사회통계연습
Association of Two or More Variables

김현우, PhD[1]

[1] 충북대학교 사회학과 조교수

September 17, 2021

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

진행 순서

1. 지난 주 리뷰
2. 과학적 이론과 변수 사이의 관계
3. 둘 이상의 숫자형 변수 사이의 관계

지난 주 리뷰

지난 주 리뷰

퀴즈 2 코멘트

- 내 잘못! 내가 직접 풀어볼 때 선택한 지역들에서는 필터링 한 뒤에도 관찰값이 적어도 5-10 개 이상은 나왔는데, 특정 지역에서는 너무 사례가 적게 나오는 경우(심지어 2 개 이하)가 있었음. 이런 경우 요약통계량을 구할수 없거나 아예 의미가 없어짐.
- 중심성향(mean, mode, median)을 통해 높고 낮음을 비교할 수 있고, 산포성향 (range, IQR, s.d.)을 통해 변동성을 비교할 수 있다. 변동성을 물어보는데 중심성향 이야기를 하면 그냥 틀린 답이다. 혼동하지 말 것!
- 소숫점은 대체로 2-3 자리 정도면 충분. 10의 자리까지 쓰는 것은 오바!
- 문제에서 "완성된 문장으로 비교하시오" 라고 쓴 이유는 그냥 요약통계량만 툭 던지면 안되고 쉬운 말로 풀이를 제시해야 하기 때문.
- 가령, "A지역은 평균은 6이고, B지역의 평균의 3이다." "A지역은 표준편차는 10 이고, B지역의 표준편차는 5이다." 로 끝내서는 안됨!
- 반드시 "A지역에서 미세먼지는 B지역보다 평균적으로 농도가 짙지만 변동성은 낮다" 하는 식의 평범한 우리말 표현이 뒤따라야함.
- 모든 독자들이 표준편차 같은 개념의 의미를 당연히 알아들을 거라고 생각해선 안됨.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

지난 주 리뷰

퀴즈 2 코멘트

- 문제에서 “세 요약통계량이 서로 상이한 결론을 내리고 있는가?” 라고 물을 때는 세 요약통계량의 값이 정확히 똑같은가를 묻고 있는 것이 아님. 결론이 똑같은가를 묻는 것임.
- 가령 range, IQR, S.D.를 비교해보니 “range에 의했을 때는 A지역 변동이 크고, IQR를 따랐을 때는 B지역 변동이 큰데, S.D.는 다시 A지역 변동이 크다. 나는 S.D.를 따라 A지역 변동이 크다고 본다. 왜냐하면 ...” 하는 식의 설명을 기대했음.
- 몇몇 학생들은 “A지역에서 PM10이 높고 B지역에서 PM2.5가 높았는데 어떻게 보고하면 좋은가?” 라는 질문을 하였음. 나에게 질문한 그대로 그대로 답하면 됨.
- PM10은 미세먼지, PM2.5는 초미세먼지임. 다른 것들을 측정하고 있음. 아마도 문제가 “당신의 생일날 미세먼지는 어느 지역에서 더 변동이 심했는가?”를 묻고 있었기 때문에 하나의 답만 써야하는 것으로 망설였나 봄. 아니었음. 그래도 이 문제로는 감점이 들어가지 않음.
- “충북 청주시와 충북 충주시의 PM10을 비교한 결과, 충북 청주시보다 제주 서귀포시의 미세먼지가 나쁘다는 결론을 내렸다.” 뭔 소리여?

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

지난 주 리뷰

총평

- 숙제를 안내면 결국 점수를 많이 잃을 수 밖에 없는 구조임.
- 질문 메일도 hxk271@cbnu.ac.kr로 보낼 것.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

과학적 이론과 변수 사이의 관계

과학적 이론과 변수 사이의 관계

저번 주에는 아래의 주제를 중심으로 연습하였다.

- 데이터의 요약(I): 중심성향
- 데이터의 요약(II): 산포성향

곰곰히 생각해보면 예전에는 단지 하나의 변수를 요약하였을 뿐이다.

- 설령 두 개의 변수(e.g., PM10과 PM2.5)를 살펴보았다고 할지라도 각각 따로 보았다.
- 두 변수 사이의 관계를 살펴보지는 않았다.

그래서 이번 주에는 둘 이상의 변수 사이의 "관계"를 보기로 한다.

- 그런데 생각처럼 하나의 변수를 볼 때와 크게 다른 것은 아니다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

‘

과학적 이론과 변수 사이의 관계

경험과학(empirical science)에서 이론(theory)의 정의와 역할에 대해 잠깐 생각해보자.

