사회통계
두 양적변수 사이의 관계

김현우, PhD[1]

[1] 충북대학교 사회학과 조교수

진행 순서

1. 변수 사이의 연관성
2. 두 양적변수 사이의 연관성

변수 사이의 연관성

변수 사이의 연관성

지난 주에는 자료 중 단지 하나의 변수를 요약했을 뿐이었다.

- 우리는 지난 주에 중심경향과 산포경향의 요약통계량을 공부했다.
- 변수 하나에 대해 각각의 요약통계량을 하나씩 제시하였고, 설령 두 개의 변수(e.g., PM10과 PM2.5)가 주어질 때도 각각 따로 요약하였다.
- 우리는 변수가 두 개 주어졌을 때 그 사이의 연관성(association) 혹은 관계(relationship)에 대해서 살펴보지 않았다.
- 그래서 이번 주에는 둘 이상의 변수 사이의 연관성 혹은 관계를 살펴보기로 한다.

변수 사이의 연관성

이론의 정의와 역할에 대해 잠깐 생각해보자.

- 경험과학(empirical science)에서 이론(theory)이란 경험적으로 검증가능한, 상호연관된 일련의 명제들(propositions)로 구성된다.
- 명제란 둘 이상의 개념들(concepts) 사이의 관계에 대한 진술이다.
- "상대적 박탈감이 증가하면 집회시위에 참여할 확률이 높아진다."
- 수학의 언어로 표현한다면 명제는 함수(function)의 꼴로 표현될 수 있다

$$Y = f(X)$$

- 명제는 수행하는 역할에 따라 공리(axiom), 정리(theorem), 가설(hypothesis), 발견(finding) 등 다양한 형태를 가질 수 있다.

변수 사이의 연관성

경험과학에서 가설과 발견이란 무엇인가?

- 가설(hypothesis)이란 (연구를 수행하기에 앞서 제시된) 둘 이상의 개념들 사이의 관계에 대한 "잠정적인(tentative)" 진술이다.
- 발견(finding)이란 (연구를 통해 어느 정도 입증된) 둘 이상의 개념들 사이의 관계에 대한 진술(=명제)이다.
- 흥미롭게도 현대의 경험과학에서는 법칙(law)이나 사실(fact)과 같은 표현이 거의 나오지 않는다. (오래된 지식을 전달하는 교과서를 제외하면) 오히려 그런 말은 황색 저널리즘이나 인터넷 커뮤니티에서나 사용된다(e.g., "이게 팩트지").

변수 사이의 연관성

- 이때 경험과학에서 가설을 테스트하기 위해 근본적으로 둘 이상의 변수 사이에 어떤 관계가 있는지를 다루게 된다.
- 하나의 변수를 요약하는 통계로는 가설을 세우거나 테스트할 수 없다.
- 제대로 된 가설이 되려면 적어도 두 개 이상의 변수 간의 연관성을 분석해야만 한다!

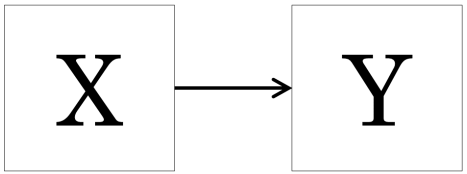
변수 사이의 연관성

종속변수와 독립변수의 역할을 명확하게 인식해야 한다.

- 종속변수(dependent variable)란 다른 무언가에 종속되어 설명되어지는 변수이고, 독립변수(independent variable)란 다른 무언가로부터 독립되어 설명하는 변수다.
- 관행상 종속변수는 Y, 독립변수는 X로 표기된다.
- 함수관계에서 독립변수는 정의역(domain), 종속변수는 치역(range)이다.

$$Y = f(X)$$

- 그림으로 나타낼 수도 있다.

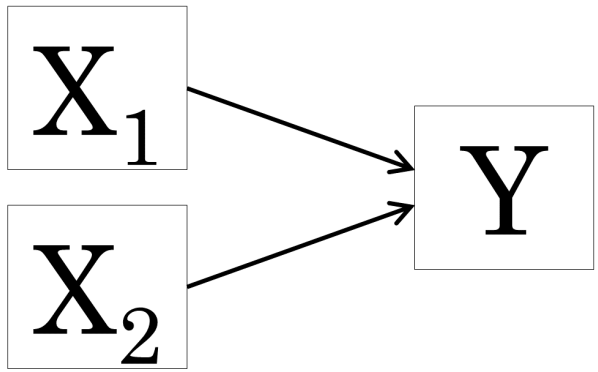


변수 사이의 연관성

하나 이상의 독립변수를 만들어 낼 수도 있다.

