사회통계
단일 변수의 시각화

김현우, PhD[1]

[1] 충북대학교 사회학과 조교수

진행 순서

1. 데이터 시각화의 시작
2. 단일변수의 시각화: 도수분포표
3. 단일변수의 시각화: 히스토그램
4. 단일변수의 시각화: 파이 차트
5. 단일변수의 시각화: 상자-수염 그림

데이터 시각화의 시작

데이터 시각화의 시작

왜 사회학도가 시각화를 배우나?

- 사람들은 보통 (숫자나 글보다) 표나 그림을 선호한다.
- (단순히 숫자를 나열할 때보다) 자료의 경향이나 특성을 이해하는데 큰 도움이 된다.
- 첫째, 변수의 분포(distribution)를 보여주어 데이터의 문제(e.g., 극단치)를 파악하거나 다음 분석을 계획하는데 용이하다.
- 둘째, 여러 변수들 사이에 혹시 어떤 연관성(association)이 있지 않나 하는 의문을 빠르게 해소하는데 유리하다.

데이터 시각화의 시작

데이터 시각화는 변수의 숫자 및 자료유형에 따라 달라진다.

- (1) 단일변수의 시각화(univariate data visualization). 이 안에서 양적변수를 시각화하는 경우와 질적변수를 시각화하는 경우가 다르다.
- (2) 둘 이상의 변수 사이 관계의 시각화(bivariate/multivariate data visualization). 다시 (1) 양적변수인 경우, (2) 모두 질적변수인 경우, (3) 양적변수와 질적변수가 뒤섞인 경우가 다르다.

단일변수의 시각화: 도수분포표

단일변수의 시각화: 도수분포표

자료의 체계적 요약은 빈도분포표를 만드는 것에서 출발한다.

- 질적변수가 주어져 있으면 각 범주별로 하나하나 세어(tally), 빈도분포표(frequency distribution table) 안에 요약하여 나타낼 수 있다.

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

X=female	Frequency	Perc.	Cum. Perc.
0	3	50.00%	50.00%
1	3	50.00%	100.00%

단일변수의 시각화: 도수분포표

- 양적변수가 주어져 있다면 먼저 계급구간(class intervals)에 따라 값을 나누고 빈도분포표로 나타낼 수 있다(Why?).

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

X=income	Frequency	Perc.	Cum. Perc.
$X \leq 100$	0	0%	0%
$100 < X \leq 200$	2	33.33%	33.33%
$200 < X <\leq 300$	1	16.67%	50.00%
$300 < X <\leq 400$	1	16.67%	66.67%
$400 < X$	2	33.33%	100.00%

