사회통계연습

More on Association of Two or More Variables

김현우, PhD[1]

[1] 충북대학교 사회학과 조교수

September 24, 2021

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

진행 순서

1. Bivariate/Multivariate Data Visualization
2. 최근의 데이터 시각화 기법들과 몇 가지 코멘트

Bivariate/Multivariate Data Visualization

Bivariate/Multivariate Data Visualization

지금까지 단변량(univariate) 데이터 시각화를 다루었다면 지금부터는 이변량(bivariate) 또는 다변량(multivariate) 데이터 시각화에 대해 이야기하자.

- 이변량/다변량 데이터 시각화 기법으로는 오로지 하나에 대해서만 배운다: 바로 산포도(scatterplot)다.
- 산포도는 둘 이상의 변수가 모두 숫자형(numerical)이라고 전제한다.
- 산포드는 사실 line fitting 과 함께 붙어다니는 기법이므로 둘 다 배워야 한다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

Bivariate/Multivariate Data Visualization

복잡하게 생각하지 말고 일단 실습부터 해보자. eCampus에서 iris.csv를 다운받아 열자.

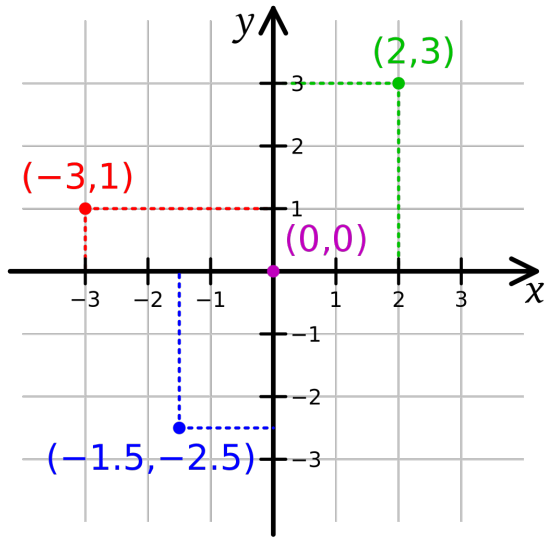
- sepal_length와 petal_length 두 변수를 살펴보자. 척도가 각각 어떻게 보이나?
- 두 변수의 columns를 하이라이트한 다음, [삽입]-[다른 분산형 차트]를 클릭하여 들어간다.
- 분산형 차트를 유심히 보고 맞는 것을 고르지 않으면 안된다. 우리는 두 변수의 사이를 관찰하는 scatterplot을 그리고 싶기 때문에 두 번째의 것을 사용한다.
- 그래프를 꾸미되 다음에 주의할 것: (1) 축 제목; (2) 축 서식에서 경계 최소최대값; (3) 추세선 삽입.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

Bivariate/Multivariate Data Visualization

산포도(scatterplot)의 기본적인 아이디어

- 산포도는 기본적으로 두 개의 변수(X와 Y라고 하자)가 주어졌을 때 하나의 관찰값(observation)을 (X, Y)로 파악한 뒤, 데카르트 좌표계(Cartesian coordinates) 위의 한 점으로 찍는 것이다.



- 수많은 관찰값들이 데이터로 주어졌다면 이를 데카르트 좌표계 위에 쫙 뿌려서 두 변수 사이의 관계를 보여준다. 그리고 fitting line은 그 대략적인 관계를 선으로 보여주는 셈이다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

Bivariate/Multivariate Data Visualization

추세선(fitting line)은 직선 $Y = b_0 + b_1 X$ 꼴로 나타낼 수도 있지만 부드럽게 꺾인 꼴로 나타낼수도 있다.

