One-way ANOVA

(일원분산분석)

숙명여자대학교 경영학부 오중산

일원분산분석 소개

일원분산분석 정의

세 개 이상 집단간에 종속변수 모평균 차이 유무를 확인하는 통계분석방법

DV

- 보통 대조군(통제군)을 하나 두고, 실험군을 2개 이상 설정
- 예) 프로모션 안한 집단(대조군), 프로모션A를 한 집단, 프로모션B를 한 집단 간의 매출평균비교
  ↳ 모평균

일원분산분석에서의 독립변수와 종속변수

독립변수(혹은 요인)는 집단을 구분하는 변수로 범주형 척도로 측정됨

종속변수는 비교 대상이 되는 변수로 실수형 /정수형 척도로 측정됨

CLT

구분	*z*-검정	*t*-검정	ANOVA
확률변수	모집단에서 정규분포를 띠어야 함 (모를 경우 표본크기 30개 이상)	모집단에서 **정규분포**를 띠어야 함	① hist : 시각적으로 확인 ② Shapiro test (엄격, 기준넘기 어려움)
모표준편차	알아야 함 : 비현실 ⇒ 거의 사용X	모름	몰라도 됨(무관함)
모분산 조건	해당사항 없음	등분산 혹은 이분산	**등분산 조건 만족**해야 함 < O : ANOVA, X : Welch ANOVA
표본크기	가급적 30개 이상	무관함(30개 미만도 가능)	**30개 이상**
비교대상 집단	2개		2개 이상(보통 **3개 이상**)

2개일 때는 t 검정 or ANOVA

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
    => 독립변수 차이 X
  - 귀무가설이 참이면, 표본평균의 평균()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 $H_0 : M_1 = M_2 = M_3$
    - $\mu_{(i = 1\sim3)}$가 모두 다른 경우 1
    - 두 개의 모평균은 동일하고, 하나만 다른 경우 2~4 => 사후 검정으로 순서 확인
      - 2 $\mu_1 = \mu_2$ & μ_3는 다름 / 3 $\mu_1 = \mu_3$ & μ_2는 다름 / 4 $\mu_2 = \mu_3$ & 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

① 표본간 편차제곱(요인분산): SSTR(sum of squares of treatments) : 서로 다른 표본 간의 차이

- 서로다른 표본간(between treatments) 표본평균 차이(편차) 제곱의 합
- SSTR이 클수록 표본간 이질성이 커져서 대립가설 채택 가능성이 커짐

② 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

SSTR 값 커진다
표본들 간의 차이 커진다
이질성이 커진다
⇓
대립가설 채택 가능성 커진다

- N: 전체측정치 개수, t: 집단 개수, n_j(j번째 집단의 표본크기) $j = 1, \cdots, t$ $(t=3)$
  - n_j가 모두 같을 필요는 없지만 30개 이상이어야 함

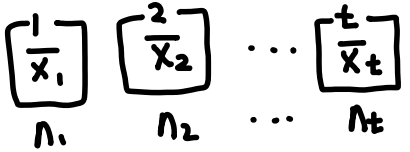
$n_1 = 57$, $n_2 = 48$, $n_3 = 63$

ANOVA Table

독립표본 t-검정 등분산

$$F = \frac{\hat{s}_1^2}{\hat{s}_2^2}$$

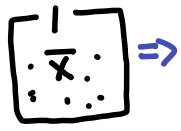
SSTR



$n_1(\bar{x} - \bar{\bar{x}})^2$

⇒ $\bar{x}$의 평균을 기준으로 $\bar{x}$들이 얼마나 퍼져있는지

SSE



⇒ $\bar{x}$와 케이스들이 얼마나 떨어져 있는지

- *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 ⇒ 대립가설 채택할 확률 높아짐
  - *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 ⇒ α보다 작거나 같아짐 ⇓ 대립가설 채택
    - SST(총편차제곱) = SSTR + SSE 이므로 (고정)

      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
  ↳ p-value > α => 등분산 조건 만족 => ANOVA
- 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
  analysis of variable
- 6단계(사후검정): 대립가설이 채택되면 Duncan Test 실시
  ↳ 사후검정
  - agricolae 패키지에 있는 duncan.test 함수사용