단일변수의 시각화: 도수분포표

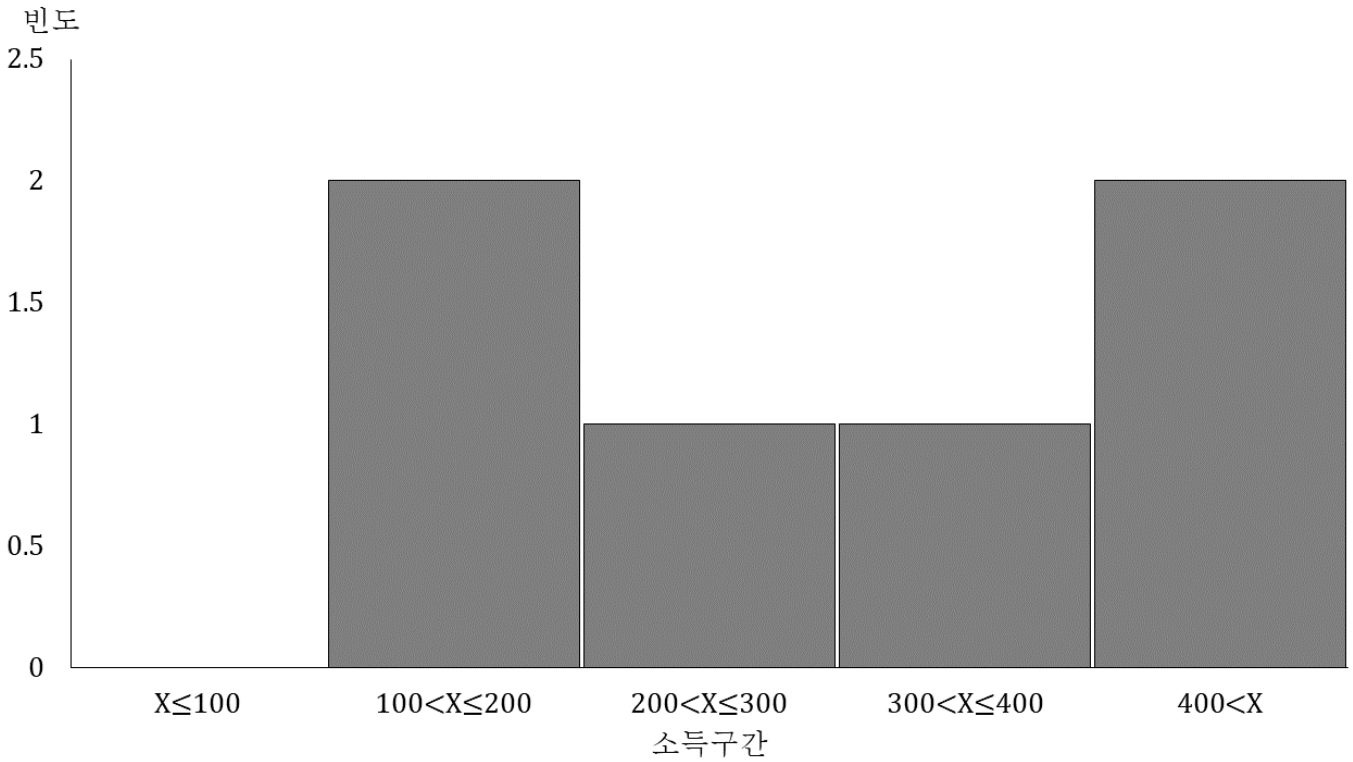
- 가장 먼저 계급구간의 수를 결정한다. 너무 적어도 너무 많아도 곤란하다(Why?).
- 계급구간의 크기를 결정할 때, 각 계급구간 간에 충분한 의미가 있도록 주의해야 한다. 하나의 구간에 충분한 사례가 들어가도록 설정한다.
- 빠진 계급구간의 범주가 있어선 안된다. 가령 [0-25), [25-50), [75-100]은 틀리다.

단일변수의 시각화: 히스토그램

단일변수의 시각화: 히스토그램

도수분포표를 그림으로 표현하면 히스토그램이 된다.

- 히스토그램(histogram)은 가장 널리 알려진 단일변수의 시각화 기법이다.



단일변수의 시각화: 히스토그램

- 도수분포표와 히스토그램은 나도 거의 매일같이 사용한다.
- "새로운 데이터의 새로운 변수다" 싶으면 일단 그것의 도수분포표와 히스토그램을 본다.
- 솔직히 말해 두 기법은 워낙 중요하기 때문에 자료유형이고 뭐고 생각하기도 전에 일단 둘 다 확인하는 편이 좋다.
- 아주 간단한 것들인데 이걸 미처 확인하지 않아서 문제가 되는 경우는 무척 많아도, 확인해서 문제가 된 경우는 들어본 적이 없다.

단일변수의 시각화: 히스토그램

원칙적으로 히스토그램은 막대 그래프와는 구분된다.

