

제3장

상관분석



제3장 상관분석

◆ 서론(introduction)

- 자연 및 사회현상의 규명에 있어서 관련된 변수들 간의 상호관련성을 갖게 되는 경우가 흔히 있음.
- 예를 들어 가계소득과 저축, 흡연량과 폐암발병률 등이 있음.
- 상관분석(correlation analysis)은 이와 같이 두 변수간의 선형관계가 존재하는지 또는 존재하지 않는지를 분석함. 즉, 상관분석은 변수들 간의 선형성의 강도에 대한 통계적 분석이라 할 수 있음.
- 따라서 변수들 간의 구체적인 함수관계를 파악하는 것이 아님.
- 그러나 한 변수의 값으로부터 다른 변수의 값을 예측하고자 하는 경우(예를 들어 가계소득으로부터 저축을 예측하는 경우), 즉 변수들 간의 구체적인 함수관계를 파악하고자 하는 경우에는 회귀분석(regression analysis)이라는 통계적 분석이 사용됨.



제3장 상관분석

◆ 공분산(covariance)

· 연관성의 척도

- 만약 두 변수가 독립이 아니라면 변수들 간에 어떤 연관성(association)이 존재할 것이고, 그 연관성 정도는 높을 수도 있고 또는 낮을 수도 있음.
- 변수들 간에 연관성 정도는 여러 가지 방법에 의하여 측정할 수 있음.
- 그 중 하나인 공분산(covariance)을 설명할 때 두 변수는 크기가 측정되는 수량변수(metric variable), 즉 질적 변수(qualitative variable)가 아닌 양적 변수(quantitative variable)어야 함.



제3장 상관분석

◆ 공분산(covariance)

- 연관성의 방향

- 두 변수간의 연관성은 방향을 갖게 됨. 즉, 한 변수의 값이 커질 때(작아질 때) 다른 변수의 값도 커지면(작아지면) 두 변수는 정(+)의 연관성이 있고, 반대로 한 변수의 값이 커질 때(작아질 때) 다른 변수의 값이 작아지면(커지면) 두 변수는 부(-)의 연관성이 있다고 함.
- 예를 들어 사람들의 키와 몸무게간에는 정(+)의 연관성이 있고, 흡연량과 기대수명간에는 부(-)의 연관성이 있을 것임.

- 공분산의 정의

- 두 변수간의 선형 연관성을 나타내는 공분산(covariance)은 다음과 같이 정의됨.

$$\text{Cov}(X, Y) = \sigma_{XY} = E(X - \mu_X)(Y - \mu_Y), \text{ 단, } \mu_X = E(X), \mu_Y = E(Y)$$

- 상관계수(correlation coefficient)는 공분산으로부터 유도됨.



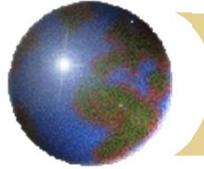
제3장 상관분석

◆ 공분산(covariance)

- 공분산의 의미

- 공분산은 두 변수 X , Y 가 서로 어떤 패턴(pattern)을 보여주는가를 나타냄.

- $\text{Cov}(X, Y) > 0$ 이면 X 가 증가(감소)할 때 Y 도 증가(감소)
 - $\text{Cov}(X, Y) < 0$ 이면 X 가 증가(감소)할 때 Y 는 감소(증가)
 - $\text{Cov}(X, Y) = 0$ 이면 두 변수는 아무런 상관이 없음.

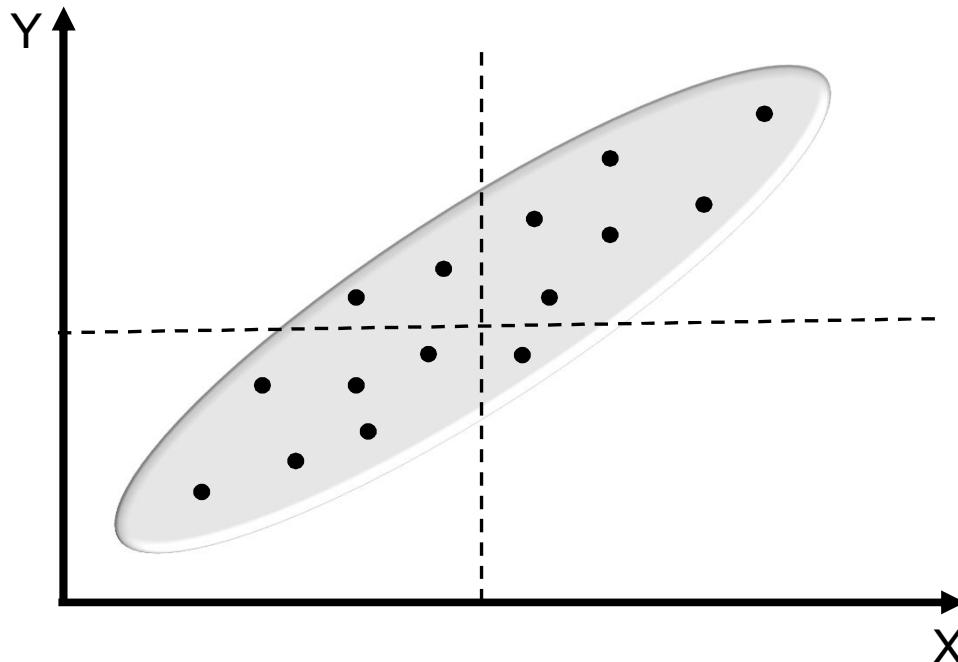


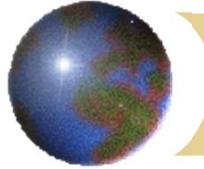
제3장 상관분석

- ◆ 공분산(covariance)

- 공분산의 의미

- 공분산이 정(+)의 값을 가지는 경우 많은 관측값치들은 1사분면과 3사분면에 분포함.



