양측검정은 2Xp-value로 나옴

t-검정

숙명여자대학교 경영학부 오중산

집단간 평균비교

세 가지 집단간 평균비교 방법

- 모표준편차를 알기 어렵기 때문에 *z*-검정은 한계가 있음
- *t*-검정의 유형 구분
  - 독립표본 *t*-검정 vs. 대응표본 *t*-검정
- ANOVA의 유형 구분 : 분산분석 (Analysis of Varience)
  - One-way ANOVA vs. Two-way ANOVA
  - ANOVA vs. MANOVA

통계량	불편추정량	모수
$\bar{X}$	$\bar{X}$	M
$\bar{P}$	$\bar{P}$	P
S^2	$\hat{S}^2$	σ^2
S	$\hat{S}$	σ

IV 1개 / IV 2개 ⇒ 상호작용 ex) 배고픔 ⇒ 음식냄새 상호작용

Multi variete ⇒ 종속변수 여러개 ex) DV_1 DV_2 DV_3 한번에 비교 (많이 안쓰임)

구분	*z*-검정	*t*-검정	ANOVA
종속변수 (DV) 확률변수	모집단에서 정규분포를 띠어야 함 (모를 경우 표본크기 30개 이상)	모집단에서 **정규분포**를 띠어야 함	
모표준편차	모수 **알아야 함** ⇒ 잘 안쓰임	모름	몰라도 됨(무관함)
모분산 조건	해당사항 없음	등분산 혹은 **이분산**	**등분산 조건 만족**해야 함
표본크기 n	가급적 30개 이상	무관함(**30개 미만**도 가능) ↳ 그래도 $n_1, n_2 \geq 30$ 좋음	30개 이상 n_1, n_2 모두
비교대상 집단	2개		2개 이상(보통 **3개 이상**)

독립표본 *t*-검정

ex) 학년에 따른 용돈 모평균 차이
IV
1.2.3.4 학년
$M_1 M_2 M_3 M_4$

t검정은 M_1, M_2 이런식으로 두개만 가능
ANOVA는 한 번에 가능

독립표본 *t*-검정 정의

- 서로 다른 두 모집단을 대상으로 모평균 차이 유무 검정
- 독립변수(IV)와 종속변수(DV)
  - 독립변수는 집단을 구분하는 변수로 범주형 척도로 측정
    - 집단은 두 개로 구분되어야 하므로, 만약 세 개 이상인 경우 두 개로 재분류

ex) 1.2.3.4학년
저학년 고

    - 회귀분석의 경우 독립변수와 종속변수가 모두 정량적 변수이므로 더 발전된 분석방법
  - 종속변수는 모평균 차이 비교의 대상이 되는 변수로 정량적 변수

num. (dbl) . int

독립표본 t-검정

1. 전제조건 확인

? 두 가지 가설과 정규성 조건

? 두 가지 가설

? $H_0: \mu_1 - \mu_2 = 0$ & $H_a: \mu_1 - \mu_2 \neq 0$

↳ 양측검정

만족해야하는데 모집단 대상으로 알 수 없음 ⇒ 표본 대상으로 검정

↱ $X_1 \sim N(M_{저}, \sqrt{M_{저}^2})$

n_1 저 n_2 고 ≥ 30

? 정규성 조건

X_1 X_2 : 저학년의 한달용돈, 고학년의 한달용돈

30개가 넘어도 확인하는 습관 들여야 하지만 현실에서 만족하기 어려움

? t-검정을 위해서는 두 확률변수가 모두 정규성 조건을 만족해야 함

? 두 개 표본의 표본크기가 모두 30개 이상이면 상관없지만, 30개 미만인 경우에는 t-검정을 시해하기에 앞서 사전에 정규성 조건을 확인해야 함