- 교과서에 따르면 (거의 같은 그림인데) 질적변수에 대해서는 막대 그래프(bar chart), 양적변수에 대해서는 히스토그램을 사용해야 한다.
- 나는 그런 구분이 솔직히 우스꽝스럽다고 생각한다.
- 히스토그램과 막대 그래프의 유일한 차이는 막대(bins)가 서로 붙어있나 붙어있지 않나 정도에 불과하다(심지어 미적인 이유로 인해 히스토그램에서도 막대를 살짝 떼놓기도 한다).
- 히스토그램이 양적변수에 대응한다는 점도 곰곰히 생각해보면 다소 어색하다. 결국 도수분포표의 계급구간을 나누면서 양적변수를 질적변수로 재부호화(recoding)하기 때문이다.
- 결론적으로 말해 양적변수건 질적변수건 히스토그램(혹은 막대 그래프)을 그릴 수 있다.

단일변수의 시각화: 히스토그램

예제 1. midterms.csv는 통계학을 수강하는 36명 학생의 중간고사 점수 자료이다. 이 자료의 변수 유형은 무엇인가? 이 자료를 10점 단위 구간으로 도수분포표와 히스토그램을 작성하시오.

71	71	73	74	76	77	77	78	79
36	53	57	62	63	65	68	69	69
86	88	88	88	89	90	92	92	93
79	80	80	81	81	91	82	83	83

단일변수의 시각화: 히스토그램

- 엑셀에서 데이터분석을 하려면 분석도구(Analysis ToolPak)를 설치해야 한다.
- 분석도구는 무료인데다 몇 가지 초보적인 데이터 분석에 유용하지만, 엑셀을 기동시킬 때마다 느려진다는 흠이 있다.
- 메뉴에서 [파일]-[옵션]-[추가 기능]으로 들어가 메뉴 화면 하단에 "이동(G)"를 고르고 "분석 도구"를 체크하면 된다.
- "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.

단일변수의 시각화: 히스토그램

- 중간고사 점수는 양적변수이고, 양적변수를 분석한다면 계급구간이 필요하다.
- 각 범주별로 상한값을 다른 칸에 먼저 입력해 두어야 한다(40부터 10씩 증가시키는 쉬운 방법이 있다).
- 데이터 분석에서 [히스토그램]을 선택한다.
- 히스토그램 메뉴에서 “입력” 섹션의 “계급 구간” 안에 미리 입력해 둔 상한값을 하이라이트하자.
- “이름표”는 하이라이트 부분 안에 변수 이름(variable name)과 계급구간 이름을 포함하고 있을 경우 체크한다.

단일변수의 시각화: 히스토그램

- 결과물로 나온 빈도분포표와 히스토그램을 반드시 꾸미자!
- 축 제목에서 “계급”은 너무 일반적인 표현이므로 “점수”와 같이 구체적인 표현으로 바꾸자.
- (필요에 따라) “기타”를 제거해야 할 수도 있다.

단일변수의 시각화: 히스토그램

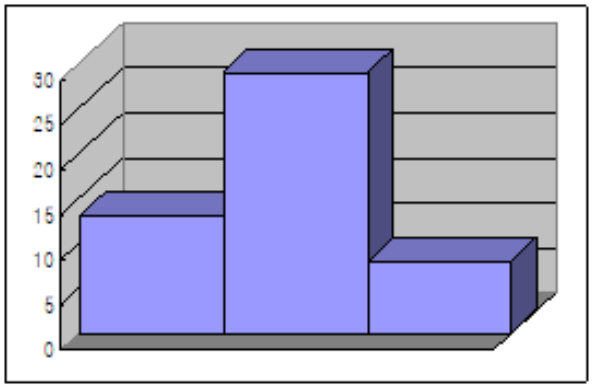
- 히스토그램만 작성하고 싶다면 사실은 좀 더 편리한 방법도 있다!
- 해당 변수를 하이라이트하고 [삽입] 메뉴에서 [통계 차트 삽입] 아이콘을 눌러 "히스토그램"을 곧바로 고르는 것이다.
- 엑셀에 사소한 오류(glitch)가 있어 구간 크기나 숫자를 조절하거나 할 때 오른쪽 윈도우를 확대해야만 보인다.
- 이 방식으로는 도수분포표를 함께 얻을 수 없다. 하지만 피벗 테이블(pivot table)로 도수분포표만 따로 얼마든지 쉽게 만들 수 있다.

