

사회통계연습

Association of Two or More Variables

김현우, PhD¹

¹충북대학교 사회학과 조교수

September 24, 2021

진행 순서

- ① 지난 주 리뷰
- ② 데이터 시각화의 시작
- ③ Univariate Data Visualization

지난 주 리뷰

지난 주 리뷰

퀴즈 3 코멘트

- 노 코멘트

데이터 시각화의 시작

데이터 시각화의 시작

“훌륭한 스케치가 긴 연설보다 낫다(A picture is worth a thousand words).”

데이터 시각화(data visualization)는 최근 수 년 사이에 폭발적인 인기를 얻고 있다.

- 데이터 외적 영역에서도 몇 가지 중요한 추세가 발견된다!
- 웹 디자인 시장이 폭발적으로 성장했고 모바일 웹 디자인 역시 지속적으로 성장하고 있다.
- 유튜브, 틱톡, 인스타그램 등의 인기에 힘입어 사진/그림/동영상 제작 및 편집 역시 중요한 스킬 중 하나가 되었다.
- 데이터와 관련해서는 데이터 저널리즘(Data Journalism)의 성장이 이 분야를 리드한 것으로 보인다.
- 미국에서는 데이터 시각화 분야만으로 학위과정이 따로 있다. 한국에서도 데이터사이언스 분야가 급성장하고 있기 때문에 자신의 장점으로 고려해 볼 만한 분야.

데이터 시각화의 시작

왜 사회학도가 시각화를 배우나?

- 데이터의 경향이나 패턴을 이해하는데 (단순히 숫자를 나열할 때보다) 큰 도움이 된다.
- 변수의 분포(distribution)를 보여주어 데이터의 문제를 파악하거나 다음 분석을 계획하는데 용이하다.
- 여러 변수들 사이에 혹시 어떤 관계가 있지 않나 하는 의문을 빠르게 해소하는데 유리하다.

사회학과를 졸업한 다음에는 무엇에 쓰이나?

- 데이터 시각화는, 물론 사회학의 연구 목적에도 유용하지만, 여러분이 정부조직, 비영리단체, 복지기관, 교육기관, 사기업 등에 자리를 잡은 뒤에도 실용적인 쓸모를 갖는다.
- 여러분의 한참 선배 세대는 오로지 글로써 의도를 전달했다. 하지만 여러분은 이미 멀티미디어 세대이고 여러분의 후배 세대는 아예 글을 읽으려 들지 않을지도 모른다. 시각화는 점진적으로 필수적인 의사소통의 수단이 되어가고 있다.

데이터 시각화의 시작

기초적인 수준에서 볼 때, 데이터 시각화는 요약 및 분석 시나리오에 따라 두 종류로 나뉜다.

- 하나의 변수를 요약하는 상황(univariate data visualization)
- 둘 이상의 변수 사이 관계를 분석하는 상황(bivariate/multivariate data visualization)

하나의 변수를 요약하는 상황

- 숫자형(numerical) 변수인 경우
- 범주형(categorical) 변수인 경우

둘 이상의 변수 사이 관계를 분석하는 상황

- 숫자형(numerical) 변수인 경우
- 범주형(categorical) 변수인 경우

Univariate Data Visualization

Univariate Data Visualization

가장 널리 알려진 단변량 데이터 시각화(univariate data visualization)는 바로 히스토그램(histogram)이다.

- 나도 거의 매일같이 사용한다. “새로운 데이터의 새로운 변수다” 싶으면 일단 히스토그램을 본다.
- 원칙적으로 보면 히스토그램은 막대차트(bar chart)와는 구별된다. 교과서에 따라서는 (거의 같은 그림인데) 범주형 자료에 대해서는 막대차트(bar chart), 숫자형 자료에 대해서는 히스토그램으로 구별하기도 한다.
- 나는 그런 구분은 우스꽝스럽다고 생각한다. 히스토그램이 숫자형 자료에 대응한다는 견해도 다소 이상하게 들린다. 결국은 히스토그램의 막대(bin)를 나누면서 설령 숫자형(numerical) 자료라도 범주형으로 recoding하기 때문이다.
- 솔직히 말해 히스토그램은 워낙 중요하기 때문에 범주형이고 뭐고 생각하기도 전에 일단 히스토그램을 보는 것이 좋다. 아주 간단한 것인데 이걸 미처 확인하지 않아서 문제가 되는 경우는 무척 많아도, 확인해서 문제가 된 경우는 들어본 적이 없다.
- 다만 히스토그램을 본격적으로 공부하기 앞서 빈도분포표(frequency distribution table)를 먼저 이해하는 것이 좋다.