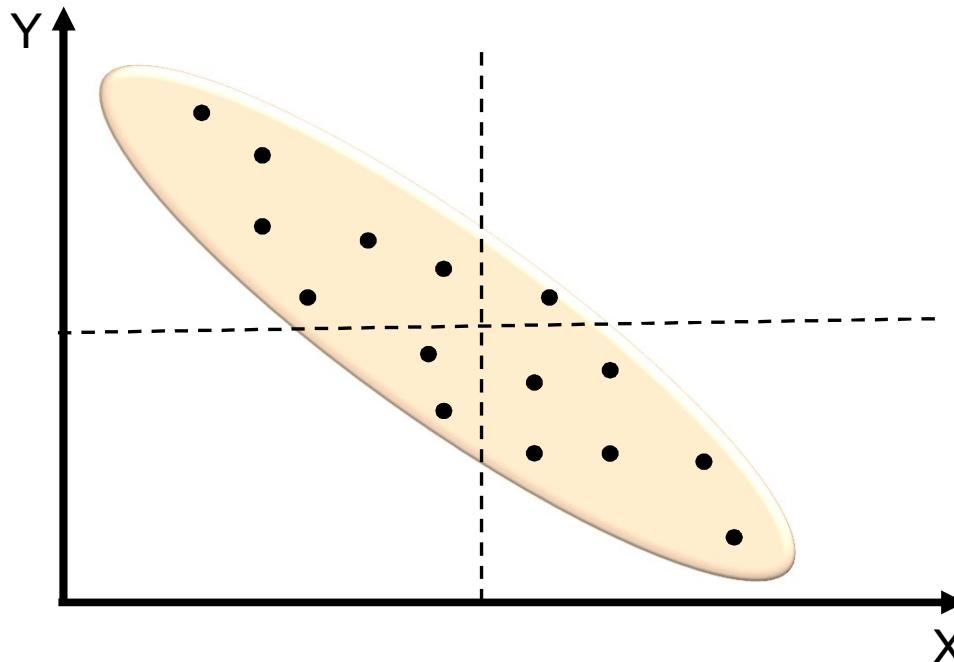


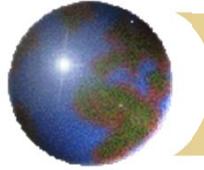
제3장 상관분석

- ◆ 공분산(covariance)

- 공분산의 의미

- 공분산이 부(-)의 값을 가지는 경우 많은 관측값들은 2사분면과 4사분면에 분포함.



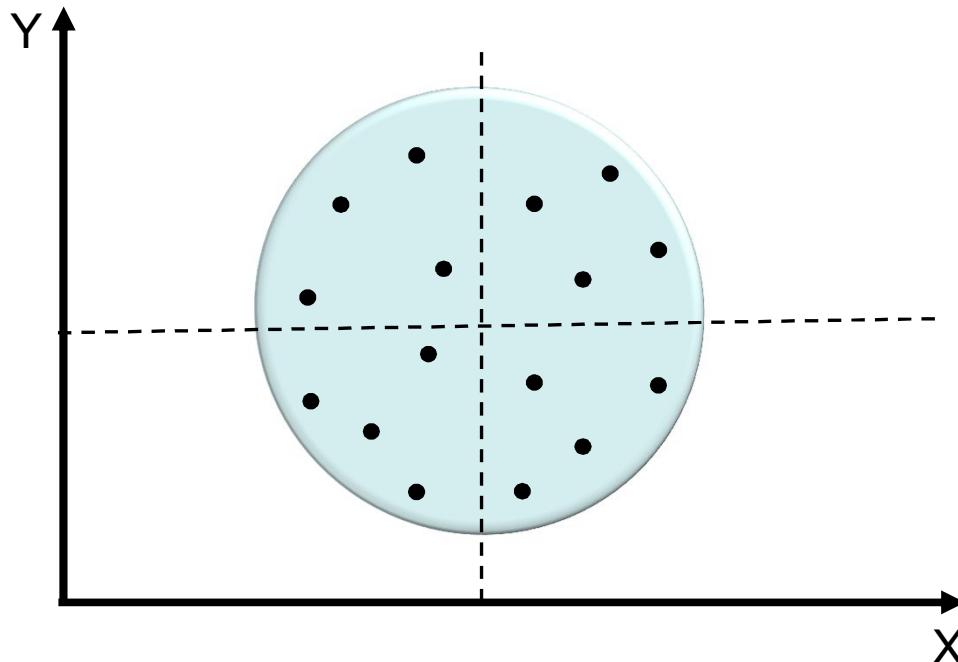


제3장 상관분석

- ◆ 공분산(covariance)

- 공분산의 의미

- 공분산이 0의 값, 즉 관측값들이 4개 면에 균일하게 분포되어 있으면 어떤 선형관계도 존재하지 않고 서로 독립임.



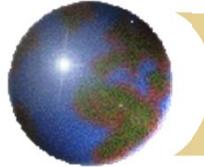


제3장 상관분석

◆ 공분산(covariance)

- 공분산의 계산

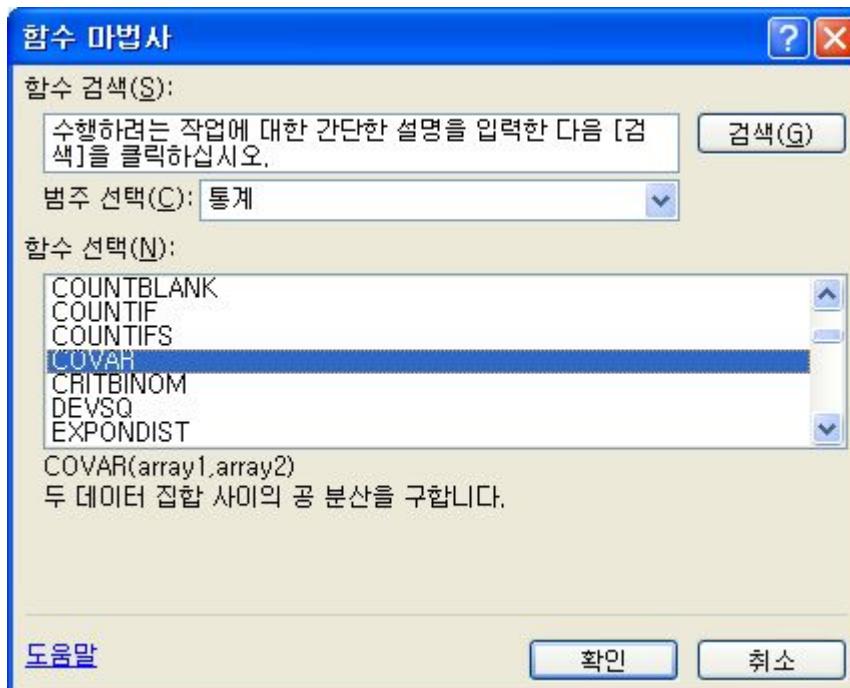
- Excel에서 공분산(COVARIANCE)을 구하는 방법은 두 가지임.
 - 함수마법사에서 ‘통계-COVAR’ 함수를 이용하는 방법
 - Excel 메뉴의 데이터-데이터 분석의 분석도구에서 ‘공분산 분석’을 이용하는 방법
- 그러나 ‘통계-COVAR’ 함수는 두 변수간에 공분산을 계산할 때만 이용할 수 있고, ‘공분산 분석’에서는 변수가 2개 이상일 때도 사용할 수 있으므로 여러 변수들 간의 공분산 값을 포함하는 공분산 행렬을 얻을 수 있음.

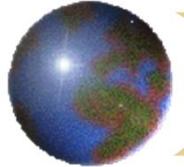


제3장 상관분석

◆ 공분산(covariance)

- 함수마법사를 클릭하고 ‘통계-COVAR’ 함수를 선택함.





제3장 상관분석

◆ 공분산(covariance)

- Array1과 Array2에 각각 X와 Y의 데이터 영역을 지정하고 확인 버튼을 누름.

	A	B	C	D	E	F	G	H	I
5	x	y							
6	1	1				COV(X, Y) [B5:B12]			
7	2	2.5							
8	3	3							
9	4	4.5							
10	5	5							
11	6	6.5							
12	7	7							
13									
14									
15									
16									
17									
18									
19									
20									
21									
22									
23									
24									

함수 인수

COVAR

Array1 A5:A12 = {"x":1;2;3;4;5;6;7}
Array2 B5:B12 = {"y":1;2.5;3;4;5;6;5;7}

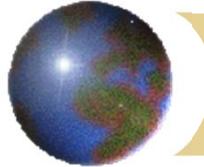
= 4

두 데이터 집합 사이의 공 분산을 구합니다.

Array2 은(는) 둘째 값들의 평균입니다. 범위는 숫자, 이름, 숫자가 들어 있는 배열이나 참조가 될 수 있습니다.

수식 결과= 4

도움말(H) 확인 취소

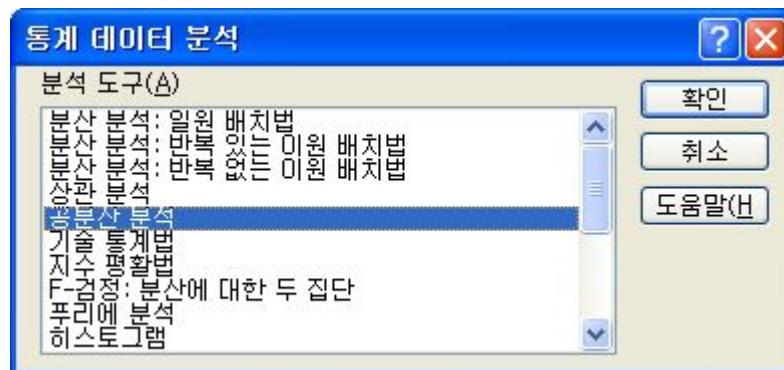


제3장 상관분석

- ◆ 공분산(covariance)

- 공분산의 계산

- Excel의 메뉴에서 데이터-데이터 분석을 클릭한 후 ‘공분산 분석’을 선택하고 확인 버튼을 누름.