과학적 이론(scientific theory)

- 과학적 이론은 경험적으로 검증가능한, 상호연관된 일련의 명제들(propositions)이다.
- 명제란 둘 이상의 개념들(concepts) 사이의 관계에 대한 진술이다(e.g., "X가 증가하면 y는 감소한다").
- 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.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

과학적 이론과 변수 사이의 관계

개념(concept)

- 추상적인 현상을 경계지어 개념으로 부르지만, 이는 경험적 연구를 위해 조작적 정의(operational definition)를 거쳐 변수(variables)로 측정(measure)되어야 한다.

가설(hypothesis)과 발견(finding)

- 가설이란 (연구를 수행하기에 앞서 제시된) 둘 이상의 개념들 사이의 관계에 대한 "잠정적" 진술이다.
- 발견이란 (연구를 통해 어느 정도 입증된) 둘 이상의 개념들 사이의 관계에 대한 진술(=명제)이다.
- 흥미롭게도 경험과학에서는 법칙(law)이나 사실(fact)과 같은 표현이 거의 나오지 않는다. 오히려 그런 말은 저널리즘에서 폭넓게 쓰인다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

과학적 이론과 변수 사이의 관계

결국 경험과학에서 이론이란 근본적으로 둘 이상의 변수 사이에 어떤 관계가 있는지를 다루는 것이다!

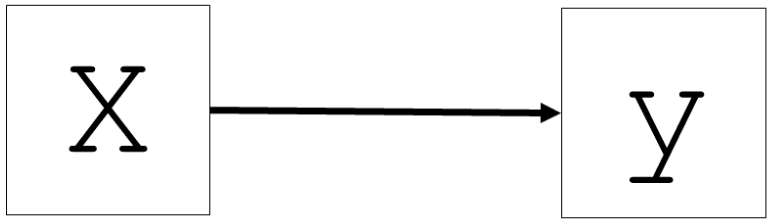
- 하나의 변수를 분석/요약하는 통계로는 이론을 세우거나 테스트할 수 없다. 그런 것은 이론이 아니다.
- 적어도 이론이라고 불리려면 두 개 이상의 변수의 관계를 분석해야만 한다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

과학적 이론과 변수 사이의 관계

종속변수와 독립변수

- 종속변수(dependent variable)란 다른 무언가에 종속되어 설명되어지는 변수다.
- 독립변수(independent variable)란 다른 무언가로부터 독립되어 설명하는 변수다.
- $y = f(x)$



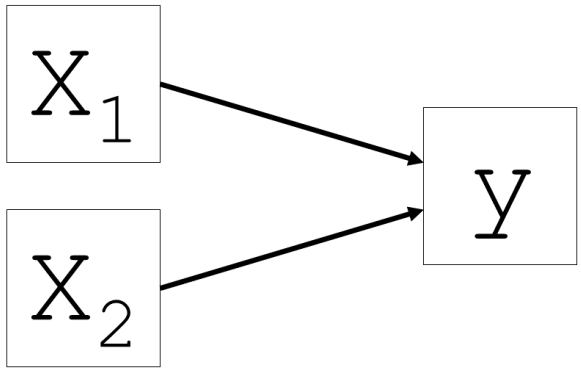
독립변수 →종속변수

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

과학적 이론과 변수 사이의 관계

독립변수는 하나 이상일수도 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 엄격하게 살펴볼 수 있다.
- 근본적으로 무엇이 관심변수이고 통제변수인가는 하는 것은 연구자의 주관에 달렸다.
- $Y = f(X_1, X_2)$



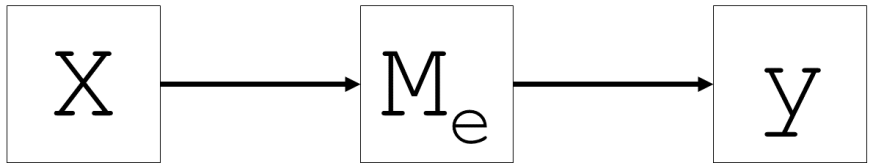
통제변수의 이용

충북대학교 CHUNGBUK NATIONAL UNIVERSITY

과학적 이론과 변수 사이의 관계

독립변수와 종속변수의 사이에 좀 더 복잡한 관계가 있을 수도 있다.