- 여러 독립변수들을 관심변수(variable of interest)와 통제변수(control variable) 그룹으로 나누어 접근할 수 있다.
- 통제변수의 동시적 영향력을 배제한 상태에서(controlled out), 관심변수와 종속변수의 관계를 좀 더 집중해서 살펴볼 수도 있다.
- 근본적으로 무엇이 관심변수이고 통제변수인가는 하는 것은 연구자의 주관에 달렸다.

$$Y = f(X_1, X_2)$$



변수 사이의 연관성

독립변수와 종속변수의 사이에 좀 더 복잡한 관계가 있을 수도 있다.

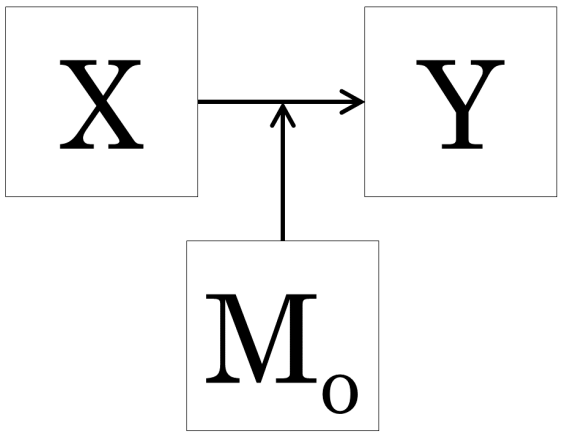
- 매개변수(mediating variables)라는 형태도 있다.
- "상대적 박탈감이 생겨나면 비슷한 위치의 사람들끼리 모이기 쉬워지고, 이렇게 모인 사람들은 집회시위에 참여하는 경향이 있다."
- 아래 식에서 M_e는 X와 Y 사이에서 매개변수 역할을 수행한다.

$$Y = f_2(M_e) = f_2(f_1(X))$$

변수 사이의 연관성

- 조절변수(moderating variables)라는 형태도 있다.
- "교육수준이 증가할수록 임금은 높아진다. 그러나 남자의 경우 그 연관성이 상대적으로 강한 반면, 여자의 경우 그 연관성이 상대적으로 약하다."
- 아래 식에서 M_o는 X와 Y 사이에서 조절변수의 역할을 수행한다.

$$Y = f(X, M_o, X \cdot M_o)$$



변수 사이의 연관성

두 변수 사이의 관계를 살펴볼 때도 자료유형에 주의를 기울여야 한다.

- 변수의 척도(scale)에 따라 조금씩 다른 분석방법을 사용해야 하기 때문이다.
- 예를 들어 두 변수 모두 등간(interval) 또는 비율(ratio)인 경우(=양적변수)에는 상관계수(correlation coefficient)를 자주 사용한다.
- 앞으로 사회통계학에서 내내 이러한 다양한 기법들에 관해 배우게 된다.
- 분석기법을 사용하기에 적절한 상황과 해석방법을 혼동하지 않도록 주의해야 한다. 결국 많은 연습이 답이다.

두 양적변수 사이의 연관성

두 양적변수 사이의 연관성

두 양적변수 사이의 연관성을 분석할 때 상관계수를 이용할 수 있다.

- 상관계수(correlation coefficient)는 두 양적변수 사이 선형관계(linear relationship)의 방향(direction)과 강도(strength)를 보여준다.
- 상관계수가 양(+)인 경우 두 변수는 같은 방향으로 변화한다("X가 증가하면 Y도 증가한다").
- 상관계수가 음(-)인 경우 두 변수는 다른 방향으로 변화한다("X가 증가하면 Y는 감소한다").
- 상관계수가 큰 값이면 두 변수는 좀 더 밀접하게 함께 움직인다("X가 변화하면 Y도 민감하게 변화한다").
- 상관계수가 작은 값이면 두 변수는 훨씬 느슨한 관계를 갖는다("X가 변화해도 Y는 둔감하게 변화한다").

두 양적변수 사이의 연관성

상관계수를 계산하려면 먼저 공분산에 대해 이해할 필요가 있다.

- 다시 한 번 X의 분산(variance) 공식을 돌이켜보자.

$$Var(X) \equiv \sigma_X^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2$$
$$= \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)(X_i - \mu)$$

- 변수가 하나(X_i) 주어져 있을 때 편차(deviation)를 제곱했는데 이렇게 바꾸면 어떨까?

$$Cov(X, Y) \equiv \sigma_{XY} = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu_X)(Y_i - \mu_Y)$$

- 이것이 공분산(covariance)이다.

