Part 2. R 통계분석 (데이터 분석 전문가 양성과정)

07

두 집단의 차이검정

경북대학교 배준현 교수
(joonion@knu.ac.kr)

07. 두 집단의 차이검정

- **평균검정:** *t-test*
  - 평균을 비교할 때 쓸 수 있는 가설검정 방법
  - 단일표본 평균검정: *one-sample* t-test
    - 표본평균을 가설로 정한 값과 비교
  - 독립표본 평균검정: *two-independent-samples* t-test
    - 두 집단간의 평균을 비교해서 집단의 차이에 대한 가설검정
  - 대응표본 평균검정: *paired-samples* t-test
    - 관측값이 서로 쌍을 이루는 경우(예: 사전-사후)에 대한 가설검정

07. 두 집단의 차이검정

- **단일표본 평균검정:** *one sample* t-test
  - **하나의 표본 데이터를 이용하여 모집단의 평균을 검정**
    - **표본집단의 평균이 모집단의 평균과 같은가를 검정하고 싶을 때**
  - R: `t.test(x, mu)`
    - x: a numeric vector of data values.
    - mu: a number indicating the *true value of the mean*.

07. 두 집단의 차이검정

- H_0: 호주 대학생 전체의 평균 키는 175cm이다.
- H_1: 호주 대학생 전체의 평균 키는 175cm가 아니다.

```
> library(MASS)
> t.test(survey$Height, mu = 175)

        One Sample t-test

data:  survey$Height
t = -3.8451, df = 208, p-value = 0.0001602
alternative hypothesis: true mean is not equal to 175
95 percent confidence interval:
 171.0380 173.7237
sample estimates:
mean of x
 172.3809
```

07. 두 집단의 차이검정

- H_0: 호주 대학생 전체의 평균 키는 172cm이다.
- H_1: 호주 대학생 전체의 평균 키는 172cm가 아니다.

```
> t.test(survey$Height, mu = 172)

        One Sample t-test

data:  survey$Height
t = 0.55913, df = 208, p-value = 0.5767
alternative hypothesis: true mean is not equal to 172
95 percent confidence interval:
 171.0380 173.7237
sample estimates:
mean of x
 172.3809
```

07. 두 집단의 차이검정

- H_0: 호주 대학생 전체의 평균 키는 171cm이다.
- H_1: 호주 대학생 전체의 평균 키는 171cm가 아니다.

```
> t.test(survey$Height, mu = 171, conf.level = 0.99)

        One Sample t-test

data:  survey$Height
t = 2.0272, df = 208, p-value = 0.04392
alternative hypothesis: true mean is not equal to 171
99 percent confidence interval:
 170.6100 174.1517
sample estimates:
mean of x
 172.3809
```

07. 두 집단의 차이검정

- H_0: 호주 대학생 전체의 평균 키는 173cm보다 크지 않다.
- H_1: 호주 대학생 전체의 평균 키는 173cm보다 크다.

```
> t.test(survey$Height, mu = 173, alternative = "greater")

        One Sample t-test

data:  survey$Height
t = -0.90894, df = 208, p-value = 0.8178
alternative hypothesis: true mean is greater than 173
95 percent confidence interval:
 171.2554      Inf
sample estimates:
mean of x
 172.3809
```

07. 두 집단의 차이검정

- H_0: 호주 대학생 전체의 평균 키는 174cm보다 작지 않다.
- H_1: 호주 대학생 전체의 평균 키는 174cm보다 작다.

```
> t.test(survey$Height, mu = 174, alternative = "less")

        One Sample t-test

data:  survey$Height
t = -2.377, df = 208, p-value = 0.00918
alternative hypothesis: true mean is less than 174
95 percent confidence interval:
     -Inf 173.5063
sample estimates:
mean of x
 172.3809
```