Univariate Data Visualization

빈도분포표(frequency distribution table) 만들기

- 범주형(categorical) 자료가 주어져 있으면 각 범주별로 하나하나 세서(tally) 즉각 빈도분포표 안에 요약하여 나타낼 수 있다.

ID	female	IQ	income	socialclass
1	1	135	250	1
2	0	110	310	2
3	1	128	1500	3
4	0	98	122	2
5	1	106	450	2
6	0	102	190	1

Female	Frequency	Cum. Perc.
0	3	50.00%
1	3	100.00%

Univariate Data Visualization

- 숫자형(numerical) 자료가 주어져 있다면 먼저 구간(intervals) 단위로 나누고 빈도분포표로 요약하여 나타낼 수 있다.

ID	female	IQ	income	socialclass
1	1	135	250	1
2	0	110	310	2
3	1	128	1500	3
4	0	98	122	2
5	1	106	450	2
6	0	102	190	1

Income	Frequency	Cum. Perc.
$x \leq 200$	2	33.33%
$200 < x < 300$	1	50.00%
$300 < x < 400$	1	66.67%
$400 \leq x$	2	100.00%

Univariate Data Visualization

엑셀에서 빈도분포표를 만들고 히스토그램을 만드려면 먼저 **분석 도구**(Analysis ToolPak)를 설치해야 한다.

- 메뉴에서 [파일]-[옵션]-[추가 기능]으로 들어가 메뉴 화면 하단에 “이동(G)”를 눌러 “분석 도구”를 체크하면 된다.
- 이제부터 데이터 분석 기능을 사용할 수 있지만 엑셀을 기동시킬 때마다 느려진다는 흄이 있다.
- “분석 도구” 밑에 “해 찾기 추가 기능”은 수리사회학 등 여러 분야에서 최적화 문제(optimization problems)를 풀 때 제법 쓸모있지만 우리 수업에서는 다루지 않는다.
- “분석 도구”가 제대로 설치되었다면 [데이터] 메뉴를 선택했을 때 우측 끄트머리에 [데이터 분석]이 새로 생겨난다.
- 안 생겨났으면 엑셀을 껐다가 다시 켜자.

Univariate Data Visualization

eCampus에서 toy.csv를 다운받아 엑셀에서 열자. 먼저 categorical 변수의 히스토그램을 그려보자.

- 메뉴에서 [데이터]-[데이터 분석]을 클릭하여 [통계 데이터 분석] 메뉴를 열자. 쭉 보면 뭔가 이런저런 통계분석을 할 수 있다!
- 일단 “히스토그램”을 골라 메뉴를 열자. “입력” 섹션과 “출력 옵션” 섹션으로 크게 나뉜다.
- “입력” 섹션 안의 “입력 범위”에 보고 싶은 변수를 하이라이트하자.
- “출력 옵션” 섹션 안의 “누적 백분율(Cum. Perc.)”와 “차트 출력”을 체크하고 “확인.”

빈도분포표와 히스토그램은 약간의 문제가 있다. 표와 그래프를 꾸미자.

- 무의미한 계급으로 0.5가 있으므로 행 삭제.
- 축 제목에서 “계급(rank)”은 지나치게 일반적인 표현. “성별”로 바꾸자.
- “기타”가 아니라 3으로 바꾸자.
- 누적 %의 [축 서식] 수정. 오른쪽 y축 경계 최대값은 당연히 100%이다.

Univariate Data Visualization

이번에는 숫자형(numerical) 변수인 IQ로 히스토그램을 그려보자.

