Correlation Analysis (상관관계분석)

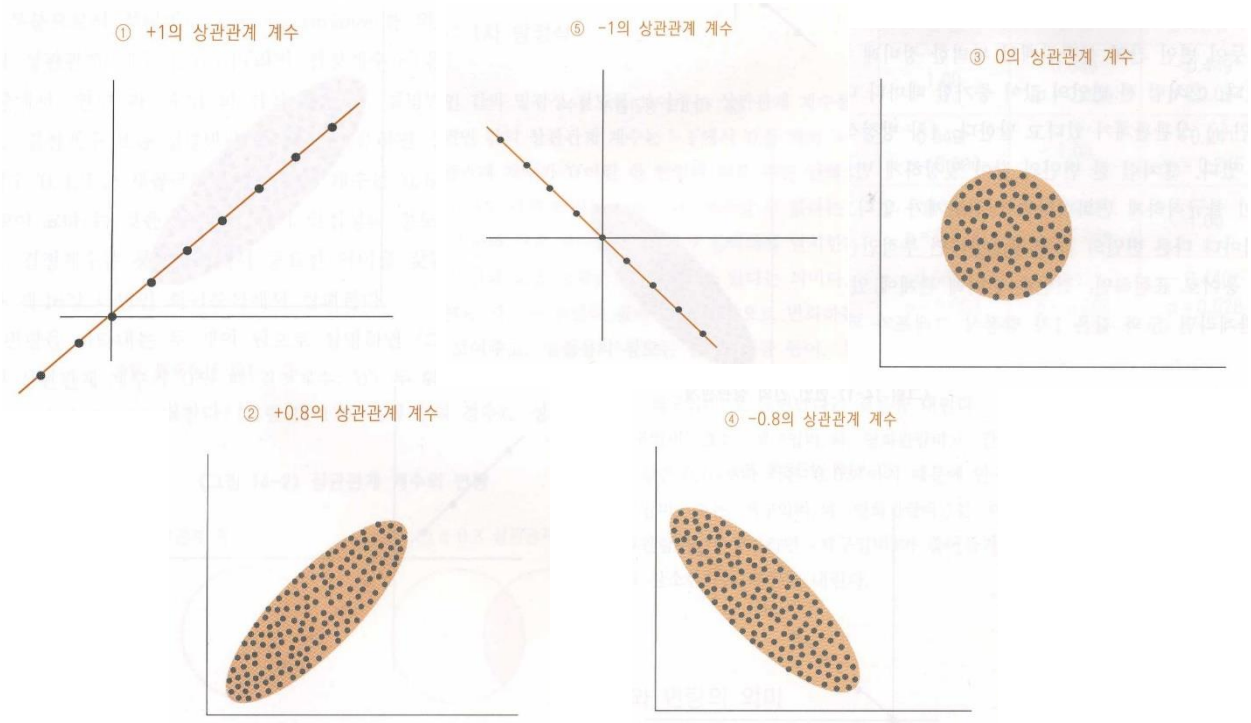
1. 소개

〈표 14-1〉 상관관계분석의 조건

1. 측정: 등간척도 (또는 비율척도)
2. 수: 두개 이상 여러 개

상관관계는 연속적 속성을 갖는 두 변인들 간 상호연관성에 대한 기술 통계를 제공할 뿐 아니라, **두 변인 간의 상호 연관성에** 대한 **통계적 유의성을** 검증해 주는 통계분석 기법

➔ Pearson 상관계수는 두 연속형 변수 사이의 **선형적인** 상관성(linear correlation)을 분석
➔ 커뮤니케이션 분야에서는 흔히 피어슨(곱적률) 상관계수 *r* 을 자주 사용: -1 +1 사이
➔ Karl Pearson➔ Pearson's product-moment coefficients of correlation = Pearson correlation coefficient = Zero order correlation coefficient ➔ *r*