두 양적변수 사이의 연관성

공분산과 분산의 아이디어는 거의 똑같다!

- 다만 (분산처럼 X_i의 편차를 제곱하는 대신) 공분산에서는 X_i와 Y_i의 편차를 서로 곱했을 뿐이다.
- 그렇게 두 변수의 편차끼리 곱한다면 어떤 값이 나올지 상상해보자.
- 만약 모두 양수(+)이거나 음수(-)이면 공분산은 양수(+)가 되고,
- 만약 한쪽이 양수(+)이고 다른 쪽이 음수(-)이면 공분산은 음수(-)가 된다.

두 양적변수 사이의 연관성

예제 1. gpa.csv는 10명의 학생들이 받은 다섯 과목의 기말시험 점수를 0점과 100점 사이에서 나타내고 있다. (1) 수학 점수와 사회 점수, (2) 수학 점수와 물리 점수 간 공분산을 각각 구하시오. 이 분석에 따를 때 수학 점수가 높은 학생은 어떤 과목을 더 잘한다고 볼 수 있는지 평가하시오.

두 양적변수 사이의 연관성

- 수학 점수와 물리 점수 사이의 공분산을 다음의 공식에 따라 계산해보자.

$$Cov(X,Y) = \sigma_{XY} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_Y)$$

- 엑셀에서 COVARIANCE.P(·) 함수를 사용해서 검산해보자.
- 이것은 모집단의 공분산이고 표본의 공분산도 계산할 수 있음을 기억하자.
- (1) 수학 점수와 사회 점수 사이의 공분산과 (2) 수학 점수와 물리 점수 사이의 공분산도 엑셀 함수로 계산해보자.
- 왜 수학와 점수와 물리 점수는 양(+)의 연관성을 가지고 있고, 그 강도가 수학과 사회의 연관성보다 더 높다고 할 수 있을까?

두 양적변수 사이의 연관성

공분산은 흥미로운 아이디어를 제시하고 있지만 명확한 단점이 있다.

- 공분산 $Cov(X, Y)$은 X 내부(within X)의 분산과 Y 내부(within Y)의 분산이 다를 수 있다는 점을 고려하지 않는다.
- 둘을 비교하려면 일단 표준화를 해야 하는데 제대로 표준화가 안되었다는 의미다.
- 표준화가 되지 않은 숫자가 그냥 계산되었기 때문에 공분산값은 직관적으로 그 의미를 파악하기 어렵다.
- "아니, 공분산이 241.12 이라고 나왔는데 대체 이게 뭘 뜻하지?"
- 이 문제는 (공분산 뿐 아니라) 분산에서도 마찬가지로 발견된다! 단위가 달라지면 분산과 공분산 모두가 달라진다.

두 양적변수 사이의 연관성

예제 2. showmethemoney.csv는 8명의 시민들이 설문에 참여하여 월 소득(income), 월급 백만원 단위 카테고리(income_cat), 그리고 거주지 평수(housesize)를 보고한 자료이다. (1) income과 housesize의 공분산과 (2) income_cat과 housesize의 공분산을 계산하시오. 공분산은 같은지 다른지 논하시오. 이를 토대로 income과 income_cat이 같은 변수인지 다른 변수인지 논하시오.

두 양적변수 사이의 연관성

Karl Pearson은 공분산의 한계를 보완하는 천재적인 접근을 제시했다.

- 그는 공분산을 두 변수 X와 Y의 각각의 표준편차로 나누어줌으로서, 공분산을 X 내부의 분산과 Y 내부의 분산에 대해 표준화했다(Why?).
- 뿐만 아니라, 일부러 분산이 아닌 두 개의 표준편차 곱으로 나누어주었기 때문에 표준화된 값은 절묘하게 -1과 1 사이에 놓이게 된다!
- 이것이 이른바 피어슨 적률상관계수(Pearson's product-moment correlation coefficient)이다. 줄여서 상관계수 ρ라고 한다(ρ는 rho라고 읽는다).

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

두 양적변수 사이의 연관성

다시 showmethemoney.csv로 돌아가 상관계수를 계산해보자.

- 먼저 income과 housesize의 공분산을 다시 계산해보자. 엑셀 함수는 COVARIANCE.P(·)이다.
- 다음으로 income의 표준편차와 housesize의 표준편차를 각각 계산해보자. 엑셀 함수는 STDEV.P(·)이다.
- 아래 식에 따라 엑셀에서 계산해보자. 엑셀에서 계산할 때는 괄호에 주의할 것!

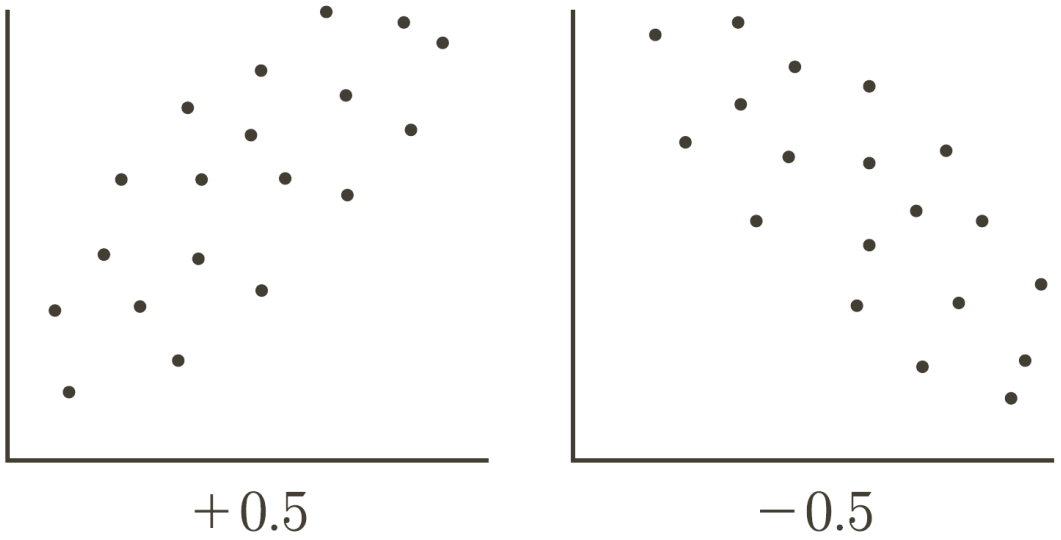
$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

- 상관계수는 엑셀에서 CORREL(·) 함수로 편리하게 구할 수 있다.

두 양적변수 사이의 연관성

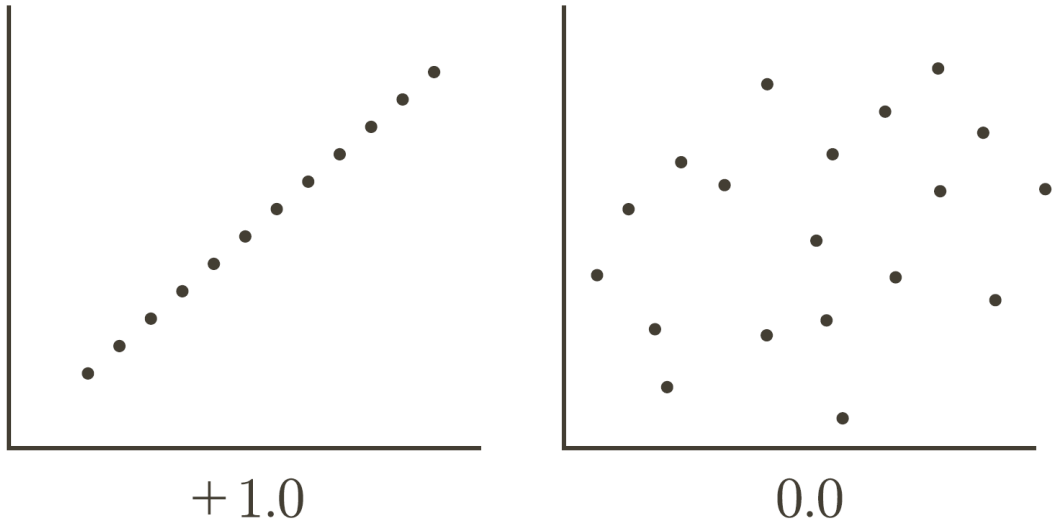
상관계수의 해석은 익숙해지기 전까지 다소 혼동스러울 수도 있다.

- 상관계수는 반드시 -1과 1사이에 놓인다.
- 상관계수가 0보다 크면 두 변수는 서로 같은 방향으로 움직인다("X가 증가하면 Y도 증가한다"). 상관계수가 0보다 작으면...
- 상관계수는 산점도(scatterplot)를 통해 시각화(visualization)할 수도 있다.