단일변수의 시각화: 히스토그램

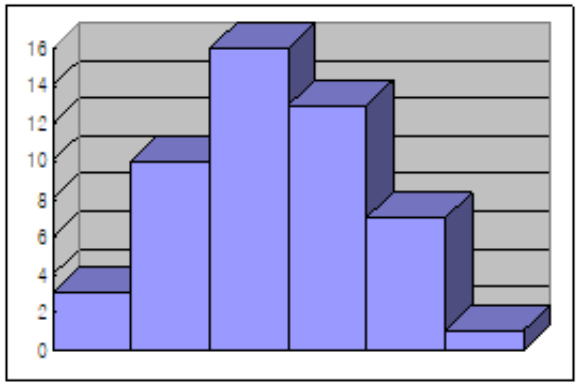
- 히스토그램을 그릴 때 빈도(frequency)를 y축으로 삼을 수도 있지만, 밀도(density) 또는 백분위(percentage)를 사용할 수도 있다.
- 그림 자체야 둘 다 비슷하지만 해석상 밀도나 백분위 쪽이 좀 더 편하다(Why?).
- 엑셀의 SUM(·) 함수를 사용하여 각 항목별 백분위를 계산한 뒤, 그쪽으로 하이라이트를 옮겨 히스토그램을 수정해보자.

단일변수의 시각화: 히스토그램

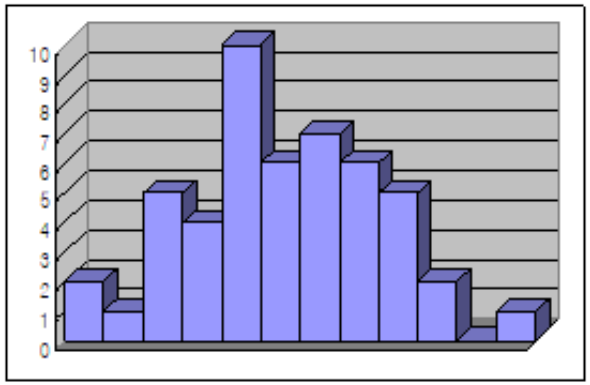
- 같은 자료라도 히스토그램의 구간 설정에 따라 완전히 다른 자료처럼 보인다.