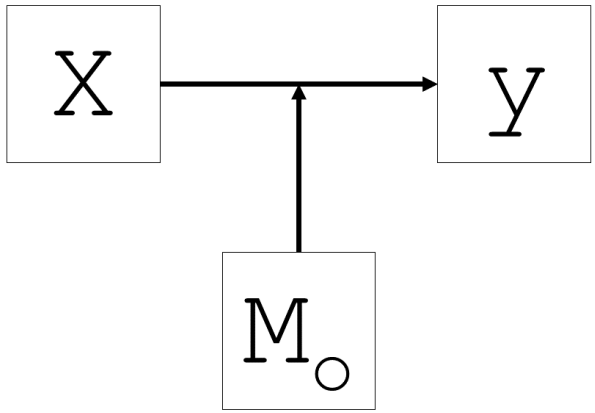
- 매개변수(mediating variables)라는 형태도 있다.
- $Y = f_2(M_e) = f_2(f_1(X))$



충북대학교
CHUNGBUK NATIONAL UNIVERSITY

과학적 이론과 변수 사이의 관계

- 조절변수(moderating variables)의 형태도 있다.
- $Y = f(X, M_o, X \cdot M_o)$



충북대학교
CHUNGBUK NATIONAL UNIVERSITY

과학적 이론과 변수 사이의 관계

둘 이상의 변수 사이의 관계를 살펴볼 때는 결국 몇 가지 방법으로 요약된다.

- 변수의 척도(scale)에 따라 조금씩 다른 방법을 필요로 한다.
- 두 변수가 등간형(interval), 비율형(ratio)인 경우(숫자형; numerical)에는 상관계수(correlation coefficient)를 자주 사용한다.
- 두 변수가 명목형(nominal), 서열형(ordinal)인 경우(범주형; categorical)에는 교차표(cross-tabulation)을 자주 사용한다.

해석 방법과 주의사항을 혼동하지 않도록 주의해야 한다.

- 결국 많은 연습이 답이다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

둘 이상의 숫자형 변수 사이의 관계

둘 이상의 숫자형 변수 사이의 관계

둘 이상의 숫자형(numerical) 변수 사이의 관계를 볼 때는 일차적으로 상관분석(correlation analysis)을 수행한다.

- 상관분석은 상관계수(correlation coefficient)를 구하는 과정이다.
- 상관계수를 이해하려면 먼저 공분산(covariance)을 이해할 필요가 있다.

다시 한 번 분산(variance)의 식을 돌이켜보자.

$$\begin{aligned} \mathrm{Var}(X) = \sigma^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)(x_i - \mu) \end{aligned}$$

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

둘 이상의 숫자형 변수 사이의 관계

변수가 하나(x_i) 주어져 있을 때 편차(deviation)를 구해 제곱했는데 이걸 이렇게 바꾸면 어떨까?

$$\text{Cov}(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)$$

공분산과 분산의 아이디어는 거의 똑같다!

- 다만 x_i 하나의 편차를 제곱하는 대신, x_i와 y_i의 편차를 서로 곱했을 뿐이다.
- 두 변수의 편차끼리 곱할 때, 모두 양수(+)이거나 음수(-)이면 양수(+)가 되고, 어느 한쪽이 양수(+)이고 다른 쪽이 음수(-)이면 음수(-)가 된다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

둘 이상의 숫자형 변수 사이의 관계

공분산을 엑셀에서 직접 계산해보자.

- eCampus에서 sociologists.csv를 다운받아 엑셀로 불러오자. 이것은 다섯 명의 사회학자에 대한 선호도를 0점과 100점 사이에서 조사한 결과이다.
- Weber 선호도의 분산(variance)을 직접 계산해보자. var.p() 함수를 사용해서 검산해보자.
- Weber와 Marx 선호도 사이의 공분산(covariance)을 계산해보자.

$$\mathrm{Cov}(X,\ Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)$$

- Weber와 Marx 선호도 사이의 공분산(covariance)을 계산해보자. covariance.p() 함수를 사용해서 검산해보자.
- Marx와 Parsons 선호도 사이의 공분산을 계산해보자. Marx와 Burawoy 선호도 사이의 공분산도 계산해보자.
- Marx를 선호하는 사람은 둘 중 누구를 선호할까?

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

둘 이상의 숫자형 변수 사이의 관계

공분산은 흥미로운 아이디어를 제시하고 있지만 명확한 단점이 있다.

- Cov(X, Y)는 X 내부(within X)의 분산과 Y 내부(within Y)의 분산이 다를 수 있다는 점을 고려하지 않는다. 제대로 표준화가 안되었다는 의미다.
- 그 결과 그 자체로는 해석이 안된다. 아니 대체 공분산이 241.12 이라고 나왔는데 그게 뭘 의미하나?

무슨 소리인지 잘 이해가 가지 않으니 직접 연습을 해보자!

- eCampus에서 showmethemoney.csv를 다운받아 열어보자.
- 8명의 시민들이 설문에 참여하여 월 소득, 월급 카테고리(백만원 단위), 그리고 자신의 방 평수를 보고하였다.
- Cov(income, housesize)를 계산해보라. 또 Cov(income_cat, housesize)를 계산해보라.
- 결과가 같은가? 다른가? 정말 income과 income_cat이 같은 변수인가, 아니면 다른 변수인가?