- 일반적 *r*=.30 약한 관계, *r*=.50 중간 관계, *r*=.70 강한 관계
- 두 변인간의 강도와 방향을 산포도(scatterplot)로 표현: 2 차원 공간에서 변인 X(가로축)와 Y(세로축)에 대한 각 케이스 값들 나타냄: *r* 이 커질수록 두 변인간의 관계를 나타내는 데이터들이 점점 조밀, 반대의 경우 변인간 데이터들은 점점 흩어짐

- *r* 의 절대값이 클수록 (1 에 가까울수록) 두 변수의 값들은 직선 가까이에 위치하며 따라서 두 변수 사이의 선형적인 상관성은 커진다
- *r* 이 절대값이 1 일 때 두 변수의 값은 모두 직선 위에 위치 (perfect correlation) → 한 변인의 값을 알면 다른 변인의 값을 정확하게 예측할 수 있음
- *r* 이 0 일 때 두 변수 사이에 선형적인 상관성은 없다 → 한 변인의 값을 알아도 다른 변인의 값을 전혀 예측할 수 없음

➔ 상관관계 계수(correlation coefficient)와 공변량(covariance)의 비교

〈표 14-5〉 공변량과 상관관계 계수

$$COVxy = \frac{\sum(X - X\text{평균})(Y - Y\text{평균})}{N-1}$$

$$r\,(\text{상관관계 계수}) = \frac{COVxy}{SxSy}$$

〈표 14-6〉 〈교육〉과 〈텔레비전시청시간〉의 가상 데이터

응답자	교육		텔레비전시청시간	
	원 점수	차이 점수	원 점수	차이 점수
1	2	2 - 3.2 = - 1.2	3	3 - 3.4 = -0.4
2	2	2 - 3.2 = - 1.2	2	2 - 3.4 = -1.4
3	3	3 - 3.2 = -0.2	4	4 - 3.4 = + 0.6
4	4	4 - 3.2 = + 0.8	3	3 - 3.4 = - 0.4
5	5	5 - 3.2 = + 1.8	5	5 - 3.4 = + 1.6
평균	3.2		3.4	
표준편차	1.30		1.14	

$$\text{공변량} = \frac{(-1.2)(-0.4) + (-1.2)(-1.4) + (-0.2)(0.6) + (0.8)(0.4) + (1.8)(1.6)}{4}$$

$$= \frac{(0.48) + (1.68) + (-0.12) + (0.32) + (2.88)}{4}$$

$$= 1.15$$

$$\text{상관관계 계수} = \frac{1.15}{1.30 \times 1.14}$$

$$= 0.775$$

- 상관관계 계수와 공변량은 둘 다 두 변인 사이의 상관관계의 강도(strength)를 나타냄
- 공변량은 측정단위에 따라 변함 (예: 170cm 과 2,000g 의 공변량 > 1.7m 와 2kg 의 공변량)

- 상관관계 계수는 공변량을 각 변인의 표준편차를 곱한 값으로 나눈 것이기에 측정단위에 따라 변하지 않음

2. 연구절차

<데이터>

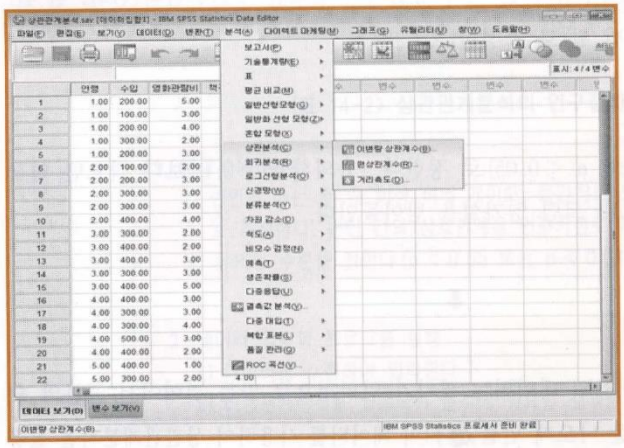
〈표 14-3〉 상관관계분석의 가상 데이터

응답자	연령	수입	영화 관람비	책 구입비	응답자	연령	수입	영화 관람비	책 구입비
1	1	200	5	2	14	3	300	3	3
2	1	100	3	2	15	3	400	5	2
3	1	200	4	1	16	4	400	3	4
4	1	300	2	4	17	4	300	3	3
5	1	200	3	3	18	4	300	4	3
6	2	100	2	2	19	4	500	3	4
7	2	200	3	1	20	4	400	2	5
8	2	300	3	2	21	5	400	1	3
9	2	300	3	4	22	5	300	2	4
10	2	400	4	3	23	5	300	3	5
11	3	300	2	4	24	5	400	2	3
12	3	400	2	3	25	5	500	1	4
13	3	400	3	3					

