3회차

가설검정

지난 수업에는 1개 집단의 평균을 기준점과 비교하거나 2개 집단의 평균을 비교하는 t-검정에 대해서 알아보았다.

이번 시간에는 3개 집단의 평균을 비교하는 ANOVA 검정과 범주형 변수를 비교하는 카이제곱 검정을 알아보자.

ANOVA(Analysis of Variance, 분산분석)

> **집단이 3개 이상일 때의 집단의 평균 차이 비교**
> 비교 대상은 평균의 차이이지만 비교하는 과정에서 분산이 쓰이므로 분산분석으로 표현

ANOVA의 원리 (일원분산분석, One-way ANOVA)

집단 내에서의 분산과 집단 간의 분산을 비교하는 방식.

집단 내에서의 분산보다 집단 간의 분산이 더 크다면, 집단 간 차이가 있다고 판단.

- 가설
  - 귀무가설(H_0) : 모든 집단의 평균이 같다
  - 대립가설(H_1) : 적어도 하나의 집단 평균은 다르다

ANOVA 검정은 대부분 F-분포를 기반으로 하는 F-검정을 사용한다. F-통계량을 기반으로 하는 F-검정을 토대로 ANOVA를 설명하겠다.

예시) 각 반의 수학점수

	A반	B반	C반
	61	54	49
	87	57	61
	79	75	69
	64	85	83
	90	62	75
반별 분산	173.7	170.3	170.8
반별 평균	76.2	66.6	67.4
그룹 내 분산	반별 분산의 평균		171.6
그룹 간 분산	반별 평균의 분산 * 5		141.87
F값	그룹 간 분산 / 그룹 내 분산		0.83
p-value			**0.46**

F값이 0.05를 훌쩍 뛰어넘는 수치를 보인다 → 귀무가설을 기각 못 함.

세 집단은 평균의 차이를 보이지 않는다는 결론이 난다.

위 예시에서 알 수 있듯이 ANOVA 검정에서는 그룹 간 데이터가 벌어진 정도가 그룹 내에서의 데이터가 벌어진 정도보다 커야 그룹 간에서 차이가 있다고 판단한다.

ANOVA 종류

한편 t 검정과 비슷하게 ANOVA에도 다양한 종류가 있다. 앞서 배웠던 t-검정과 같이 케이스를 나눠서 비교해보자.

	2개 이하 비교 (t-검정)	3개 이상 비교 (F-검정)
집단 평균 vs 기준값	단일표본 t-검정	
집단 간 비교 **(A vs B / A vs B vs C)**	이표본 t-검정 (독립표본 t-검정)	일원 ANOVA
집단 간 비교 **(A vs A' vs B vs B')**		이원 ANOVA
동일 집단의 전후 비교	대응표보 t-검정	반복측정 ANOVA

참고로 우리가 앞서봤던 예시는 일원ANOVA(one-way ANOVA)이다.

이제 이원 ANOVA (two-way ANOVA)에 대해서 살펴보자.

이원 ANOVA(two-way ANOVA)

눈치가 빠른 사람이라면 위 표에서 대충 이원 ANOVA는 그룹을 나누는 기준이 2가지라는 걸 짐작할 수 있다.

이원 ANOVA의 경우, 그룹을 2가지 기준으로 나눈 뒤 각 기준으로 인해 차이가 발생하는지, 두 기준의 상호작용효과로 차이가 발생하는지를 확인할 수 있다. 아래 예시를 통해 확인해보자.

예시2) A반 남,여 / B반 남,여 수학성적비교

	A반(남)	A반(여)	B반(남)	B반(여)
	61	87	85	57
	79	91	87	85
	83	83	86	35
	64	80	93	90
	62	75	82	62
반별 평균	69.8	83.2	86.6	65.8

위 예시에서 반,성별 2가지 기준으로 그룹을 나누었다. 이원 ANOVA 검정을 이용하면 우리는 3가지 질문에 대답할 수 있다.

1. 반에 따라 성적의 차이가 나타나는가
2. 성별에 따라 성적의 차이가 나타나는가
3. 반,성별 상호작용의 효과로 차이가 나타나는가

이원 ANOVA 검정에서는 두 기준의 상호작용 효과에 따른 차이를 같이 살펴본다는 점이 일원 ANOVA와 차이가 나타나는 점이라고 할 수 있다.