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

둘 이상의 숫자형 변수 사이의 관계

Karl Pearson은 이를 보완하는 천재적인 접근을 제시했다.

- 그는 두 변수 X와 Y의 각각의 표준편차(분산이 아니고!)를 분모로 각각 나누어줌으로서 X 내부의 분산과 Y 내부의 분산이 다를 수 있는 가능성을 제거하고 표준화를 이루었다.
- 뿐만 아니라, 일부러 분산이 아닌 표준편차로 나누어주었기 때문에 표준화된 값은 절묘하게 -1과 1사이로 두 변수가 얼마나 강한 상관관계를 가지고 있는지 보여준다.
- 이것이 이른바 피어슨의 적률상관계수(Pearson's product-moment correlation coefficient)이다. 줄여서 상관계수(ρ)다. ρ는 rho라고 읽는다.
- 전에 언급한 바 있듯, σ는 sigma라고 읽고 표준편차를 의미한다.

$$\rho = \frac{\mathrm{Cov(X,Y)}}{\sigma_x \sigma_y}$$

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

둘 이상의 숫자형 변수 사이의 관계

다시 showmethemoney.csv로 돌아가 상관계수를 계산해보자.

- 먼저 Cov(income, housesize)를 계산해보자. 엑셀 함수는 covariance.p()이다.
- 다음으로 income의 표준편차와 housesize의 표준편차를 각각 계산해보자. 엑셀 함수는 stdev.p()이다.
- 아래 식대로 엑셀에서 계산해보자. 괄호에 주의할 것!

$$\rho = \frac{\mathrm{Cov(X,Y)}}{\sigma_x \sigma_y}$$

- 상관계수를 편리하게 구하는 엑셀함수는 correl()이다. 검산을 해보자.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

둘 이상의 숫자형 변수 사이의 관계

상관계수에 관한 혼동하기 쉬운 상식!

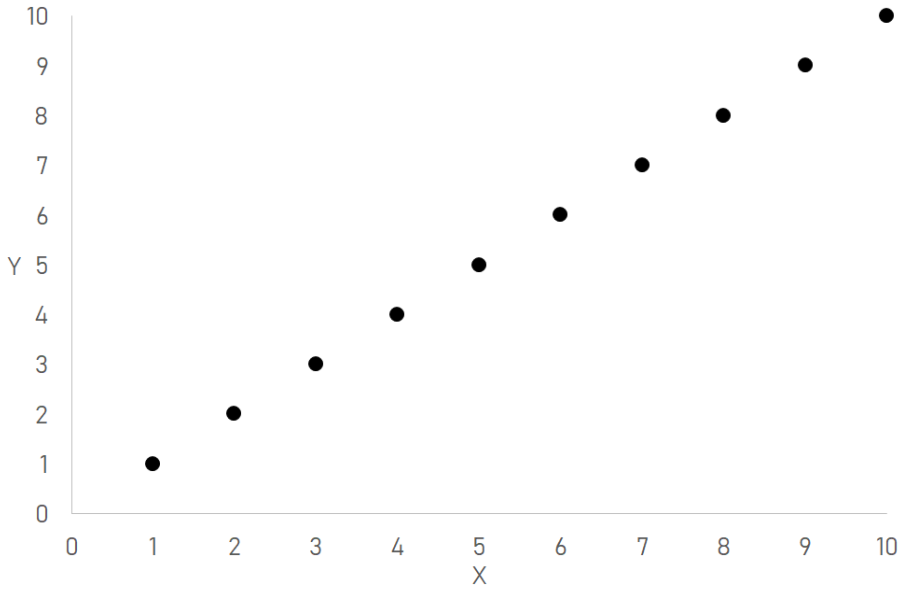
- 상관계수는 반드시 -1과 1사이에 놓인다.
- 0과 1 사이를 적당히 사분위수로 나눈 뒤, 각각 리커트 4점 척도와 같이 해석할 수 있다. 물론 0과 -1 사이도 마찬가지이다.
- 상관계수가 0보다 크면 두 변수는 서로 같은 방향으로 움직인다. 0보다 작으면...
- 상관계수가 0보다 크면 클수록(그리고 0보다 작으면 작을수록) 두 변수는 더욱 서로 긴밀하게 발맞추어 움직인다.
- 산포도(scatterplot)를 통해 대략 상관계수가 어떻게 나올지 짐작할 수도 있다.
- 하지만! 의외로 기울기는 상관계수의 본질이 아니다. 관찰값 사이의 흩어짐(산포경향)이 상관계수의 본질이다.