07. 두 집단의 차이검정

```
> t.height <- t.test(survey$Height, mu = 172)
> str(t.height)
List of 10
 $ statistic  : Named num 0.559
  ..- attr(*, "names")= chr "t"
 $ parameter  : Named num 208
  ..- attr(*, "names")= chr "df"
 $ p.value    : num 0.577
 $ conf.int   : num [1:2] 171 174
  ..- attr(*, "conf.level")= num 0.95
 $ estimate   : Named num 172
  ..- attr(*, "names")= chr "mean of x"
 $ null.value : Named num 172
  ..- attr(*, "names")= chr "mean"
 $ stderr     : num 0.681
 $ alternative: chr "two.sided"
 $ method     : chr "One Sample t-test"
 $ data.name  : chr "survey$Height"
 - attr(*, "class")= chr "htest"
```

```
> t.height$statistic
        t
0.5591299
> t.height$p.value
[1] 0.5766746
> t.height$conf.int
[1] 171.0380 173.7237
attr(,"conf.level")
[1] 0.95
```

07. 두 집단의 차이검정

- **독립표본 평균검정:** *two-independent-samples* t-test
  - 두 개의 독립표본 데이터를 이용하여
    - 각각 대응되는 두 개의 모집단평균이 서로 동일한지 검정
  - 두 집단이 서로 차이가 있는지를 검정하는 것과 같은 의미
    - 남녀 간의 영어시험 점수에 차이가 있는가?
    - 흡연자와 비흡연자 간의 폐질환 발생률은 차이가 있는가
    - 고학력자와 저학력자 간의 텔레비전 시청 시간에 차이가 있는가?
  - R: `t.test(formula, data, …)`
    - *formula*: a formula of the form lhs ~ rhs.
    - data: a data frame containing the variables in the formula.

- MASS 패키지: *cats* 데이터셋
  - 고양이의 심장 무게와 몸무게에 대한 데이터(1974년)
  - 변수 3개와 관측값 144개:
    - *Sex*: 고양이의 성별 (F, M)
    - *Bwt*: 고양이의 몸무게 (kg)
    - *Hwt*: 고양이의 심장 무게 (g)

```
> library(MASS)
> ?cats
> str(cats)
'data.frame':  144 obs. of  3 variables:
 $ Sex: Factor w/ 2 levels "F","M": 1 1 1 1 1 1 1 1 1 1 ...
 $ Bwt: num  2 2 2 2.1 2.1 2.1 2.1 2.1 2.1 2.1 ...
 $ Hwt: num  7 7.4 9.5 7.2 7.3 7.6 8.1 8.2 8.3 8.5 ...
```

07. 두 집단의 차이검정

```
> library(stargazer)
> stargazer(cats, type = "text")
==========================================
Statistic  N   Mean  St. Dev.  Min   Max
------------------------------------------
Bwt        144 2.724   0.485   2.000 3.900
Hwt        144 10.631  2.435   6.300 20.500
------------------------------------------

> with(cats, tapply(Bwt, INDEX=list(Sex), FUN = mean))
       F        M
2.359574 2.900000

> with(cats, tapply(Hwt, INDEX=list(Sex), FUN = mean))
        F         M
 9.202128 11.322680
```

07. 두 집단의 차이검정

- H_0: 고양이의 몸무게는 성별에 따른 차이가 없다.
- H_1: 고양이의 몸무게는 성별에 따라 차이가 있다.

```
> t.test(formula = Bwt ~ Sex, data = cats)

        Welch Two Sample t-test

data:  Bwt by Sex
t = -8.7095, df = 136.84, p-value = 8.831e-15
alternative hypothesis: true difference in means between group F and group M is
not equal to 0
95 percent confidence interval:
 -0.6631268 -0.4177242
sample estimates:
mean in group F mean in group M
       2.359574        2.900000
```

07. 두 집단의 차이검정

- H_0: 고양이의 심장 무게는 성별에 따른 차이가 없다.
- H_1: 고양이의 심장 무게는 성별에 따라 차이가 있다.

```
> t.test(formula = Hwt ~ Sex, data = cats)

        Welch Two Sample t-test

data:  Hwt by Sex
t = -6.5179, df = 140.61, p-value = 1.186e-09
alternative hypothesis: true difference in means between group F and group M is
not equal to 0
95 percent confidence interval:
 -2.763753 -1.477352
sample estimates:
mean in group F mean in group M
       9.202128       11.322680
```

