사회통계연습
상관분석과 산점도

김현우, PhD[1]

[1] 충북대학교 사회학과 부교수

진행 순서

1. 상관분석 톺아보기
2. 유의성 검정
3. 상관분석 연습
4. 심슨의 역설

상관분석 톺아보기

상관분석 톺아보기

두 양적변수 사이의 관계를 볼 때는 일차적으로 상관분석을 수행한다.

- 상관분석은 상관계수(correlation coefficient)를 구하는 기법이고, 상관계수를 이해하려면 먼저 분산(variance)과 공분산(covariance)을 돌이켜 볼 필요가 있다.
- 먼저 다시 분산의 식을 돌이켜보자.

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

- 변수가 하나만 주어져 있을 때 편차(deviation)를 구해 제곱하여 분산을 구했는데, 아래처럼 살짝 바꾸면 곧바로 공분산이 된다.

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

상관분석 톺아보기

- 공분산과 분산의 아이디어는 거의 똑같다!
- 두 변수의 편차끼리 곱할 때, 모두 양수(+)이거나 음수(-)이면 공분산은 양수(+)가 되고, 어느 한쪽이 양수(+)이고 다른 쪽이 음수(-)이면 공분산은 음수(-)가 된다.
- 공분산은 흥미로운 아이디어를 제시하고 있지만 명확한 단점이 있었다. $\mathrm{Cov}(X, Y)$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 어려웠기 때문에 새로운 대안이 필요했다.

상관분석 톺아보기

Karl Pearson은 공분산의 단점을 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) ρ이다.

$$\rho = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

상관분석 톺아보기

상관계수의 해석은 매우 간단하지만 혼동하지 않도록 주의해야 한다.

- 상관계수는 반드시 -1과 1사이에 놓인다. 상관계수가 1에 가까울수록 (그리고 -1에 가까울수록) 두 변수는 높은 연관성을 가지고 있다고 할 수 있다(Why?).
- 상관계수가 0보다 크면 두 변수는 서로 같은 방향(정방향)으로 움직인다. 즉 "X가 증가하면 Y도 증가한다."
- 상관계수가 0보다 작으면 두 변수는 서로 다른 방향(역방향)으로 움직인다. 즉 "X가 증가하면 Y는 감소한다."
- 왜 이렇게 해석되는지 공분산의 분자 부분을 잘 들여다보자.

상관분석 톺아보기

- 해석은 이런 식이 편리하다: 0과 1 사이를 사분위수로 나누고 각각 리커트 4점 척도로 의미를 부여한다. 물론 0과 -1 사이에서도 마찬가지이다.

상관계수	상관관계의 해석
[-1, -0.75]	매우 강한 역방향
[-0.75, -0.5]	다소 강한 역방향
[-0.5, -0.25]	다소 약한 역방향
[-0.25, 0]	매우 약한 역방향
[0, 0.25]	매우 약한 정방향
[0.25, 0.5]	다소 약한 정방향
[0.5, 0.75]	다소 강한 정방향
[0.75, 1]	매우 강한 정방향

유의성 검정

유의성 검정

상관계수에 대해서도 다음과 같이 유의성 검정을 할 수 있다.

- 상관계수에 대해서는 무조건 양측검정을 수행하며, 주어진 가설 구조는 다음과 같다:

$$H_0 : \rho = 0$$
$$H_a : \rho \neq 0$$

- t 검정을 통해 유의확률(p-value)을 구한다. 단, t 분포의 꼴은 $(n-2)$의 자유도로 결정된다.
- 검정통계량 t 값은 다음과 같다:

$$t = \frac{\hat{\rho} - \rho}{SE_\rho} = \frac{\hat{\rho}}{\sqrt{\frac{1-\hat{\rho}^2}{n-2}}} = \hat{\rho}\sqrt{\frac{n-2}{1-\hat{\rho}^2}}$$

- 즉 (1) 추정된 상관계수 $\hat{\rho}$가 커지고 (2) 사례수 n가 많아질수록 t 값이 커져 귀무가설을 기각하기 쉬워진다.

유의성 검정

소표본만 벗어나도 상관계수의 유의성 검정에 사실 큰 의미가 없다.