일원분산분석 실습1

1단계: 가설수립

데이터프레임 만들기

e 커머스 데이터

- 기존에 만들어서 전처리 과정을 거친 ttest 데이터프레임을 복사해서 anova1 데이터프레임 생성 `anova1 <- ttest`

`anova1 %>% group_by(priority) %>% summarise(mean(price, na.rm=T))`

- priority에 따른 price 평균값 비교하기

5개

- 이상치 검토 후 제거하여 anova_new 데이터프레임 만들기

* 이상치 제거 언제? 초기 or 서브데이터 맘대로

* 종속변수 이상치 제거

① `library(psych)` ② `descr <- describe(anova1$price)`

③ `descr <- descr %>% mutate(UL = mean + 2*sd)`, LL도 만들기

④ `table(anova1$price > descr$UL)`

⑤ `anova1_new <- anova1 %>% filter(price <= descr$UL)`

- priority 측정값 중에서 Critical을 High로 통합하여 새로운 변수 prior 만들기

`anova1_new <- anova1_new %>% mutate(prior = fct_collapse(priority, "High" = c("Critical", "High")))`

- 두 가지 가설 수립

t=4

- 독립변수: prior / 종속변수: price

: 세 집단 간에 price 모평균은 모두 동일하다

- 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.

* forcats 패키지

변수의 척도가 범주형으로 되어있을 때 씀

일원분산분석 실습1

```
anova1_H <- anova1_new %>% filter(prior == "High")
anova1_M <- anova1_new %>% filter(prior == "Medium")
anova1_L <- anova1_new %>% filter(prior == "Low")
anova1_N <- anova1_new %>% filter(prior == "Not Specified")
```

- 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
  - ① summary 함수로 min, max 확인
  - histogram을 통한 시각적 검토와 shapiro.test 함수를 활용한 통계적 검토
  - 정규성 조건 만족을 위한 표본크기 검토

min, max 확인

```
① summary(anova1_H $ price)
② hist(anova1_H $ price, breaks = seq(0, 600, 10))
```

↳ 정규분포인지 확인

```
shapiro.test(anova1_H $ price)
```

: 정규성 만족 X but 샘플사이즈 많음

- 4단계: 등분산성 검토

t-test에서는 var.test 함수 씀

- car 패키지에 있는 leveneTest 함수 사용

```
leveneTest(price ~ prior, data = anova1_new)
```

$\sigma^2_H = \sigma^2_M = \sigma^2_L = \sigma^2_N$

```
library(car)
```

- 기본 명령문: leveneTest(DV~IV, data = df)

종속변수 독립변수

  - 주의! df는 서브 데이터프레임이 아니라, 통합 데이터프레임

```
> leveneTest(price~prior, data = anova1_new)
Levene's Test for Homogeneity of Variance (center
 = median)
       Df F value Pr(>F)
group   3  1.8795 0.1307
      8288
```

↳ 유의수준보다 큼 ⇒ 등분산성 만족

일원분산분석 실습1

```
> summary(anova1_result)
             Df   Sum Sq Mean Sq F value
prior         3    55875   18625   1.863
Residuals  8288 82839853    9995
            Pr(>F)
prior        0.133
Residuals
```

↳ α보다 큼 ⇒ 귀무가설 채택

⇒ 네 집단간 price 모평균 같다

5단계: 일원분산분석 실시

① anova1_result <- aov(price ~ prior, data = anova1_new)

② summary(anova1_result)

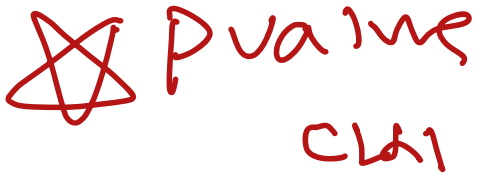
- 등분산조건 만족시에는 내장함수인 aov 함수 사용
- 등분반조건 만족하지 못할 경우에는 내장함수인 oneway.test 함수 사용
  - 이분산 가정 t-검정과 마찬가지로 Welch's ANOVA를 시행함
  - 기본 코드는 aov와 동일하며, var.equal = F가 default 상태
    ↳ 이분산 가정 ⇒ Oneway.test 함수에서
  - 대립가설을 엄격하게 검정함

II종 오류 : H_a 참인데 기각 . 발생할 확률 B 커짐 , 검정력 1-B 작아짐

(대립가설 참 → 채택할 확률)

- 참고: NA가 있으면 자동적으로 이를 제외하고 실행함

anova.oneway test



일원분산분석 실습2 ✡payment에 따른 expense✡

- 다음과 같은 one-way ANOVA를 실행하시오.
  - 데이터: pttest
  - IV: payment
    3가지 (간편. 신카. 계좌이체)
  - DV: expense
    한달동안 지출 금액
  - 유의수준(α) = 0.01

H_0: payment 별 expense는 같다

STEP 1: 가설수립

① anova2 <- pttest

② anova2 %>% group_by(payment) %>% summarise(mean(expense, na.rm = T))

H_0 : 세 집단 간에 expense 모평균은 모두 동일하다

H_a : 적어도 한 집단의 expense 모평균은 다른 집단과 다르다

STEP 2 : 이상치 검토 및 제거

① library(psych)
② descr <- describe(anova2$expense)
③ descr <- descr %>% mutate(UL = mean + 2 * sd), descr <- descr %>% mutate(LL = mean - 2 * sd)
④ table(anova2$expense > descr$UL)
⑤ anova2_new <- anova2 %>% filter(expense <= descr$UL)

STEP 3 : 서브데이터 프레임 만들기

anova2_simple <- anova2_new %>% filter(payment = "간편결제")
anova2_account <- anova2_new %>% filter(payment = "계좌이체")
anova2_credit <- anova2_new %>% filter(payment = "신용카드")

STEP 4 : 정규성 검토 (서브 데이터 전부)

① summary(anova2_simple$expense)
② hist(anova2_simple$expense, breaks = seq(0,2000,40))
③ shapiro.test(anova2_simple$expense)

STEP 5 : 등분산 조건 (통합된 df)

① library(car)
② leveneTest(expense ~ payment, data = annova2_new)

```
> leveneTest(expense ~ payment, data = anova2_ne
w)
Levene's Test for Homogeneity of Variance (center
 = median)
       Df F value  Pr(>F)
group   2  3.5664 0.02866 *
      887
```

> α=0.01
=> 등분산 만족

만약 α=0.05 => p-value 유의 => 등분산 조건 만족X
=> 이분산 가정
one-way ANOVA 시행

#STEP6. one way ANOVA 시행 : 등분산

anova2_result <- aov(expense ~ payment, data = anova2_new)

```
> summary(pttest_result)
             Df    Sum Sq Mean Sq F value
payment       2   8844779 4422389   24.74
Residuals   887 158531041  178727
             Pr(>F)
payment     3.49e-11 ***
Residuals
```

α=0.01 보다 작음

=> 대립가설 채택

STEP 7: 사후분석

① library(agricolae) ▷ 독립변수 (IV)
② duncan.test(anova2_result, "payment", console = T)

```
          expense groups
신용카드 662.2990      a
계좌이체 465.4400      b
간편결제 460.2562      b
>
```

=> 신용카드 expense 모평균 > 계좌이체 expense 모평균 = 간편결제 expense 모평균

추가작업 : α=0.05, 이분산 가정 oneway ANOVA

oneway.test(expense ~ payment, data = anova2_new)
종속 독립

```
data:  expense and payment
F = 24.281, num df = 2.00, denom df =
280.52, p-value = 1.881e-10
```

α=0.05 보다 작음 => 유의 => 대립가설 채택

사후분석

① library(dunn.test)
② dunn.test(anova2_new$expense, anova2_new$payment, method = "bonferroni")
종속(DV) IV(독립)

```
data: x and group
Kruskal-Wallis chi-squared = 59.0641, df = 2, p-v
alue = 0



                          Comparison of x by gro
up
                                (Bonferroni)
Col Mean-|
Row Mean |       간편결제        계좌이체
---------+----------------------
    계좌이체 |  -0.005656
         |     1.0000
         |
    신용카드 |  -7.301456  -4.627767
         |    0.0000*     0.0000*

alpha = 0.05
Reject Ho if p <= alpha/2
```

열 기준 : 간편을 계좌와 비교 => 간편 M가 조금 작음

p-value > α=0.05 유의 X => 계좌이체 모평균 = 간편결제 모평균

- 신용카드 모평균 > 계좌이체 모평균

↳ 유의 => 신용카드 모평균 > 간편결제 모평균

왜?

=> 신용카드 > 계좌이체 = 간편결제