변인
- 연령: 5 점 척도(1=10 대, 2=20 대, 3=30 대, 4=40 대, 5=50 대 이상)로 측정된 나이
- 수입: 월 평균소득(단위: 만원)
- 영화관람비: 월평균 영화관람비(단위: 만원)
- 책구입비: 월평균 책구입비(단위: 만원)

<SPSS 실행방법>

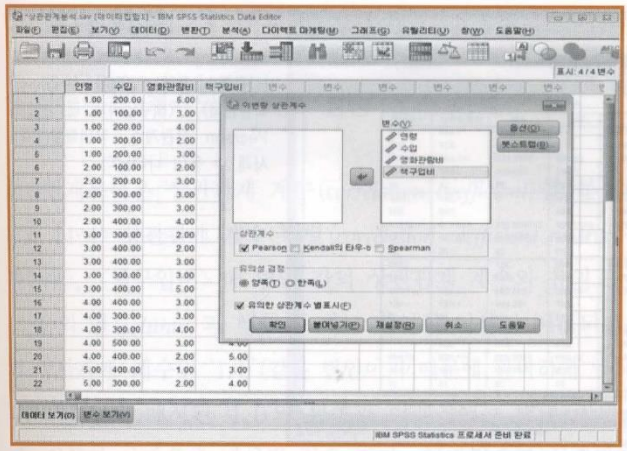
1)



[실행방법 1]

메뉴판의 [분석(A)]을 선택하여 [이변량 상관계수(B)]를 클릭한다.

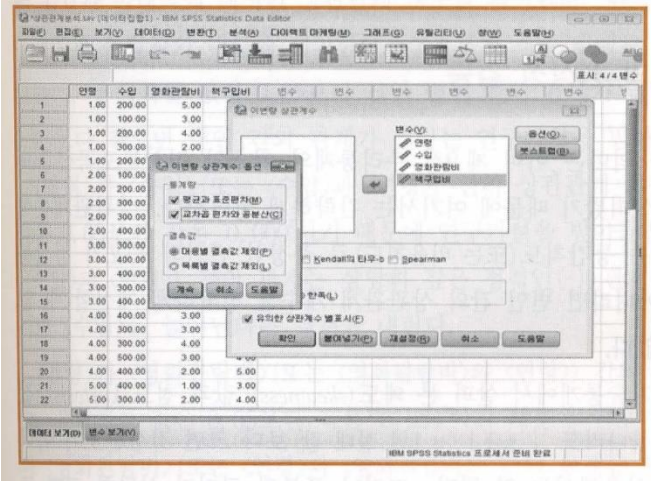
2)



[실행방법 2]

[이변량 상관계수] 창이 나타나면, 왼쪽 칸에서 오른쪽 [변수(V)] 칸으로 분석하고자 하는 변인을 클릭하여 이동시킨다(➡). [상관계수]의 [☑ Pearson], [유의성 검정]의 [◉ 양쪽(T)], [☑ 유의한 상관계수 별표시(F)]는 기본으로 설정되어 있다. 오른쪽의 [옵션]을 클릭한다.

3)



[실행방법 3]

[이변량 상관계수: 옵션] 창이 나타나면, [통계량]의 [☑ 평균과 표준편차(M)], [☑ 교차곱 편차와 공분산(C)]을 클릭한다. [결측값]의 [◉ 대응별 결측값 제외(P)]는 기본으로 설정되어 있다. 아래의 [계속]을 클릭한다. [실행방법 2]의 [이변량 상관계수]창으로 다시 돌아가 [확인]을 클릭한다.