제3장 상관분석

◆ 공분산(covariance)

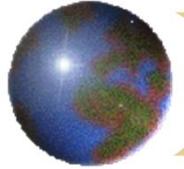
- 입력범위(I)에 X와 Y의 모든 변수를 지정, 데이터 방향에서 열(C) 선택, 그리고 첫째 행 이를 표 사용(L)을 클릭함.

The screenshot shows a Microsoft Excel spreadsheet with data in columns A and B. The data is as follows:

	A	B	C	D	E	F	G	H	I
5	x	y							
6	1	1			COV(X, Y)	4			
7	2	2.5							
8	3	3							
9	4	4.5							
10	5	5							
11	6	6.5							
12	7	7							
13									
14									
15									
16									
17									
18									
19									

A dialog box titled "공분산 분석" (Covariance Analysis) is open. It contains the following settings:

- Input Range: \$A\$5:\$B\$12
- Data Orientation: Column (C) (selected)
- Checkmark: 첫째 행 이름표 사용 (L) (First Row Headers Used)
- Output Options:
 - Output Range: \$E\$8
 - Other options: New Worksheet (P), New Workbook (W)



제3장 상관분석

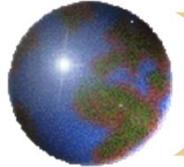
◆ 공분산(covariance)

- Excel의 출력결과는 다음과 같이 2×2 행렬로 나타남.

	A	B	C	D	E	F	G	H	I
5	x	y							
6		1	1		COV(X, Y)	4			
7		2	2.5						
8		3	3			x	y		
9		4	4.5		x		4		
10		5	5		y		4	4.061224	
11		6	6.5						
12		7	7						

- 위의 출력결과는 다음을 의미함.

$$\begin{bmatrix} \text{Var}(X) & \text{Cov}(X, Y) \\ \text{Cov}(X, Y) & \text{Var}(Y) \end{bmatrix} = \begin{bmatrix} 4 & 4 \\ 4 & 4.061224 \end{bmatrix}$$



제3장 상관분석

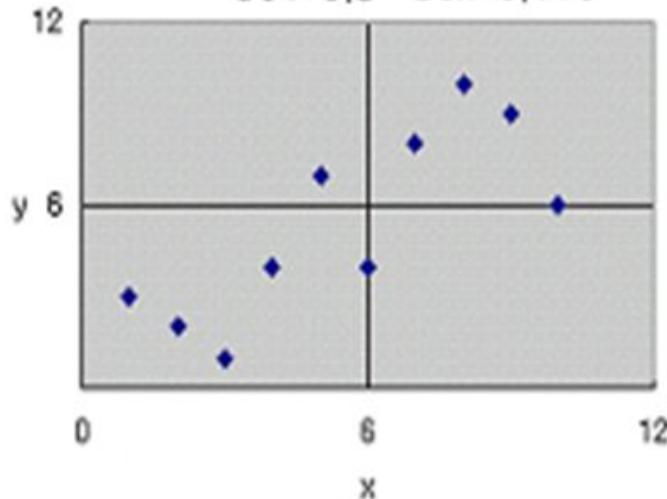
◆ 공분산(covariance)

▪ 공분산의 문제점

- 공분산이 크다고 반드시 두 변수간 연관성이 높지 않음.
- 공분산은 변수의 측정단위와 범위에 영향을 받음.

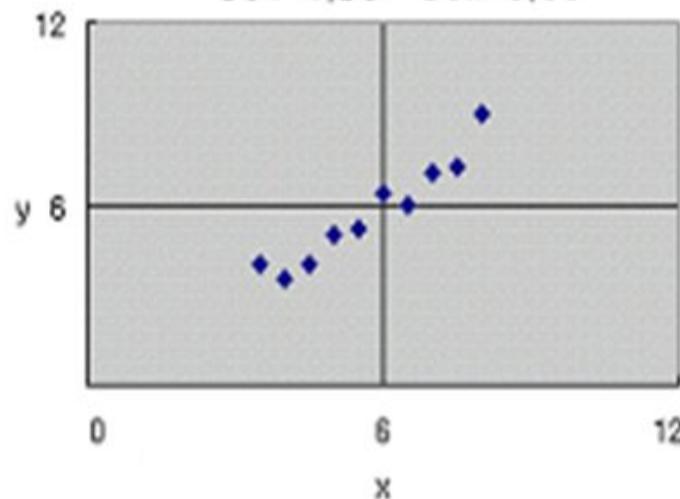
(1) 범위가 큰 경우

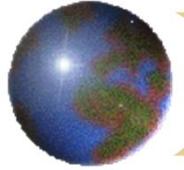
Cov=6.5 Corr=0.779



(2) 범위가 작은 경우

Cov=1.59 Corr=0.96





제3장 상관분석

◆ 상관분석(correlation analysis)