- 아까 언급하였듯 숫자형 변수는 범주형 변수로 먼저 recoding한 다음에 히스토그램을 그려야 한다.
- 각 범주별로 상한값(upper limit)을 먼저 입력하자. 90부터 10씩 증가시키는 쉬운 방법이 있다.
- 히스토그램 메뉴 안에 “입력” 섹션 안에 “계급 구간” 안에 상한값을 하이라이트하자.
- “이름표”는 하이라이트 부분 안에 label을 포함하고 있을 경우 체크한다.
- 표와 그래프를 좀 더 꾸미자.

Univariate Data Visualization

그 다음으로는 또 다른 numerical 변수인 income으로 히스토그램을 그려보자. 표와 그래프도 좀 더 꾸미자.

사실은 이렇게 하지 않고 좀 더 쉽게 하는 방법도 있다.

- 해당 변수를 하이라이트하고 [삽입] 메뉴에서 [통계 차트 삽입] 아이콘을 눌러 “히스토그램”을 곧바로 고르는 것이다.
- 엑셀에 사소한 글리치(glitch)가 있어 bin 숫자를 조절하거나 할 때 오른쪽 윈도우를 확대하지 않으면 보이지 않는다.
- 어찌되었든 이렇게 하면 만들 때는 편하긴 한데, 도수분포표를 얻을 수 없고 히스토그램을 꾸밀 때 좀 더 귀찮다는 단점이 있다.

Univariate Data Visualization

eCampus에서 KGSS_religions.xlsx를 다운받자. 이것은 〈한국종합사회조사〉 2018년 데이터로 간단한 인구학적 변수와 종교 관련 문항 변수만을 담고 있다.

- 각 변수의 구체적인 의미는 KGSS_labels.pdf 파일에서 확인할 수 있다.
- 나는 특히 RELIGSPOS가 궁금하기 때문에 이제 이 변수에 대해 빈도분포표와 히스토그램을 만들기로 한다.
- 근데 아까 배운대로 엑셀에서 히스토그램이 만들어지지 않는다. 왜냐하면 엑셀 “분석 도구”에서는 숫자로만 빈도분포표를 만들 수 있기 때문이다(불편!).
- 하지만 빈도분포표를 피벗 테이블로 간단히 만들수 있고 그걸로 다시 히스토그램을 쉽게 만들수 있다.
- 만들어진 피벗 테이블을 하이라이트한 다음, [삽입] 메뉴에서 [2차원 세로 막대형]을 골라 히스토그램을 만들고 좀 더 예쁘게 꾸며보자.

Univariate Data Visualization

히스토그램을 그릴 때 y축은 빈도(frequency)로 삼을 수도 있지만, 밀도(density) 또는 percentage로 삼을 수도 있다.

- 그림 자체야 둘 다 비슷하지만 해석상 밀도 쪽이 좀 더 편하다.
- 엑셀의 sum 함수를 사용하여 각 항목별 비율(proportion)을 계산한 뒤, 그쪽으로 하이라이트를 옮겨 표를 만들어보자.

Univariate Data Visualization

히스토그램만큼 중요하진 않지만 파이 차트(pie chart)도 범주형(categorical) 자료에 대해서 제법 쓰인다.

- 몇몇 전문가들은 pie chart를 제한적으로 사용하도록 충고한다!
- 사실 pie chart는 아주 조금만 복잡해져도 정보전달이 안되고, 정보전달이 쉽게 될 정도로 단순하게 만들면 그림을 그리는 의미가 없어진다. 게다가 histogram 같이 더 나은 대안도 있다.
- 다만 pie chart도 꾸미기에 따라서는 나름 예술적인 요소와 정보 전달 요소를 함께 갖출 수 있다. 여러분의 센스에 달렸다.

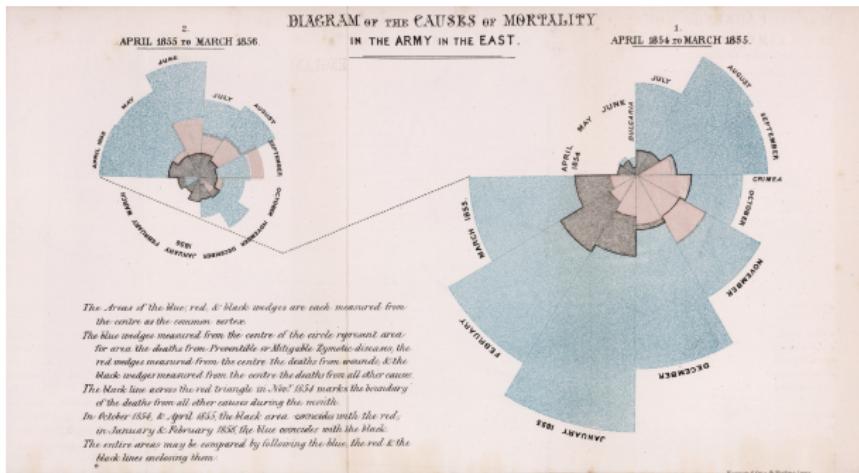
eCampus에서 국내대학현황.xlsx을 다운받아 열자. “지역별 대학의 수”를 pie chart로 나타내자.

- 피벗 테이블로 도수분포표(frequency distribution table)을 만들자. 하이라이트 한 영역을 가지고 [삽입]-[원형 또는 도넛형 차트 삽입]을 선택한다.
- 꾸밀 때는 다음에 주의하자: (1) 슬라이스 색깔; (2) 데이터 레이블; (3) “쪼개진 원형”; (4) “쪼개진 조각” 등.

Univariate Data Visualization

Pie chart는 사실 응용통계학 분야에서 독특한 역사를 가지고 있다.

- 19세기 중반 Florence Nightingale은 평범한 간호사("the lady with the lamp") 이상의 역할을 수행했다. 그녀는 역사상 최초의 실무형 보건통계학자 중 한 명이였고 직접 병원 자료에 근거해 통계를 구축한 뒤 분석을 수행하였다. 특히 Pie chart를 능숙하게 사용했다(그녀는 Rose Diagram이라고 불렸다).



충북대학교
CHUNGBUK NATIONAL UNIVERSITY

Univariate Data Visualization

히스토그램만큼 중요하진 않을지라도 **상자-수염** 그림(Box-Whisker plot)도 나름 교과서 레벨에서는 유명하다.

- 상자-수염 그림은 숫자형(numerical) 자료에 대응한다.
- 이 그림은 중심성향 및 산포성향 주요 요약통계량을 그림 하나로 전달한다는 장점이 있다

eCampus에서 서울시수돗물 수질검사.csv를 다운받자. 이것은 서울특별시 상수도사업본부 수질과에서 2020년에 발표한 자료이다.

- 먼저 산도(pH)의 상자-수염 그림을 그려보자.
- 산도(pH)를 하이라이트하고 [삽입] 메뉴에서 [통계 차트 삽입] 아이콘을 눌러 “상자 수염”을 고르자.
- Minimum, 1st Quartile, 2nd Quartile (=Median), Mean, 3rd Quartile, 4th Quartile (=Maximum)을 함수로 구해 확인하자.
- 이번엔 검사 실적의 상자-수염 그림을 그리자.
- 튀어나온 그 점들이 바로 극단치(outliers)다. 이 점들은 서울시 어느 구의 검사 실적인가?

Univariate Data Visualization

트리맵(treemap)은 위계적(hierarchical) 자료를 나타내기에 좋다.

- 자료에 위계성이 뚜렷하면 뚜렷할수록 트리맵은 독자에게 신선한 충격을 안겨준다 (e.g., “불균형 문제가 이렇게 심했어?”)
- 최근 계산과학(computational science) 및 계산 알고리즘의 혁신 웨이브에 힘입어 비교적 최근 인기 몰이를 하고 있다.
- 각 사각형의 면적은 개별 항목별로 주어진 값에 비례한다(area-based visualization).

eCampus에서 온실가스 에너지 목표관리 명세서 주요정보.xlsx을 열자.
이것은 환경부 온실가스종합정보센터에서 공표한 정부통계이다.