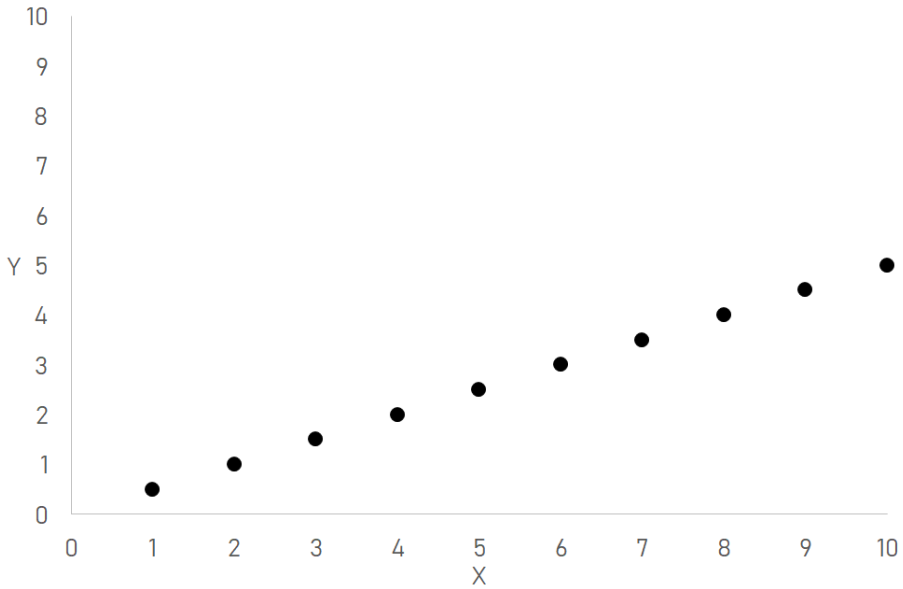
충북대학교
CHUNGBUK NATIONAL UNIVERSITY

둘 이상의 숫자형 변수 사이의 관계

- 왼쪽과 오른쪽 산포도 둘 다 상관계수(ρ)는 1이다.

X(공통)	Y(왼쪽)	Y(오른쪽)	X(공통)	Y(왼쪽)	Y(오른쪽)
1	1	0.5	6	6	3
2	2	1	7	7	3.5
3	3	1.5	8	8	4
4	4	2	9	9	4.5
5	5	2.5	10	10	5