↱ 전제조건

? 두 확률변수가 정규분포를 띠면, 두 표본평균 각각의 표본분포도 정규성 조건 만족

$X_{저}$ $X_{고} \sim N$ $\bar{X} \sim N$

? 따라서 $\bar{x}_1 - \bar{x}_2$ 도 정규분포를 띰

* $\bar{x} \rightarrow M$

$\bar{x}_1 - \bar{x}_2 \Rightarrow M_1 - M_2$

독립표본 *t*-검정

σ_1^2 (1) = σ_2^2 (2) M, σ, σ^2

$\hat{s}_1^2 - \hat{s}_2^2 = 0$

X_1 X_2

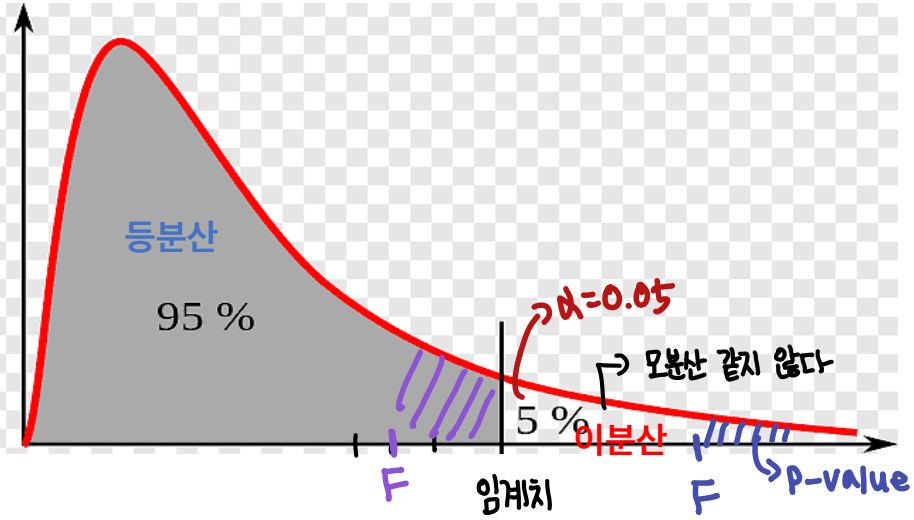
등분산 검정

homogeneity of variance 분산이 같다 ⇒ t 검정을 하기 위한 전제조건 확인

1. 정규성 조건 만족하는지
2. 등분산 조건 만족하는지

- *t*-검정에 앞서 두 모분산이 같은지 확인해야 함
  - $H_0: \sigma_1^2 - \sigma_2^2 = 0$ & $H_a: \sigma_1^2 - \sigma_2^2 \neq 0$
    - ↳ 채택 되어야 함
  - *F*-통계량 활용 검정
    - $\hat{s}_1^2 = \dfrac{\sigma_1^2 \text{ (큰 값)}}{\sigma_2^2 \text{ (작은 값)}}$, $df = n-1$ (자유도)
  - *F*-통계량이 1에 가까워야 등분산 조건 만족 (H_0 채택)

⇒ H_0 채택



* p-value ≤ α : 대립가설 채택 ⇒ 등분산 X (유의하다)

p-value ≥ α : 귀무가설 채택 ⇒ 등분산 O

① 등분산 만족 ⇒ 등분산 t-검정
② 이분산 t-검정

독립표본 t-검정

2. 등분산 가정 독립표본 t-검정

- $H_0: \sigma_1^2 - \sigma_2^2 = 0$ 조건 만족 (등분산 검정에서)
- 아래와 같은 절차에 따라 t값을 구한 후, 양측검정

$t \rightarrow z \sim N(0,1)^2$
샘플 사이즈 많으면 표준정규분포 수렴
$df = n-1$ ↑

$$\bar{X}_1 - \bar{X}_2 \sim N(\mu_1 - \mu_2(=0), \sigma^2_{\bar{X}_1 - \bar{X}_2}(= s_p{}^2(\frac{1}{n_1} + \frac{1}{n_2})))$$

$$s_p = \sqrt{\{(n_1-1)\hat{s}_1^2 + (n_2-1)\hat{s}_2^2\} \div df} \quad (df = n_1 + n_2 - 2)$$

$$\bar{X}_1 - \bar{X}_2 \sim N(0, (\sqrt{\frac{s_1{}^2 + s_2{}^2}{n-1}})^2) \text{ (만약, } n_1 = n_2 = n)$$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_{\bar{X}_1 - \bar{X}_2}}$$

2X p-value ≤ α : 대립가설 채택 ⇒ 모평균 차이 존재

2X p-value ≥ α : 귀무가설 채택 ⇒ 모평균 같다

독립표본 t-검정

이분산 가정 독립표본 t-검정

- $H_a: \sigma_1^2 - \sigma_2^2 \neq 0$ 조건 만족 (등분산 검정에서)
- Welch 검정을 실시하며, 아주 엄밀하게는 t값이 아니므로 t'으로 표기
- 아래와 같은 절차에 따라 t'값을 구한 후, 양측검정

$$t' = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_{\bar{X}_1 - \bar{X}_2}}$$

