사회통계연습
시각화(II)

김현우, PhD[1]

[1] 충북대학교 사회학과 부교수

진행 순서

1. 둘 이상 변수간 연관성의 시각화
2. 상자-수염 그림과 막대차트 비교
3. 산점도와 적합선
4. 시계열 자료의 시각화
5. 공간자료의 시각화

둘 이상 변수간 연관성의 시각화

둘 이상 변수간 연관성의 시각화

둘 이상 변수 사이의 관계를 시각화한다면 분석 방법도 조금 달라진다.

- 지금까지 단일변수(univariate)가 있는 자료의 시각화를 다루었다면, 지금부터는 둘 이상의 변수(bivariate or multivariate)가 있는 자료의 시각화에 대해 공부하자.
- "주어진 모든 변수들의 척도를 정확하게 판단하고" 어떤 시각화를 할 것인지 결정한다.

둘 이상 변수간 연관성의 시각화

변수가 두 개인 경우를 예로 들어보자.

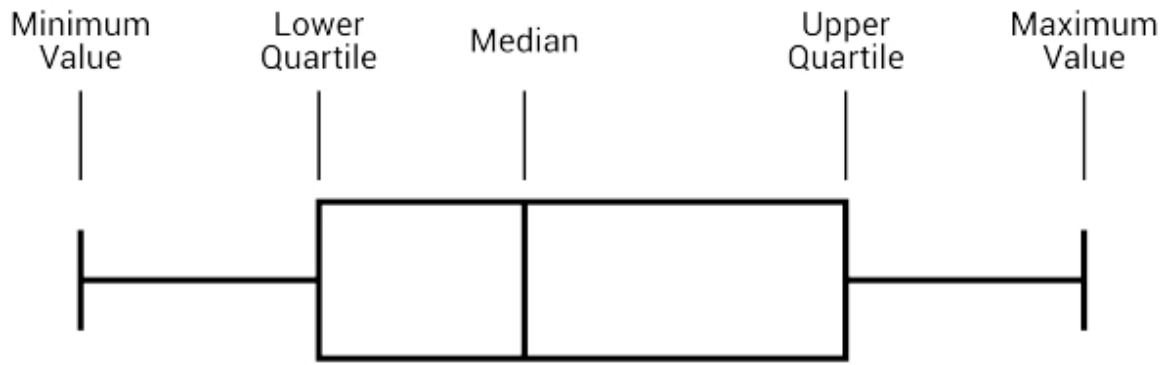
- (1) 두 변수 모두 질적변수(e.g., 리커트 척도 문항 2개)라면 교차표(cross-tabulation)와 막대차트를 그려 연관성을 표현할 수 있다. 유의성 검정을 위해 χ^2 분석(χ^2 analysis)을 사용할 수 있다.
- (2) 한 변수는 질적변수(e.g., 성별), 나머지 변수는 양적변수(e.g., 월평균 소득)라면 상자-수염 그림(box-whisker plot)을 그려 비교할 수 있다. 유의성 검정을 위해 t-검정(t-test)과 일원분산분석(ANOVA)을 사용할 수 있다.
- (3) 두 변수 모두 양적변수(e.g., 월평균 소득과 수학연수)라면 산점도(scatterplot)를 그려 비교할 수 있다. 유의성 검정을 위해 상관분석(correlation analysis)과 회귀분석(regression analysis)을 사용할 수 있다.