- 위 식을 잘 살펴보면 금방 눈치챌 수 있는 부분인데, 예컨대 $n = 50$ 정도로 작은 샘플에서 $\hat{\rho} = 0.3$ 정도의 값만 나와주어도 이미 95% 신뢰수준에서 통계적으로 유의하게 귀무가설을 기각할 수 있다(Why?).
- 그런데 오늘날 경험적 사회과학 연구에서 $n = 50$ 짜리 연구는 없다. 게다가 $\hat{\rho} = 0.3$ 는 거의 없는 수준의 상관관계에 불과하다.
- 많은 보고서와 논문에서는 기술통계(descriptive statistics)의 일환으로 상관계수행렬을 제시해왔다. 보다 근래에는 상관계수행렬은 보고서 또는 논문의 한 페이지를 통째로 잡아먹기 때문에 보고하지 않는 경우도 많아졌다. 여기에 더해 (상관계수 옆에) 별까지 잔뜩 붙이다보면 페이지 공간을 쓸떼없이 차지하고 별 의미도 없다.

유의성 검정

연습 1. CENSUS13.SAV에서 모든 변수들의 연관성을 살펴보고 이를 시각화하시오.

상관분석 연습

상관분석 연습

연습 2. SOUNDTEST.SAV에서 두 변수 balance와 eval의 상관관계에 관해 논하시오.

상관분석 연습

상관관계를 볼 때 꼭 시각화하여 살펴보아야 한다.

- 설령 $\rho \approx 0$이라 하더라도 두 변수 간 관계가 없다고 결론지을 수 없다.
- 상관계수는 기본적으로 두 변수간 선형적 관계의 강도(strength of the linear relationship)를 나타내 보인다.
- 다시 말해, 두 변수 사이에 선형적이지 않은 관계, 즉 비선형적 관계(nonlinear relationship)가 있는 경우에는 상관계수가 오해를 불러온다. 애시당초 두 변수 사이가 U자형, 역U자형, W자형 등등이 아니라는 보장이 어디에 있는가?
- 게다가 상관계수는 극단치(outliers)를 가진 경우 여기에 민감하게 영향받을 수 있다.
- 이런 문제를 방지하기 위해 반드시 산점도를 그려보아야 한다.

상관분석 연습

질적변수 문항이 여럿 주어졌을 때도 경우에 따라 상관분석을 할 수 있다.

- 많은 사회조사에서는 양적척도가 좀처럼 사용되지 않고 대체로 질적척도인 경우가 많다. 반면 통계학의 분석기법(상관분석이나 회귀분석 등)은 대체로 양적변수일 때 편리하다.
- 그러므로 사회조사에서는 질적변수로 물어보고 이를 합성지수로 (재)구축하여 대체로 등간(approximately interval)인 양적변수로 바꾸는 전략이 종종 사용된다!
- 종종 대체로 등간인 양적변수들의 관계를 산점도(scatterplot)로 시각화하면 점이 나란히 줄지어서 나타나기 쉽다(Why?).

상관분석 연습

- 코드북 등을 잘 살펴보고 행렬식 문항이 있으면 내적 일관성 확인을 고민해볼 수 있다.

2 작년(2018년) 연말 기준으로 우리 동네 사람들에 대해 느낀 점을 응답해 주십시오.

항목	전혀 그렇지 않다	그렇지 않은 편이다	보통 이다	그런 편이다	매우 그렇다
1) 서로서로 잘 알고 지내는 편이었다	①	②	③	④	⑤
2) 동네에서 일어나는 일에 대해 자주 이야기했다	①	②	③	④	⑤
3) 어려운 일이 있으면 서로 잘 도왔다	①	②	③	④	⑤
4) 동네의 각종 행사와 모임에 적극적으로 참여했다	①	②	③	④	⑤
5) 동네 아이가 낯선 아이들에게 괴롭힘을 당하면 도와줄 것 같았다	①	②	③	④	⑤
6) 범죄사건이 발생하면 경찰에 신고할 것 같았다	①	②	③	④	⑤
7) 범죄예방을 위해 순찰을 해야 한다면 이 활동에 참여할 것 같았다	①	②	③	④	⑤

- 신뢰도 계수를 계산해보고 만일 특정 문항을 제외했을 때 신뢰도 계수가 크게 상승한다면 이를 제외해야 한다.
- 색깔 바꾸기와 적합선 그리기를 확실히 해야 한다.