우리는 차이의 여부를 앞선 예시와 같이 F값과 p-value로 판단할 수 있다. 아래 결과를 보고 해석해보자.
이번 예시에서는 F값이 세개나 나오므로 계산과정은 생략하겠다.

	sum_sq	df	F	p-value
반	0.45	1.0	0.002724	0.959025
성별	68.45	1.0	0.414284	0.528922
반:성별	1462.05	1.0	8.848842	0.008939
Residual	2643.60	16.0	NaN	NaN

위 결과에서 맨 오른쪽 컬럼(p-value)을 보면, 반:성별의 상호작용에서만 p-value가 0.05보다 작은 수치를 보인다. 즉, 반, 성별에 각각에 의한 차이는 나타나지 않았지만 반:성별의 상호작용에 의한 평균 차이는 발생한다고 결론을 내릴 수 있다.

반복측정 ANOVA(RM ANOVA, Repeated Measures ANOVA)

이번에는 대응표본 t-검정과 비슷하게 한 집단을 대상으로 여러 번 데이터를 구해서 차이를 비교하는 검정을 방법을 알아보겠다.

예시3) 식단별 몸무게 감소(kg) 비교

참가자	식단 A	식단 B	식단 C
은중	1.2	2.5	3.1
상연	0.9	2.2	2.8
상학	1	2.4	3.3
재준	1.4	2.1	2.9
세리	1.1	2.6	3.2

이번 예시에서는 각 행이 한 사람에서 나오는 데이터라는 차이를 보인다. 첫 행을 두고 얘기하면 은중이가 식단 A,B,C를 했을 때 감소한 무게를 각각 기록한 것임을 알 수 있다. 이 때 우리는 반복측정 ANOVA 검정을 수행하게 되는데 앞선 예시처럼 결과치를 바로 보겠다.

Anova

	F Value	Num DF	Den DF	Pr > F
diet	140.5631	2.0000	8.0000	0.0000

결과를 보면 맨 오른쪽 Pr > F (= p-value) 를 보면 0에 근사한 값을 갖는다. 따라서 식단별로 몸무게 감소에 차이가 있었음을 알 수 있다.

ANOVA 검정의 전제조건

t-검정과 같이 ANOVA 검정에도 전제조건이 있다.

정규성, 등분산성의 경우 t-test일 때와 동일한 방식으로 검정하며, 독립성은 데이터를 수집하는 단계에서 확보되어야 하는 조건이다. 구형성은 반복측정 ANOVA일 때만 검정하는 것으로 자세한 내용은 표 아래의 설명을 참고하자.

전제조건	설명	검정방법	위반시 대체방법
정규성	각 그룹 데이터가 정규분포를 따름	Shapiro-Wilk test Kolmogorov-Smirnov test Q-Q plot	Kruskal-Wallis test
등분산성	각 그룹의 데이터가 동일한 분산값을 가짐	**Levene's test** Bartlett's test	Welch ANOVA
독립성	각 데이터는 서로 독립적 (단, 반복측정 ANOVA에서 한 참가자의 데이터는 서로 독립적이지 않음. 대신 참가자 간은 독립적이어야 함.)	-	-
구형성	각 조건들 간 차이의 분산이 같음. **반복측정 ANOVA만 해당**	Mauchly's test	Greenhouse-Geisser 보정 Huynh-Feldt 보정

구형성(Sphericity)& 보정방법.

구형성이란 **모든 조건 간 차이의 분산이 같아야 한다**는 것을 의미한다.

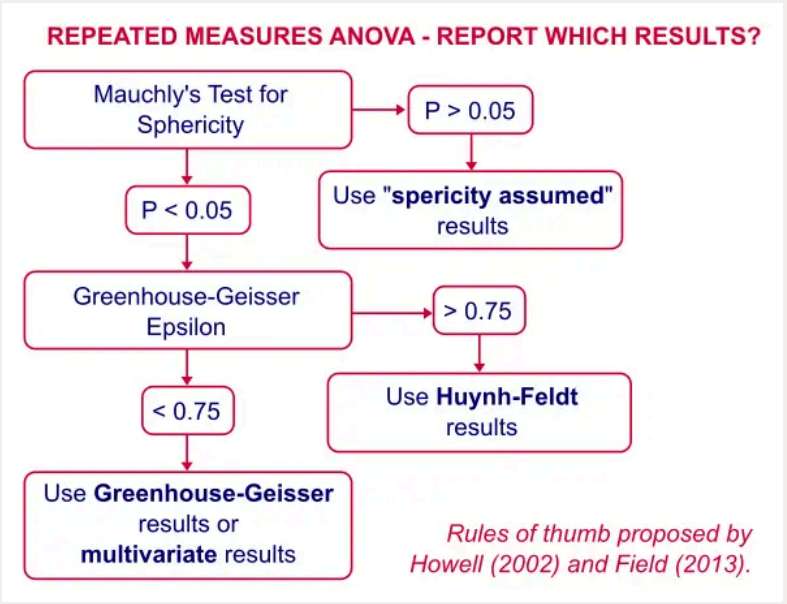
아래 예시를 보면 time2-time3 분산이 다른 조합에 비해 유난히 작으므로 구형성 조건이 깨졌다.

Subject	Time 1	Time 2	Time 3	Time 1 – Time 2	Time 1 – Time 3	Time 2 – Time 3
1	45	50	55	-5	-10	-5
2	42	42	45	0	-3	-3
3	36	41	43	-5	-7	-2
4	39	35	40	4	-1	-5
5	51	55	59	-4	-8	-4
6	44	49	56	-5	-12	-7
			Variance:	**13.9**	**17.4**	**3.1**

출처: Laerd Stastitcs 구형성 조건이 깨진 사례 (**3.1**)

구형성을 만족하지 않을 경우, F값이 실제보다 크게 계산되는 문제를 야기한다. 이러한 문제를 해결하기 위해 구형성 위반 시, 자유도를 보정해서 F값을 조정하는 방식을 사용한다.

Greenhouse-Geisser 보정, Huynh-Feldt 보정 방법의 경우 아래의 플로우차트처럼 Greenhouse-Geisser 보정에서 값이 0.75 이하일 때까지 사용하고 이상일 때에는 Huynh-Fledt를 사용한다.



출처: SPSS tutorial https://www.spss-tutorials.com/spss-repeated-measures-anova-example-2/

사후검정

> ANOVA 검정에서는 차이의 유무만 확인할 수 있다. 여러 개의 집단 중 어느 집단끼리 차이가 발생했는지 알아내기 위해서는 사후 검정의 절차가 필요하다.

- F검정 결과 : A,B,C 그룹 간 차이가 있다.
- 사후검정 : 아래 각각의 쌍에서 차이를 보이는가?
  - A vs B
  - A vs C
  - B vs C

→ 각각의 조합에 대해서 t-검정이나 유사한 방법으로 비교

사후검정 방법

방법	설명	전제 조건	적합한 상황
Tukey HSD	**가장 대표적인 방법** 모든 그룹 쌍 간 평균 비교	정규성, 등분산성, 각 표본의 크기가 유사해야 함	그룹 수 많고 균형 잡힌 데이터
Bonferroni (사후용)	쌍별 비교 후 p값 보정	정규성, 등분산성	비교 수 적고 오류 제어 우선
Scheffé	모든 선형 조합 비교	정규성, 등분산성	각 표본의 크기가 다른 경우 복잡한 비교
Dunnett	기준 집단 vs 나머지 그룹 비교 (모든 쌍에 대한 비교는 불가)	정규성, 등분산성	기준집단이 있는 경우
Games-Howell	등분산성 불필요, 표본 크기가 달라도 됨	정규성	Welch ANOVA와 같이 사용가능 분산과 표본 수가 달라도 안정적, 실제 실험 데이터에 적합

다중검정

> 여러 개의 가설을 동시에 검정(test) 하는 것

다중검정의 고질적인 문제

여러 개의 동시에 검정할 때, 제 1종 오류(잘못 기각할 확률) 가 누적되어 전체 오류율이 커지는 현상이 발생할 수 있습니다. **제1종 오류 : 귀무가설이 참인데 기각*

예시

- 가설 검정이 1개일 경우: 전체오류율$= \alpha = 0.05$
- 가설 검정이 2개일 경우: 전체오류율$= 1-(1-\alpha)^2 = 1-(0.95)^2 \approx 0.0975$
- 가설 검정이 10개일 경우: 전체오류율$= 1-(1-\alpha)^{10} = 1-(0.95)^{10} \approx 0.401$

10개의 비교만 해도,약 40% 확률로 우연히라도 하나는 유의하다고 잘못 판단할 수 있는 상황이 생기게 된다.

이를 방지하기 위해, 다음과 같은 보정방법을 사용해서 전체 오류율을 줄인다.

방법	설명	특징	적합한 상황
Bonferroni	유의수준을 test 수로 나누는 방식 - α = 0.05이고, test 수가 10개라면, α/10 =0.05/10 = 0.005로 조정	단순하고 직관적, 보수적	비교 수가 적을 때
Holm	p값을 정렬한 후 순차적으로 보정 - α = 0.05이고, test 수가 10개라면, - 제일 작은 p-value ≤ α/10, 그 다음 작은 p-value ≤ α/9 …	Bonferroni보다 강력	일반적인 다중 비교
Sidak	가설이 서로 독립일 때 사용 $\alpha_{\text{보정}} = 1-(1-\alpha)^{1/m}$	독립 비교 시 사용	독립 가설 검정 시
Hochberg	Holm의 역방향 방법 - 제일 큰 p-value ≤ α/10, 그 다음 큰 p-value ≤ α/9 …	Holm보다 보수적	일부 상황에서 대체 사용

방법	설명	특징	적합한 상황
Benjamini-Hochberg (FDR)	거짓 발견 비율 제어 : 기각한 가설 중에서 잘못 기각된 비율을 줄임	대규모 검정에 적합	유전자 분석, 대량 비교

사실 이미 눈치 챈 사람도 있겠지만 사후 검정은 다중 검정의 한 케이스로 사후 검정은 ANOVA 검정에서 차이가 있을 때 시행하는 다중검정이라고 보면 된다.

카이제곱 검정

범주형 변수를 비교할 때 사용되는 검정방법으로 기대와 일치하는지 검정하는 적합도 검정과 두 범주형 변수의 독립성을 확인하는 독립성 검정이 있다.

카이제곱 적합도 검정 (Goodness-of-Fit Test)

> 한 집단의 여러 범주 분포가 기대와 일치하는지 검정할 때 사용

- 사용 조건
  - 하나의 범주형 변수에 3개 이상의 범주가 있고
  - 각 범주가 예상된 비율(기대값)과 다른지를 검정
- 예시
  - “고객이 A/B/C 브랜드를 고른 비율이 모두 1:1:1일까?”
  - A: 40명, B: 30명, C: 30명 → 기대값은 33.3명씩
- 귀무가설(H_0)
  - “관측된 분포는 기대 분포와 같다”

카이제곱 독립성 검정 (Test of Independence)

> 두 범주형 변수 간에 관련이 있는지(독립인지)를 검정할 때 사용

- 사용 조건
  - 교차표(Contingency Table)로 표현 가능한 두 범주형 변수
  - 행과 열 변수 간 관계(연관성)이 있는지를 봄
- 예시
  - “성별과 구매 여부가 관련이 있을까?”

성별	구매함	구매 안 함
남성	40	60
여성	55	45

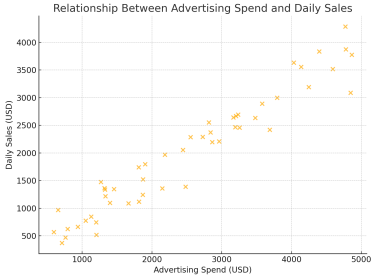
- 귀무가설(H_0)
  - “성별과 구매 여부는 서로 독립이다” (즉, 성별은 구매 여부에 영향을 주지 않는다)

상관관계(correlation)

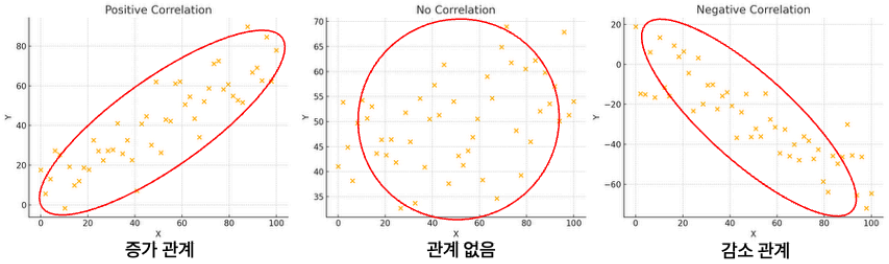
수치형 변수(numerical variable)일 때

산점도(scatter plot)

- 수치형 변수 2개는 x축과 y축으로 이루어진 산점도로 표현할 수 있다.



- 2개 변수의 관계성을 상관(correlation)이라고 하며, 2개의 확률변수 또는 데이터 사이의 관계성을 의미함



→ 상관이 있다고 해서 원인과 결과를 뜻하는 인과관계가 있는지까지는 알 수 없음!!

상관계수의 종류

피어슨 상관계수(Pearson Correlation Coefficient)

> 두 변수 사이 선형 관계의 정도와 방향을 수치로 표현하는 지표

두 변수를 scatter plot에 그려봤을 때 직선형태의 관계가 나왔을 때 사용하는 것이 적합 (비선형인 경우, 아래의 스피어만 상관 계수 혹은 켄달의 타우 상관계수를 사용)

▼ 수식

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- 분자 : 공분산(x, y)
- 분모 : x_i, y_i 각각의 표준편차

스피어만 순위상관계수 p (Spearman's rank correlation coefficient p)

> 두 변수의 순위 간 상관관계를 측정하는 지표 → 값 자체보다는 순위차이에 집중

순위형 데이터이거나 연속형 데이터가 비선형일 때 사용

▼ 수식

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)}$$

- d_i: 각 관측치의 순위 차이
- n: 데이터 개수

켄달의 타우

> 두 변수 간의 **순위 일치 정도**를 측정하는 지표 → 즉, 관측치 쌍 간의 **순서가 서로 일치하는지, 불일치하는지**를 비교하여 계산

순위형 데이터이거나 연속형 데이터가 비선형일 때 사용

▼ 수식

$$D$$

- C: **일치하는 순서 쌍(concordant pairs)** 개수
- D: **불일치하는 순서 쌍(discordant pairs)** 개수
- T: 변수 X 내에 묶인 순위(tied pairs) 개수
- U: 변수 Y 내에 묶인 순위 개수

상관계수 해석

- 세 지표 모두 -1 ≤ r ≤ 1 범위를 갖게 됨

상관 계수	상관 정도
0.7 < |r| ≤ 1	강한 상관
0.4 < |r| ≤ 0.7	중간 정도 상관
0.2 < |r| ≤ 0.4	약한 상관
0.0 < |r| ≤ 0.2	거의 상관 없음

범주형 변수일 때

Cramér’s V

> 범주형 변수 간의 연관성 정도를 측정하는 지표

범주형 변수의 교차표를 기반으로 (contingency table) 계산

▼ 수식

$$V = \sqrt{\frac{\chi^2}{n(k-1)}}$$

- χ^2 : 카이제곱 통계량
- n : 전체 표본 수
- k : 둘 중 더 작은 범주의 수

Cramér’s V 해석

Cramér’s V 값	연관성 강도 (일반적 가이드)
0.00 ~ 0.10	매우 약함
0.10 ~ 0.30	약함
0.30 ~ 0.50	중간
0.50 이상	강함

카이제곱 독립성 검정이 범주형 변수의 독립성 유무를 확인하는 거라면, Cramers'V 는 독립적이지 않은 두 범주형 변수의 상관관계 정도를 확인

다음 수업 안내

- 실전 연습 : 10/2(목) 2시
- 통계 4회차 : 10/13(월) 10시
    - 수업내용 : 회귀