상자-수염 그림과 막대차트 비교

상자-수염 그림과 막대차트 비교

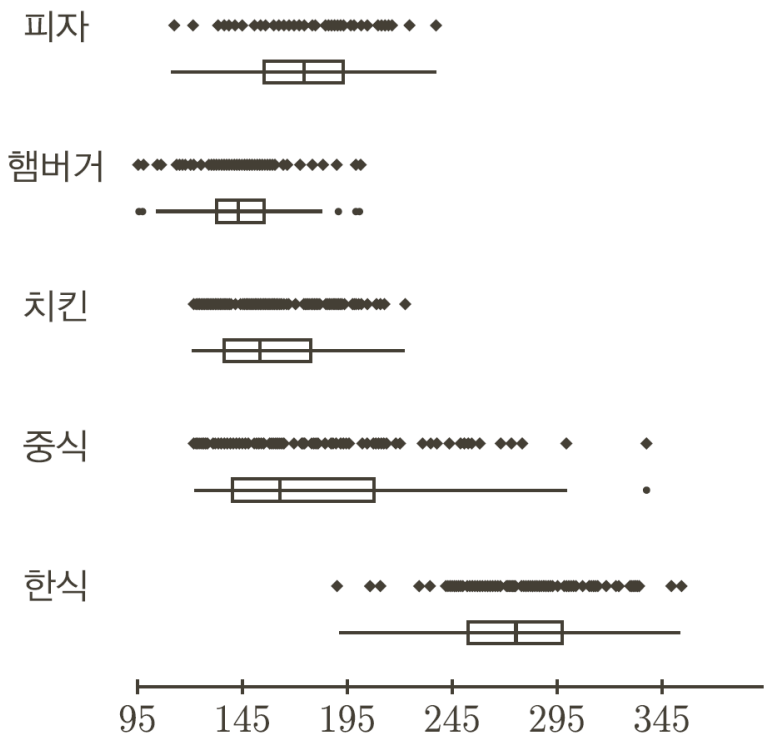
상자-수염 그림도 단일한 양적변수 시각화에서 자주 사용된다.

- 상자-수염 그림(Box-Whisker plot)은 중심경향과 산포경향의 주요 요약통계량을 그림 하나로 전달한다는 장점이 있다.
- 특히 상자-수염 그림을 통해 이상점(outliers)도 한눈에 파악할 수 있다.



상자-수염 그림과 막대차트 비교

- 사실 겨우 하나의 변수를 요약하기 위해 상자-수염을 그리는 경우는 거의 없고 (Why?), 여러 변수의 상자-수염 그림을 나란히 비교할 수는 있다.
- 이때는 사실 (여러 개의 양적변수를 비교하는 것이 아니라) 양적변수인 배달시간과 질적변수인 음식 종류 사이의 연관성을 살펴본 것임에 주의해야 한다(Why?).



상자-수염 그림과 막대차트 비교

연습 1. NHANES.SAV 자료에서 인종(race)과 이완기 혈압(bpdiast)의 연관성이 성별(sex)에 따라 어떻게 달라지는가를 상자-수염 그림으로 시각화하시오.

상자-수염 그림과 막대차트 비교

질적변수와 질적변수의 관계는 어떻게 시각화할 수 있을까?

