데이터과학을 위한 R프로그래밍

6주차. R을 이용한 통계분석

이혜선 교수

포항공과대학교 산업경영공학과

목차

6주차. R을 이용한 통계분석

6주차

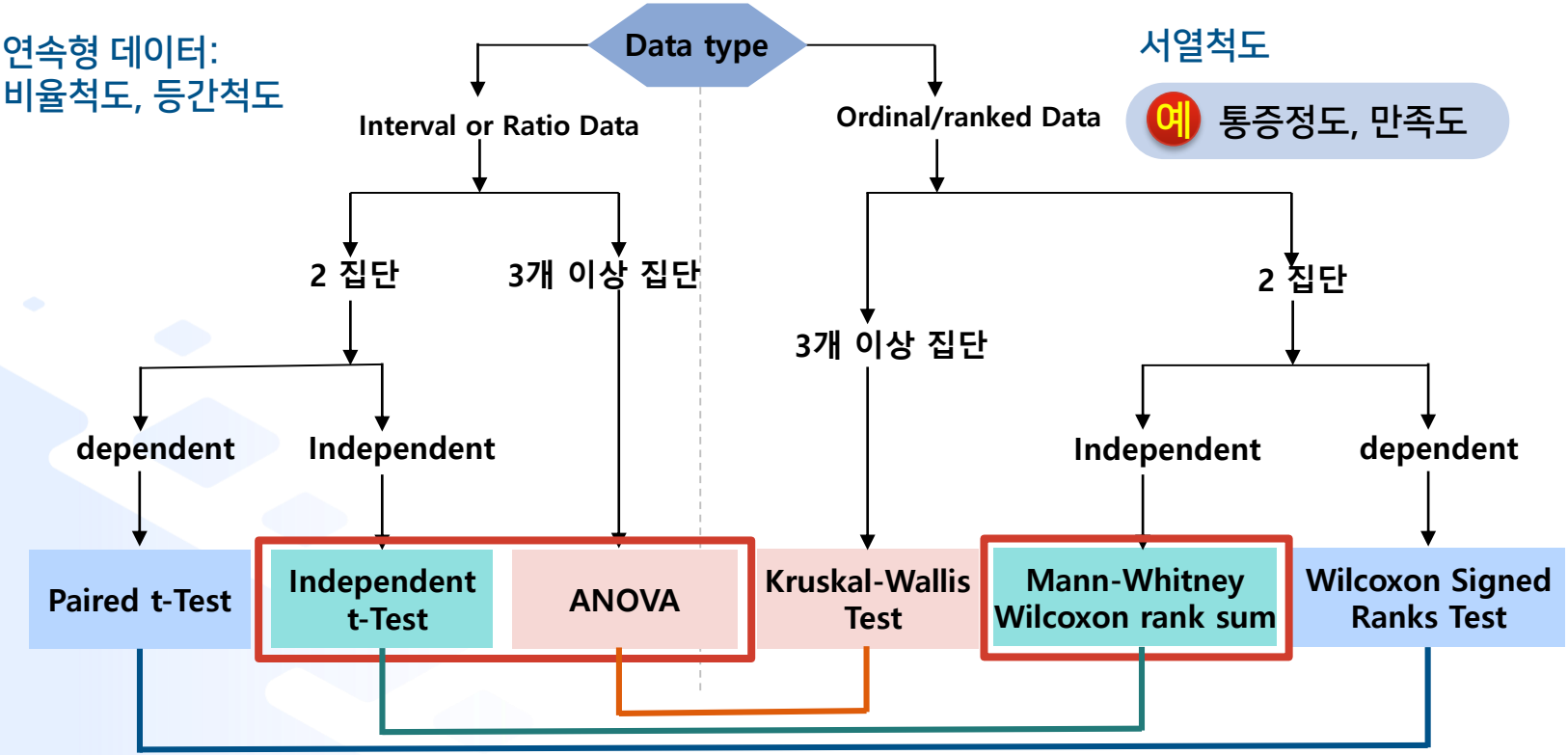
1차시

두 그룹간 평균 비교 분석

두 그룹간 평균비교분석

POSTECH

모집단간 차이에 대한 검정 (모수/비모수 검정)



두 그룹간 평균비교분석

POSTECH

단일표본의 평균검정

- 단일표본의 평균검정 :t.test (변수, mu =검정하고자 하는 평균값)
  - 가설 1: G3(최종성적)의 평균은 10인가? H_0 (null Hypothesis: 귀무가설) : μ=10

```
# t-test for two sample means

# set working directory
setwd("D:/tempstore/moocr")

### student math grade data ###
stud<-read.csv("stud_math.csv")

head(stud)
dim(stud)
str(stud)

attach(stud)
```

```
# single t-test : to test whether or not mean of G3 is 10
t.test(G3, mu=10)
```

```
> t.test(G3, mu=10)

        One Sample t-test

data:  G3
t = 1.8011, df = 394, p-value = 0.07245
alternative hypothesis: true mean is not equal to 10
95 percent confidence interval:
 9.961992 10.868388
sample estimates:
mean of x
 10.41519
```