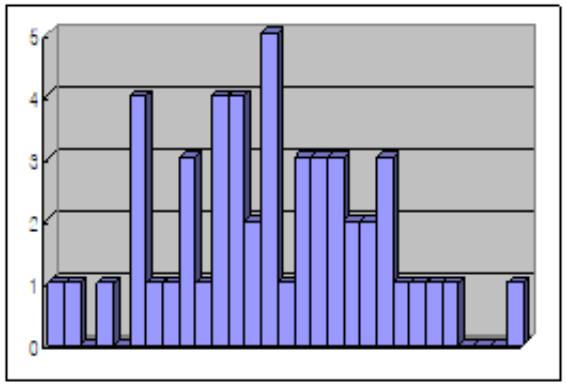
계급 구간=3



계급 구간=6



계급 구간=12



계급 구간=30

단일변수의 시각화: 히스토그램

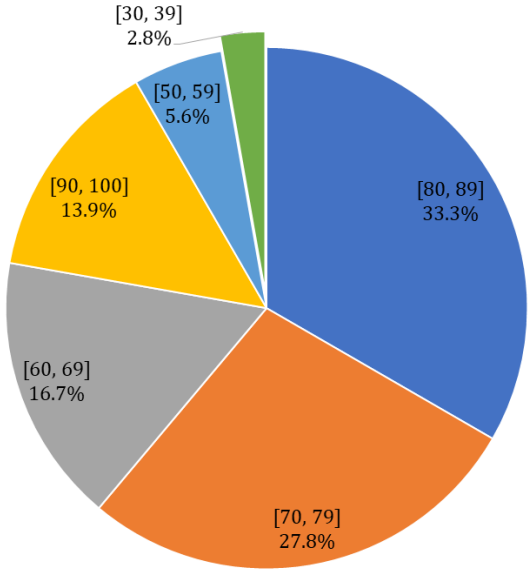
예제 2. WHR.csv는 국가별로 사회적 지지(social support)와 건강기대수명(healthy life expectancy) 통계를 보고하고 있다. 사회적 지지의 도수분포표와 히스토그램을 작성하시오. 이때 적절한 간격(interval)을 주의깊게 선택하시오.

단일변수의 시각화: 파이 차트

단일변수의 시각화: 파이 차트

질적변수를 나타낼 때는 파이 차트도 제법 쓰인다.

- 앞서 예제 1에서 사용한 중간고사 점수 자료 변수는 아래와 같은 파이 차트(pie chart)로 나타낼 수 있다.



단일변수의 시각화: 파이 차트

- 양적변수건 질적변수건 파이 차트를 그릴 수 있다.
- 파이 차트는 큰 조각부터 배열하는 것이 보통이다. 이를 위해 자료를 정렬(sort)해야 한다.
- 특정 조각을 강조하기 위해 살짝 돌출시킬 수도 있다.
- 범례를 추가하여 정확한 비율을 표시할 수도 있다.

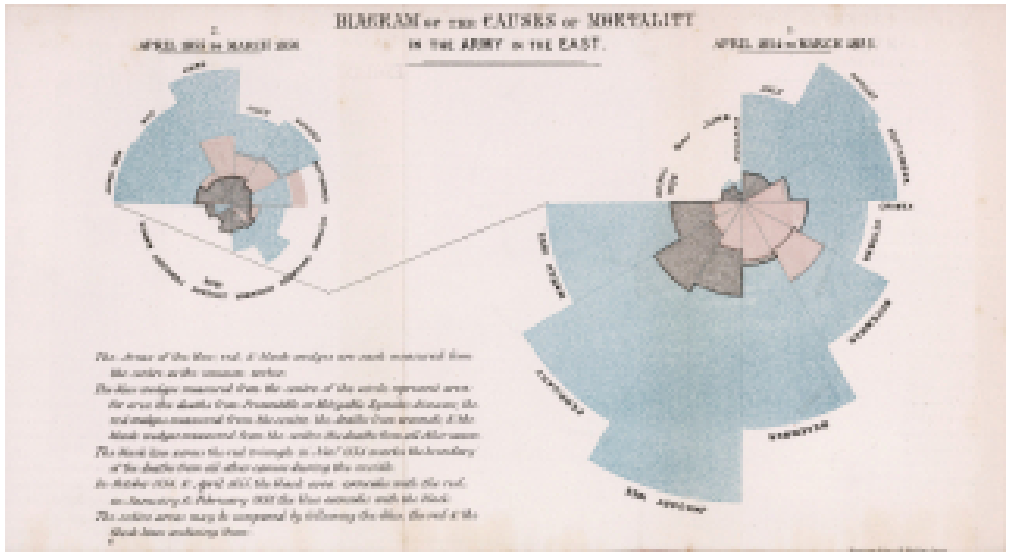
단일변수의 시각화: 파이 차트

- 하지만 몇몇 전문가들은 파이 차트를 제한적으로 사용하도록 충고한다.
- 사실 파이 차트는 아주 조금만 복잡해져도 정보전달이 안되고, 정보전달이 쉽게 될 정도로 단순하게 만들면 그림을 그리는 의미가 없어진다.
- 사실 히스토그램보다 특별히 더 나은 것 같지도 않다.
- 다만 파이 차트도 꾸미기에 따라서는 나름 예술적인 요소와 정보 전달 요소를 함께 갖출 수 있다. 여러분의 센스에 달렸다.

단일변수의 시각화: 파이 차트

파이 차트는 사실 통계학 분야에서 독특한 역사를 가지고 있다.

- 19세기 중반 Florence Nightingale은 평범한 간호사("the lady with the lamp")가 아니었다.
- 그녀는 최초의 실무형 보건통계학자 여성으로, 직접 병원 자료를 수집하고 통계를 구축한 뒤 분석을 수행하였다.
- Adolphe Quetelet에 영향을 받았으므로 사회학과도 간접적인 인연이 있었다.



단일변수의 시각화: 파이 차트

예제 3. employed.csv는 지역주민 600명의 취업상태를 조사한 결과를 정리한 도수분포표이다. 이를 토대로 취업 상태를 파이 차트로 나타내시오.

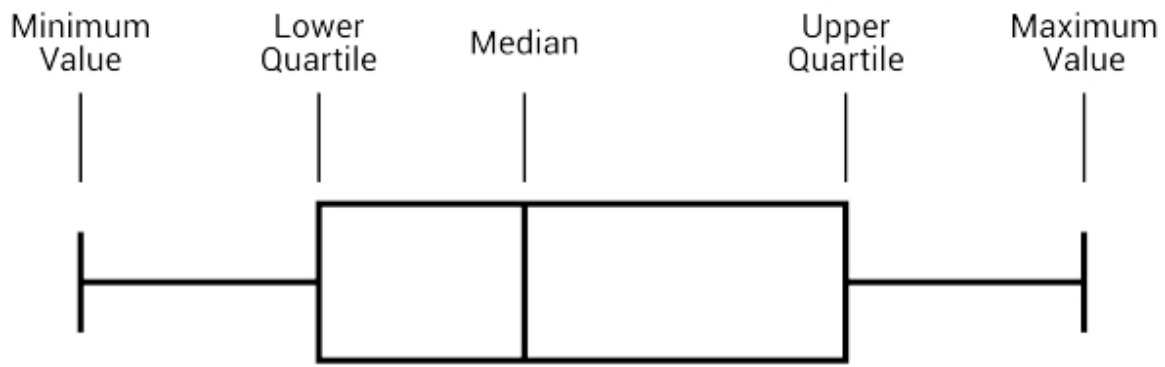
코드	취업상태	도수	상대도수
1	정규직	243	0.41
2	비정규직	98	0.16
3	휴직/실업	32	0.05
4	은퇴	76	0.13
5	학생	45	0.08
6	주부	83	0.14
7	기타	23	0.04
	합계	600	1.00

단일변수의 시각화: 상자-수염 그림

단일변수의 시각화: 상자-수염 그림

상자-수염 그림은 양적변수를 요약하는데 사용될 수 있다.

- 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.



단일변수의 시각화: 상자-수염 그림

예제 4. 서울시수돗물 수질검사.csv는 서울특별시 상수도사업본부 수질과에서 발표한 자료의 일부이다. “검사실적” 변수의 상자-수염 그림을 그리고 해석하시오.

단일변수의 시각화: 상자-수염 그림

- 검사실적을 하이라이트하고 [삽입] 메뉴에서 [통계 차트 삽입] 아이콘을 눌러 “상자 수염”을 고르면 된다.
- 해석할 때는 최소값, 첫번째 사분위수, 두번째 사분위수(=median), 평균, 세번째 사분위수, 네번째 사분위수(=최대값)를 언급하면 된다.
- 각각을 엑셀 함수로 구해 답을 확인해 볼 수 있다.
- 튀어나온 점들은 극단치(outliers)라고 볼 수 있다. 이 점들은 서울시 어느 구의 검사 실적인가?