 두 경우에 대해서 공분산 계산

```
> cov(htcm, wtkg)
```

```
[1] 17.5
```

```
> cov(htin, wtlb)
```

```
[1] 15.356
```

* 서로 다른 공분산 값이 얻어짐

- + 같은 방법으로 상관계수 계산

```
> cor(htcm, wtkg)
```

```
[1] 0.7373406
```

```
> cor(htin, wtlb)
```

```
[1] 0.7481781
```

- 두 상관계수의 값이 약간 달라 보이지만 이는 cm -> inch, kg -> lb로 단위변경을 할 때 반올림이 발생하여 난 차이이며 실제로는 같은 상관계수 값이어야 함



R-언어의 함수



예시 1

+ 정확한 값

```
> htin <- htcm*.394  
> wtlb <- wtkg*2.2  
> htin  
[1] 66.980 70.526 68.556 65.798 68.950  
> wtlb  
[1] 136.4 147.4 154.0 132.0 158.4
```

+ 정확한 값 사용 결과

```
> cov(htcm, wtkg)  
[1] 17.5  
> cov(htin, wtlb)  
[1] 15.169  
> cor(htcm, wtkg)  
[1] 0.7373406  
> cor(htin, wtlb)  
[1] 0.7373406
```

* 공분산의 값은 다르지만 상관계수의 값은 두 단위 모두에 대해서 같은 값을 얻음



R-언어의 함수



예시 2



결측치 포함된 자료의 처리

다음의 자료는 결측치를 포함한 자료이다.



결측치를 포함한 자료

```

> x1 <- c(1, NA, 2, 3, NA, 4, 5, NA, 6, 7, NA, 8, 9, 10, 11)
> x2 <- c(NA, 1, 2, NA, 3, 4, NA, 6, 7, NA, 8, 9, 9, 11, 12)
> x3 <- c(1, 2, NA, 3, 4, NA, 5, 6, NA, 7, 8, NA, 9, 10, 11)
> x <- cbind(x1, x2, x3)
> t(x)

```

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]	[,11]	[,12]	[,13]	[,14]	[,15]
x1	1	NA	2	3	NA	4	5	NA	6	7	NA	8	9	10	11
x2	NA	1	2	NA	3	4	NA	6	7	NA	8	9	9	11	12
x3	1	2	NA	3	4	NA	5	6	NA	7	8	NA	9	10	11



R-언어의 함수



예시 2



결측치를 포함한 자료

```
> cor(x, use="everything")
```

하나라도 결측치이면 NA

	x1	x2	x3
x1	1	NA	NA
x2	NA	1	NA
x3	NA	NA	1

```
> cor(x, use="all.obs")
```

하나라도 결측치이면 에러 메시지

Error in cor(x, use = "all.obs") : cov/cor에 결측치들이 있습니다

```
> cor(x, use="pairwise.complete.obs")
```

각(x1, x2), (x1, x3), (x2, x3) pair 들끼리 계산

	x1	x2	x3
x1	1.0000000	0.9931966	1.0000000
x2	0.9931966	1.0000000	0.9973115
x3	1.0000000	0.9973115	1.0000000



R-언어의 함수



예시 2



결측치를 포함한 자료

```
> cor(x, use="complete")
```

모든 x1, x2, x3가 NA가 아닌 경우만 계산(없으면 에러)

	x1	x2	x3
x1	1.0000000	0.9819805	1.0000000
x2	0.9819805	1.0000000	0.9819805
x3	1.0000000	0.9819805	1.0000000

```
> cor(x, use="na.or.complete")
```

모든 x1, x2, x3가 NA가 아닌 경우만 계산(없으면 NA)

	x1	x2	x3
x1	1.0000000	0.9819805	1.0000000
x2	0.9819805	1.0000000	0.9819805
x3	1.0000000	0.9819805	1.0000000



R-언어의 함수



예시 2



결측치를 포함한 자료

```
> cor(x[1:12,], use="complete")
```

Error in cor(x[1:12,], use = "complete") : no complete element pairs

← # 모든 x1, x2, x3가 NA가 아닌 경우만 계산(없으면 에러)

```
> cor(x[1:12,], use="na.or.complete")
```

	x1	x2	x3
x1	NA	NA	NA
x2	NA	NA	NA
x3	NA	NA	NA

← # 모든 x1, x2, x3가 NA가 아닌 경우만 계산(없으면 NA)



R-언어의 함수



예시 3



자료

BMI 자료에서 키와 몸무게의 상관계수 및 공분산을 계산해보고, 이 결과를 상관계수를 공식을 적용하여 확인해보자.

사용할 자료

```
> BMI <- read.table(url("http://jupiter.hallym.ac.kr/ftpdata/data/bmi.txt"),  
  col.names=c("height", "weight", "year", "religion", "gender", "marriage"))
```



내용 : 2000년, 177명에 대한 조사 결과

- 키, 몸무게, 출생년도
- 종교(Bu=불교, C1=개신교, C2=가톨릭, No=없음)
- 성별(F=여자, M=남자)
- 결혼여부(N=미혼, Y=기혼)



R-언어의 함수



예시 3

+ 키와 몸무게의 공분산

```
> cov(BMI$height, BMI$weight)  
[1] 25.42062
```

+ 키와 몸무게의 상관계수

```
> cor(BMI$height, BMI$weight)  
[1] 0.6473976
```

+ 상관계수의 공식 적용

```
> cov(BMI$height, BMI$weight)/(sd(BMI$height)*sd(BMI$weight)) # 공식적용  
[1] 0.6473976
```

R-언어의 함수

예시 4



독립인 경우의 상관계수

표준정규분포로부터 100개씩의 난수를 세 번 얻어 이를 100×3행렬에 저장하여 이의 상관계수를 얻어보자.