- 법인명과 온실가스 배출량(tCO₂eq)만을 개별적으로 복사하여 새 탭에 붙여넣자.
- 여러분이 실무나 연구 상황에서 종종 마주치게 될 단순한 예외 상황을 여기서 볼 수 있다. 온실가스 배출량이 엑셀에서 제대로 숫자로 인식되고 있지 않다. 숫자가 left-aligned 되어 있고 숫자 앞에 아포스트로피(apostrophe)가 붙어 있음에 주의할 것!
- 숫자로 재인식시킨 뒤, 자료를 이리저리 소팅(sorting)해보자. 이상한 값도 끼어있으니 삭제해야 한다.
- 온실가스 배출량 변수를 하이라이트하여 [삽입]-[트리맵]을 선택하자. 꾸미기는 역시 필요하다.

Univariate Data Visualization

약간 데이터의 성격을 바꾸어 시계열 자료(time-series data)도 시각화 할 수 있다.

- 시계열 자료란 “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).
- 분석 목적에 따라 이들 요소를 정밀하게 분해(decompose)할 필요가 있지만, 우리 수업에서는 (1) 일계차분(first-order differencing)에 의한 추세 제거(de-trending) 와 (2) 이동평균(moving average)에 의한 잡음 제거(de-noising)만 해보자.

Univariate Data Visualization

eCampus에서 dow.csv를 다운받자. 이것은 1953년에서 1990년 초 사이 미국 다우 주가지수 데이터이다.

- 이것은 일간(daily) 데이터인데, 휴일 등 데이터가 축적되지 않은 날도 있어 갭(gap)이 있다.
- 가장 먼저 date 변수와 dowclose 변수의 위치를 바꾸자. 두 변수를 하이라이트하고 [메뉴]-[2차원 꺾은선형]을 고르자.
- 이 그림은 종종 라인 차트(line chart)라고도 불리우며 시계열 자료에 대해 쓰인다. (명확한 이유없이) 시계열 자료가 아닌데 line chart를 쓰는 것은 삼가하자.
- 시계열 차트를 보면 다우지수는 계속 성장하다가 80년대 중반 이후 급성장하고 있다. 다시 말해, 이 시계열 자료에는 상이하게 성장하는 추세들(trends)이 있다.

Univariate Data Visualization

먼저 일계차분(first-differencing)을 해보자.

- 추세를 일계차분으로 제거(detrending)하기 위해 $\Delta \text{dowclose}_t = \text{dowclose}_t - \text{dowclose}_{t-1}$ 를 옆 column에 계산하자.
- 이렇게 계산된 column으로 다시 시계열 차트를 그려보자. 원래의 시계열 차트와 이 시계열 차트 간 어떤 차이가 있는지 눈으로 살펴보자.
- 다만 일계차분의 결과 이 자료에서 추세가 제대로 제거되었는가는 별개의 문제다. 그것은 본격적인 시계열 분석 수업의 주제가 된다.

Univariate Data Visualization

다음으로 이동평균(moving average)을 해보자.

- 처음 그린 시계열 차트를 들여다보면 큰 사이클이 그려지는 동안에도 조그마한 잡음(noise)이 계속 요동치는 것을 확인할 수 있다.
- 엑셀의 average 함수를 이용하여 1-month, 2-month, 3-month 구간별로 “이동하는(moving)” 하는 평균(average)을 구하자. 이때 셀 참조(cell reference)가 자연스럽게 이동하도록 그대로 내버려두면 된다.
- 시계열 차트에서 잡음이 평균 속에서 사라지는(averaged out) 것을 관찰할 수 있다. 이런 방식을 활용하여 의사결정이 잡음에 의해 혼란을 겪지 않고 전반적인 추세(trends)를 관찰하도록 돋는다.

Univariate Data Visualization

우리 수업에서 단변량(univariate) 시각화의 마지막 기법은 인구학적 주제에 관한 것이다.