- 상관계수(correlation coefficient)
 - 공분산의 경우 두 변수간 관계의 방향은 알 수 있지만, 관계의 정도는 알 수 없음. 왜냐하면 공분산은 두 변수의 측정단위에 따라 그 값이 달라지기 때문임.
 - 이러한 문제를 해결하기 위해 측정단위에 관계없이 관계의 정도를 비교할 수 있도록 표준화한 것이 상관계수임.
 - 공분산을 두 변수 X 와 Y 의 각 표준편차로 나누면 다음과 같은 모상관계수를 구할 수 있음.

$$\rho = \text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) \text{Var}(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

- 이때 상관관계가 $0 < \rho \leq 1$ 이면 정(+)의 상관관계, $-1 \leq \rho < 0$ 이면 부(-)의 상관관계, $\rho = 0$ 이면 상관관계가 없다는 의미가 아니라 선형의 상관관계가 아니라는 의미임.



제3장 상관분석

◆ 상관분석(correlation analysis)

- 상관계수(correlation coefficient)
 - 상관계수의 경우에도 모집단(population)과 표본(sample)으로 엄격히 구분됨. 모상관계수는 상수인 반면, 표본상관계수는 변수임. 그러나 모집단에서 어떤 표본이 추출되느냐에 따라 표본의 상관계수는 달라짐.
 - 표본상관계수는 표본공분산을 각각의 표본표준편차로 나누어 표준화한 값을 나타내며 ‘ r ’로 표기함.
 - 표본상관계수는 피어슨(Karl Pearson)에 의하여 제안되었기 때문에 ‘피어슨의 표본상관계수’라고도 함.

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



제3장 상관분석

- ◆ **상관분석(correlation analysis)**

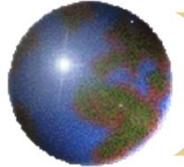
- **상관계수(correlation coefficient)**
 - 상관계수의 경우에도 단순히 두 개의 변수가 어느 정도 강한 관계에 있는가를 측정하는 단순상관분석(simple correlation analysis), 3개 이상의 변수들 간의 관계에 대한 강도를 측정하는 다중상관분석(multiple correlation analysis)이 있음.
 - 다중상관분석에서 다른 변수들과의 관계를 고정하고 두 변수만의 관계에 대한 강도를 나타내는 것을 편상관관계분석(partial correlation analysis)이라고 함.



제3장 상관분석

◆ 상관분석(correlation analysis)

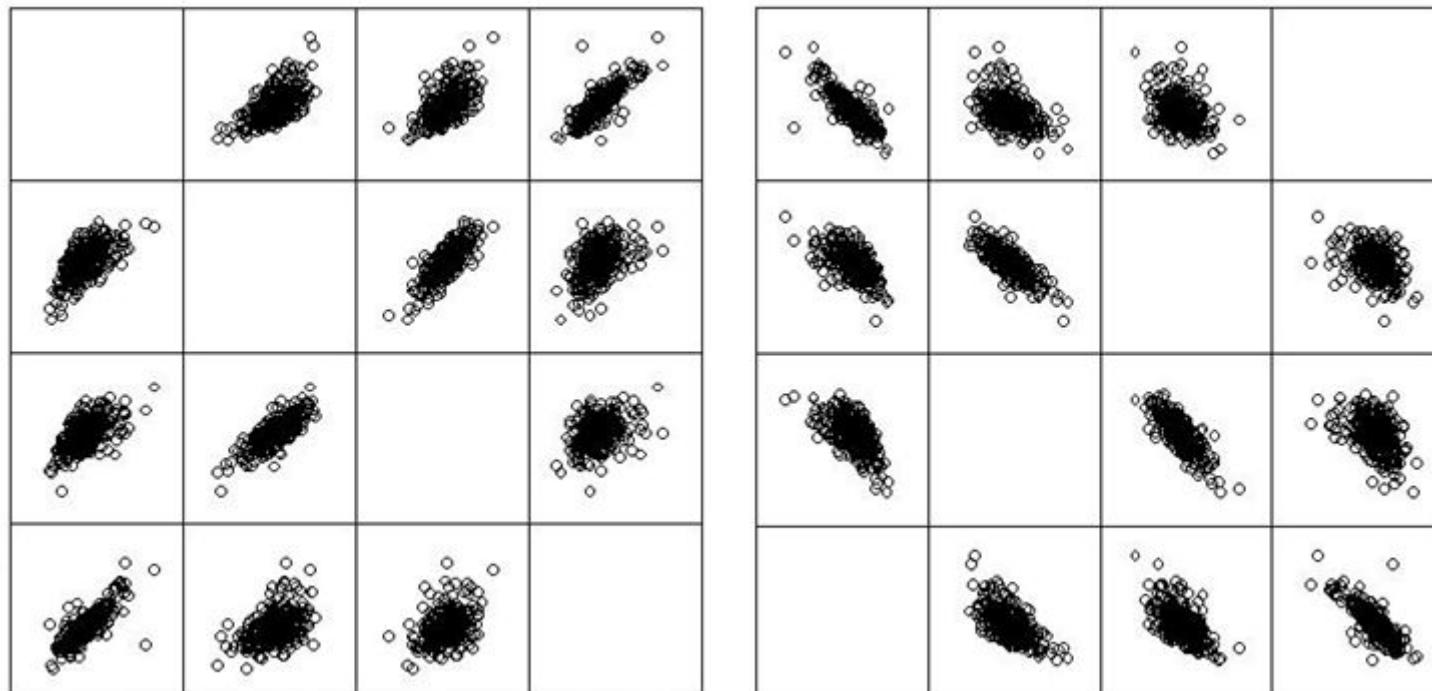
- 표본상관계수는 두 변수의 선형(직선)관계 정도를 나타내는데 다음과 같은 특징이 있음.
 - 표본상관계수(r)는 항상 -1 과 1 사이에 있음 ($-1 \leq r \leq 1$).
 - 표본상관계수의 절대값의 크기는 선형(직선)관계에 가까운 정도를 나타내고, 표본상관계수의 부호는 선형(직선)관계의 방향을 나타냄.
 - $r > 0$: 산점도에서 점들이 좌하방에서 우상방으로 띠를 형성(우상향의 형태)
 - $r < 0$: 산점도에서 점들이 좌상방에서 우하방으로 띠를 형성(우하향의 형태)
 - $r = 1$: 모든 점들이 기울기가 양수인 직선상에 위치
 - $r = -1$: 모든 점들이 기울기가 음수인 직선상에 위치