세 개의 열을 가진 자료를 x로 얻기

```
> x <- cbind(rnorm(100), rnorm(100), rnorm(100)) # 100X3 행렬
```



상관계수 얻기

```
> cor(x)
```

	[,1]	[,2]	[,3]
[1,]	1.00000000	0.04526954	-0.01003914
[2,]	0.04526954	1.00000000	-0.08316823
[3,]	-0.01003914	-0.08316823	1.00000000

➤ 이 행렬의 (i,j) 번째 원소는 x의 i번째 열과 j번째 열의 상관계수임

- (1,2)번째 와 (2,1)번째 원소인 0.045는 모두 첫 번째 열과 두 번째 열의 상관계수임

➤ 대각선은 (i,i)번째 원소이고 이들은 같은 변수의 상관계수이므로 모두 1임



R-언어의 함수



예시 4

- + 난수는 모두 독립적으로 얻어지므로 이론적인 상관계수는 0이며 자료에서 상관계수를 계산할 때는 0에 가까운 값을 얻게 됨
 - 모집단 평균이 0인 모집단에서 100개를 뽑아 표본평균을 계산하면 0이 아니라 0에 가까운 값을 얻는 것과 같은 원리임



상관계수에 대한 이해


📖 변수와 상관계수와의 관계

📖 상관계수는 0이지만 두 변수 사이에 관계가 있는 경우



변수와 상관계수와의 관계



 두 변수가 독립이면 상관계수가 0(자료에서는 0에 가까운 값)이지만 상관계수가 0이라고 해서 두 변수가 독립이라고 할 수 없는 경우가 많음

 이 경우는 두 변수의 관계가 직선이 아닌 관계에 있는 경우에 많이 발생함





상관계수는 0이지만 두 변수 사이에 관계가 있는 경우



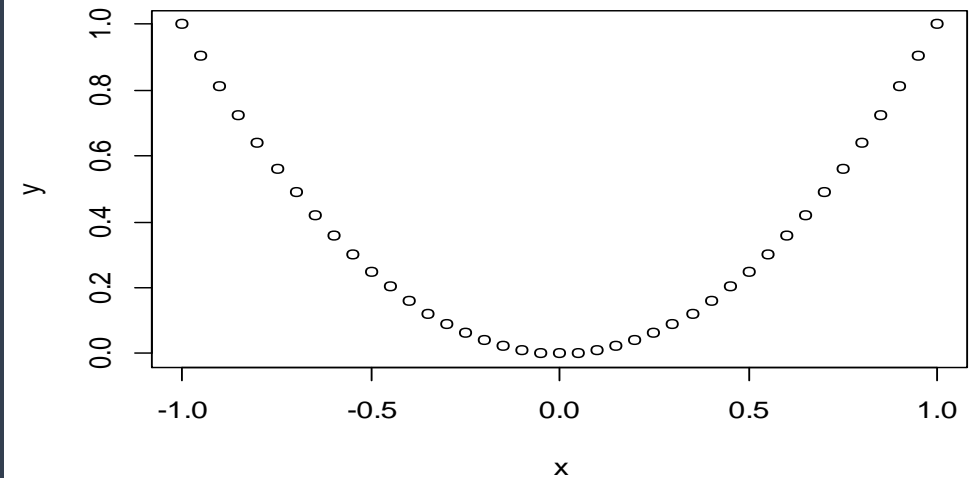
예시



자료

다음과 같이 x 와 y 를 정의하고 두 변수 사이의 상관계수를 계산해보고 산점도로 두 변수 사이의 관계를 확인해보자.

```
> x <- seq(-1,1, by=0.05)
> y <- x^2 #  $y = x^2$ 인 관계
> cor(x,y)
[1] 1.775033e-16
> plot(x,y)
```



- x 는 -1에서 1 사이의 값이고, $y=x^2$ 으로 $y = x^2$ 인 포물선을 이루는 관계가 있음
- x 와 y 사이 상관계수는 0임
- 위 산점도에서 보는 것과 같이 x 가 증가하면 y 는 감소하다가 증가하는 패턴이 뚜렷하게 보임



상관계수는 0이지만 두 변수 사이에 관계가 있는 경우



피어슨(Pearson) 선형 상관계수

- + 선형(linear)은 '직선'라는 의미로 이 상관계수는 직선관계는 수치로 잘 표현할 수 있지만 직선이 아닌 관계에 대해서는 잘 표현할 수 없음을 말함
- + 상관에 대한 분석에서는 상관계수만 보면 잘못된 판단을 내릴 수 있으므로 반드시 산점도를 확인하여 직선 이외의 관계가 있는지 확인할 필요가 있음

