사회통계연습
시각화(I)

김현우, PhD[1]

[1] 충북대학교 사회학과 부교수

진행 순서

1. 도수분포표와 히스토그램
2. 트리맵과 오자이브

도수분포표와 히스토그램

도수분포표와 히스토그램

가장 중요한 단일변수 시각화 기법은 빈도분포표와 히스토그램이다.

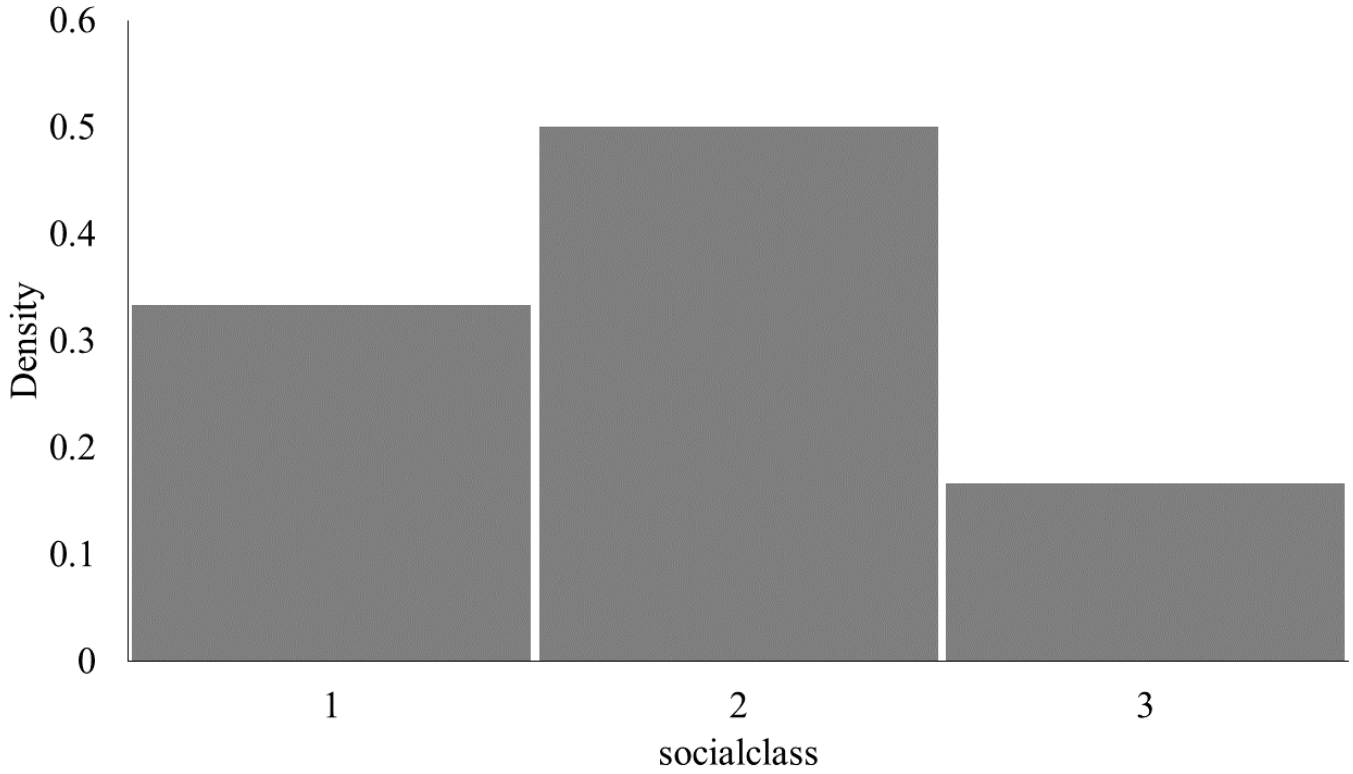
- (명목척도나 서열척도로 측정한) 질적변수가 주어져 있으면 각 범주별로 하나하나 세서(tally) 빈도분포표 안에 요약할 수 있다.

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

socialclass	Frequency	Perc.	Cum. Perc.
1	2	33.33%	33.33%
2	3	50.00%	83.33%
3	1	16.67%	100.00%

도수분포표와 히스토그램

- 이렇게 만들어진 빈도분포표를 그림으로 나타내면 곧 막대차트(bar chart) 또는 히스토그램(histogram)이 된다.
- 도수분포표와 히스토그램 그리고 막대차트는 본질적으로 같은 정보를 전달한다 (Why?).



도수분포표와 히스토그램

연습 1. NHANES.SAV 자료에서 주관적 건강(health)의 빈도분포표와 히스토그램(또는 막대차트)을 작성하고 이를 해석하시오.

도수분포표와 히스토그램

연습 2. NHANES.SAV 자료에서 연령집단(agegrp)의 빈도분포표와 히스토그램(또는 막대차트)을 작성하고 이를 해석하시오.