유의성 검정

연습 3. 2019년 전국범죄피해조사 자료(CRIME.SAV)에서 응답자들이 인지한 동네 주위 환경(2번)과 범죄 두려움(7번) 사이의 연관성을 살펴보고자 한다. 적절한 가설을 세운 뒤, 유의성 검정을 실시하고 이를 적절하게 시각화하시오.

심슨의 역설

심슨의 역설

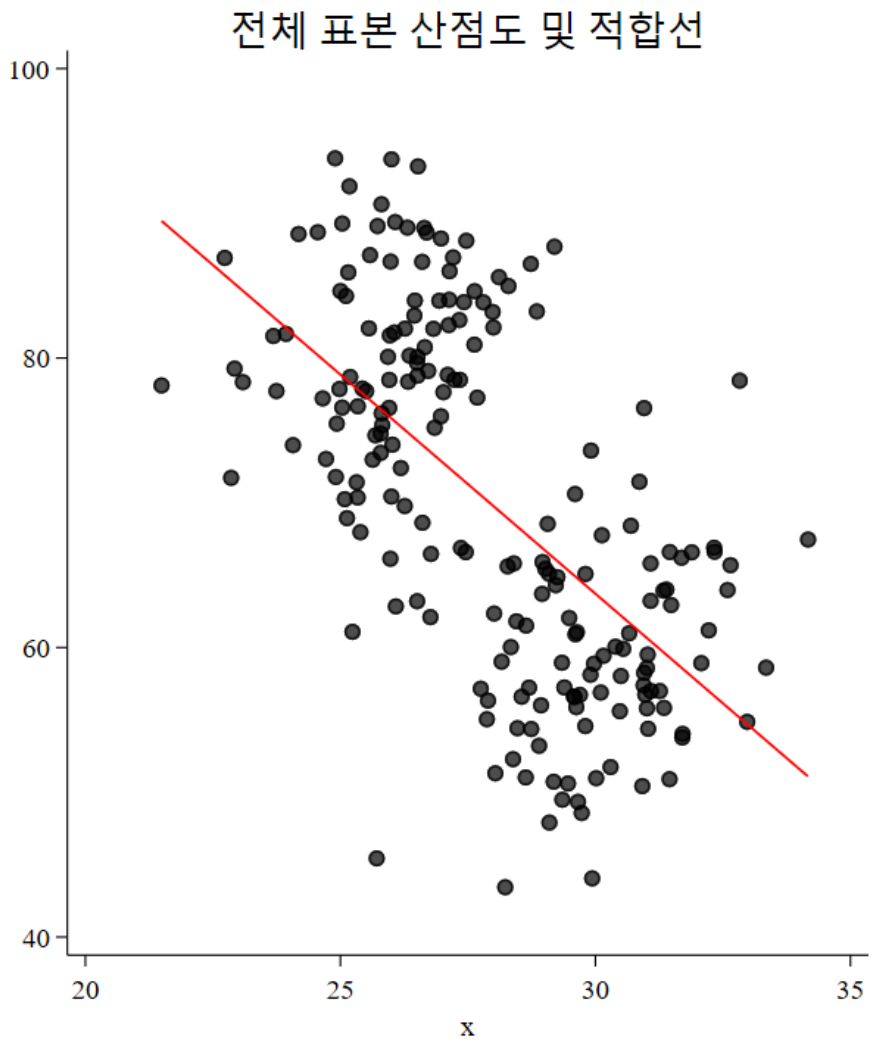
모든 통계분석에서 집단 내부의 차이를 심각하게 고려해야 한다.