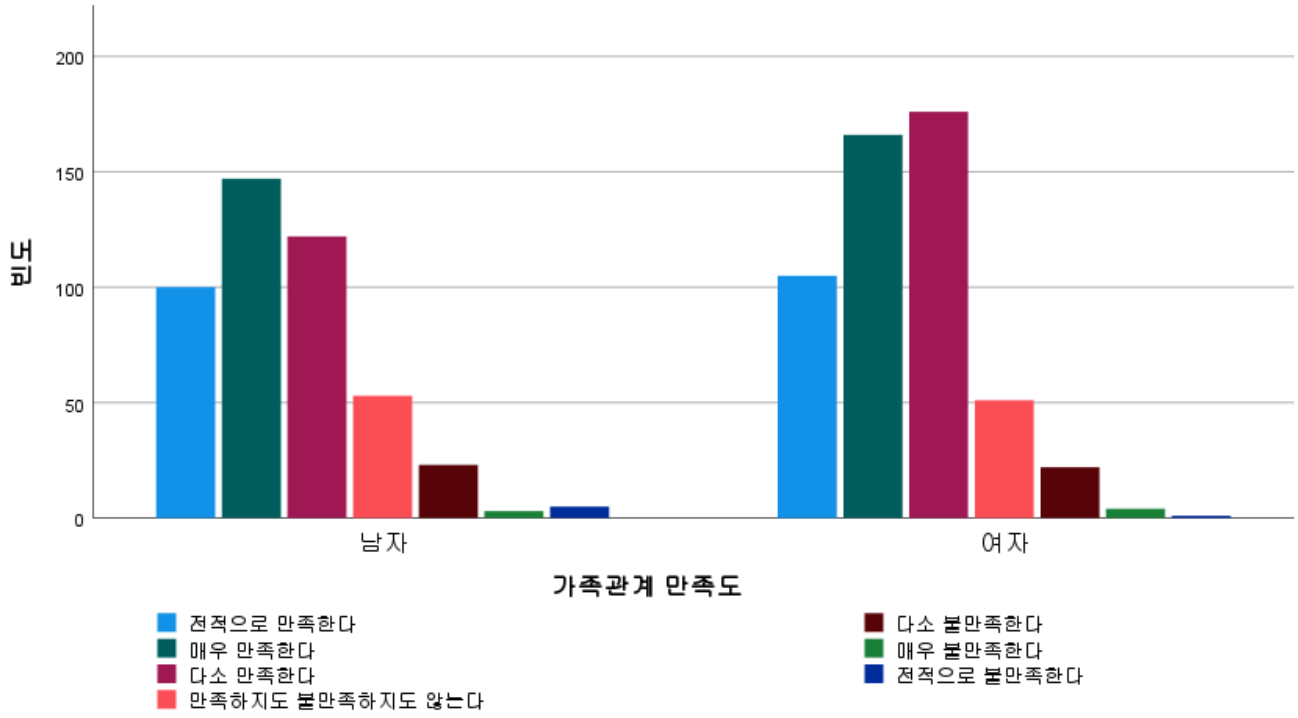
- 곰곰이 생각해보자. 이완기 혈압은 양적변수인 반면, 인종과 성별은 질적변수였다!
- 상자-수염 그림이 질적변수와 양적변수의 관계를 시각화한다면, 막대차트는 질적변수와 질적변수의 관계를 시각화하는데 사용될 수 있다.
- 어떻게 이런 그림을 그릴 수 있을까?

상자-수염 그림과 막대차트 비교

연습 2. VALUE.SAV 자료에서 가족관계 만족도(FAMSATIS)와 성별(SEX)의 연관성을 시각화하시오.

상자-수염 그림과 막대차트 비교

- '수평누적 막대도표'를 선택한다. 'X에 군집: 색상 선정'에 FAMSATIS를 넣고 x 축에는 SEX를 넣는다.
- 그래프를 더블 클릭하고 레이블을 클릭한 뒤, [특성]-[범주형]에서 불필요한 레이블은 삭제해야 그림이 예뻐진다.
- 'X에 군집: 색상 선정'에 SEX를 넣고 x축에 FAMSATIS를 넣는다면 그림이 달라진다(Why?).



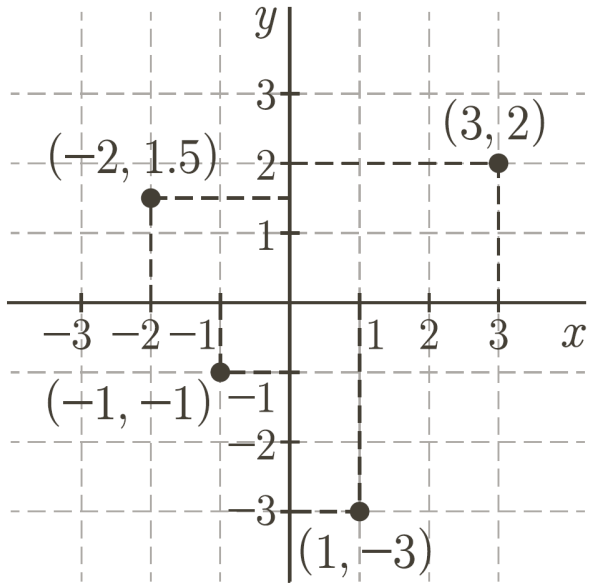
산점도와 적합선

산점도와 적합선

두 양적변수의 연관성을 살펴보기 위해 산점도를 활용할 수 있다.