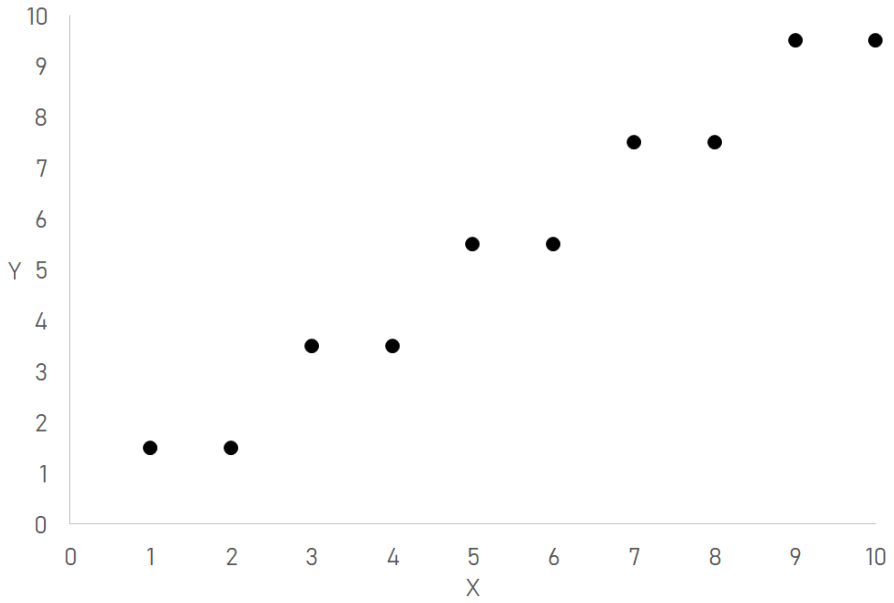


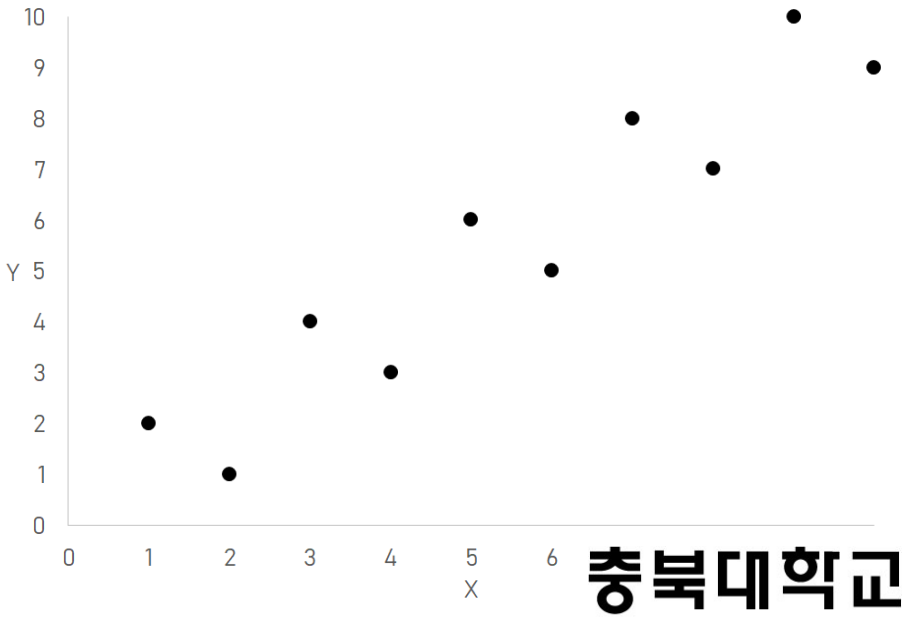
충북대학교
CHUNGBUK NATIONAL UNIVERSITY

둘 이상의 숫자형 변수 사이의 관계

- 왼쪽 산포도에서 Y는 X로부터 ± 0.5씩 더해 계산되었고, 오른쪽 산포도에서 Y는 X로부터 ± 1씩 더해 계산되었다.
- 왼쪽 상관계수($\rho = .98$)가 오른쪽 상관계수($\rho = .94$)보다 크다.

X(공통)	Y(왼쪽)	Y(오른쪽)	X(공통)	Y(왼쪽)	Y(오른쪽)
1	1.5	2	6	5.5	5
2	1.5	1	7	7.5	8
3	3.5	4	8	7.5	7
4	3.5	3	9	9.5	10
5	5.5	6	10	9.5	9