두 양적변수 사이의 연관성

- 상관계수가 1에 가까울수록 두 변수 사이 양(+)의 상관성이 높아진다. 상관계수가 -1에 가까울수록...
- 극단적으로 상관계수가 -1 혹은 1 인 경우를 완전상관(perfect correlations), 상관계수가 0인 경우를 무상관(no correlations)이라고 부른다.
- (기울기가 아니라) 관찰값 사이의 흩어짐(산포경향)이 상관계수의 본질이다.



두 양적변수 사이의 연관성

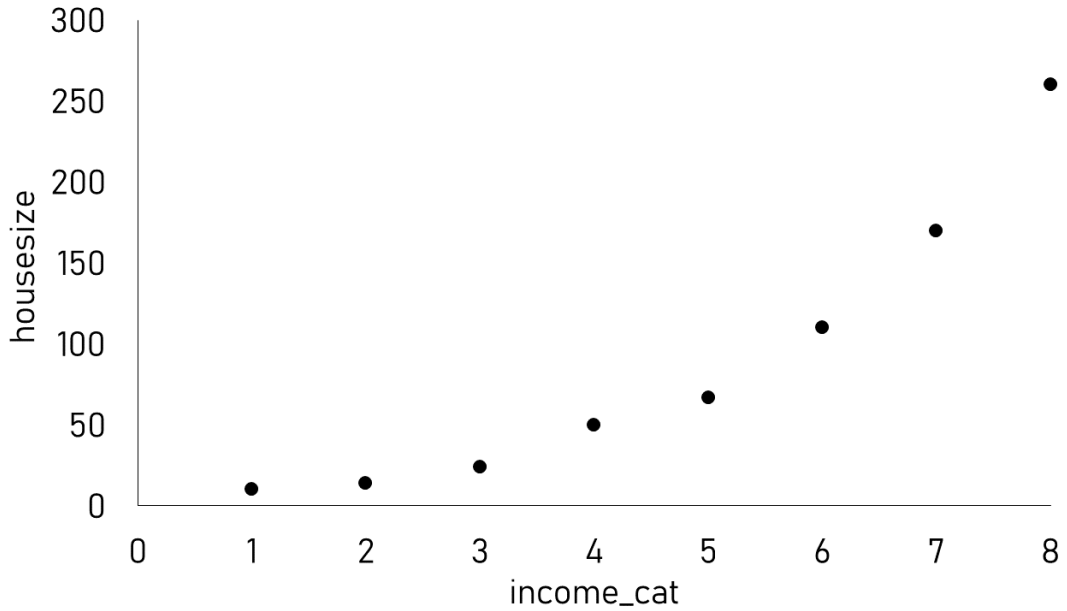
- 상관계수의 절대값(absolute value)도 몇 가지 방식으로 해석할 수 있다.
- 0과 1 사이를 적당히 사분위수(quartiles)로 나눈 뒤, 각각 리커트 4점 척도(1 사분위수=매우 약한 상관성; 2사분위수=다소 약한 상관성; 3사분위수=다소 강한 상관성; 4사분위수=매우 강한 상관성)처럼 해석할 수 있다.

|상관계수|	상관성의 해석
0 ~ 0.25	매우 약한 상관성
0.25 ~ 0.5	다소 약한 상관성
0.5 ~ 0.75	다소 강한 상관성
0.75 ~ 1	매우 강한 상관성

두 양적변수 사이의 연관성

상관계수는 두 변수 사이의 선형관계를 전제로 한다.

- 그런데 아까 showmethemoney.csv에서 income (또는 income_cat)과 housesize 사이의 관계는 사실 비선형적(non-linear)이다.



두 양적변수 사이의 연관성

- X와 Y의 관계는 경우에 따라서 U자형, 역U자형, W자형 등등 다양할 수도 있다.
- 그러면 어떻게 상관계수를 확인하기 전에 두 변수의 관계가 선형적인지 알 수 있나?
- 산점도를 그려보면 알 수 있다. 그러므로 반드시 상관계수를 계산할 때는 산점도를 함께 확인해야 한다(산점도 그리는 법은 다음 주에 배운다).

두 양적변수 사이의 연관성

예제 3. 유동인구수.xlsx를 엑셀로 불러오자. 남자10대가 다니는 (혹은 다니지 않는) 시간대 및 거리의 패턴이 가장 비슷한 인구집단이 누구인지 식별하고자 한다. 이 자료에서 주어진 유동인구수 변수는 어떤 척도인지 답하시오. 성별-연령별 유동인구수 변수의 연관성을 파악하기 위해 어떠한 분석기법이 가장 적절한지 고르고 그 이유를 설명하시오. 적절한 분석을 수행하고 결론과 해석을 내리시오.