- 산점도란 두 양적변수 X와 Y가 주어졌을 때 하나의 관측치(observation)를 하나의 벡터(vector) (X, Y)로 파악한 다음, 이것들을 데카르트 좌표계(Cartesian coordinates) 위에 뿌린 것이다.

ID	MATH	ENG
1	-2	1.5
2	3	2
3	-1	-1
4	1	-3



산점도와 적합선

연습 3. NHANES.SAV 자료에서 키(height)와 몸무게(weight)의 연관성을 시각화하시오.

산점도와 적합선

- 산점도와 더불어 “여러 점들 사이의 추세를 나타내는 선”, 즉 적합선(fitting line)을 그려 두 변수 사이의 관계를 또렷하게 시각화할 수 있다. SPSS에서는 ‘선형 적합선’을 체크한다.
- 여러 개의 연속변수 사이의 연관성은 산점도 행렬(scatterplot matrix)로 나타낼 수 있으며 이때는 SPSS에서 [차트 작성기]의 ‘산점도 행렬’을 사용한다.
- SPSS나 Jamovi 보다는 엑셀에서 산점도를 그리는 편이 좀 더 예쁜 것 같다.
  아쉽게도 엑셀에서는 산점도 행렬을 그리기 어렵다(여러번 그려 합쳐도 된다).

산점도와 적합선

연습 4. CORN.SAV 자료에서 옥수수와 콩의 실제 헥타르당 재배면적과 위성사진 픽셀 사이의 연관성을 시각화하고 해석하시오.

시계열 자료의 시각화

시계열 자료의 시각화

자료가 시간에 따라 변화하는 추세도 시각화 할 수 있다.

- 시계열 자료(time-series data)란 "시간에 따라 관측된 데이터"이다. 대표적인 예로 일별 주가지수나 연간 강수량 등을 생각해 볼 수 있다.
- 시계열 자료를 분석하는 시계열 분석(time-series analysis)은 이미 경제학 분야에서 엄청나게 발전하여 특히 금융공학 쪽에서 고도의 기법이 연구되고 있다(베이지안 시계열 분석이나 머신러닝 금융공학 등).
- 하지만 사회학에서는 시계열 자료나 시계열 분석이 거의 다루어지지 않는다(Why?).
- 관찰된 시계열 데이터는 사실 몇 가지 요소가 결합된 혼합물이다: 추세성(trends), 계절성(seasonality)/주기성(cyclicity), 잡음(noise).