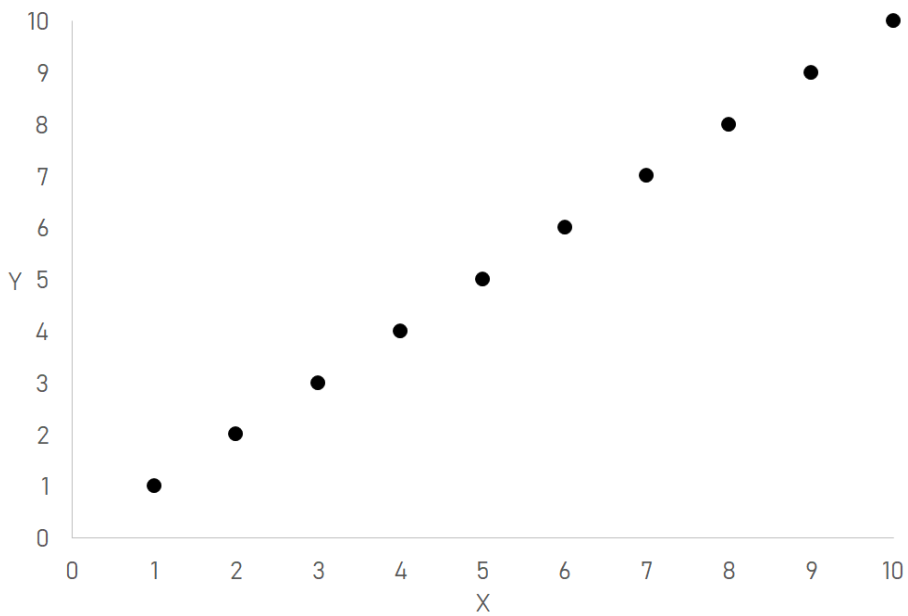


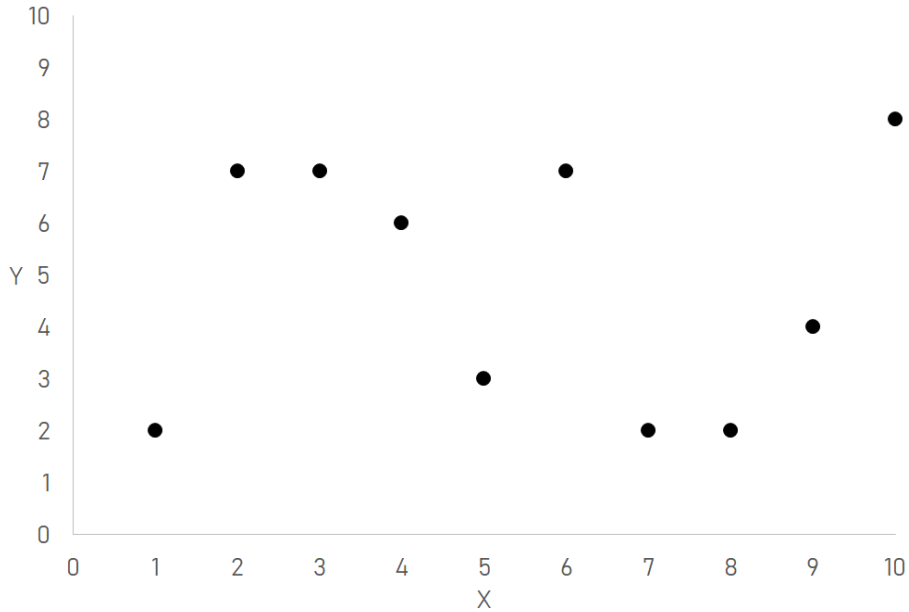
충북대학교 CHUNGBUK NATIONAL UNIVERSITY

둘 이상의 숫자형 변수 사이의 관계

왼쪽 산포도의 상관계수(ρ)이지만 (규칙없이 퍼진) 오른쪽 산포도의 ρ는 0이다.

X(공통)	Y(왼쪽)	Y(오른쪽)	X(공통)	Y(왼쪽)	Y(오른쪽)
1	1	2	6	6	7
2	2	7	7	7	2
3	3	7	8	8	2
4	4	6	9	9	4
5	5	3	10	10	8