- 직선 $Y = b_0 + b_1 X$ 꼴로 표현되어 있을 때 b_1은 직선의 기울기(slope)를 나타낸다. b_1이 양수(+)이면 직선은 우상향하고, 음수(-)이면 직선은 우하향한다. b_0은 직선의 절편(intercept)을 나타낸다. 절편이란 이 직선과 y축이 만나는 자리를 의미한다. $b_0 = 0$인 경우 이 직선은 원점(origin)을 통과한다.
- 부드러운 선은 이른바 다항식 피팅(polynomial fitting)이라고 불리운다. 2차 다항식은 $Y = b_0 + b_1 X + b_2 X^2$ 꼴이고, 3차 다항식은 $Y = b_0 + b_1 X + b_2 X^2 + b_3 X^3$ 꼴이다. 엑셀에서는 6차까지 사용할 수 있으며 차수가 높아질수록 선은 더욱 부드럽게 변한다.
- 분석툴에 따라서는 fitting line을 회귀선(regression line)이라고 부를 수도 있다. 이유는 나중에 알게 된다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

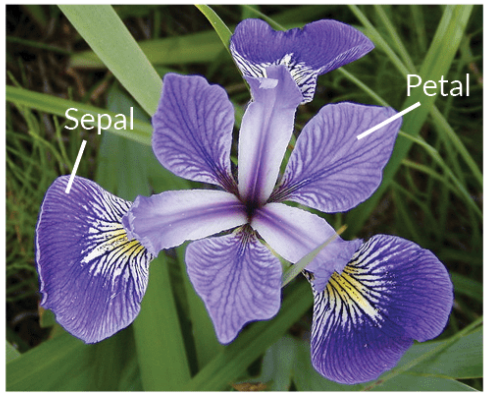
Bivariate/Multivariate Data Visualization

그런데 이 데이터 iris.csv는 좀 나름의 역사가 있다.

- 이것은 거의 백년 쯤 전 통계학자 Ronald Fisher가 사용한 이래 매우 고전적인 자료로 여겨진다.

  Fisher, Ronald. A. 1936. "The Use of Multiple Measurements in Taxonomic Problems." Annual Eugenics 7(II): 179-188.

- 이후 수많은 통계기법이나 알고리즘을 테스트하거나 교육할 때 이 데이터가 종종 사용되어왔다.



Iris Versicolor


Iris Setosa


Iris Virginica

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

Bivariate/Multivariate Data Visualization

다변량(multivariate) 데이터 시각화의 경우 이제 슬슬 엑셀로는 귀찮아진다.

- 엑셀에서는 여러 개의 그래프를 개별적으로 그린 뒤 합쳐야 하지만, 전문적인 통계 분석툴에서는 이를 쉽게 할 수 있다.

그러니 엑셀은 이제 그만 끄고 JASP를 켜자.

- JASP 화면 왼쪽 상단의 三 자 모양의 아이콘을 누른 뒤, [Open]-[Computer]-[Browse]를 선택하자. eCampus에서 미리 다운받은 iris.csv 파일을 선택하자. 만일 iris.csv 파일을 JASP 아이콘 위에 drop하면 JASP의 구동과 동시에 파일이 열린다.
- 자료를 엑셀처럼 인터렉티브하게 살펴볼 수 있다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

Bivariate/Multivariate Data Visualization

- 이제 화면 윗쪽 메뉴의 첫번째 [Descriptives]를 클릭하자.
- 왼쪽 pane에 보면, 원하는 변수를 골라서 이동시킬 수 있다. 아까 엑셀에서 했듯, sepal_length와 petal_length를 Variables 창 아래로 이동시켜보자.
- 오른쪽 pane에 저절로 실시간 기술통계량(descriptive statistics)이 보고되는 것을 확인할 수 있다(편리!). 결과표 안에서 결측치(missing), 평균(mean), 표준편차(std. deviation), 최소값(minimum), 최대값(maximum)을 확인하자.
- 왼쪽 pane 아래쪽에 Statistics을 눌러보자. 이제 우리에게 친숙한 용어들이 많이 보인다!
- 사분위수(Quartiles), 중심성향(central tendency)의 중위값(median), 최빈치(Mode), 그리고 산포성향(Dispersion)의 분산(Variance), 범위(Range), IQR 등등... 이것저것 눌러보자. 실시간으로 업데이트 된다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