도수분포표와 히스토그램

- 교과서에 따르면 히스토그램은 양적변수에 대응하고, 막대차트는 질적변수에 대응하지만 현실적으로는 거의 같다.
- 양적변수는 질적변수로 바꿀 수 있으므로 결국 범주형 자료로 재부호화하면 쉽게 시각화할 수 있기 때문이다.
- 재부호화를 하려면 SPSS에서는 [변환]-[다른 변수로 코딩변경]이고, Jamovi에서는 [데이터]-[변환]에서 작업할 수 있다.
- 혹시 나중에 프로그래밍을 배우면 자료구조 측면에서 히스토그램과 막대차트의 차이를 이해할 수 있다.
- 만약 양적변수가 주어졌다면 (재부호화하기 앞서) 합리적인 기준점(cutpoint)을 잡기 위해 기술통계표를 먼저 확인보아야 할 필요가 있다(Why?).

도수분포표와 히스토그램

연습 3. NHANES.SAV 자료에서 이완기 혈압(bpdiast)의 빈도분포표와 히스토그램(또는 막대차트)을 작성하고 이를 해석하시오. 자신이 선택한 구간설정의 근거에 대해서도 논의하시오.

도수분포표와 히스토그램

- 일단 양적변수라도 빈도분포표와 히스토그램을 그려보고 어디에 숫자가 집중되어 있는지 살펴보자.
- 이러한 집중 전후로 구간을 나누는 것도 괜찮은 전략이 된다.
- 약간의 시행착오를 거쳐 70미만, 70이상-80미만, 80이상-85미만, 85이상-90미만, 90이상으로 나누어보았다.
- 각 범주별로 관측치가 적당히 균형을 갖춰 분배되도록 주의하자.

도수분포표와 히스토그램

응답의 척도가 같다면 여러 변수들을 하나의 표 안에 정리할 수 있다.

- 만일 여러 변수들이 동일한 부호화 기준(coding scheme)을 사용하고 있다면(e.g., 리커트 척도), 변수를 행(rows)에 놓고 퍼센트를 열(columns)에 놓는 하나의 도수분포표로 요약할 수도 있다.
- 가령 리커트 척도로 측정된 질문 10개가 있다고 하자. (10개의 빈도분포표를 따로 만들기보다) 하나의 표로 10개의 질문을 일목요연하게 정리하는 것이 낫다.

	A	B	C	D	E	F
1						
2	성관계에 대한 태도	전적으로 옳지 않다	대부분 옳지 않다	때에 따라 옳지 않다	전혀 잘못되지 않았다	
3	혼전성교	17.60%	13.40%	28.90%	40.10%	
4	혼외성교	70.70%	18.90%	7.40%	3.10%	
5	동성성교	51.60%	15.60%	13.80%	18.90%	
6	(참고: 반올림으로 인해 합이 100%와 다를 수 있음)					
7						
8						

도수분포표와 히스토그램

연습 3. VALUE.SAV 자료를 불러와 성관계에 대한 태도 세 변수에 관한 통합형 빈도분포표와 막대차트를 작성하고 그 결과를 해석하시오.

도수분포표와 히스토그램

- SPSS나 Jamovi에서 세 변수들의 빈도분포표를 출력하고 이를 엑셀로 옮겨서 편집하자.
- 표는 예쁘게 꾸며야 한다. 필요에 따라 전치행렬(transposed matrix)을 사용하자.
- Jamovi에서는 VijPlots이라는 모듈을 추가로 설치하고 Likert plot을 골라 쉽게 그릴 수 있다.
- SPSS에서는 [분석]-[표]-[사용자 정의 표]로 간다. 도수분포표로 요약할 변수를 "행"에 넣고 "요약통계량"에서 '빈도' 대신 '행 퍼센트' 안의 '행 N %'을 넣는다. "선택한 항목에 적용"하고 빠져나와 "범주 위치"는 '기본값' 대신 '행 레이블을 열로 표시'한다.

트리맵과 오자이브

트리맵과 오자이브

트리맵은 불균등한 현실을 폭로하기에 좋다.

- 자료에 위계성이 뚜렷하면 뚜렷할수록 트리맵(treemap)은 독자에게 신선한 충격을 안겨준다(e.g., “불균형 문제가 이렇게 심했어?”)
- 최근 계산과학(computational science) 및 계산 알고리즘의 혁신 웨이브에 힘입어 인기몰이를 하고 있다.
- 각 사각형의 면적은 개별 항목별로 주어진 값에 비례한다(area-based visualization).
- 트리맵을 그리기 위해 엑셀을 사용하자.

트리맵과 오자이브

연습 4. 온실가스 에너지 목표관리 명세서 주요정보.xslx를 엑셀에서 불러들여 온실가스 배출량 트리맵을 작성하시오.