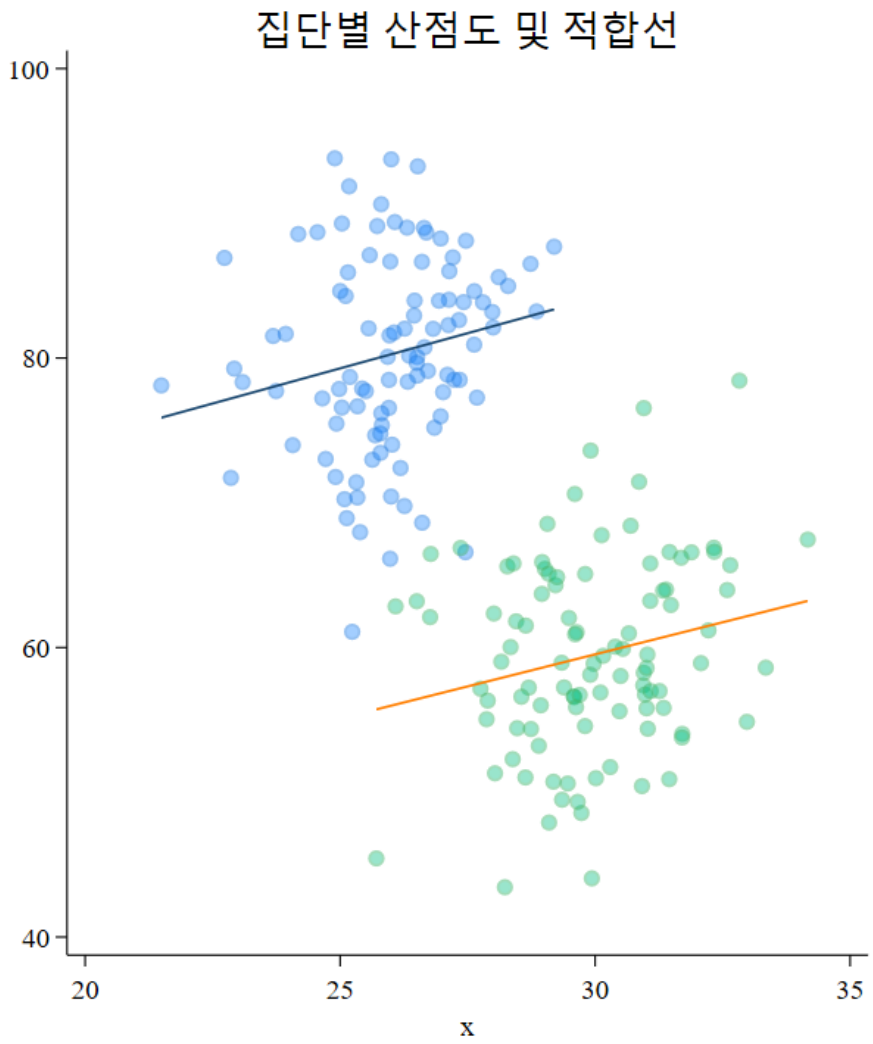
- x와 y의 관계를 살펴보고 양(+)의 상관관계를 발견했을 때, 그 상관관계는 집단 내부의 이질성(heterogeneity)은 고려하지 않은 결론으로만 받아들여야 한다 (Why?).
- 만약 집단 내부에 아주 주목할 만한 이질성이 존재한다면, 그 상관관계가 부분적으로 달라질 수 있기 때문이다.
- 이것이 이른바 심슨의 역설(Simpson's Paradox)이다.
- 특히 집계자료(aggregate data)를 분석하여 내린 결론을 성급하게 그보다 작은 단위, 가령 개인자료(individual data)로도 확대 적용하게 되면서 이런 문제가 발생한다.

심슨의 역설

연습 4. AGING.SAV에서 연령과 체중의 사이의 연관성을 살펴보기 위해 통계 검정을 수행하고 시각화하시오. 같은 검정과 시각화를 성별에 따라 별도로 수행하고 발견점에 관해 논의하시오.

심슨의 역설


전체 표본 산점도 및 적합선
100
80
60
40
20
25
30
35
x


집단별 산점도 및 적합선
100
80
60
40
20
25
30
35
x

심슨의 역설

심슨의 역설은 상관분석 뿐 아니라 모든 통계분석에서 주의해야 한다!

- 우리는 종종 두 개의 범주형 변수 사이의 관계를 넘어 제3의 변수를 고려해야 한다.
- 화재 피해와 출동한 소방관의 수에 관한 패러독스(?)가 제법 알려져 있다.