결론: α=0.05에서는 G3의 평균이 10이라고 할 수 있는 근거가 있다

t검정통계량, 자유도, p-value

Ha (대립가설) : 모평균은 10이 아니다
95% 신뢰구간 : (9.96, 10.86)
표본평균값 : 10.415

POSTECH

두 집단의 평균검정 (t-test)

두 집단 표본평균 비교검정 : t.test (타겟변수~범주형변수, data=)

가설 2 : 거주지역(R, U)에 따른 G3(최종성적) 평균에 차이가 있는가? (양측검정)

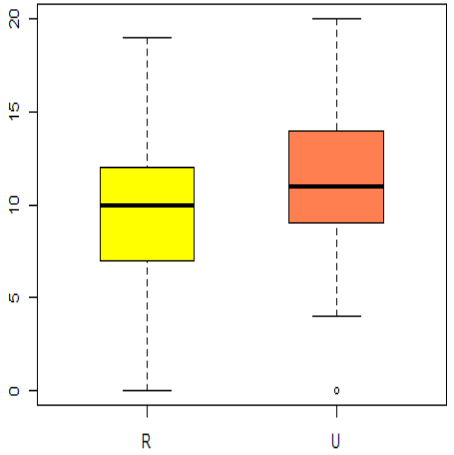
```
# 2. two sample t-test
## example 1
# to test whether or not mean of G3 is same between Urban and Rural
t.test(G3~address, data=stud)
boxplot(G3~address, boxwex = 0.5, col = c("yellow", "coral"))
```

```
> t.test(G3~address, data=stud)

        Welch Two Sample t-test

data:  G3 by address
t = -2.1101, df = 140.91, p-value = 0.03661
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.25240320 -0.07340373
sample estimates:
mean in group R mean in group U
       9.511364       10.674267
```

양측검정



p-value=0.03으로 유의수준 0.05 (α=0.05)에서
거주지역에 따라 G3는 유의한 차이가 있다고 할 수 있다.

POSTECH

두 집단의 평균검정(t-test)

- 두 집단 표본평균 비교검정 : t.test(타겟변수~범주형변수, data=)
  - 단측검정 : 기각역이 한쪽에만 있는 경우,
    alternative=c("greater") 혹은 alternative=c("less")

```
# alternative H : mu(Rural)< mu(Urban)
t.test(G3~address, data=stud, alternative = c("less"))

help(t.test)
```

$H_0 : \mu_R = \mu_U \ (\mu_R - \mu_U = 0)$
$H_1 : \mu_R < \mu_U \ (\mu_R - \mu_U < 0)$

```
> t.test(G3~address, data=stud, alternative = c("less"))

        Welch Two Sample t-test

data:  G3 by address
t = -2.1101, df = 140.91, p-value = 0.01831
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
       -Inf -0.2504199
sample estimates:
mean in group R mean in group U
       9.511364        10.674267
```

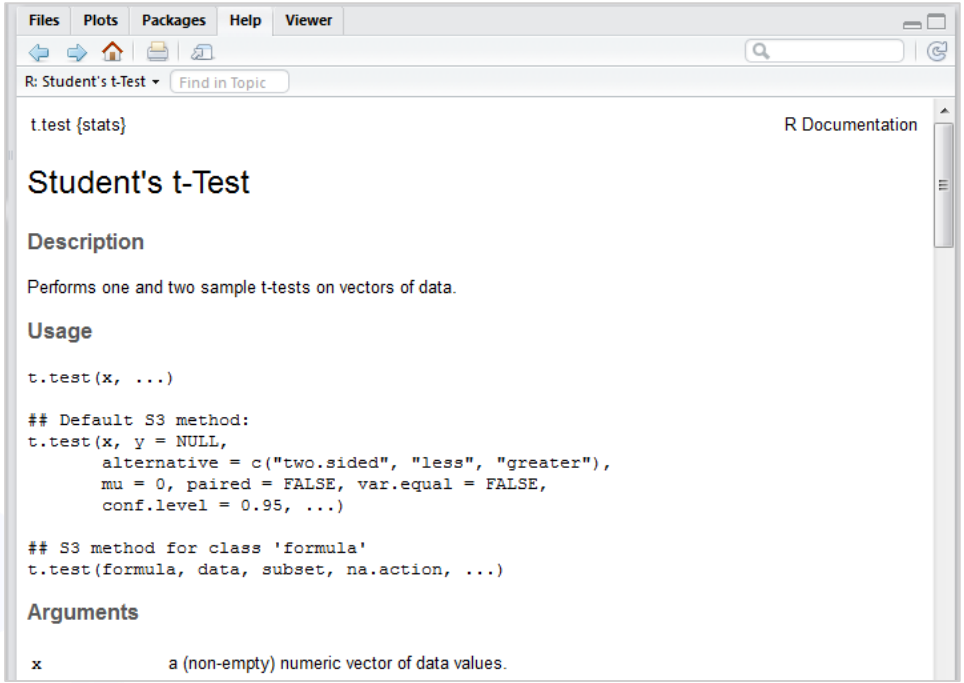
p-value=0.018로 유의수준을 0.05로 했을 때 성적(Rural)<성적(Urban)이라고 할 수 있다