시계열 자료의 시각화

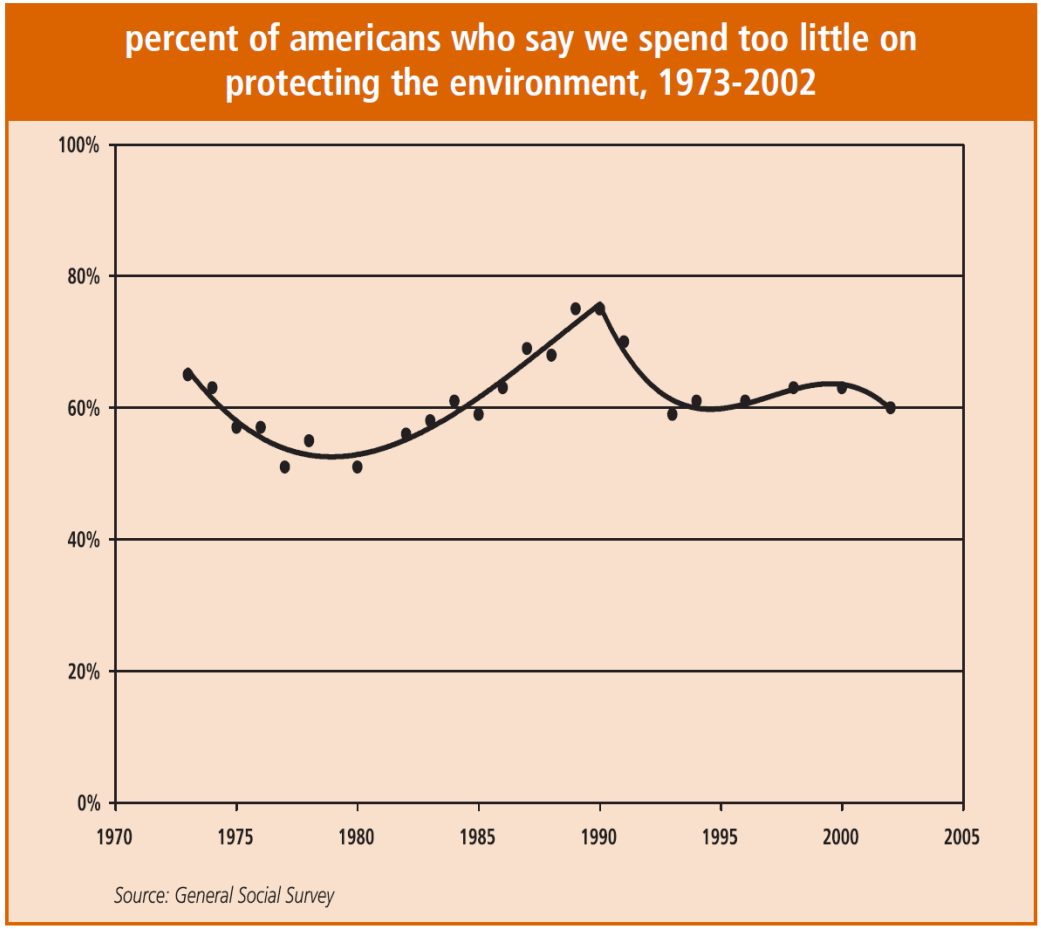
연습 5. KLEIN.SAV 자료에서 1920년에서 1941년 사이 민간부문 임금지불액(wagepriv)과 공공부문 임금지불액(wagegovt)의 시계열도표를 각각 시각화하고 서로 비교하시오.

시계열 자료의 시각화

- SPSS에서 하나 이상의 시계열 변수를 한 그림에 집어넣으려면 [차트 작성기]가 아니라 [레거시 대화 상자]에서 [선형 차트]를 골라야 한다.
- “단순”이 아니라 “다중” 그리고 “케이스 집단들의 요약값”이 아니라 “개별 변수의 요약값”을 선택해야 한다.
- ‘선 표시’에는 그리려고 하는 시계열 변수를 집어넣고, ‘범주축’에는 시간 변수를 집어넣으면 된다.
- Jamovi에서는 vijPlots 모듈에서 [Line Chart] 기능을 활용하자.

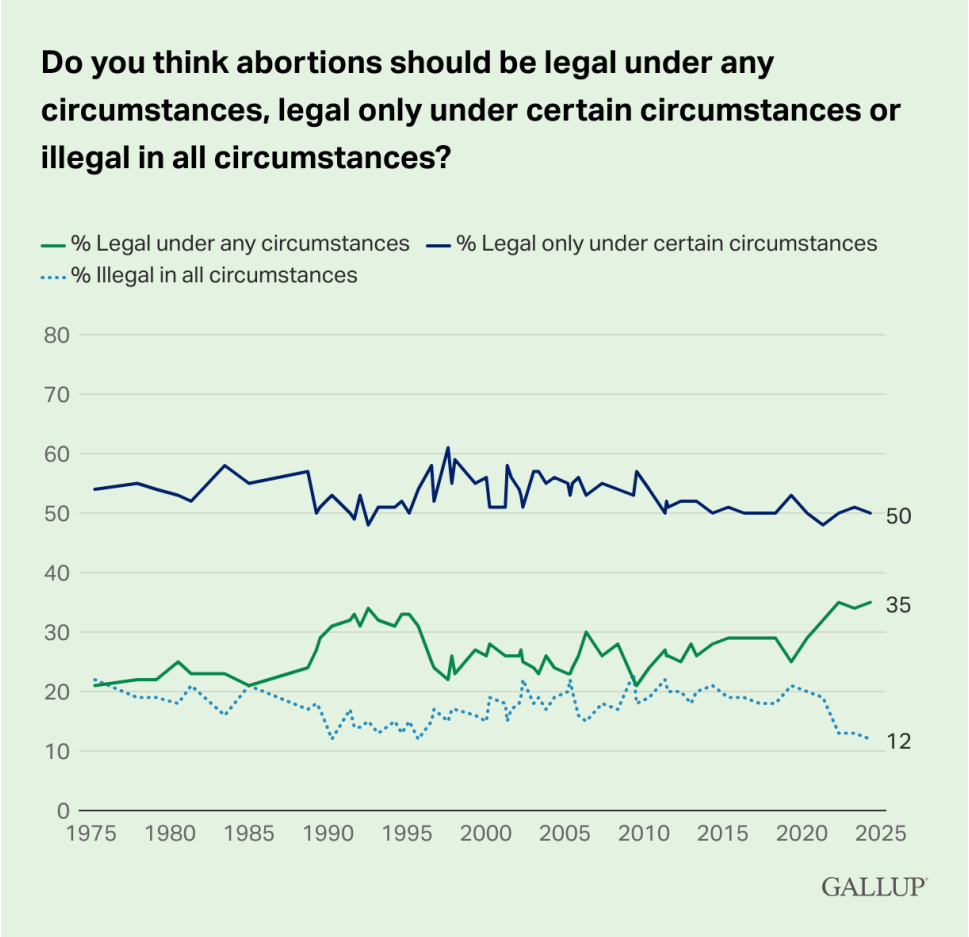
시계열 자료의 시각화

- 아래 그래프는 1973년부터 2002년까지 환경을 보호하는데 돈을 너무 적게 쓴다고 믿는 미국인의 비율이 어떻게 증감하는지 보여준다.



시계열 자료의 시각화

- 아래 그래프는 어떤 상황에 따라 임신중절이 허용될 수 있는가에 대한 사회조사를 담고 있다. 임신중절 찬반 여부의 시각화가 어떻게 이루어졌는지 살펴보자.



시계열 자료의 시각화

이런 시계열 도표를 어떻게 만들 수 있을까?

- 먼저 (1) 연도별로 위 명제에 동의 여부를 나타낸 가변수(dummy variable)를 평균으로 집계한다(aggregate).
- (2) 연도별 평균값(%)을 구한 자료로 시계열 시각화를 수행한다. 필요에 따라 두 개 이상의 선을 그릴 수도 있다.
- 가변수의 평균은 곧 비율이라는 점을 다시 떠올리자(Why?).