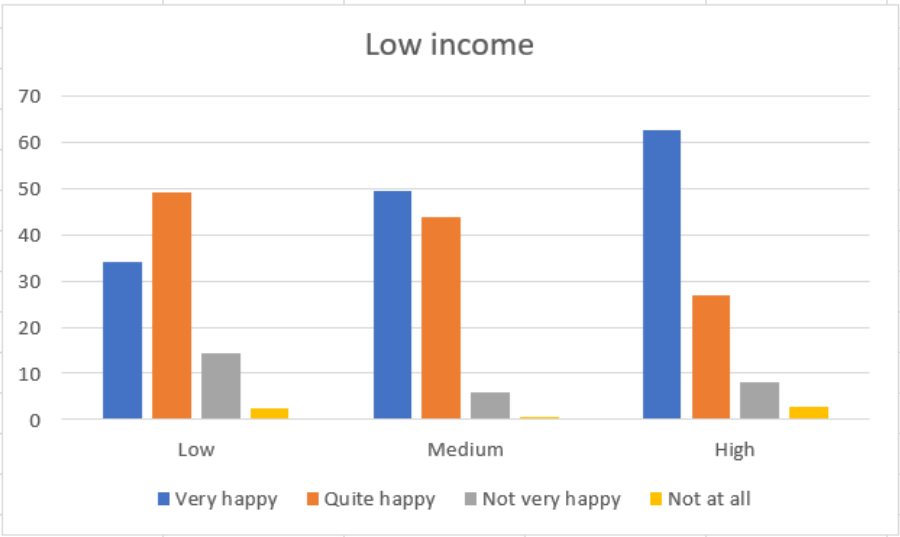
	피해 적음	피해 큼	합계
소규모 출동	97 (69.8%)	49 (30.2%)	146 (100%)
대규모 출동	42 (32.2%)	103 (67.8%)	145 (100%)
합계	139 (50.2%)	152 (49.8%)	291 (100%)

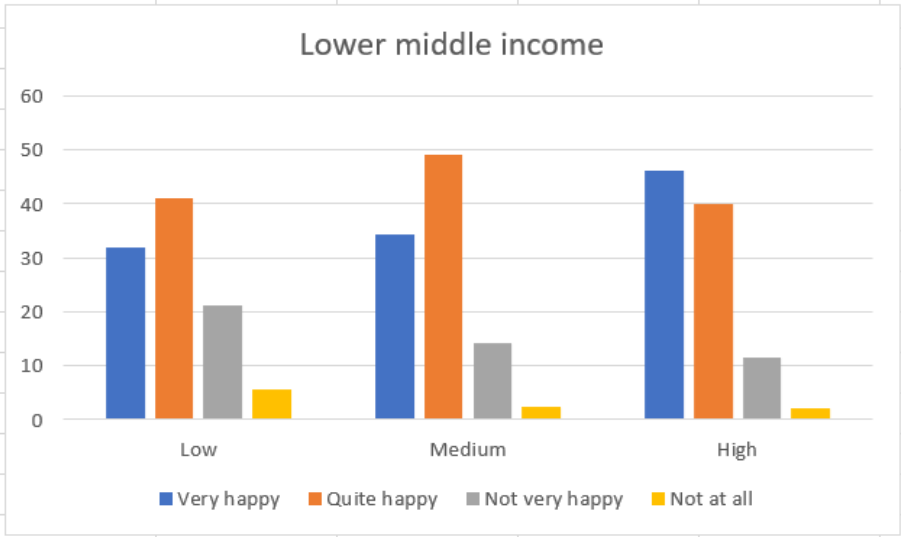
- 왜 이런 해석이 나올까? 화재의 규모라는 제3의 변수를 고려하지 않았기 때문에 생기는 허구적 인과성(spurious causality) 때문이다.
- 독립변수와 종속변수 사이에 인과관계가 있다고 믿고 χ^2 검정을 수행하지만, (단순한) 교차표나 상관계수는 이를 증명해주지 않는다!

