6. 상관계수에 의한 검정

6-1. 상관계수

6-1-1. 상관계수

① 상관(correlation)란 두 변수의 관계이다.

㉠ 한 변수가 변할 때 다른 변수가 어떻게 변하는가 또는 그 역은 어떠한가?

㉡ 두 변수가 동시에 변하는 것

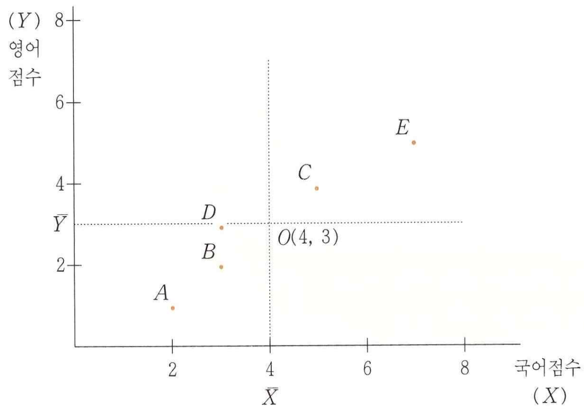
㉢ 상호관계(interaction)

[[예]] 자동차 수와 교통사고 수가 상관이 있다

; 자동차 수의 증가와 교통사고 수에 관계가 있음

[[예]] 중학생 5명의 국어와 영어 점수

; 국어와 영어점수는 정적 관계를 보임



② 상관을 인과관계(어떤 변수는 원인을 제공하고 어떤 변수는 영향만을 받는 원인과 결과관계)로 해석해서는 안 된다.

[[예]] 자동차 수가 늘어났을 때 교통사고가 늘었다고 해서 자동차 수의 증가가 교통사고를 유발하는 직접적인 원인이라고 판정할 수 없다.

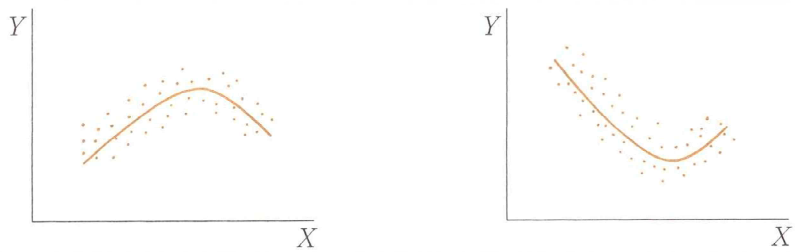
③ 단! 다른 매개변수가 통제된 실험설계의 경우라면 상관관계가 인과관계로 해석이 가능하다.

6-1-2. 상관계수의 기본 가정

① 상관연구를 위해 선형성, 등분산성, 이상치의 존재 여부, 자료의 절단여부를 확인해야 한다.

② 기본 가정을 확인하기 위한 가장 쉬운 방법은 산점도(scatter plot)를 그려 연구의 자료가 상관연구를 위한 기본 가정을 충족하는지 점검해야 한다.

㉠ 상관관계는 선형성(linearity)을 만족해야 한다.

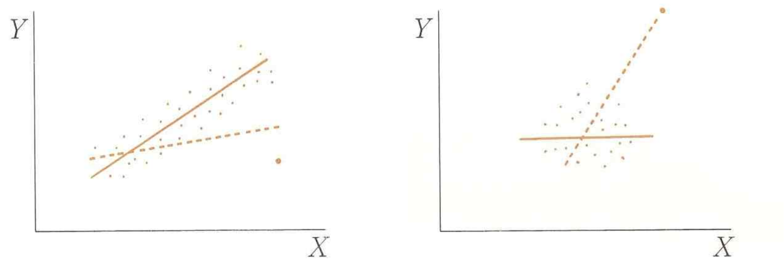


(a) (b)

따라서 우선 산점도를 그려서 선형성이 있는지를 확인한 후, 선형성이 존재하지 않는다면 자료 변화의 경향성에 비추어 세밀한 해석을 해야 한다.

㉡ 상관관계는 등분산성이어야 한다. 따라서 상관연구를 위하여 두 변수의 등분산성을 산점도로 확인해야 한다.

㉢ 상관관계는 이상치(outlier)가 없어야 한다.



㉠ 상관계수가 낮아지는 경우 ㉡ 상관계수가 높아지는 경우

따라서 산점도를 그려 이상치가 존재하는지를 밝히고 그 이상치의 자료가 의미가 있는지를 고려해야 한다.

㉣ 상관연구에서 자료는 절단(truncation)되어 있지 않아야 한다.

6-1-3. 상관계수의 종류

① Karl Pearson의 적률상관계수(product-monet correlation coefficient) r은 두 변수가 모두 연속변수일 때 상관의 정도를 분석하기 위하여 사용한다.