Bivariate/Multivariate Data Visualization

본격적으로 JASP에서 산포도(scatterplot)를 그려보자

- 왼쪽 pane에 있는 Plots 메뉴를 클릭하여 확장하면, 우리가 오늘 배운 scatterplot이 여기 있는 것을 알 수 있다.
- 이걸 체크하면 아까 엑셀에서 그렸던 것과 매우 비슷한 산포도를 볼 수 있다.
- 몇 가지 추가적인 기능도 있다. 가령 그림의 윗쪽과 오른쪽에 개별 변수들의 히스토그램을 보여준다. 피팅 라인(fitting line)을 회귀선(regression line)이라는 이름 아래 부드럽게(smooth) 또는 직선형(linear)으로 나타낼 수 있도록 한다. 이리저리 눌러보자. 실시간으로 업데이트 된다. 그림의 크기 조절도 가능하다.
- 그림은 다른 워드 프로세서로 복사하여 붙어넣을 수 있다. 그냥 그림파일로 저장할 수도 있다.

엑셀에서 했던 것처럼 상자-수염 그림(Box-Whisker Plot)도 그릴 수 있다.

- JASP에서는 이걸 상자 그림(Boxplots)이라고 해놓았다. 이걸 체크하자.
- "Boxplot element"만을 체크하고 나머지는 끄자.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

Bivariate/Multivariate Data Visualization

JASP를 켠 우리의 본래 목적은 multivariate data visualization 이였으므로 이제 왼쪽 pane 상단에서 모든 변수를 다 Variables 창 안으로 옮겨보자.

- 다시 아랫쪽 Plots 메뉴에서 "Correlation plots"를 체크해보자.
- 이 그림은 고른 모든 변수들의 있을 수 있는 모든 짝(pair) 사이에서 scatterplot을 그린 것이다.
- 우리의 예제에서 고른 모든 변수는 총 4개이다. 그것들이 왼쪽과 오른쪽 끄트머리에 레이블(label)로 표시되어 있다.
- 관점을 크게 보면 이 그래프는 결국 작은 그래프들의 행렬(matrix)이다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

최근의 데이터 시각화 기법들과 몇 가지 코멘트

최근의 데이터 시각화 기법들과 몇 가지 코멘트

시각화를 프로처럼 제대로 활용하려면 거의 필수적으로 자바스크립트(JavaScript)를 습득해야 한다.

- Java 계열 자체가 플랫폼 독립적인 범용 스크립트 언어다.
- 세계 최대 규모의 개발자 커뮤니티. 국비지원 무료교육 프로그램, 취업 1순위 언어
- 웹 프론트엔드(web front-end)를 디자인할 때 자바스크립트를 모르고는 이야기가 안된다.

이 수업은 자바스크립트를 다루지 않았다.

- 애초에 다룰 수도 없다. 별도로 한 학기 수업이 필요하기 때문이다. 게다가 나도 자바스크립트를 모른다.
- 대규모 기관/조직에서는 대체로 자바스크립트 전문가를 따로 고용하지만 상황이 꼭 그렇게 여의치 않을수도 있다.
- 대안적으로 다른 앱(app)을 사용하여 어느 정도 시각화를 흉내낼 수는 있다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

최근의 데이터 시각화 기법들과 몇 가지 코멘트

지리정보시스템(Geographical Information System: GIS)은 수십 년 사이 학술적 · 정책적 유용성을 폭넓게 증명해왔다.

