One-way ANOVA

(일원분산분석)

숙명여자대학교 경영학부 오중산

일원분산분석 소개

- 일원분산분석 정의
  - ◆ 세 개 이상 집단간에 종속변수 모평균 차이 유무를 확인하는 통계분석방법
    - ▪ 보통 대조군(통제군)을 하나 두고, 실험군을 2개 이상 설정
    - ▪ 예) 프로모션 안한 집단(대조군), 프로모션A를 한 집단, 프로모션B를 한 집단 간의 매출평균비교
- 일원분산분석에서의 독립변수와 종속변수
  - ◆ 독립변수(혹은 요인)는 집단을 구분하는 변수로 범주형 척도로 측정됨
  - ◆ 종속변수는 비교 대상이 되는 변수로 실수형/정수형 척도로 측정됨

구분	*z*-검정	*t*-검정	ANOVA
확률변수	모집단에서 정규분포를 띠어야 함 (모를 경우 표본크기 30개 이상)	모집단에서 **정규분포**를 띠어야 함	
모표준편차	알아야 함	모름	몰라도 됨(무관함)
모분산 조건	해당사항 없음	등분산 혹은 이분산	**등분산 조건 만족**해야 함
표본크기	가급적 30개 이상	무관함(30개 미만도 가능)	**30개 이상**
비교대상 집단	2개		2개 이상(보통 **3개 이상**)

일원분산분석 소개

- 두 가지 가설
  - H_0: $\mu_1 = \mu_2 = \ldots = \mu_t$ (t: 집단 개수로 $t \geq 2$)
    - 집단 간의 표본평균 차이는 우연의 결과이며, 요인효과는 없음
    - 귀무가설이 참이면, 표본평균의 평균($\bar{\bar{X}}$)이 최적의 모평균 추정치
  - H_a: 적어도 한 집단의 모평균은 다른 집단들의 모평균과 같지 않다.
    - 집단 간의 표본평균 차이는 우연의 결과가 아니며, 요인효과가 있음
    - $t = 3$이고 H_a가 채택되었을 때 경우의 수
      - $\mu_{i(i = 1\sim3)}$가 모두 다른 경우1
      - 두 개의 모평균은 동일하고, 하나만 다른 경우 2~4
        - ✓ $\mu_1 = \mu_2$ & μ_3는 다름 / $\mu_1 = \mu_3$ & μ_2는 다름 / $\mu_2 = \mu_3$ & μ_1는 다름

일원분산분석 소개

- 분산분석을 위한 세 가지 전제조건
  - 독립성: 표본 간에 종속변수 측정은 서로 독립적이어야 함
    - 어떤 표본의 임의의 사례가 다른 표본의 임의의 사례에 대한 측정에 영향을 미쳐서는 안됨
  - 정규성: 모든 모집단에서 종속변수는 정규분포를 띠어야 함
    - 표본별로 크기를 최소 30개 이상으로 해야 함
  - 등분산: 모집단 간에 종속변수 모분산은 동일해야 함
    - $H_0: \sigma_1^2 = \sigma_2^2 = \ldots. = \sigma_t^2$ (t: 집단 개수로 $t \geq 2$)

ANOVA Table

- 두 가지 편차제곱의 합
  - 표본간 편차제곱(요인분산): SSTR(sum of squares of treatments)
    - 서로 다른 표본간(between treatments) 표본평균 차이(편차) 제곱의 합
    - SSTR이 클수록 표본간 이질성이 커져서 대립가설 채택 가능성이 커짐
  - 표본내 편차제곱(오차분산): SSE(sum of squares of error)
    - 동일한 표본 내(within treatments) 측정값 차이(편차) 제곱의 합
    - SSE가 작아질수록 표본내 동질성이 커져서 대립가설 채택 가능성이 커짐
  - N: 전체 측정치 개수, t: 집단 개수, n_j(j번째 집단의 표본크기)
    - n_j가 모두 같을 필요는 없지만 30개 이상이어야 함

ANOVA Table

- F-통계량의 의미
  - 분산 간의 비율은 F-분포를 띰
    - 분자(MSTR)가 커지고, 분모(MSE)가 작을수록 F-통계량이 커짐
      - 서로 다른 표본간에는 이질성이 커야 하고, 동일한 표본 안에서는 동질성이 커야 함
  - F-통계량의 바깥 쪽 넓이(p-value)가 유의수준(α) 보다 작으면 대립가설 채택
    - F-통계량이 커질수록 유유상종(類類相從)하게 되고, p-value는 작아짐
    - SST(총편차제곱) = SSTR + SSE 이므로 SSTR과 SSE는 zero-sum 관계

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F-Ratio
Treatments, TR	$SSTR = \sum_{j=1}^{t} n_j(\bar{x}_{\cdot j} - \bar{\bar{x}})^2$	$t - 1$	$MSTR = \frac{SSTR}{t-1}$	$F = \frac{MSTR}{MSE}$
Sampling Error, E	$SSE = \sum_{j=1}^{t}\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_{\cdot j})^2$	$N - t$	$MSE = \frac{SSE}{N-t}$	
Total, T	$SST = \sum_{j=1}^{t}\sum_{i=1}^{n_j}(x_{ij} - \bar{\bar{x}})^2$	$N - 1$		

일원분산분석 검정 절차

- 1단계: 가설수립
- 2단계: 집단간 데이터프레임 생성
- 3단계: 정규성 조건 확인: shapiro.test 함수 사용
- 4단계: 등분산성 조건 확인: car패키지에 있는 leveneTest 함수 사용
  - 등분산 조건을 만족하지 못하면 Welch Test를 해야 함
- 5단계: 일원분산분석 실시 및 가설검정: 내장함수인 aov 함수 사용
- 6단계(사후검정): 대립가설이 채택되면 Duncan Test 실시
  - agricolae 패키지에 있는 duncan.test 함수 사용

일원분산분석 실습1

- 1단계: 가설수립
  - 데이터프레임 만들기
    - 기존에 만들어서 전처리 과정을 거친 ttest 데이터프레임을 복사해서 anova1 데이터프레임 생성
  - priority에 따른 price 평균값 비교하기
  - 이상치 검토 후 제거하여 anova_new 데이터프레임 만들기
  - priority 측정값 중에서 Critical을 High로 통합하여 새로운 변수 prior 만들기
  - 두 가지 가설 수립
    - 독립변수: prior / 종속변수: price
    - H_0: $\mu_H = \mu_M = \mu_L = \mu_N$ (μ_t : 해당 집단의 price 모평균)
    - H_a: 적어도 한 집단의 price 모평균은 다른 집단과 다르다.

일원분산분석 실습1

- 2단계: 집단간 데이터프레임 생성하기
  - 새로 만든 prior 변수 측정값 네 개에 따라 네 개의 서브 데이터프레임 생성
- 3단계: 네 개의 서브 데이터프레임에 대해 종속변수 정규성 검토
  - histogram을 통한 시각적 검토와 shapiro.test 함수를 활용한 통계적 검토
  - 정규성 조건 만족을 위한 표본크기 검토
- 4단계: 등분산성 검토
  - car 패키지에 있는 leveneTest 함수 사용
  - 기본 명령문: leveneTest(DV~IV, data = df)
    - 주의! df는 서브 데이터프레임이 아니라, 통합 데이터프레임

일원분산분석 실습1

- 5단계: 일원분산분석 실시
  - 등분산조건 만족시에는 내장함수인 aov 함수 사용
  - 등분반조건 만족하지 못할 경우에는 내장함수인 oneway.test 함수 사용
    - 이분산 가정 t-검정과 마찬가지로 Welch’s ANOVA를 시행함
    - 기본 코드는 aov와 동일하며, var.equal = F가 default 상태
    - 대립가설을 엄격하게 검정함
  - 참고: NA가 있으면 자동적으로 이를 제외하고 실행함

일원분산분석 실습2

- 다음과 같은 one-way ANOVA를 실행하시오.
  - 데이터: pttest
  - IV: payment
  - DV: expense
  - 유의수준(α) = 0.01