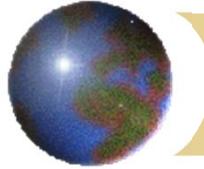


제3장 상관분석

◆ 상관분석(correlation analysis)

- 표본상관계수의 절대값이 클수록 산점도의 띠 폭은 좁아지고, 표본상관계수의 절대값이 작을수록 산점도의 띠 폭은 넓어짐.

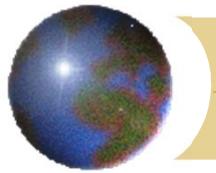




제3장 상관분석

- ◆ **상관분석**(correlation analysis)
 - 표본상관계수의 크기에 따른 해석

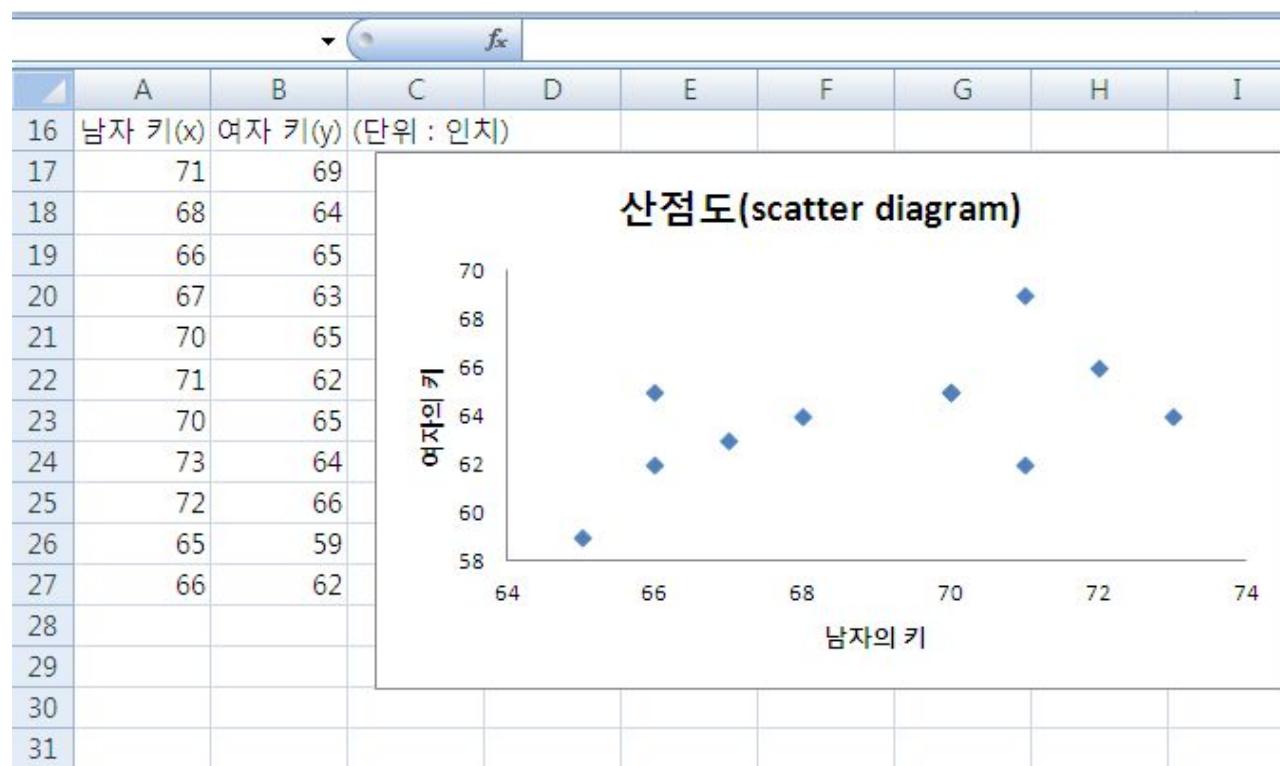
표본상관계수의 절대값	해석
0.2 이하	상관관계 거의 없음
0.2 ~ 0.4	낮은 상관관계
0.4 ~ 0.6	보통 관계
0.6 ~ 0.8	높은 상관관계
0.8 이상	매우 높은 상관관계