SE

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\left(\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}\right)}$$

$$df = \frac{\left(\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}\right)^2}{\frac{\hat{s}_1^4}{n_1^2(n_1 - 1)} + \frac{\hat{s}_2^4}{n_2^2(n_2 - 1)}}$$

독립표본 t-검정

t-검정 절차

- 1단계: 가설수립 ※ 대립가설 잘 세우기
- 2단계: 집단간 데이터프레임 생성 ex) 남.여 데이터프레임
- 3단계: 정규성 조건 확인 => 샘플 사이즈 충분히 크면 안해도 됨
- 4단계: 등분산성 조건 확인
- 5단계: 독립표본 t-검정 실시 및 가설검정 => 2x p-value 와 α 비교

독립표본 *t*-검정

t-검정 실습: 준비단계

데이터 소개(ttest.csv)

- E-commerce 업체에서의 고객 주문 관련 데이터

데이터 프레임 만들기

- 변수 중에서 name의 측정값 중에 특수문자(®)가 존재하므로 read.csv로 불러와 데이트 프레임 형성

변수명	변수 설명
priority	배송 우선순위
quantity	주문 물량
sales	판매금액
shipping	배송방법
price	단가
cost	주문처리비용
customer	고객유형
category	물품유형
name	물품명
container	포장크기(유형)
margin	순이익

독립표본 *t*-검정

t-검정 실습: 준비단계

* 데이터 확인
① library (dplyr) → glimpse (ttest) : 척도확인
② n_distinct (ttest $ priority) : 범주 확인
③ table (ttest $ priority) : 빈도확인

데이터 전처리

척도 변경하기

척도확인 glimpse ( )

ttest $ priority <- as.factor (ttest $ priority)

- 문자형 척도로 측정된 변수 중에서 name을 제외한 5개의 척도를 범주형으로 변경

범주형 척도 변수에 대한 빈도수 확인

library (descr) → freq (ttest $ priority)

변수 위치 조정: 범주형/수치형/문자형 척도 측정 변수 순서로 정리

ttest <- ttest %>% relocate (where (is.factor))

ttest <- ttest %>% relocate (margine , .before = name)

이상치 검토 및 빈도수 확인

① library(psych) ② descr<- describe (ttest [c(6:10)])

r> 이부터 세는거 아님

- 수치형 척도로 측정된 5개 변수에 대해 표본평균 ± 2표본표준편차를 상 /하한으로 설정

③ descr <- descr %>% mutate (LL= mean - 2*sd)

descr <- descr %>% mutate (UL= mean + 2*sd)

- 5개 변수 측정값 각각에 대해 상한값을 넘어서는 이상치 빈도수 확인

④ table (ttest $ price > 670.06) table (ttest $ cost > 47.37)
table (ttest $ sales >8945.98) table (ttest $ margin > 0.7837)

이상치를 제외한 데이터 프레임 생성

ttest _ new <- ttest %>% filter (sales <= 8945.98 , price <= 670.06 , cost <= 47.37 , margin <= 0.7837) : UL 보다 작은 것들

독립표본 *t*-검정

t-검정 실습

	customer	`n()`	`mean(sales)`
	<fct>	<int>	<dbl>
1	Consumer	1393	990.
2	Corporate	2605	1001.
3	Home Office	1723	1073.
4	Small Business	1396	1010.

83만큼의 차이가 있는데 모집단으로서 이 차이가 유의미한가?

```
ttest_new %>% group_by(customer) %>% summarise(n(), mean(sales))
```

1단계: 가설수립

고객유형 4가지 중 HO, CS 선택

- 독립변수(customer)와 종속변수(sales) 설정
  - 네 가지 customer 유형에 따른 sales 평균 비교
- 가설수립을 위한 문제제기: "Home Office(HO) 판매금액 평균과 Consumer(CS) 판매금액 평균은 서로 동일한가?"

=> 모평균

X

- 가설수립

같다 / 다르다

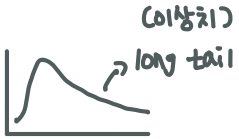
$H_0: \mu_{HO} - \mu_{CS} = 0$ & $H_a: \mu_{HO} - \mu_{CS} \neq 0$

2단계: 집단별로 데이터 프레임 만들기

- HO와 CS로 구분된 서브 데이터 프레임 생성

```
ttest_HO <- ttest_new %>% filter(customer == "Home Office")
ttest_CS <- ttest_new %>% filter(customer == "Consumer")
```

독립표본 t-검정



정리

엄격하게 정규성 조건 만족하지 않아도 괜찮음

① 샘플 사이즈 굉장히 크다

② 자연로그 ⇒ 정규분포 비슷하게 만듦

t-검정 실습

↳ 만족 어려움 ⇒ 시각적으로 왜도 / 첨도 확인

3단계: 정규성 조건 검토

① summary(ttest$sales) ⇒ seq( ) 적을 것 확인

- 두 개 서브 데이터 프레임 각각에 대해 sales 관련 히스토그램 그리기 ② hist(ttest_H0$sales, breaks = seq(0,9000,5))

  ⇒ 굉장히 long tail임

  - 히스토그램 형태를 통해 시각적으로 정규성 검토

- shapiro.test 함수를 이용하여 sales의 정규성에 대한 통계적 검정 shapiro.test(ttest_H0$sales)

  ↳ 왜도를 기준으로 정규성 검토

  ⇒ p-value = 2.2e-16 < α

  ↳ 10^{-16}

  - *p*-value가 유의하지 않아야 정규성 조건 만족

    ↳ p-value가 0으로 유의하게 나옴 (α보다 작게 나옴)

- 정규성 조건 만족하지 못할 경우 대응 방안

  ↱ n_{H0} & $n_{CS} \geq 30$ ⇒ $\bar{x}_{H0} - \bar{x}_{CS} \sim N(\quad)$: 중심극한정리

  - 두 개 표본의 크기가 모두 크기 때문에 정규성 조건을 만족하지 못하더라도 표본평균과 표본평균 차이는 정규성 조건 만족

  - 정규성 조건에 조금 더 부합하도록 종속변수를 변환하되, 오른편으로 꼬리가 길면 자연로그 변환

```
ttest_H0 <- ttest_H0 %>% mutate(lnsales = log(sales))
ttest_CS <- ttest_CS %>% mutate(lnsales = log(sales))
```

)⇒ 그나마 정규분포와 가까워짐

shapiro.test(ttest_CS$lnsales) 하면 p-value 값은 여전히 0에 가깝지만 W값 개선

독립표본 *t*-검정

t-검정 실습

$\sigma_{H0}^2 = \sigma_{CS}^2 \rightarrow F = 1$이 좋음

0.05

p-value ≤ α ⇒ 이분산 (p-value 유의하다)

p-value > α ⇒ 등분산 t 검정

4단계: 등분산성 조건 검토

```
var.test(ttest_H0$lnsales, ttest_CS$lnsales)
```

- var.test 함수를 이용한 등분산성 검토
  - *p*-value가 유의하지 않아야 등분산성 조건 만족

α 0.05 < 0.5536

```
data:  ttest_H0$lnsales and ttest_CS$lnsales
F = 1.0308, num df = 1722, denom df = 1392, p-value = 0.5536
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.9324847 1.1388074
sample estimates:
ratio of variances
          1.030772
```

↳ 유의미 X

1과 가까움

5단계: 독립표본 *t*-검정 실시 및 가설검정

- t.test 함수를 이용하여 4단계에서 등분산성 조건을 만족하면 등분산 가정 *t*-검정을 실시하고, 만족하지 못하면 이분산 가정 *t*-검정 실시

```
t.test(ttest_H0$lnsales, ttest_CS$lnsales, alternative = "two.sided", var.equal = T)
```

⇒ sales 보다 lnsales가 그나마 정규분포니까

↳ 이분산이면 F

- *p*-value가 유의하면 대립가설 채택

2X p-value

```
data:  ttest_H0$lnsales and ttest_CS$lnsales
t = 1.1265, df = 3114, p-value = 0.2601
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.0448496  0.1659631
sample estimates:
mean of x mean of y
 5.976144  5.915587
```

0.05보다 큼 ⇒ 유의미 X ⇒ 귀무가설 채택

표본평균

t-검정 실습 추가

sales의 모평균

$H_0: M_{TN} - M_{FN} = 0$ $H_1: M_{TN} - M_{FN} \neq 1$

- category에서 Technology와 Furniture 간에 sales 평균 차이 존재 여부 확인

⇒두 집단 간의 lnsales 모평균은 같다

* 독립표본 t 검정
두 개의 표본, 동일한 내용의 종속변수에 모평균 차이 있는지 없는지 검정

대응표본 *t*-검정

? 대응표본(paired sample) *t*-검정이란?

차이 있는지 없는지

? 하나의 표본에서 서로 다른 종속변수 모평균 차이 여부 비교하는 통계분석

? 표본이 하나이므로 집단을 구분할 필요가 없고, 독립변수 없음 ⇒ 종속변수만 2개 존재

? 대응표본 *t*-검정 예시

$X_A \rightarrow M_A$ $X_B \rightarrow M_B$

? (사건 없음) 동일한 소비자 집단이 한 달 동안 품목유형 A를 구매한 금액평균과 품목유형 B를 구매한 금액평균간의 차이

유무 $H_0 : M_A - M_B = 0$ $H_1 : M_A - M_B \neq 0$

* X_1 : 키 (cm) X_2 : 몸무게 (kg) ⇒ 이절적 비교 못함

? (사건 있음) 직원들에 대해 업무몰입에 대한 동기부여 교육을 시행하기 전후의 평균업무시간 차이 유무

ex) 이커머스 : promotion 전.후
⇒ 차이 있으면 promotion 효과있다

$X_{전}$: 업무시간
$X_{후}$: 후 업무시간

$M_{전} - M_{후} = 0$
$\neq 0$

대응표본 *t*-검정 pttest ← read_csv("pttest.csv", locale = locale('ko', encoding = 'euc-kr'))

대응표본 *t*-검정 절차

Xm : 아침 배송 주문 횟수 ⇒ Mm

Xw : 주말 배송 주문 횟수 ⇒ Mw

STEP1: 가설수립

$H_0: M_1 - M_2 = 0 \Rightarrow M_d$

$H_0: M_m - M_w = 0$ ($= M_d$)

$H_1: M_m - M_w \neq 1$

- H_0: μ_d (두 모평균의 차이) $= 0$ & H_a: $\mu_d \neq 0$ ②*
- ① 두 종속변수(확률변수)가 모두 정규분포를 띠거나, 표본크기가 최소 30개 이상이면 표본평균 차이는 아래와 같은 정규분포를 띰 (DV, RV)

  $n \geq 30$ (CLT: 중심극한정리)
- s_d는 차이(X_{11}-X_{12})의 표준편차이고, 자유도는 n-1

같은 샘플

$$\overline{X}_{11} - \overline{X}_{12} \sim N(\mu_d\ (= \mu_{11} - \mu_{12} = 0),\ \sigma^2_{\bar{X}_{11} - \bar{X}_{12}}\ (= \frac{s_d^2}{n}))$$

$\overline{X}_d$

↳ 표준오차 SE

$$t = \frac{\overline{X}_{11} - \overline{X}_{12}}{\sigma_{\bar{X}_{11} - \bar{X}_{12}}}$$

① → X_{11} 첫번째 RV (DV)

→ X_{12} 두번째 RV (DV)

대응표본 *t*-검정

대응표본 *t*-검정 절차

ptest <- ptest %>% mutate(d = morning - weekend)

STEP2: 차이 변수 만들기 : 파생변수 $X_{11} - X_{12}$

- 두 종속변수 차이에 대한 변수(d) 생성

STEP3: d에 대한 정규성 검토

$X_{11} - X_{12} (=d) \sim N(\ )$

$\Rightarrow \bar{X}_{11} - \bar{X}_{12} \sim N(\ )$

① shapiro.test(ptest$d)

② summary(ptest$d)
hist(ptest$d, breaks = seq(-15, 9, 1))

이상치 확인 ① describe(ptest$d) LL = -9.54, UL = 5.74

② table(ptest$d < -9.54)
table(ptest$d > 5.74

- 표본이 하나이므로 등분산 조건은 확인하지 않음

STEP4: 대응표본 *t*-검정

- t.test 함수에서 paired = TRUE 조건 추가
  : 대응표본 t 검정 해줘

t.test(ptest$morning, ptest$weekend, alternative = "two.sided", paired = T)

양측 검정 2X P-value

```
data:  pttest$morning and pttest$weekend
t = -16.192, df = 1056, p-value < 2.2e-16
alternative hypothesis: true mean difference is n
ot equal to 0
95 percent confidence interval:
 -2.132059 -1.671158
sample estimates:
mean difference
      -1.901608
```

$2.2 \times 10^{-16} < \alpha$ ⇒ 대립가설 채택

: 두 종속변수의 모평균 차이 0이 아니다

⇒ 한달동안 주말 배송 주문 평균이 새벽 배송 주문 평균보다 더 많다

↳ 통계적으로 유의미