사회통계연습
평균비교의 실제 활용

김현우, PhD[1]

[1] 충북대학교 사회학과 부교수

진행 순서

1. 등분산 가정
2. 평균비교를 위한 표와 시각화

등분산 가정

등분산 가정

t 검정의 결과를 해석할 때 그 가정에 대해 살짝 주의해야 한다.

- 두 모집단의 분산이 같다는 가정을 등분산(homogeneity of variance)이라고 부른다.
- 결과표에는 이 가정의 성립 여부에 따라 다른 결과를 제시한다.
- 만일 실험 전후로 같은 사람을 짝지워 평균을 비교하는 경우라면 이 가정은 (상대적으로) 크게 문제되지 않을 수 있다. 가령 같은 초등학생을 1년 전후로 추적하여(before vs after) 사회적 자아를 조사한다면 이 가정의 타당성이 별로 의심스럽지는 않다(Why?).
- 하지만 (쌍체표본 t 검정과는 달리) 독립표본 t 검정을 사용한다면 반드시 같은 사람끼리 짝지어 평균을 비교하지 않을 수도 있다!
- 다른 사람이라면 이 가정의 타당성이 특히 의문스럽고 이 가정은 아예 완화되어야 할 필요가 있다.

등분산 가정

연습 1. INCOME.SAV에서 가구소득과 대학교 재학 이상의 최종학력에 관한 가설을 적절하게 세우고 이를 95% 신뢰수준에서 검정하시오. 그 결과에 대해서도 간단히 논하시오.

등분산 가정

- 먼저 EDUC을 가변수(dummy variable)로 바꾸어 고졸 이하를 0, 대재 이상을 1로 재부호화(recoding)한다(Why?).
- 이제 새로운 변수를 비교 대상으로 삼아 독립표본 t 검정을 수행한다(Why?).
- SPSS에서는 Levene's Test for Equality of Variances 결과를 제공하고 있다. 이것의 귀무가설은 "두 집단은 등분산이다"라는 것이다. 귀무가설을 기각하면 아랫쪽 표를 해석하면 된다(Why?).
- Jamovi에서는 [동질성 검증]을 체크하고, 필요에 따라 [Welch's] 결과를 확인한다.
- 다행히 표본 크기가 커지면 이 가정의 위배 문제는 사실 크게 걱정하지 않아도 된다.

등분산 가정

연습 2. comedy.sav는 KBS에서 2023년 조사한 〈코미디 프로그램에 대한 시청자 인식조사〉 자료이다. (개승자에서 빠더너스까지) 최근 1년 동안 시청한 코미디 프로그램의 평균 갯수가 (1) 남녀에 따라, (2) 대졸 여부에 따라, (3) 2030 또는 그 윗세대 여부에 따라 차이가 있는지 여부를 5% 유의수준에서 검정하시오.

평균비교를 위한 표와 시각화

평균비교를 위한 표와 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 기술통계도 살펴보고 시각화도 해보자.
- 분석기법은 무엇이 적절한가? 어떤 자료 구조인가? 가정은 위배되지 않았을까?

평균비교를 위한 표와 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 표와 시각화

연습 3. 새우는 “1학년(frosh) 때는 보통 정신 못차리고 놀다가 성적(termgpa)을 망치기 쉽다”고 늘 믿어왔다. ATTEND.SAV를 활용하여 학점(termgpa), 숙제제출 비율(hwrte), 결석 수(skipped)가 1학년인지 2, 3, 4학년인지 여부에 따라 정말 상이한지 검정하시오.

평균비교를 위한 표와 시각화

- SPSS나 Jamovi에서 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(p-value)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 표와 시각화

평균비교는 기본적인 그룹 차이를 드러낼 때 보편적으로 활용된다.

- 기술통계(descriptive statistics) 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 표와 시각화

- 가령 미국의 공화당을 지지하는 주(적색)와 민주당을 지지하는 주(청색) 사이에 살인과 자살 등 폭력이 어떻게 다른지 살펴본 연구를 참고하자(별 붙이는 갯수가 관례와 다름에 주의!).

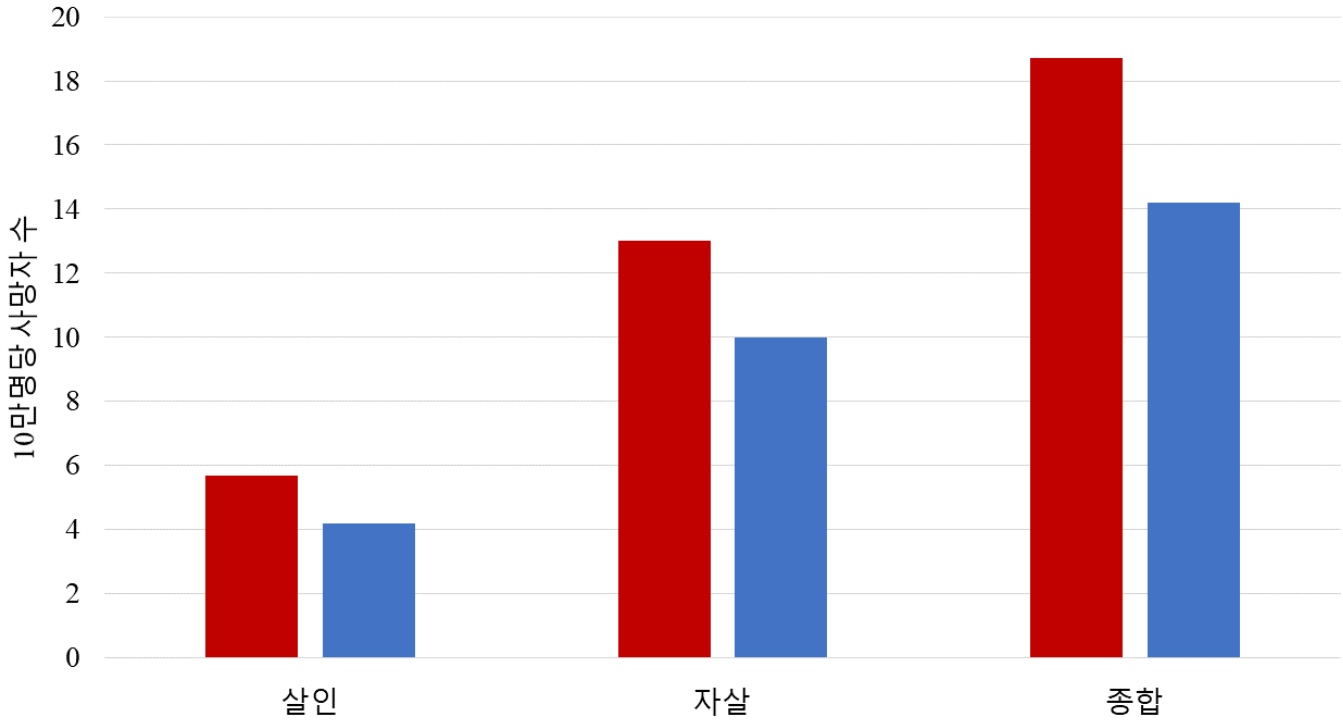
2000 사망자 (10만 명당)	적색 주(30개)		청색 주(20개)		통계적 유의수준	
	평균	표준편차	평균	표준편차	t	p
살인	5.7	2.85	4.2	2.43	1.90	0.064
자살	13.0	2.89	10.0	2.95	3.57	0.001*
종합	18.7	3.80	14.2	4.02	4.01	0.000*
2004 사망자 (10만 명당)	적색 주(31개)		청색 주(19개)		통계적 유의수준	
	평균	표준편차	평균	표준편차	t	p
살인	5.7	2.67	4.0	2.15	2.38	0.021*
자살	13.9	3.19	10.2	2.70	4.28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

(〈왜 어떤 정치인은 다른 정치인보다 해로운가?〉 239페이지)

평균비교를 위한 표와 시각화

시각화도 매우 유용할 수 있다.

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



평균비교를 위한 표와 시각화

연습 4. ATTEND.SAV에서 숙제제출 비율(hwrte) 및 출석 비율(atndrte)가 1학년인지 여부에 따라 상이한 정도를 시각화하시오.

평균비교를 위한 표와 시각화

- SPSS에서 히스토그램을 선택하거나 상자-수염 그림을 선택한다.
- 두번째 아이콘 [수평누적 막대도표]를 고르고 x축에는 frosh를 놓고, y축에 두 변수들을 모두 놓는다.
- 이때 y축에서 두번째로 놓을 때 변수는 반드시 (+) 부분에 놓아야 한다(Why?).
- 그래프를 어느 정도 꾸며야 한다.

평균비교를 위한 표와 시각화

연습 5. 새우는 청년층(30세 이하)와 비청년층(30세 초과) 있는 사람이 (그렇지 않은 경우보다) KBS 프로그램에 대한 만족도가 높을 것이라고 예측하고 있다. KBS.SAV에서 “KBS 프로그램이 기대에 부합하는 정도”에 관한 11개의 문항을 모두 합산하여 만족도를 계산할 수 있다. 자녀 유무와 만족도 사이의 연관성을 판단하기 위한 귀무가설과 대립가설을 세우고 95% 신뢰수준에서 이를 통계적으로 검정하시오.

평균비교를 위한 표와 시각화

- 귀무가설은 “자녀 유무에 따른 KBS 프로그램에 대한 만족도에 차이가 없다”.
- 자녀가 없는 집단은 0, 자녀가 있는 집단은 1로 재부호화하고, 만족도 문항들은 모두 합산하여 합성지수를 만든다.
- 독립표본 t 검정에서 두 집단 간에 등분산 가정은 위배되므로 이에 맞게 읽어야 한다.
- “99.9% 신뢰수준에서 통계적으로 유의하게 귀무가설을 기각하고 대립가설을 채택할 수 있다.”