POSTECH

두 집단의 평균검정 (t-test)

☑ 두 집단 표본평균 비교 도움말 보기 : help(t.test)

```
help(t.test)
```

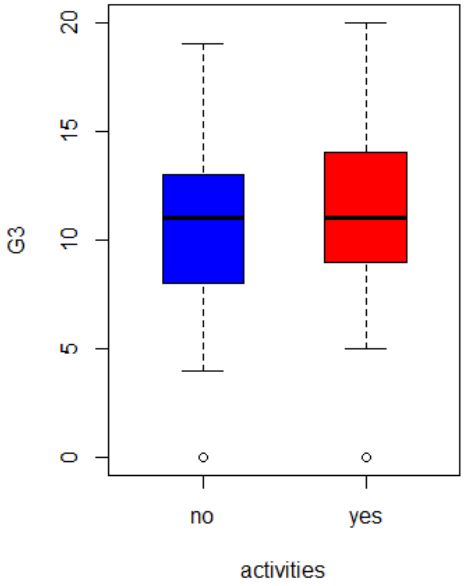


POSTECH

두 집단의 평균검정 (t-test)

- 두 집단 표본평균 비교검정 : t.test(타겟변수~범주형변수, data=)
  - 가설 3 : 방과후 활동여부(yes, no)에 따른 G3(최종성적) 평균에 차이가 있는가?

```
## example 2
# to test whether or not mean of G3 is equal for activities
t.test(G3~activities, data=stud)
boxplot(G3~activities, boxwex = 0.5, col = c("blue", "red"))
```



상자그림(Boxplot)에서 보면 방과후
활동여부는 G3(성적)에 뚜렷한 차이를 볼 수 없음

두 그룹간 평균비교분석

POSTECH

두 집단의 평균검정 (t-test)

- 두 집단 표본평균 비교검정 : t.test(타겟변수~범주형변수, data=)
  - 가설 3 : 방과후 활동 여부(yes, no)에 따른 G3(최종 성적) 평균에 차이가 있는가?
  - t-test 검정통계량에 의한 검정결과

```
> t.test(G3~activities, data=stud)

        Welch Two Sample t-test

data:  G3 by activities
t = -0.31944, df = 392.98, p-value = 0.7496
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.0542623  0.7595503
sample estimates:
 mean in group no mean in group yes
         10.34021          10.48756
```

p-value

양측검정

p-value=0.75는 유의수준 0.05보다 큽니다.

즉 검정통계량의 값이 기각역에 있지 않다.

⇒ 귀무가설(평균이 같다)를 기각 불가

⇒ 방과 후 활동 여부는 G3에 유의한 영향이 없다!

평균(G3(방과 후 활동 없음)-G3(방과 후 활동))
차이에 대한 신뢰구간 = (-1.05, 0.79)

신뢰구간 사이에 0값이 있다는 것은
차이가 없음을 의미!!

POSTECH

두 집단의 비모수적 비교검정

두 모집단의 비모수적 방법 (Wilcoxon rank sum Test) : wilcox.test(x,y)

- wilcox.test는 타겟변수가 서열척도(통증정도, 만족도, ..)일 때 사용할 수 있다

```
# Wilcoxon signed-rank test
# wilcox.test(G3, mu=10)
wilcox.test(G3~address)
```

```
> wilcox.test(G3~address)

        Wilcoxon rank sum test with continuity correction

data:  G3 by address
W = 11278, p-value = 0.01776
alternative hypothesis: true location shift is not equal to 0
```

wilcox.test(타겟변수~범주형변수)

```
help(wilcox.test)
```

Wilcoxon Rank Sum and Signed Rank Tests

Description

Performs one- and two-sample Wilcoxon tests on vectors of data; the latter is also known as 'Mann-Whitney' test.

Usage

```
wilcox.test(x, ...)

## Default S3 method:
wilcox.test(x, y = NULL,
            alternative = c("two.sided", "less", "greater"),
            mu = 0, paired = FALSE, exact = NULL, correct = TRUE,
            conf.int = FALSE, conf.level = 0.95, ...)
```