- 불평등, 범죄, 사회운동, 경제, 정치, 역사 등에 걸쳐 각종 테마에 관한 지도를 그리고 그 위에 나타내고자 하는 지리적 정보를 표현한다.
- 최근 웹 리소스나 API의 발전 및 대중화에 힘입어 과거 어느 때보다 수준높은 GIS 데이터 시각화가 가능하게 되었다.

GIS 데이터 시각화 학습은 어떻게 할까?

- 기능이 폭넓지만 유료인 ArcGIS에서 기능은 제한적이지만 무료인 QGIS까지 다양한 도구가 있다.
- 충북대학교(또는 인근 대학)에서도 잘 찾아보면 GIS를 깊이있게 배울 수 있는 수업이 어딘가에서 개설될 것 같긴 한데...

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

최근의 데이터 시각화 기법들과 몇 가지 코멘트

인터렉티브 그래픽스(Interactive Graphics)의 예

- 미국 사회에서 인종격리(racial segregation)과 소수자 중개상들(middlemen minority)의 GIS
- 영화 〈8 Mile (2003)〉에서 Eminem의 “Lose Yourself”
- http://racialdotmap.demographics.coopercenter.org/
- 이 인터렉티브 맵(interactive map)은 자바스크립트로 만들어졌다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

최근의 데이터 시각화 기법들과 몇 가지 코멘트

데이터 시각화는 비즈니스 인텔리전스(Business Intelligence; BI)의 핵심적인 도구 중 하나다.

- 경영지원 혹은 비즈니스 인텔리전스(BI) 차원에서 수많은 툴이 출시되었다.
- 잘 알려진 툴 중에서는 (1) 태블로(Tableau), (2) 마이크로소프트의 Power BI, (3) 마이크로스트래티지(Microstrategy) 등이 있다. 대부분 고가의 툴이라 개인은 라이선스를 구매하기 어렵고 기관/조직에서 구매해야 한다.
- 다만 일부 프로그램은 기능제한부로 무료다. 어떤 프로그램은 대학생에게 특별히 무료로 풀려있다. 나도 대학(원)생 때는 무료로 사용했었다.
- 이런 것들은 주로 리포트(report) 또는 대시보드(dashboard)를 만들어 수집된 데이터의 요약 및 분석 결과를 실시간으로 조직 구성원과 공유하려는 목적을 갖는다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

최근의 데이터 시각화 기법들과 몇 가지 코멘트

Tableau Public을 기동하고 eCampus에서 다운받은 MortalityDataWHR2021C2.xlsx을 열고 다음의 report 및 dashboard를 간단히 만들어보자.

- Population 2020의 histogram
- 두 변수, “COVID-19 deaths per 100,000 population in 2020”와 “Index of exposure to COVID-19 infections in other countries as of March 31” 사이의 scatterplot
- Gini coefficient of income의 mapping

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

최근의 데이터 시각화 기법들과 몇 가지 코멘트

데이터 시각화는 최근 들어 급성장하고 있는 분야.

- 학술적인 측면에서는 다소 평가절하되고 있는 느낌(학자들은 대체로 그림보다는 글을 강조하기 때문)이다.
- 그럼에도 시각화는 나름의 긴 역사를 가지고 있고 또 그 중요성에 대해서도 사람들이 인지는 하고 있다. 다만 무엇을 어떻게 시각화를 하는지 진지하게 파고드는 사람이 생각보다 매우 적다.
- 최근 데이터 저널리즘의 발전을 보면 앞으로 이 분야에서 지속적으로 전문가에 대한 수요가 있을 것임을 짐작케 한다.
- Google Analytics 등이 개인 또는 조직/기관이 운영하는 웹페이지의 방문자 데이터를 수집 및 분석하는데 중요한 도구로 급부상하였는데, 그 분석 결과를 어떻게 멋진 그래프로 정리하고 대시보드를 만들거나 리포트를 제작하는가 하는 것은 데이터 분석가(Data Analyst)의 중요한 직무가 되었다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