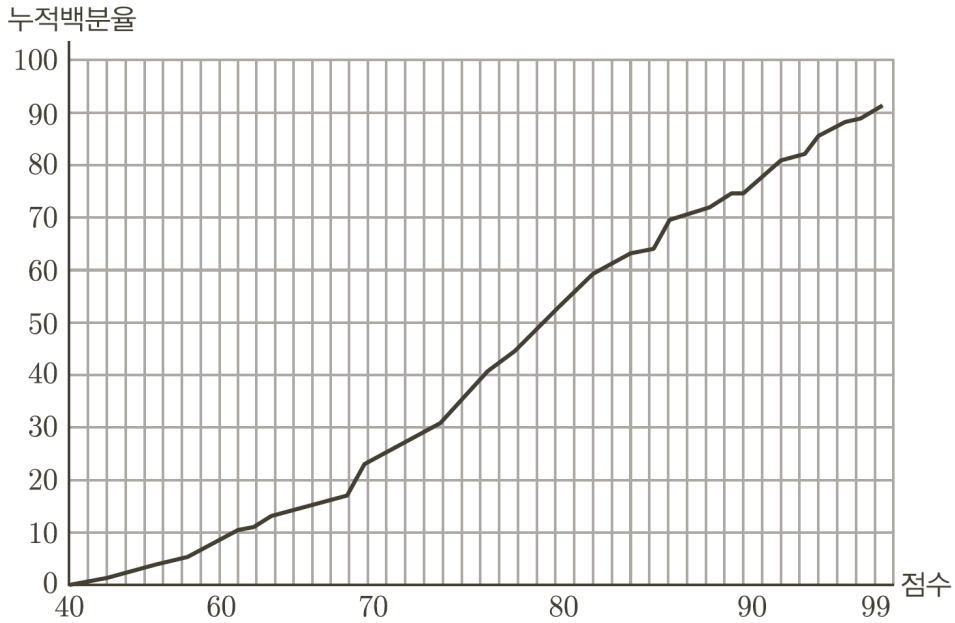
트리맵과 오자이브

- 법인명과 온실가스 배출량만을 개별적으로 복사하여 새 탭에 붙여넣자. (필요에 따라) 자료를 이리저리 정렬(sort)하여 이상한 값을 제거해야 한다.
- 온실가스 배출량 변수를 하이라이트하여 [삽입]-[트리맵]을 선택한다.
- 법인명이 레이블로 나오는 편이 보기 좋으므로 [차트 요소]에서 [데이터 레이블]을 선택해야 한다.
- 약간의 꾸미기는 역시 필요하다.

트리맵과 오자이브

오자이브는 구간별 누적백분율을 시각화한다.

- 오자이브(ogive)의 x축은 (나타내고자 하는 변수의) 도수(frequency) 혹은 계급구간(class interval)을 나타내고, y축은 누적백분율(cumulative percentage)을 나타낸다.

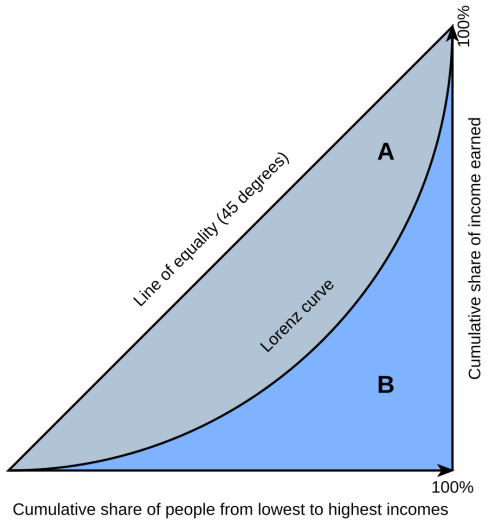


트리맵과 오자이브

연습 5. INCOME.SAV 자료에서 월평균 가구소득을 히스토그램과 오자이브로 각각 시각화하고 그 함의를 해석하시오.

트리맵과 오자이브

- SPSS에서 오자이브를 그리기 위해서 [그래프]-[차트 작성기]에서 '선형 차트'를 먼저 고르고 x축에 월평균 가구소득을 넣는다. "통계량"을 '누적 퍼센트'로 바꾸어야 한다.
- Jamovi에서는 scatr이라는 모듈을 설치하고 [분석]-[탐구]에 있는 Pareto 차트를 선택하여 그린다(단 해당 변수를 서열척도로 바꾸어야 한다). 차라리 Excel에서 그리는 것이 나을 수도 있다.
- 히스토그램과 비교해보면서 오자이브의 의미를 해석해보자.
- 오자이브와 로렌츠 곡선(Lorenz curve)은 (소득불평등을 나타내는) 이른바 지니계수(gini coefficient)와 직결된 시각화 기법이다.
- 두 시각화 기법의 공통점과 차이점을 생각해보자!



트리맵과 오자이브

연습 6. INCOME.SAV 자료에서 월평균 가구소득을 트리맵으로 시각화하고 그 함의를 해석하시오.