공간자료의 시각화

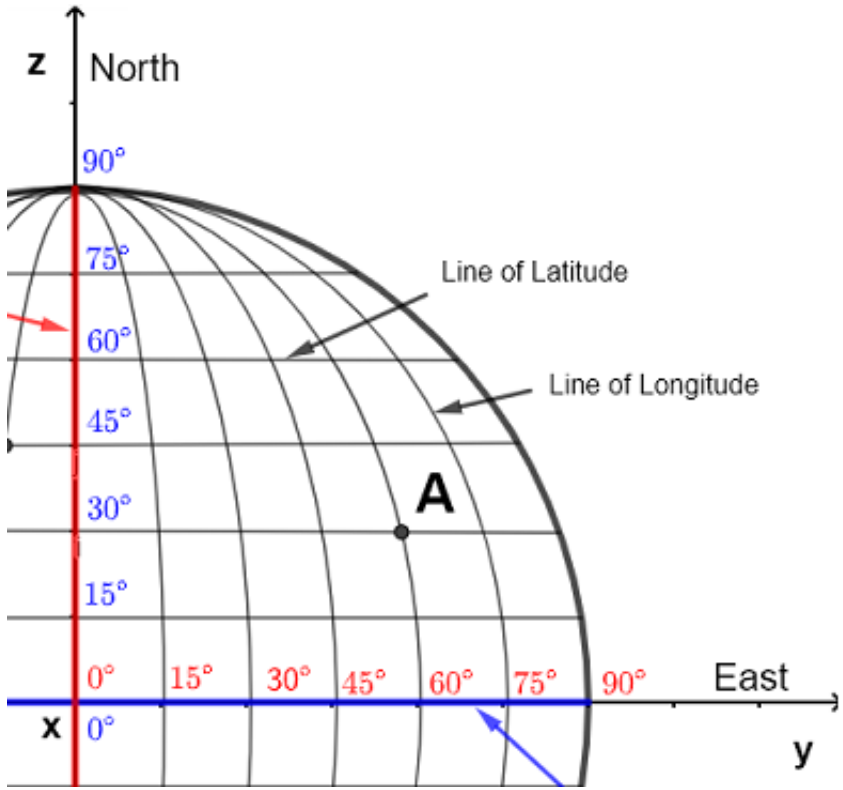
공간자료의 시각화

자료가 공간에 따라 어떻게 분포하는가도 시각화 할 수 있다.

- 공간자료(spatial data)란 "공간에 따라 관측된 데이터"이다. 대표적인 예로 지역별 인구수 등을 생각해 볼 수 있다.
- 공간자료는 이른바 지리정보시스템(geographical information system; GIS)의 발전에 힘입어 큰 관심을 받고 있다. 지리학 분야 뿐 아니라 사회학에서도 중요하게 다루어져 왔다.

공간자료의 시각화

- 공간자료는 적어도 세 구성요소로 이루어져 있다: (1) 위도(latitude), (2) 경도(longitude), (3) 자료값(data value).



공간자료의 시각화

연습 6. PHARMACY.SAV 자료에 주어진 위도와 경도로 산점도를 그려보고 공간자료의 구조를 파악하시오.

공간자료의 시각화

- 위도와 경도가 주어져 있지 않고 지명이 텍스트(text)로 주어졌다면 이른바 지오코딩(geocoding) 단계를 거쳐야 한다.
- e.g., "충북 청주시 서원구 충대로 충북대학교" $\rightarrow$ 36.6287° N, 127.4606° E
- 공간자료란 결국 위도와 경도와 함께 주어지는 정보라고 할 수 있다(Why?).

공간자료의 시각화

연습 7. KOSIS 국가통계포털에서 [행정구역(시군구)별, 성별 인구수]를 다운로드받고 엑셀에서 불러들이시오. 남녀인구비율을 나타내는 공간자료를 시각화하시오.

공간자료의 시각화

- 아쉽지만 엑셀은 본래 지리정보시스템이 아니므로 한국의 고유한 지명을 제대로 인식하지 못하는 문제가 있다(엑셀 버전이 낮으면 아예 지오코딩이 지원되지 않는다).
- 가령 세종시는 이름과 경계를 제대로 인식하지 못해 지오코딩에 실패하였다(Why?).
- 지도에서 레이블이 나타나도록 설정하고 제대로 지명이 인식되고 있는지 여부를 재확인해야 한다.
- 본격적인 실무나 연구에서는 공간자료분석을 위해 (1) ArcGIS나 QGIS같은 지리정보시스템 소프트웨어를 사용하거나, (2) 최소한 Tableau와 같은 비즈니스 인텔리전스(Business Intelligence; BI)를 사용한다.