〚예〛 키와 몸무게 사이의 상관관계

② 두 변수 모두가 연속변수가 아닌 서열척도에 의한 질적변수일 경우는 상관계수의 특수한 형태인 Spearman의 등위상관계수(rank correlation coefficient)를 사용한다.

〚예〛 국어성적 등수와 영어성적 등수 사이의 상관관계
작품평가에서 평가자가 작품에 대한 등수를 선정하였을 때, 평가자간의 상관성

③ 독립변수가 명명척도에 의하여 두 종류로 명확히 구분된 질적변수이며, 종속변수가 연속적인 양적변수일 때는 양류상관계수(point-biserial correlation coefficient)를 사용한다.

〚예〛 성별과 국어점수의 상관관계

(현대통계학에서는 양류상관계수를 계산하지 않음. 명명척도에 의하여 이분된 독립변수에 연구자가 임의의 수를 부여하여 Karl Pearson의 단순적률상관계수로 두 변수의 상관 정도를 추정)

④ 독립변수와 종속변수 모두 연속적인 양적변수이나 연구자가 독립변수를 어떤 특정 점수에 의하여 양분하여 가상적인 질적점수로 변환한 후 가상적으로 양분된 독립변수와 연속적인 양적변수인 종속변수와의 상관관계를 분석하기 위하여 양분상관계수(biserial correlation coefficient)를 사용한다.

〚예〛 지능점수와 학업성취도 관계 연구, 인위적으로 저능아와 정상아로 양분, 정상아와 저능아인 독립변수와 학업성취도 와의 상관관계

⑤ 독립변수와 종속변수 모두 이분화 된 질적변수일 때 두 변수의 상관 정도를 알기 위해서는 ϕ계수(phi coefficient)를 사용한다. 독립변수와 종속변수가 모두 질적변수로 크기의 정도를 의미하지 않기 때문에 ϕ계수에서 +, -의 기호는 의미를 가지지 않는다.

〚예〛 어떤 사건에 대한 찬반 여부와 성별과의 관계
학교현장에서 두 개의 교수법에 따른 학업 성공 여부와의 관계

6-2. 공분산(covariance)

① 공분산 : 두 변수가 동시에 변하는 정도, 한 변수가 그 평균으로부터 변화할 때 다른 변수가 그 평균으로부터 변하는 정도이다.
② X와 Y변수의 공분산 : X변수가 변할 때 Y변수가 동시에 얼마나 변하는가를 나타내기 위하여 X변수의 편차와 Y변수의 편차의 곱을 이용한다.
③ 모수치의 공분산 σ_{XY}, 통계치의 공분산 s_{XY}
④ $s_{XY}=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})=\frac{1}{n-1}\left[\sum_{i=1}^{n}x_iy_i-\frac{\left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}y_i\right)}{n}\right]$

6-3. 상관계수(correlation coefficient)

Q. 두 변수의 동시 변화에 대하여 공분산 사용은 어떠한 문제점을 담고 있는가?
A. 공분산을 이용해 두 변수의 상관을 추정한다면 두 변수에 어떤 값을 더하거나 뺄 경우 공분산은 변하지 않지만, 곱하거나 나누었을 경우 공분산은 변화한다. 특히 단위를 달리하여 공분산을 구하면 다른 값이 얻어진다. 따라서, 이 문제점을 해결하기 위하여 피어슨(K. Pearson, 1896)이 상관계수 계산 공식을 유도하였다.

$$\rho_{XY}=\frac{\sigma_{XY}}{\sigma_X\sigma_Y}$$

① 상관계수 : 두 변수 x와 y의 동시에 변하는 정도를 나타낸 지수
② 산점도를 통해 두 변량 x, y 사이의 증가나 감소의 비율이 일정하다는 것을 확인할 수 있다.
③ 모수치의 상관계수는 ρ(로우)라 표기하고 통계치의 상관계수는 r로 표기한다.
④ 상관계수는 정적이든 부적이든 공분산의 양과 비례한다.
⑤ 공분산의 절댓값이 크면 상관이 높을 것이라고 예측할 수 있다.
⑥ 상관계수의 범위 : -1.0이상 +1.0 이하
⑦ 피어슨의 적률상관계수(Pearson's product-moment correlation coefficient)

상관계수 공식 : $r_{XY}=\frac{s_{XY}}{s_Xs_Y}$

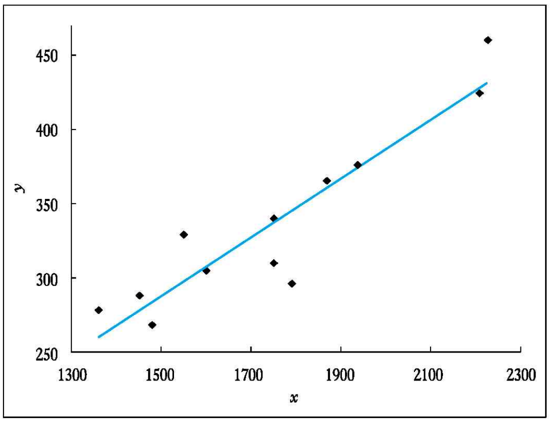
(s_X=x의 표준편차, s_Y=y의 표준편차, s_{XY}=x와 y의 공분산(covariance))