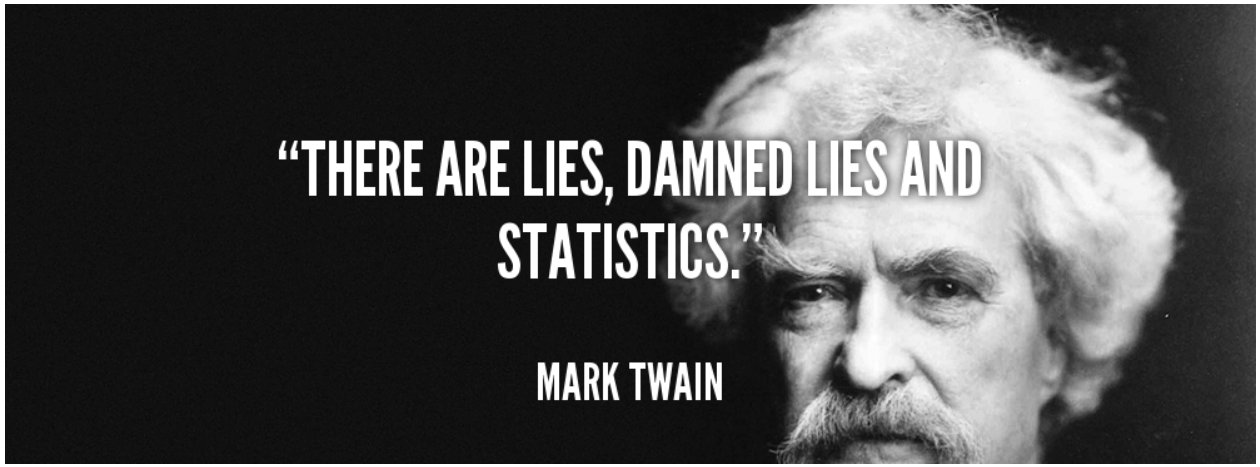
최근의 데이터 시각화 기법들과 몇 가지 코멘트

통계학에 관한 가장 유명한 격언 가운데 하나는 통계를 사용한 거짓말에 대한 경고다.

- 이 분야에서 가장 유명한 입문서 중 하나는 아마도 이것.

  허프, 대럴. 2004.『새빨간 거짓말, 통계』. 더불어책.

- 아마도 가장 흔한 형태의 통계적 거짓말은 그래프로 사기를 치거나 편협한 결론으로 유도하는 것.
- 데이터 시각화가 급부상하고 있는 만큼 "시각적 정보"를 비판적으로 읽고 생산할 수 있는 능력이 중요한 스킬이 될 것.



충북대학교
CHUNGBUK NATIONAL UNIVERSITY

최근의 데이터 시각화 기법들과 몇 가지 코멘트

말할 필요도 없지만 이 수업에서 시각화를 전부 다룬게 아니다.

- 반드시 다루어야 하는 기초적인 그래프를 몇 개 다루었고, 수업의 일정 부분은 최신의 기법 몇가지를 소개하려고 노력하였다.
- 하지만 여전히 못 다룬 중요한 그래프들, 예컨대 소셜 네트워크(social network) 이나 애니메이션(animation) 뿐 아니라, AR/VR/XR 등 차세대 기법들은 중요하지 않아서 안 다룬게 아니라 시간이 없어서 못 다룬 것이다.

나중에 조금 시간을 내서 큰 서점이나 도서관에 가보자.

- 인터넷 검색도 나쁘지 않지만 별로 체계적이진 않다.
- "시각화"로 검색해서 이 주제에 관해 어떤 책들이 나와있는지를 쭉 살펴보면 대략적으로 이 분야가 어떤 느낌인지 파악할 수 있다.
- 다만 책을 보고 괜히 압도당하지는 말자. 배우지 않은 뒷부분을 미리 보면 겁을 먹기 쉽다. 막상 배우면 기초 정도는 생각보다 쉽게 한다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY