사회통계연습
집계화와 기술통계표 구성

김현우, PhD[1]

[1] 충북대학교 사회학과 부교수

진행 순서

1. 기술통계표
2. 표준화
3. 집계화

기술통계표

기술통계표

기술통계를 확인했다면 이를 보고해야 한다.

- 기술통계를 표(table)와 그림(figure)으로 잘 요약해야 한다!
- 표 안에 들어가야 하는 정보는 대체로 정해져 있다: (1) 평균(또는 중앙값), (2) 표준편차(또는 분산), (3) 최소값 및 최대값.
- 먼저 표를 깔끔하게 만들어 보고하고 이를 언어적으로 표현한다. “이 변수의 평균은 얼마이고 표준편차는 얼마이다. 최소값과 최대값은 각각 얼마와 얼마이다.”
- 리커트 척도의 평균이나 표준편차 따위를 그대로 보고하는 것은 좋지 않다(Why?).

기술통계표

연습 1. INCOME.SAV 자료를 불러와 성별, 연령, 학력, 월평균 가구소득에 관한 기술통계표를 만들고 해석하시오.

기술통계표

	A	B	C	D
1			빈도	비율
2	성별	남자	479	40.70%
3		여자	699	59.30%
5	연령대	29세 이하	97	8.20%
6		30세~39세	174	14.80%
7		40세~49세	215	18.30%
8		50세~59세	265	22.50%
9		60세~69세	274	23.30%
10		70세 이상	153	13.00%
12	최종학력	중졸 이하	236	20.00%
13		고졸	385	32.70%
14		전문대학(2 · 3년제)	201	17.10%
15		대학교(4년제) 이상	356	30.20%
17	월평균 가구소득	99만원 이하	80	6.80%
18		100만원~299만원	288	24.40%
19		300만원~499만원	305	25.90%
20		500만원~699만원	274	23.30%
21		700만원 이상	231	19.60%
22		전체 응답자	1,178	100%

기술통계표

기술통계표는 깔끔하게 어느 정도 꾸며져야 한다.

- 기술통계표를 작성할 때, 양적변수를 기준으로 할 것인지, 질적변수를 기준으로 작성할 것인지 먼저 결정해야 한다(물론 혼합할 수도 있다).
- 기술통계표를 꾸밀 때는 엑셀을 활용하는 것이 바람직하다. 적절한 폭과 넓이를 갖추어야 한다. 과도한 칸막이는 피하자.
- 기존 문헌에서 표를 어떻게 꾸몄는가를 보고 흉내내서 연습해야 한다.
- 최종적으로는 한글이나 마이크로소프트 워드에 복사하여 붙여넣자.

기술통계표

기술통계표는 표본의 성격을 묘사하는데도 흔히 사용된다.

- 여러분의 연구에 참여한 표본(sample) 혹은 실험 참가자(subjects)의 인적 특성을 요약한다.
- 이때 주요 변수에 따라 표본 또는 실험 참가자 간의 차이점을 부각시킬 수도 있다.

기술통계표

〈표 2〉 응답자 특성

구분		사례수	%
전체		1,203	100.0
지역별	서울/인천/경기	592	49.2
	강원	36	3.0
	대전/충청	120	10.0
	대구/경북	127	10.6
	광주/전라	120	10.0
	부산/울산/경남	195	16.2
	제주	13	1.1
지역크기별	대도시	581	48.3
	중소도시	527	43.8
	군(읍/면)	95	7.9
성별	남자	606	50.4
	여자	597	49.6
연령별	20대	282	23.4
	30대	304	25.3
	40대	290	24.1
	50대	196	16.3
	60세이상	131	10.9
학력별	초등졸이하	158	13.1
	중졸	513	42.6
	고졸	530	44.1
	모름/무응답	2	0.2
직업별	농/수/축산업	28	2.3
	자영업	293	24.4
	블루칼라	248	20.6
	화이트칼라	217	18.0
	전업주부	240	20.0
	학생/무직/기타	176	14.6
	모름/무응답	1	0.1
가구소득 수준	100만원이하	87	7.2
	101-200만원	281	23.4
	201-300만원	347	28.8
	301-500만원	376	31.3
	501-1천만원	97	8.1
	1천만원초과	12	1.0
	모름/무응답	3	0.2

4) 전국인구현황(2006년 12월말 기준 주민등록인구통계, 행정자치부)을 모집단으로 하여 지역, 성, 연령별 구성비에 따라 표본을 배분하였다. 최종조사지점은 지역을 고려한 읍면동의 통반리로, 각 최종조사지점에서 성×연령 할당에 따라 조사대상자를 추출하였다. 이 과정에서 응답자의 성, 연령 특성이 다른 변인(교육수준, 가구 소득수준, 직업 등)에 우선하여 고려된 관계로 지역, 성, 연령 외 다른 특성들은 모집단 구성비와 차이가 있을 수 있다.

김춘애·서원식. 2013. "병원조직 구성원의 교육훈련 만족도 결정요인." 한국산학기술학회 논문지 14(10): 5012-5019.

기술통계표

<표 1> 분석대상자의 일반적 특성 및 성별에 따른 t-검정 결과

변수	전체			여성			남성			T-검정
	표본	평균/비율	표준편차	표본	평균/비율	표준편차	표본	평균/비율	표준편차	
종속변수										
우울 (2차 조사)	3,246	18.14	5.78	1,419	19.10	5.96	1,827	17.40	5.52	***
독립변수										
우울 (1차 조사)	3,248	16.73	4.88	1,423	17.39	5.23	1,825	16.22	4.53	***
사별여부 (사별=1)	3,265	0.03	0.17	1,433	0.05	0.22	1,832	0.01	0.11	***
연령 (1차 조사)	3,265	68.43	6.10	1,433	67.58	5.53	1,832	69.09	6.43	***
성별 (여성=1)	3,265	0.44	0.50							
교육수준 초등학교이하	3,263	0.58	0.49	1,432	0.75	0.43	1,831	0.45	0.50	***
중학교	3,263	0.16	0.37	1,432	0.14	0.35	1,831	0.17	0.38	*
고등학교	3,263	0.17	0.38	1,432	0.09	0.28	1,831	0.24	0.43	***
대학이상	3,263	0.09	0.28	1,432	0.02	0.15	1,831	0.14	0.35	***
자가소유 (2차 조사)	3,265	0.86	0.35	1,433	0.85	0.35	1,832	0.86	0.35	
경제적 만족도 (2차 조사)	3,265	4.66	2.38	1,433	4.46	2.33	1,832	4.81	2.41	***
주관적 건강상태 (2차 조사)	3,265	2.07	0.88	1,433	1.91	0.84	1,832	2.20	0.89	***
배우자와의 관계만족도 (1차 조사)	3,265	6.99	2.06	1,433	6.67	2.18	1,832	7.25	1.93	***
자녀와의 관계만족도 (1차 조사)	3,230	7.31	1.93	1,418	7.33	1.96	1,812	7.29	1.92	
사회활동 참여여부 (1차 조사)	3,265	0.57	0.49	1,433	0.50	0.50	1,832	0.63	0.48	***

주: +p<.1, *p<.05, **p<.01, ***p<.001

이민아. 2014. "사별과 우울에 대한 종단분석: 성차와 배우자와의 관계만족도를 중심으로." 한국인구학 37(1): 109-130.

표준화

표준화

어떤 변수를 비교분석하기 앞서 표준화가 필요할 수 있다.

- 분산(variance) σ^2은 편차의 제곱합(sum of squared deviations)이다. 그러므로 본래 단위가 cm라면 분산의 단위는 cm^2가 된다(Why?).

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

- 분산과 공분산(covariance)은 공통적으로 단위민감성(unit sensitive)이라는 이슈가 있다(즉 단위가 달라지면 분산도 달라진다).
- 변수를 무척도(scale-free)로 만들기 위해 표준화(standardization)가 필요하다.

표준화

- 확률분포(probability distribution)에서 이미 원점수(raw score) X를 Z 점수(Z score)로 표준화하는 법을 배웠다.

$$Z = \frac{X - \mu}{\sigma}$$

- 어떤 변수를 Z 점수로 표준화하면 자동적으로 그 변수의 평균은 0, 표준편차는 1이 된다(Why?).
- (다른 표준화 기법도 많이 있지만) Z 점수는 가장 널리 쓰이는 표준화 기법 중 하나로 종종 정규화(normalization)라고도 한다.
- SPSS에서는 [분석]-[기술통계량]-[기술통계]에서 '표준화 값을 변수로 저장'을 활성화한다.
- Jamovi에서는 [데이터]-[계산]에서 '$Z(\cdot)$' 함수를 활용한다.

표준화

연습 2. INCOME.SAV 자료를 불러와 월평균 가구소득 INCOM0의 Z 점수를 계산하시오. 이 표준화 점수의 평균 및 표준편차를 원점수의 평균 및 표준편차와 비교하시오.

집계화

집계화

자료는 가공되기 전 아무런 정보도 전달하지 못한다.

- 어떤 가공도 가해지기 전 (원석같은) 자료를 원자료(raw data)라고 부른다.
- 원자료는 많은 가능성을 내포하고 있지만 그 자체로는 구체적인 정보를 보여주지 않는다.
- 특정한 정보를 캐내려면 그에 맞는 방향으로 점차 가공을 해나가야하며 그 가공은 결국 집계화(aggregation)의 방식에 따라 이루어진다.

집계화

흔히 집계화는 선거 결과에 비유된다.

- 수십만 장의 투표용지 뭉치로는 선거 결과를 알 수 없고, 하나하나 개표하여 각 후보에 던져진 표의 수를 세어야(count) 한다.
- 최종 결과표에는 전체 투표수를 분모로, 후보별 득표수를 분자로 한 비율(proportion)이 명시되어야 한다(Why?).
- (선거 결과에서 그러하듯) 집계화는 변수를 하나의 숫자로 요약하는 과정이다.
- 어떤 경우에 집계화는 요약통계(=기술통계)를 계산하는 과정에서 자연스럽게 이루어진다.
- 기술통계는 반드시 결과표를 제시하려는 목적으로만 수행되는 반면, 우리는 집계화를 별도로 수행하여 인위적으로 집계자료(aggregate data)를 생산할 수 있다.

집계화

우리나라 통계청에서 제공하는 통계 자료는 거의 대부분 집계자료이다.

- 가령 [의료기관 시군구(대전,세종,충북,충남)별 진료실적 현황-전체]를 보자.
- 현재 이 자료의 관찰단위(unit of observations)는 무엇인가?
- 화면 상단 [주석정보]를 클릭하여 집계되기 이전 자료의 관찰단위는 무엇인가?
- 집계화가 높게 올라갈수록 프라이버시 침해(infringement of privacy) 위험이 낮아진다.
- 가령 개별 환자 자료를 인터넷에 올릴 수는 없지만 집계자료 뿐이라면 프라이버시 걱정없이 필요한 정보만을 제공할 수 있다.
- 하지만 한 번 집계화된 자료는 집계화되기 이전의 정보를 상실하므로, 집계화된 목적 자체와 상이한 목적의 분석에는 더이상 사용되기 어렵다.
- 따라서 "집계자료 말고 원자료를 달라"는 연구자와 이를 거부하는 자료생성기관 사이의 씨름이 종종 벌어진다.

집계화

연습 3. INCOME.SAV 자료에서 월 가구소득(INCOM0)의 평균 및 표준편차를 최종학력별로 집계하시오.

집계화

집계화를 수행할 때는 분석의 목적을 미리 고려해야 한다.

- 집계화는 SPSS에서 [데이터]-[데이터 통합]으로 쉽게 할 수 있다(데이터 통합은 aggregation의 이상한 번역이다).
- Jamovi에서는 VMEAN() 함수를 사용하여 평균을 구할 수는 있다(다만 집계를 직접 수행하기는 불편하므로 SPSS를 이용하자).
- 진짜 수행하려고 하는 분석의 목적을 고려하여 분석단위(unit of analysis)를 결정해야 하므로 그에 걸맞는 수준으로 집계화를 수행한다.
- 분석을 개인 단위로 수행할 예정이라면(즉 분석단위가 개인이라면) 집계화할 필요가 없다(Why?).
- 그러나 지역별 혹은 국가별로 분석을 수행하고 싶다면 그에 걸맞게 원자료를 집계화해야 한다.

집계화

연습 4. GOD.SAV 자료를 불러와 신에 대한 믿음을 다음과 같은 단위로 집계하시오: (1) 국가별(B_COUNTRY), (2) 국제지역별(regionWB), 마지막으로 (3) 전세계. 각각의 집계자료를 다른 이름으로 저장하시오(미리 재부호화 및 결측치 처리를 할 것!).