- 우리 과에서는 (아직) 인구학(demography) 내지 인구학 방법론(demographic methods)이 개설되지 않고 있지만, 종합적으로 판단할 때 인구학은 사회학에서 가장 큰 연구 분야다.
- 미국 인구조사국(Census Bureau) 뿐 아니라 Facebook과 같은 SNS 기업에서도 인구학을 전공한 사회학자를 고용하는 경우가 제법 있다.
- 우리나라의 저출산/저출생, 고령화, 각종 유병률(prevalence rate) 변화 등을 고려하면 사회학에서 다루어야 하는 수많은 인구 문제가 여럿 있는 셈이다.

Univariate Data Visualization

가장 보편적인 형태인 성별-연령별 인구 피라미드(population pyramid)를 그려보자.

- 인구 피라미드는 일종의 막대 차트(bar chart)다. 다만 가로로 변형되었기에 알아보기 어려울 수는 있다.
- 인구학적 데이터를 찾는 것 자체가 굉장히 중요한 노하우(또는 know-where)이기 때문에 이것도 가볍게 연습하자.
- eCampus에 일부러 자료를 올리지 않았다. 국가통계포털(<https://kosis.kr>)에 직접 가자.
- 메뉴에서 [국내통계]-[주제별 통계]를 클릭하자. 다시 [인구]를 고르고 [장래인구추계]-[전국(2017년 기준)]-[성 및 연령별 추계인구(1세별,5세별)]을 눌러 들어가자. 표가 나오면 [통계표조회]를 클릭하고 [다운로드]하자.
- 다운받은 자료에서 2017년 성별-연령별로 자료를 복사하여 새 템에 붙여넣자. 이 때 columns 두 개는 성별에 할당하고, 여러 개의 rows에는 연령별에 할당하자.
- “합계” row는 지우자.

Univariate Data Visualization

	남자	여자
0 - 4세	1102364	1048855
5 - 9세	1188607	1120380
10 - 14세	1173575	1090502
15 - 19세	1554471	1427984
20 - 24세	1882066	1647591
25 - 29세	1791765	1576396
30 - 34세	1825165	1666377
35 - 39세	2089929	1964371
40 - 44세	2083256	1995148
45 - 49세	2278705	2228237
50 - 54세	2107628	2056212
55 - 59세	2102223	2123676
60 - 64세	1554982	1615386
65 - 69세	1098706	1187813
70 - 74세	800334	956466
75 - 79세	620799	873516
80세이상	482218	1046208

Univariate Data Visualization

- 전체 자료를 하이라이트해서 [삽입]-[2차원 가로 막대형]을 선택하자.
히스토그램처럼 세로가 아니라 가로다.
- 벌써 그럴듯하게 그려졌지만 왼쪽이 없다. 왼쪽에 남자를 보내고 싶다면 남자의 모든 관찰값에 -1을 곱하는 새로운 column을 만들자.
- 차트에서 남자쪽 bar를 클릭하여 활성화하고 하이라이트를 -1을 곱한 새로운 column으로 옮기자.
- 상당히 그럴듯하게 완성되었지만 뭔가 정렬(align)이 잘 되어있지 않아 어색하다.
이건 bar를 우클릭한 뒤, “데이터 계열 서식”을 선택하고 “계열 겹치기”를 양수(+)로
“간격 너비”를 적절히 좁히면 해결되는 문제다.
- 놓치기 쉽지만 남자쪽의 x축이 (-)로 처리되어 있어 어색하다. 이쪽도 x축을
활성화한 뒤, 우클릭해서 “축 서식”을 선택하자. “축 옵션”에서 아이콘들을 잘
살펴보자. “표시 형식”을 바꿀 수 있는 부분이 있다. “숫자”로 바꾼 뒤, “음수(N)”를
어떻게 처리할지 선택하는 부분이 있다. (-)가 없는 숫자로 고르면 되는데 빨간색이
마음에 들지 않는다. 그런데 밑에 서식 코드를 보니 [빨강]이라는 부분이 있다. 이걸
지우고 “추가(A)”를 누르면 해결된다.