제3장 상관분석

● 상관분석(correlation analysis)

- 데이터 전체 영역을 지정한 후 Excel의 메뉴에서 삽입-차트-분산형을 클릭



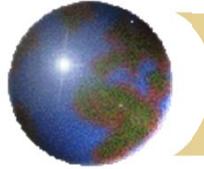


제3장 상관분석

- ◆ **상관분석(correlation analysis)**

- **상관계수의 계산**

- Excel에서 상관계수(correlation coefficient)를 구하는 방법은 두 가지임.
 - 함수마법사에서 ‘통계-CORREL’ 함수를 이용하는 방법
 - Excel 메뉴의 데이터-데이터 분석의 분석도구에서 ‘상관 분석’을 이용하는 방법



제3장 상관분석

◆ 상관분석(correlation analysis)

- 두 변수의 상관계수는 Excel의 함수마법사에서 ‘통계-CORREL’ 함수를 이용하여 구할 수 있음.

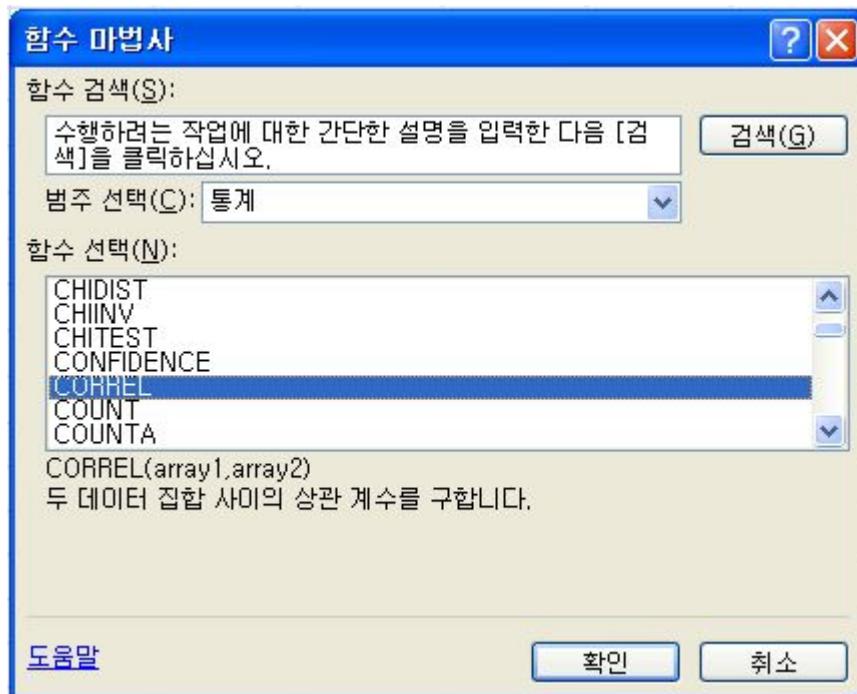
A screenshot of Microsoft Excel showing the ribbon menu and a blank worksheet. The formula bar at the top has the cell reference 'A1' and the function button 'fx' circled in red. The rest of the screen shows a grid of columns A through M and rows 1 through 20.

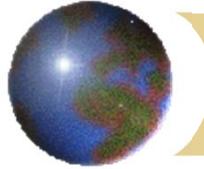


제3장 상관분석

◆ 상관분석(correlation analysis)

- 함수마법사를 클릭하고 ‘통계-CORREL’ 함수를 선택함.



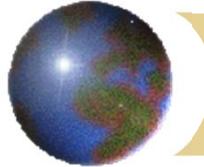


제3장 상관분석

◆ 상관분석(correlation analysis)

- Array1과 Array2에 데이터 영역을 선택, 확인 버튼을 누름.

	A	B	C	D	E	F	G	H	I
16	남자 키(x)	여자 키(y)	(단위 : 인치)						
17	71	69			CORREL	:16:B27)			
18	68	64							
19	66	65							
20	67	63							
21	70	65							
22	71	62							
23	70	65							
24	73	64							
25	72	66							
26	65	59							
27	66	62							
28									
29									
30	함수 인수							?	X
31	CORREL								
32	Array1	A16:A27	= {"남자 키(x)":71:68:66:67:70:71:70...}						
33	Array2	B16:B27	= {"여자 키(y)":69:64:65:63:65:62:65...}						
34	= 0.558054712								
35	두 데이터 집합 사이의 상관 계수를 구합니다.								
36	Array2 를(는) 둘째 값들의 셀 범위입니다. 범위는 숫자, 이름, 숫자가 들어								
37	있는 배열이나 참조가 될 수 있습니다.								
38	수식 결과= 0.558054712							확인	취소
39	도움말(H)								

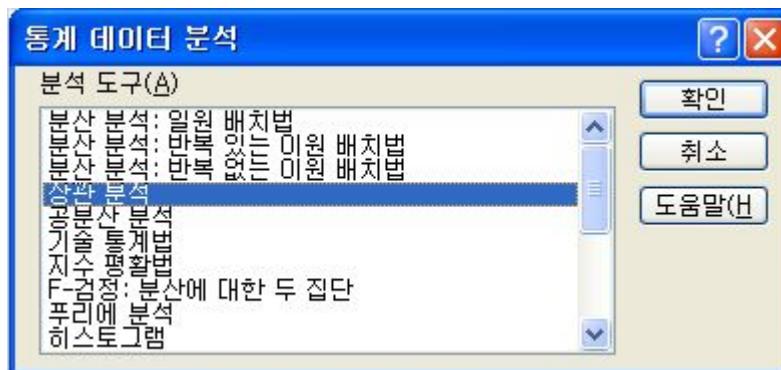


제3장 상관분석

- ◆ **상관분석(correlation analysis)**

- **상관계수의 계산**

- Excel의 메뉴에서 데이터-데이터 분석을 클릭한 후 ‘상관 분석’을 선택하고 확인 버튼을 누름.