⑧ 상관계수 r의 제곱인 r^2을 결정계수(coefficient of determination)라고 한다.
⑨ 상관관계의 언어적 표현:
㉠ 정적(+)인 관계= 상관이 있다, 낮다 등으로 서술
㉡ 부적(-)인 경우= 정적인 관계의 언어적 표현 앞에 '부적인 상관이 있다' 혹은 '상관이 부적으로 높다'라고 표현

상관계수 범위	상관관계의 언어적 표현
0.00~0.20	상관이 거의 없다.
0.20~0.40	상관이 낮다.
0.40~0.60	상관이 있다.
0.60~0.80	상관이 높다.
0.80~1.00	상관이 매우 높다

[예] 다음 표는 12개 주거용 건물의 주거면적 x와 판매가격 y를 나타낸 것이다.

㉠ 산점도 그리기



건물	x (제곱피트)	y (천만원)
1	1360	278.5
2	1940	375.7
3	1750	339.5
4	1550	329.8
5	1790	295.6
6	1750	310.3
7	2230	460.5
8	1600	305.2
9	1450	288.6
10	1870	365.7
11	2210	425.3
12	1480	268.8

㉡ 상관관계 구하기

[풀이]

x와 y의 합과 표준편차를 구하면 각각

$$\sum_{i=1}^{n} x = 20.980,\ s_x = 281.4842,\ \sum_{i=1}^{n} y = 4043.5,\ s_x = 59.7592$$

이다. 공분산은

$$s_{xy} = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)/n\right]$$

$$= \frac{1}{11}[7,240,383 - (20,980)(4043.5)/12] = 15,545.19697$$

이므로

$$r = \frac{s_{xy}}{s_x s_y} = \frac{15,545.19697}{(281.4842)(59.7592)} = 0.9241$$

이다.

㉢ 주거용 건물의 주거면적과 판매가격은 상관이 매우 높다.

6-4. 상관계수 유의성 검정

① 상관계수의 유의성 검정은 상관계수의 수치가 애매할 경우에 사용된다. 예를 들어 상관계수가 0.85로 수치가 크면, 가설검정을 하지 않고도 상관관계가 있다고 할 수 있다.

② 두 변수 X, Y 사이에 유의미한 선형관계가 존재하는가를 판단하는 유의성 검정

㉠ 가설설정

H_0 : 두 변수 간에 상관관계가 없다. $\rho=0$

H_1 : 두 변수 간에 상관관계가 있다. $\rho\neq 0$

㉡ 검정통계량 $t_{n-1}=r\dfrac{\sqrt{n-2}}{\sqrt{1-r^2}}$ $(df=n-2)$

③ 검정통계량은 자유도 $n-2$인 t분포를 따른다.

〚예〛 A 농구팀으로부터 10명을 랜덤하게 추출하여 키와 몸무게를 측정하였을 때, 키(㎝)와 몸무게(㎏)의 상관계수를 구하여라. 그리고 이 상관계수 r은 0과 유의적인 차이가 있는가?

선수	몸무게X	키 Y
1	73	185
2	71	175
3	75	200
4	72	210
5	72	190
6	75	195
7	67	150
8	69	170
9	71	180
10	69	175

[풀이]

$S_{xy}=328$ $S_{xx}=60.4$ $S_{yy}=2610$이 되므로

$$r=\frac{328}{\sqrt{(60.4)(2610)}}=0.8261\text{이다.}$$

r의 값이 거의 1에 가깝고 매우 크므로 키와 몸무게 사이에는 강한 양의 상관관계를 가진다고 할 수 있다.

㉠ 가설 H_0 : $\rho=0$, H_1 : $\rho\neq 0$

㉡ 검정통계량을 계산하면

$$T=r\sqrt{\frac{n-2}{1-r^2}}=0.8261\sqrt{\frac{10-2}{1-(0.8261)^2}}=4.15$$

자유도 8인 t 분포

㉢ 검정통계량은 $t_{0.005}=3.355$보다 크고, 양쪽꼬리 $p-$값은 $2(0.005)=0.01$보다 작기 때문에 상관계수는 유의수준 $\alpha=0.01$에서 유의하다고 할 수 있음

㉣ $r^2=(0.8261)^2=0.6824$는 총 변동중의 약 68%가 회귀직선에 의해서 설명되는 것을 의미