- **대응표본 평균검정:** *paired-samples* t-test
  - 독립표본의 가정: 두 개의 표본이 서로 독립인 모집단으로부터 표본추출
  - 두 개의 표본이 서로 독립이 아닌 모집단으로부터 추출
  - 대응표본: 두 표본의 값이 쌍(pair)을 이루고 있는 경우
  - 예) 아침식사가 IQ 테스트 점수에 미치는 영향
    - 독립표본: 무작위로 실험 대상자를 선정하여 두 개의 집단으로 나눔
      - 한 집단은 아침식사를 하고, 다른 집단은 아침식사를 거르고 테스트
    - 대응표본: 무작위로 실험 대상자를 선정: IQ 테스트를 두 차례 실시
      - 한 번은 아침식사를 하고, 다른 한 번은 아침식사를 하지 않고 테스트

07. 두 집단의 차이검정

- **표준 패키지: *sleep* 데이터셋**
  - 약물이 수면 시간에 미치는 영향에 대한 실험 데이터(1959년)
  - 변수 3개와 관측값 20개:
    - *extra*: 수면시간의 증가
    - *group*: 처방된 약물
    - *ID*: 환자의 ID

```
> ?sleep
> str(sleep)
'data.frame':  20 obs. of  3 variables:
 $ extra: num  0.7 -1.6 -0.2 -1.2 -0.1 3.4 3.7 0.8 0 2 ...
 $ group: Factor w/ 2 levels "1","2": 1 1 1 1 1 1 1 1 1 1 ...
 $ ID   : Factor w/ 10 levels "1","2","3","4",..: 1 2 3 4 5 6 7 8 9 10 ...

> sleep
extra group ID
1    0.7     1  1
2   -1.6     1  2
3   -0.2     1  3
4   -1.2     1  4
5   -0.1     1  5
6    3.4     1  6
7    3.7     1  7
8    0.8     1  8
9    0.0     1  9
10   2.0     1 10
…… (이하생략)
```

```
> library(tidyr)
> spread(sleep, key = group, value = extra)
   ID    1    2
1   1  0.7  1.9
2   2 -1.6  0.8
3   3 -0.2  1.1
4   4 -1.2  0.1
5   5 -0.1 -0.1
6   6  3.4  4.4
7   7  3.7  5.5
8   8  0.8  1.6
9   9  0.0  4.6
10 10  2.0  3.4
```

07. 두 집단의 차이검정

- H_0: 약물 복용 전과 복용 후에 환자의 수면시간에는 차이가 없다(약물이 효과가 없다).
- H_1: 약물 복용 전과 복용 후에 환자의 수면시간에는 차이가 있다(약물이 효과가 있다).

```
> t.test(extra ~ group, data = sleep, paired = TRUE)

        Paired t-test

data:  extra by group
t = -4.0621, df = 9, p-value = 0.002833
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.4598858 -0.7001142
sample estimates:
mean of the differences
                  -1.58
```

07. 두 집단의 차이검정

- H_0: 약물 복용 전과 복용 후에 환자의 수면시간에는 차이가 없다(약물이 효과가 없다).
- H_1: 약물 복용 전과 복용 후에 환자의 수면시간에는 차이가 있다(약물이 효과가 있다).

```
> sleep.wide <- spread(sleep, key = group, value = extra)
> names(sleep.wide) <- c("ID", "group.1", "group.2")

> t.test(sleep.wide$group.1, sleep.wide$group.2, paired = T)

        Paired t-test

data:  sleep.wide$group.1 and sleep.wide$group.2
t = -4.0621, df = 9, p-value = 0.002833
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.4598858 -0.7001142
sample estimates:
mean of the differences
                  -1.58
```

Any Questions?