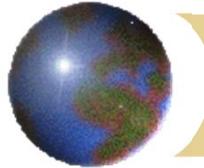


제3장 상관분석

◆ 상관분석(correlation analysis)

▪ 상관계수의 계산

- 입력범위는 두 데이터의 전체 영역을 선택함.
- 데이터 방향은 자료들이 세로로 정렬되어 있으면 ‘열’(column)을, 가로로 정렬되어 있으면 ‘행’(row)을 선택
- 데이터의 제목을 영역에 포함시키려면 첫째 행 이름표 사용을 선택, 그렇지 않으면 선택하지 않음.
- 현재 작업 시트(sheet)에 출력하고자 할 경우 출력범위를 선택하여 출력하고자 하는 셀을 지정한 후 확인 버튼을 누름.

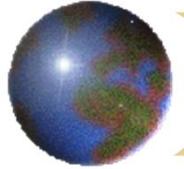


제3장 상관분석

- **상관분석(correlation analysis)**
 - **상관계수의 계산**

	A	B	C	D	E	F	G	H	I
16	남자 키(x)	여자 키(y) (단위 : 인치)							
17	71	69			CORREL	0.558055			
18	68	64							
19	66	65							
20	67	63							
21	70	65							
22	71	62							
23	70	65							
24	73	64							
25	72	66							
26	65	59							
27	66	62							
28									
29									
30									
31									
32									

The screenshot shows a Microsoft Excel spreadsheet with data in columns A and B. Column A contains male height values (x), and column B contains female height values (y). The formula bar shows E19. A correlation analysis dialog box is open over the spreadsheet, titled '상관 분석'. It has two main sections: '입력' (Input) and '출력 옵션' (Output Options). In the 'Input' section, '입력 범위(I):' is set to \$A\$16:\$B\$27, '데이터 방향:' (Data direction) is set to '열(C)' (Column), and '첫째 행 미룹표 사용(L)' (Use first row as labels) is checked. In the 'Output Options' section, '출력 범위(O):' is set to \$E\$19, and '새로운 워크시트(P)' (New worksheet) is selected.



제3장 상관분석

◆ 상관분석(correlation analysis)

▪ 상관계수의 계산

Screenshot of Microsoft Excel showing correlation analysis results.

The Excel ribbon shows the "Data" tab selected. The "Data Tools" group contains icons for Access, Web, Text, Other, and External Data. The "Connections" group contains icons for New Connection, Refresh, Delete, Filter, and Sort & Filter.

The worksheet displays the following data:

	A	B	C	D	E	F	G	H	I
16	남자 키(x)	여자 키(y)	(단위 : 인치)						
17	71	69			CORREL	0.558055			
18	68	64							
19	66	65				남자 키(x) 여자 키(y)			
20	67	63				남자 키(x) 1			
21	70	65				여자 키(y) 0.558055	1		
22	71	62							
23	70	65							
24	73	64							
25	72	66							
26	65	59							
27	66	62							

남매의 신장간에는 선형관계가 0.558 정도라는 것을 의미함.

상관계수의 부호가 정(+)이라는 것은 남자의 키가 크면 여자의 키도 크다는 것을 시사함.



제3장 상관분석

◆ 상관분석(correlation analysis)

▪ 상관계수의 측정단위

- 상관계수의 또 하나의 특징은 어느 한 변수 또는 두 변수의 모든 값에 0이 아닌 상수가 더해지거나 곱해지더라도 그 값이 변하지 않는다는 것임. 즉, 상관계수는 측정척도의 원점과 단위의 변환으로 변경되지 않음.
- 이러한 결과는 상관계수의 사용에 대해 중요한 의미를 가짐. 즉, 측정값들이 단위가 센티미터나 인치 또는 분이나 초로 되어있든 변수들 간의 상관계수 값은 항상 일정함.
- 실제로 변수들 모두에 대하여 측정값의 원점 또는 단위가 변할 때에도 상관계수 r 은 변하지 않는다는 사실은 상관계수의 활용범위를 크게 해줌.



제3장 상관분석

- ◆ **상관분석(correlation analysis)**

- **상관계수의 한계**

- 상관계수는 수학적인 관계일 뿐 속성의 관계로 확대 해석해서는 안됨.
 - 상관계수는 선형관계의 척도임. 상관계수가 낮더라도 비선형(곡선)관계가 있을 수 있으므로 반드시 산점도(scatter diagram)로 확인해야 함.
 - 상관계수는 자료분석의 초기단계일 뿐 결론단계에 사용되는 통계량은 아님.