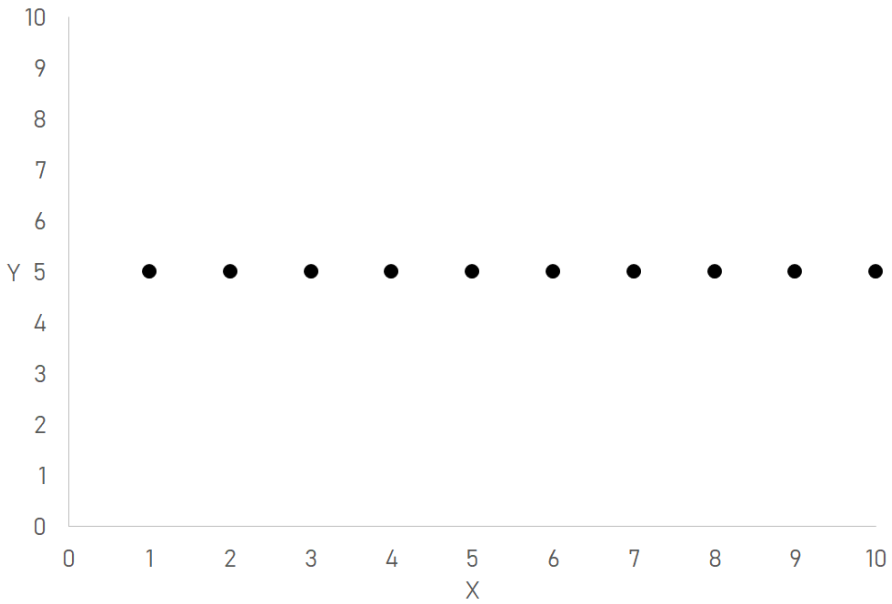


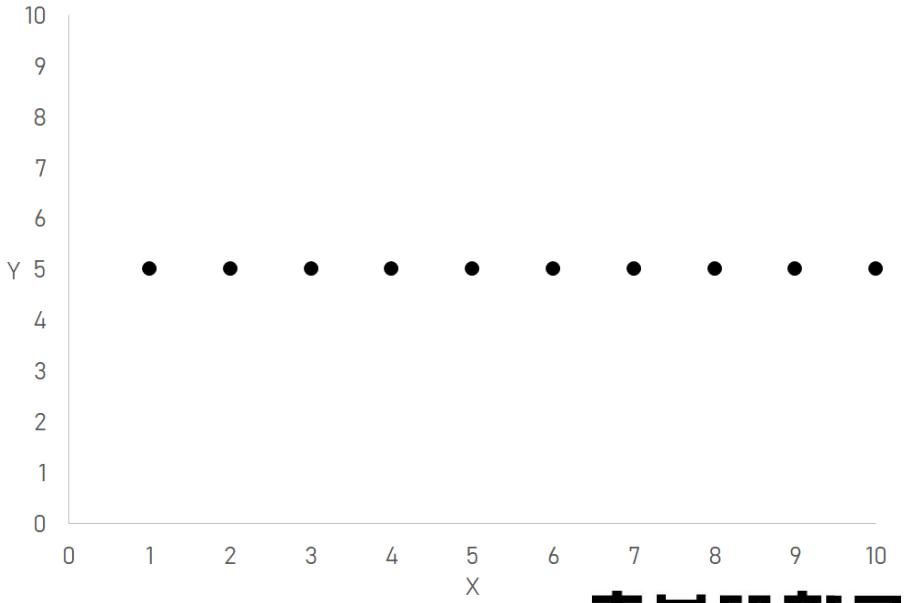
충북대학교
CHUNGBUK NATIONAL UNIVERSITY

둘 이상의 숫자형 변수 사이의 관계

- 왼쪽 산포도에서 Y는 .0001씩이라도 커져서 $\rho = 1$이지만, 오른쪽 산포도는 전혀 변동하지 않아 수평선이고 (ρ)는 아예 계산되지 않는다(Why?).

X(공통)	Y(왼쪽)	Y(오른쪽)	X(공통)	Y(왼쪽)	Y(오른쪽)
1	5	5	6	5.0005	5
2	5.0001	5	7	5.0006	5
3	5.0002	5	8	5.0007	5
4	5.0003	5	9	5.0008	5
5	5.0004	5	10	5.0009	5



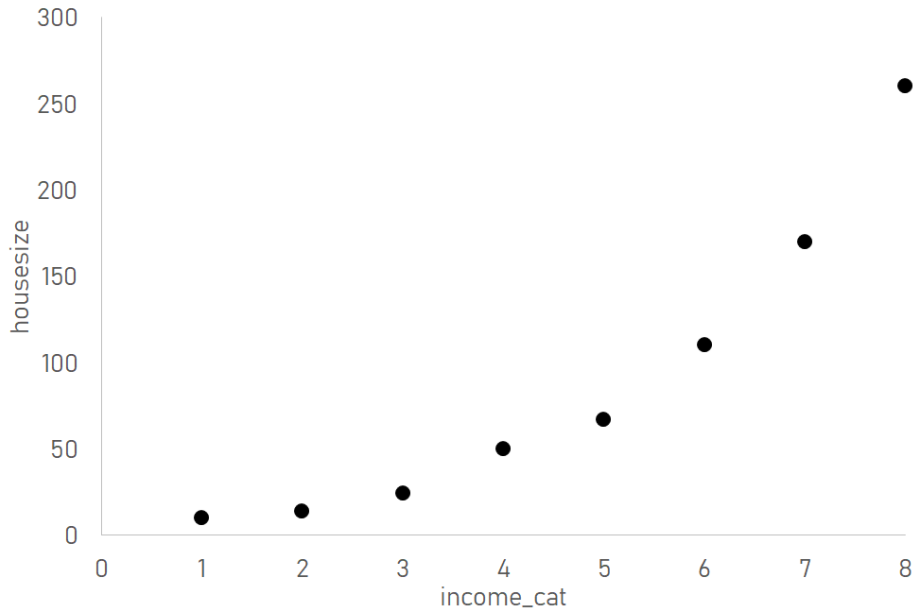


충북대학교
CHUNGBUK NATIONAL UNIVERSITY

둘 이상의 숫자형 변수 사이의 관계

그런데 아까 showmethemoney.csv에서 사실 income (또는 income_cat)과 housesize 사이의 관계는 선형적(linear)이지 않다.

- 이것은 비선형적(non-linear)이다. 보다 구체적으로는 곡선형(curvilinear)이다.



- X와 Y의 관계는 경우에 따라서 U자형, 역U자형, W자형 등등 다양할 수도 있다.
- 상관분석은 기본적으로 X와 Y의 관계가 선형적일 것으로 가정하기 때문에 만일 데이터가 그렇지 않으면 문제를 일으킨다.
- 나중에 이에 관해 더 깊이있게 검토한다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

둘 이상의 숫자형 변수 사이의 관계

eCampus에서 유동인구수.xlsx를 다운받아 엑셀에서 열자.

- 저명한 한국의 사회학자 시예진은 유유상종의 원리(homophily principle)에 따라 거리를 활보할 때도 남녀가 또래끼리(20대는 20대끼리, 30대는 30대끼리 등) 함께 다닐 것이라는 가설(hypothesis)을 세웠다.
- 당신 생각에 어떤 두 변수를 고르면 좋을까?
- 두 변수는 각각 어떤 척도라고 생각되나?
- 두 변수 사이의 관계를 알고 싶을때 어떤 분석이 적합할까?
- 해봐.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY