계량분석
Mean Comparisons in Practice

김현우, PhD[1]

[1] 충북대학교 사회학과 조교수

October 10, 2024

진행 순서

1. 평균비교를 위한 표/시각화
2. 연습문제

평균비교를 위한 표/시각화

평균비교를 위한 표/시각화

데이터 분석에 앞서 꼼꼼한 준비가 필요하다.

- 평균비교를 할 때 사실 하나는 양적변수, 다른 하나는 질적변수이다(Why?).
- 가설을 세울 때는 먼저 사회학적 상상력이 필요하다. 이 순간에는 컴퓨터 화면에서 시선을 떼고 곰곰히 생각에 생각을 거듭해야 한다. 독립변수와 종속변수는 무엇인가?
- 결측치 처리에 주의해야 한다. 신중하게 기술통계(descriptive statistics)도 살펴보고 시각화도 해보자.
- 분석기법은 무엇이 적절한가? 어떤 자료 구조인가? 가정은 위배되지 않았을까?

평균비교를 위한 표/시각화

유의성 검정에서 α 값을 미리 정해놓는 경우는 그다지 없다.

- 실무나 연구 상황에서는 먼저 t 값을 써놓고, 유의성 검정을 수행한 다음에 독특한 표식을 남긴다.
- 그 대신 신뢰수준이 99.9%일 때는 별 3개(***), 99%일 때는 별 2개(**), 95%일 때는 별 1개(*)를 통계량 뒤에 덧붙여 표기한다.
- 좀 더 구차해지고 싶을 때는 90% 신뢰수준에 대해 대거 1개(†)를 붙인다.
- 주의할 것은 이것이 관례에 지나지 않는다는 점이다!

평균비교를 위한 표/시각화

- SPSS에서 t 검정을 수행하고 결과를 요약하는 표 하나를 만들 수 있다.
- 먼저 격차 $\bar{X}_1 - \bar{X}_2$를 나타낼 수 있다. 그리고 그 격차가 표본을 넘어 모집단 수준에서 일반화될 수 있는가를 판정하기 위해 t 검정을 수행한다(Why?).
- t 값을 쓰고 그 뒤에 유의확률(=Sig.)을 표시할 수도 있지만, 읽는 사람 입장에서 조금 귀찮기 때문에 별이나 대거로 요약할 수 있다.

	1학년		2·3·4학년		격차	t 값
	평균	표준편차	평균	표준편차		
학점	2.4	0.7	2.7	0.7	0.2	3.4***
숙제제출 비율	84.8	21.3	88.8	18.6	4.0	2.3**
결석 수	6.1	5.3	5.8	5.5	-0.3	-0.5

평균비교를 위한 표/시각화

평균비교는 기본적인 그룹 차이를 드러낼 때 보편적으로 활용된다.

- 기술통계 수준으로 여러 변수들이 두 개의 사회적 집단 사이에 어떻게 다른지 간단히 보여줄 때도 매우 유용하다(Why?).

<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

- 반드시 표 밑의 각주에 별의 갯수에 대한 설명을 달아야 한다!

평균비교를 위한 표/시각화

- 가령 미국의 공화당을 지지하는 주(적색)와 민주당을 지지하는 주(청색) 사이에 살인과 자살 등 폭력이 어떻게 다른지 살펴본 연구를 참고하자.

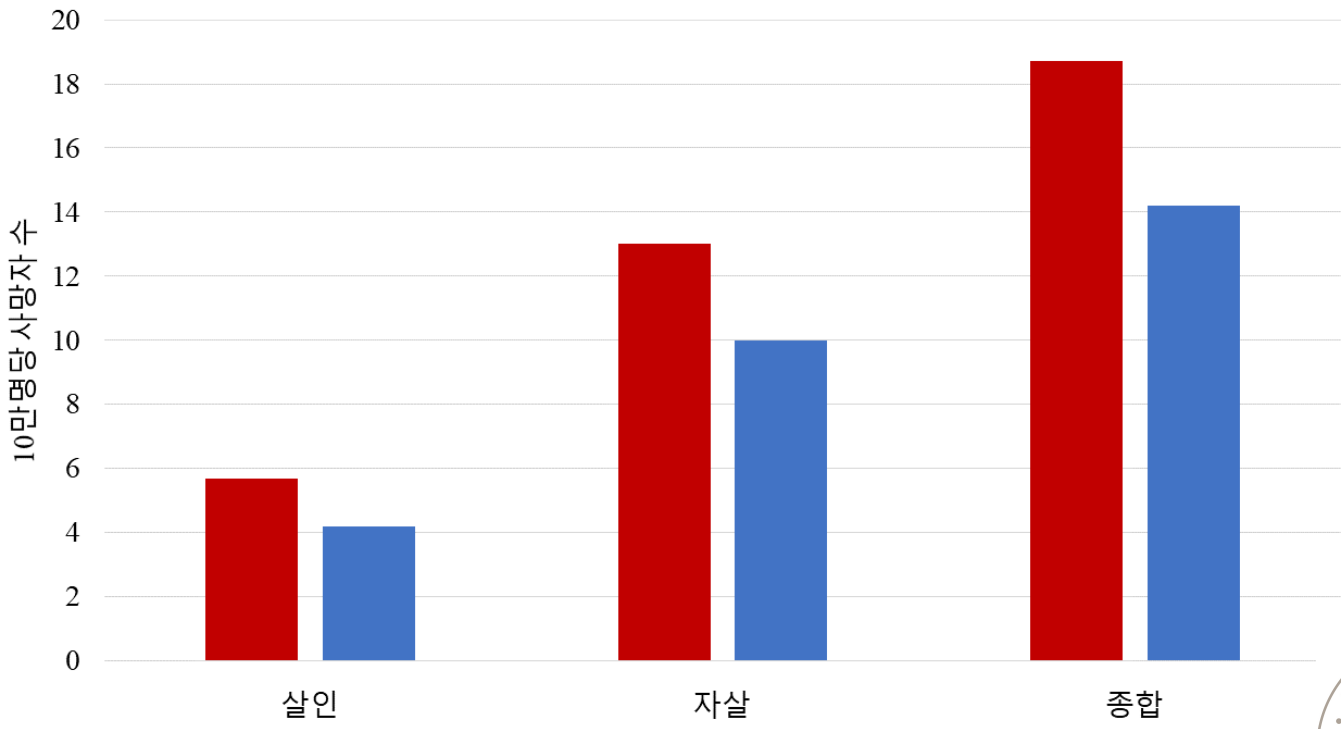
2000 사망자 (10만 명당)	적색 주(30개)		청색 주(20개)		통계적 유의수준	
	평균	표준편차	평균	표준편차	t	p
살인	5.7	2.85	4.2	2.43	1.90	0.064
자살	13.0	2.89	10.0	2.95	3.57	0.001*
종합	18.7	3.80	14.2	4.02	4.01	0.000*
2004 사망자 (10만 명당)	적색 주(31개)		청색 주(19개)		통계적 유의수준	
	평균	표준편차	평균	표준편차	t	p
살인	5.7	2.67	4.0	2.15	2.38	0.021*
자살	13.9	3.19	10.2	2.70	4.28	0.000*
종합	19.6	4.04	14.2	2.90	5.16	0.000*

(〈왜 어떤 정치인은 다른 정치인보다 해로운가?〉 239페이지)

평균비교를 위한 표/시각화

시각화도 매우 유용할 수 있다.

- 특히 평균비교를 문항별로 여러번 반복한다면 히스토그램이나 상자-수염 그림을 통해 보여주는 편이 복잡한 표보다 전달력이 높다!



평균비교를 위한 표/시각화

Stata에서 평균비교를 위한 표를 만들어보자.

- ttest 결과물을 복사하여 붙여넣기 할 수도 있다. 반복문(foreach)을 사용하면 좀 더 편리하게 할 수 있다.
- 기술통계표를 만들때는 아무래도 dtable이 편리하다.
- 최종적으로는 엑셀로 옮겨 붙여 어느 정도는 편집해야 한다.

평균비교를 위한 표/시각화

경험적 연구에서 t 검정은 크게 두 부분에서 주로 활용된다.

- 첫번째는 아까 설명한 표본에 관한 기술통계(descriptive statistics)를 제시하는 부분이고, 두번째는 회귀분석(regression analysis)에서 계수(coefficients)의 유의성 검정(significance test) 부분이다.
- 먼저 연구자는 자신의 표본 안의 "핵심이 되는 이분형(dichotomous) 관심변수 내지 종속변수에 따라" 다른 여러 변수들이 어떻게 달라지는지 t 검정을 통해 살펴볼 수 있다.
- 기술통계가 훨씬 풍성해지고 볼만한 내용을 담게 된다.

평균비교를 위한 표/시각화

- 기술통계와 밀접하게 잇닿아있는 주제인데, t 검정은 관찰자료를 사용한 인과추론(causal inference with observational data)과도 잇닿아 있다.
- 이른바 성향점수(propensity scores)를 활용한 인과분석에서 통제집단(control group)과 처치집단(treatment group) 사이에 사전적으로 밸런스(balance)가 맞는지 살펴볼 수 있다.
- 그리고 매칭(matching), 가중치 부여(weighting), 또는 계층화(stratification) 이후에 "관찰가능한 변수들에 대한 조건부(conditional on observables)"로 밸런스가 맞추어졌는지를 사후적으로 확인하기 위해서도 쓰인다.
- 다만 이것은 이 수업의 수준을 한참 벗어난 것이므로 더이상 다루지 않는다.

평균비교를 위한 표/시각화

Table 1. Covariate Imbalance Prior to Matching

Covariate	Description of original covariate	As used for estimating the propensity score	Differences in covariate means prior to matching: Two-sample t statistic	Differences in covariate means prior to matching: Standardized difference in %*
Child characteristics				
Sex	Female/male	0, 1	−1.02	−7
Twin	Single/multiple birth	0, 1	−1.28	−10
Sibpos	Oldest child (no, yes)	0, 1	−2.33	−16
C-age	Age at start of study	Months	.46	3
Mother characteristics				
SES	Socioeconomic status (9 ordered categories)	Integers 1–9	3.66	26
Education	Mother's education (4 ordered categories)	Integers 1–4	2.09	15
Single	Unmarried (no, yes)	0, 1	−5.70	−43
M-age	Age (years)	Years	8.99	59
Height	Mother's height (5 ordered categories)	Integers 1–5	2.55	18
Characteristics of the pregnancy				
WGTHGT3	(Weight gain)/height^3 (30 values based on category midpoints)	30 values	−.00	−0
PBC415	Pregnancy complications (an index)	Index value and its square	2.61	17
PRECLAM	Preeclamsia (no, yes)	0, 1	1.82	9
RESPILL	Respiratory illness (no, yes)	0, 1	1.73	10
LENGEST	Length of gestation (10 ordered categories)	$(10 - i)^{1/2}$ and i for $i = 1, 2, \ldots, 10$	.72	6
Cigarette	Cigarette consumption, last trimester (0 = none, plus 4 ordered categories)	Integers 0–4 and their squares	−.48	−3
Other Drugs				
Antihistamine	No. of exposures to antihistamines (0–6)	Integers 0–6 and their squares	1.76	10
Hormone	No. of exposures to hormones (0–6)	Integers 0–6 and their squares	8.41	28
HRMG1	Exposed to hormone type 1 (no, yes)	0, 1	2.67	15
HRMG2	Exposed to hormone type 2 (no, yes)	0, 1	3.75	19
HRMG3	Exposed to hormone type 3 (no, yes)	0, 1	3.46	18

*The standardized difference in percent is the mean difference as a percentage of the average standard deviation: $100(\bar{x}_1 - \bar{x}_{0R})/[(s_1^2 + s_{0R}^2)/2]^{1/2}$, where for each covariate, $\bar{x}_1$ and $\bar{x}_{0R}$ are the sample means in the treated group and the control reservoir and s_1^2 and s_{0R}^2 are the corresponding sample variances.

평균비교를 위한 표/시각화

Stata에서 평균비교를 위한 시각화를 연습해보자.

- graph bar 또는 graph box를 사용할 수 있다(Why?).
- 원하는 형태로 자료의 꼴을 변경해야 할 수도 있다. 바뀐 뒤의 꼴을 상상할 수 있어야 하는데 경험이 부족하면 쉽지 않다.
- 경우에 따라서는 엑셀을 사용하는 편이 오히려 더 편리할 수도 있다.

평균비교를 위한 표/시각화

- t 검정은 회귀분석 계수의 유의성 검정에도 사용된다.
- 이 맥락에서 귀무가설은 "k 번째 회귀계수가 0이다($H_0 : b_k = 0$)"로, 다시 말해 해당 독립변수는 종속변수를 설명하는데 의미가 없다는 뜻이다.
- 물론 연구자는 이 귀무가설을 기각하고 싶기 마련이다($H_a : b_k \neq 0$)!
- 이에 관해 더 자세한 내용은 몇 주 뒤에 다루게 된다.

연습문제

연습문제

연습 1. MEDCOND.SAV에서 각종 제한사항의 비율을 성별에 따라 비교하는 표를 작성하고 유의성 검정 결과를 보고하시오. 이때 표에는 (노령을 제외한) 모든 제한사항의 총합계 역시 성별로 비교하는 내용을 추가하시오.

연습문제

연습 2. 2021년 언론수용자 조사 자료를 활용하여 언론 역할별 중요성(문79)과 언론 역할별 수행 정도(문80)를 각각 합산하여 합성지수(composite index)를 만드시오. 두 변수의 차이로 계산하여 기대 미충족 변수를 만드시오. 이제 세 변수가 최종학력 대졸 여부에 따라 어떻게 다른지 비교하는 표와 그래프를 작성하고 유의성 검정 결과를 보고하시오.