<결과분석>

1) 상관관계 계수와 유의도 검증 결과 해석

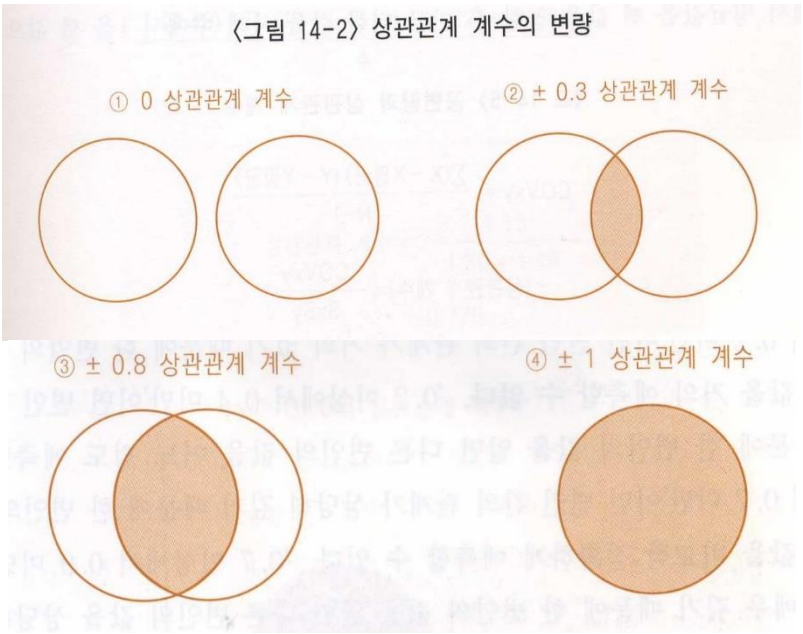
〈표 14-4〉 변인 간의 상관관계 계수 행렬

구 분	연 령	수 입	영화관람비	책구입비
연 령	1.00	0.649 p = 0.000	- 0.449* p = 0.024	0.563* p = 0.003
수 입	0.649* p = 0.000	1.00	- 0.241 p = 0.245	0.532* p = 0.006
영화관람비	- 0.449* p = 0.024	-0.241 p = 0.245	1.00	- 0.440* p = 0.028
책구입비	0.563* p = 0.003	0.532* p = 0.006	- 0.440* p = 0.028	1.00

*p < 0.05

2) 결정계수의 의미: 설명변량
- 상관관계 계수(*r*)를 제곱한 값(r^2)을 결정계수(coefficient of determination)라 부름
- 이 값은 두 변인이 겹친 부분으로서 설명변량(explained variance)의 비율을 의미함 (예: 연령과 수입의 *r* 이 0.649 라면 r^2는 0.421 → 이는 두 변인 변량(variance)의 42.1%가 겹친 부분으로 설명될 수 있음을 의미 → 이는 한 변인의 변화량에 따라 다른 한 변인의 변화량이 42.1% 만큼 설명될 수 있음을 의미)

〈그림 14-2〉 상관관계 계수의 변량



표, 그림 출처: 최현철. (2013). *사회과학 통계분석*. 나남

3. 상관관계 계수 해석 시 주의할 점
- 두 변인간 관계를 설명하는 피어슨 *r* 이 높게 나타났지만, 실제로는 두 변인 간에 이론적으로 아무런 연관성이 없을 수 있음 → 이를 "거짓관계(spurious relationship)"

(예: 한도시의 아이스크림 판매량을 보면 수영장에서 익사율이 높을 때 아이스 크림 판매량 역시 증가하는 것을 알 수 있음 → 피어슨 r 높을것 임 → 그러나 실제로는 여름에는 아이스크림 판매량이 증가할 뿐만 아니라 무더위를 피하기 위해 수영장을 찾는 이용객도 많을 것이므로 익사율이 높아 질 것 → 익사율의 원인은 아이스크림 판매량이 아니라 무더위를 피하기 위한 수영장 이용임)