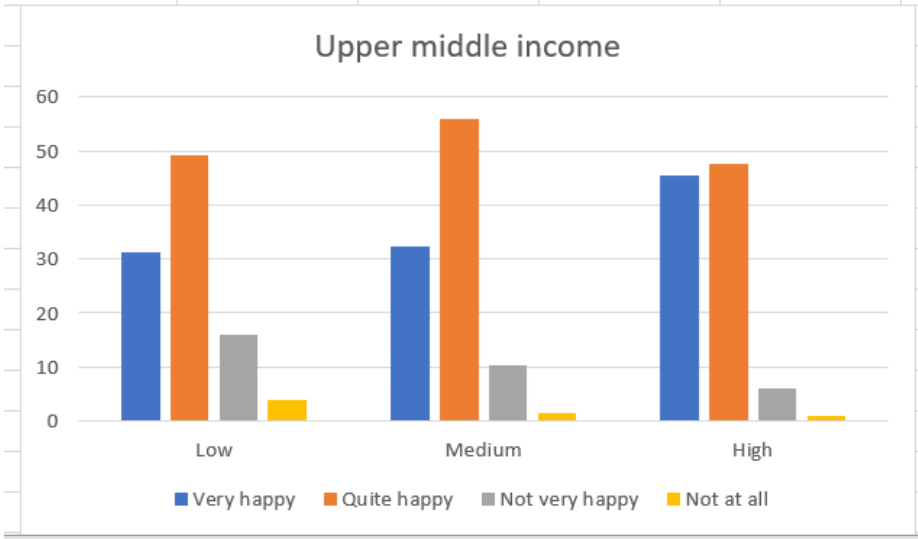
심슨의 역설

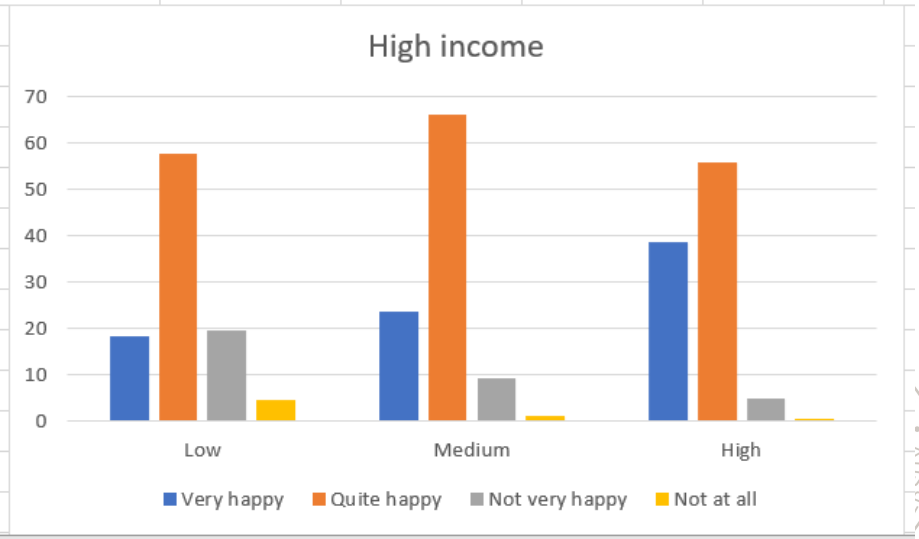
연습 5. EASTERLIN.SAV에서 소득수준 Q288R과 행복감 Q46 사이의 연관성을 살펴보기 위해 표를 작성하고 시각화하시오. 같은 표와 그래프를 국가 소득수준(incomeWB) 별로 각각 작성한 뒤, 발견점을 논의하시오.

심슨의 역설


Low income
70
60
50
40
30
20
10
0
Low
Medium
High
Very happy
Quite happy
Not very happy
Not at all


Lower middle income
60
50
40
30
20
10
0
Low
Medium
High
Very happy
Quite happy
Not very happy
Not at all


Upper middle income
60
50
40
30
20
10
0
Low
Medium
High
Very happy
Quite happy
Not very happy
Not at all


High income
70
60
50
40
30
20
10
0
Low
Medium
High
Very happy
Quite happy
Not very happy
Not at all

심슨의 역설

- 전체 자료로 x와 y의 관계를 살펴보았을 때와 소득수준별로 쪼개어 그 연관성을 살펴보았을 때가 다르다.
- 소득수준이 높은 나라에서는 개인의 소득이 높아져도 행복감이 그만큼 증가하지 않는다. 이른바 이스털린의 역설(Easterlin's Paradox)이 이와 관련한 현상을 지적한다.
- 그런데 자료를 얼마나 그리고 어떤 식으로 쪼갤 것인가는 상당히 애매한 문제이다 (Why?).
- 항상 그래프를 그리는 문제에 신경써야 한다. 종종 정답이 없으므로 다양한 방식으로 그려보고 어떤 방식이 가장 직관적인 해석을 제공하는지 고민해야 한다.