Paired Sample\_t-test

Youngjin\_Lee

2022 2 13

Table of Contents

# 대응표본 (Paired Sample) t-test

문제 K제약회사의 신제품 개발부서에서는 3개월 안에 살이 빠지는 다이어트 약을 개발하였다. 회사 경영진에게 새롭게 개발한 다이어트약이 효과가 있는지를 보고하기 위하여 약의 효능을 검증하였다. 약을 먹기 전의 체중과 약을 먹은 후 3개월 후의 체중을 조사하였다. 과연 새로운 약은 다이어트에 효과가 있는가

# 1.기본 package 설정, library 로드

# 2.데이터 불러오기

pst\_tb <- read\_csv('data\\pst.csv',   
 col\_names = TRUE,  
 locale=locale('ko', encoding='euc-kr'), # 한글  
 na=".") %>%  
 round(2) %>% # 소수점 2자리로 반올림  
 mutate\_if(is.character, as.factor)

## Rows: 20 Columns: 3  
## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## dbl (3): id, 섭취전, 섭취후  
##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

str(pst\_tb)

## spec\_tbl\_df [20 x 3] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ id : num [1:20] 1 2 3 4 5 6 7 8 9 10 ...  
## $ 섭취전: num [1:20] 82 54 74 75 71 76 70 62 77 75 ...  
## $ 섭취후: num [1:20] 75 50 74 71 69 73 68 62 68 72 ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. id = col\_double(),  
## .. 섭취전 = col\_double(),  
## .. 섭취후 = col\_double()  
## .. )  
## - attr(\*, "problems")=<externalptr>

pst\_tb

## # A tibble: 20 x 3  
## id 섭취전 섭취후  
## <dbl> <dbl> <dbl>  
## 1 1 82 75  
## 2 2 54 50  
## 3 3 74 74  
## 4 4 75 71  
## 5 5 71 69  
## 6 6 76 73  
## 7 7 70 68  
## 8 8 62 62  
## 9 9 77 68  
## 10 10 75 72  
## 11 11 72 70  
## 12 12 83 77  
## 13 13 78 71  
## 14 14 74 74  
## 15 15 68 67  
## 16 16 76 73  
## 17 17 75 77  
## 18 18 75 71  
## 19 19 75 76  
## 20 20 71 74

### long형으로 변형

pst\_tb\_long <- pst\_tb %>%  
 pivot\_longer(c("섭취전","섭취후"), #c("1999, 2000")에러남   
 names\_to = "시간",  
 values\_to = "몸무게")  
  
pst\_tb\_long

## # A tibble: 40 x 3  
## id 시간 몸무게  
## <dbl> <chr> <dbl>  
## 1 1 섭취전 82  
## 2 1 섭취후 75  
## 3 2 섭취전 54  
## 4 2 섭취후 50  
## 5 3 섭취전 74  
## 6 3 섭취후 74  
## 7 4 섭취전 75  
## 8 4 섭취후 71  
## 9 5 섭취전 71  
## 10 5 섭취후 69  
## # ... with 30 more rows

### 차이 추가

pst\_tb <- pst\_tb %>%  
 mutate(차이 = 섭취후-섭취전)  
  
pst\_tb

## # A tibble: 20 x 4  
## id 섭취전 섭취후 차이  
## <dbl> <dbl> <dbl> <dbl>  
## 1 1 82 75 -7  
## 2 2 54 50 -4  
## 3 3 74 74 0  
## 4 4 75 71 -4  
## 5 5 71 69 -2  
## 6 6 76 73 -3  
## 7 7 70 68 -2  
## 8 8 62 62 0  
## 9 9 77 68 -9  
## 10 10 75 72 -3  
## 11 11 72 70 -2  
## 12 12 83 77 -6  
## 13 13 78 71 -7  
## 14 14 74 74 0  
## 15 15 68 67 -1  
## 16 16 76 73 -3  
## 17 17 75 77 2  
## 18 18 75 71 -4  
## 19 19 75 76 1  
## 20 20 71 74 3

# 3.기술통계분석

skim(pst\_tb)

Data summary

|  |  |
| --- | --- |
| Name | pst\_tb |
| Number of rows | 20 |
| Number of columns | 4 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| numeric | 4 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: numeric**

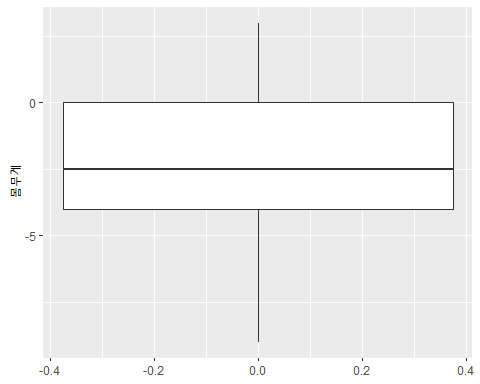
| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| id | 0 | 1 | 10.50 | 5.92 | 1 | 5.75 | 10.5 | 15.25 | 20 | ▇▇▇▇▇ |
| 섭취전 | 0 | 1 | 73.15 | 6.43 | 54 | 71.00 | 75.0 | 76.00 | 83 | ▁▁▃▇▂ |
| 섭취후 | 0 | 1 | 70.60 | 6.10 | 50 | 68.75 | 71.5 | 74.00 | 77 | ▁▁▁▆▇ |
| 차이 | 0 | 1 | -2.55 | 3.14 | -9 | -4.00 | -2.5 | 0.00 | 3 | ▂▁▇▃▂ |

pst\_tb %>%   
 get\_summary\_stats(섭취전, 섭취후, 차이)

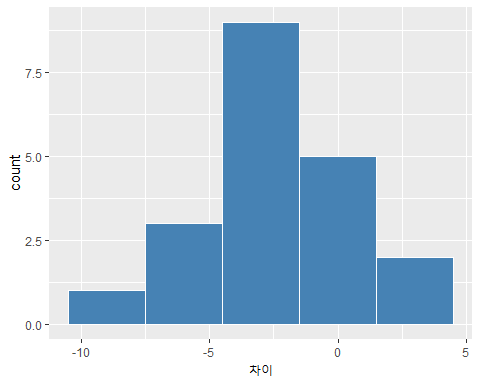
## # A tibble: 3 x 13  
## variable n min max median q1 q3 iqr mad mean sd se  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 섭취전 20 54 83 75 71 76 5 3.71 73.2 6.43 1.44   
## 2 섭취후 20 50 77 71.5 68.8 74 5.25 3.71 70.6 6.10 1.36   
## 3 차이 20 -9 3 -2.5 -4 0 4 2.96 -2.55 3.14 0.701  
## # ... with 1 more variable: ci <dbl>

# 4.그래프 그리기(박스그래프,히스토그램)

pst\_tb %>%   
 ggplot(aes(y = 차이)) +  
 geom\_boxplot() +  
 labs(y = "몸무게")



pst\_tb %>%   
 ggplot(aes(x = 차이)) +  
 geom\_histogram(binwidth = 3,   
 color = "white",   
 fill = "steelblue")

 # 5.이상치 제거, 이상치 확인

pst\_tb %>%  
 identify\_outliers(차이)

## [1] id 섭취전 섭취후 차이 is.outlier is.extreme  
## <0 행> <또는 row.names의 길이가 0입니다>

# 6.정규분포 검정

pst\_tb %>%  
 shapiro\_test(차이)

## # A tibble: 1 x 3  
## variable statistic p  
## <chr> <dbl> <dbl>  
## 1 차이 0.977 0.883

## # 7.paired t-검정

two-sided test: alternative = c(“two.sided”) right-sided test: alternative = c(“greater”) left-sided test: alternative = c(“less”) — paired = TRUE

pst\_tb\_long %>%   
 t\_test(formula = 몸무게 ~ 시간,  
 ref.group = "섭취후",  
 paired = TRUE,  
 alternative = "less",  
 detailed = TRUE)

## # A tibble: 1 x 13  
## estimate .y. group1 group2 n1 n2 statistic p df conf.low  
## \* <dbl> <chr> <chr> <chr> <int> <int> <dbl> <dbl> <dbl> <dbl>  
## 1 -2.55 몸무게 섭취후 섭취전 20 20 -3.64 0.00088 19 -Inf  
## # ... with 3 more variables: conf.high <dbl>, method <chr>, alternative <chr>

Cohen’s d(effect size) 0.2 (small effect), 0.5 (moderate effect) and 0.8 (large effect)

pst\_tb\_long %>%   
 cohens\_d(formula = 몸무게 ~ 시간,   
 paired = TRUE)

## # A tibble: 1 x 7  
## .y. group1 group2 effsize n1 n2 magnitude  
## \* <chr> <chr> <chr> <dbl> <int> <int> <ord>   
## 1 몸무게 섭취전 섭취후 0.813 20 20 large

# 8.추론(infer)을 이용한 가설검정 및 그래프

## 8.1 표본평균(x)을 이용한 검정그래프

###표본평균(x) 계산

x\_bar <- pst\_tb %>%  
 specify(response = 차이) %>% # hypothesize 없음  
 calculate(stat = "mean") %>% # stat = "mean"  
 print()

## Response: 차이 (numeric)  
## # A tibble: 1 x 1  
## stat  
## <dbl>  
## 1 -2.55

### Bootstrapping을 이용한 귀무가설 분포 생성

set.seed(123)   
null\_dist\_x <- pst\_tb %>%  
 specify(response = 차이) %>%  
 hypothesize(null = "point",   
 mu = 0) %>%  
 generate(reps = 1000,   
 type = "bootstrap") %>%  
 calculate(stat = "mean") %>%  
 print()

## Response: 차이 (numeric)  
## Null Hypothesis: point  
## # A tibble: 1,000 x 2  
## replicate stat  
## <int> <dbl>  
## 1 1 0.8   
## 2 2 -0.95   
## 3 3 0.200  
## 4 4 1.05   
## 5 5 1.7   
## 6 6 -0.150  
## 7 7 1.1   
## 8 8 -0.55   
## 9 9 0.100  
## 10 10 0.200  
## # ... with 990 more rows

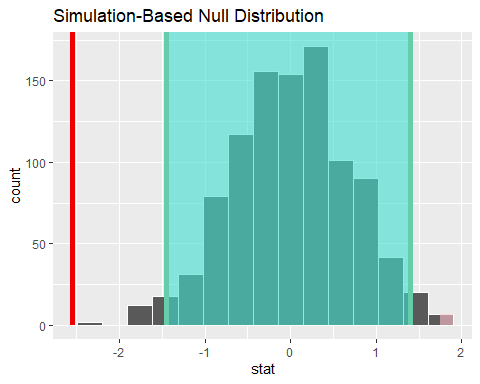
### 신뢰구간 생성

null\_dist\_ci <- null\_dist\_x %>%  
 get\_ci(level = 0.95,   
 type = "percentile") %>%  
 print()

## # A tibble: 1 x 2  
## lower\_ci upper\_ci  
## <dbl> <dbl>  
## 1 -1.45 1.4

### 그래프 그리기

null\_dist\_x %>%  
 visualize() + # method 없음  
 shade\_p\_value(obs\_stat = x\_bar,  
 direction = "two-sided") + # x\_bar  
 shade\_confidence\_interval(endpoints = null\_dist\_ci) # CI

 ### p\_value

null\_dist\_x %>%  
 get\_p\_value(obs\_stat = x\_bar,   
 direction = "two-sided") %>%  
 print()

## Warning: Please be cautious in reporting a p-value of 0. This result is an  
## approximation based on the number of `reps` chosen in the `generate()` step. See  
## `?get\_p\_value()` for more information.

## # A tibble: 1 x 1  
## p\_value  
## <dbl>  
## 1 0

## 8.2 t값을 이용한 검정그래프

### t\_cal 계산

t\_cal <- pst\_tb %>%  
 specify(response = 차이) %>%  
 hypothesize(null = "point", # hypothesize 필요  
 mu = 0) %>%   
 calculate(stat = "t") %>% # stat = "t"   
 print()

## Response: 차이 (numeric)  
## Null Hypothesis: point  
## # A tibble: 1 x 1  
## stat  
## <dbl>  
## 1 -3.64

### Bootstrapping을 이용한 귀무가설 분포 생성

set.seed(123)   
null\_dist\_t <- pst\_tb %>%  
 specify(response = 차이) %>%  
 hypothesize(null = "point",   
 mu = 320) %>%  
 generate(reps = 1000,   
 type = "bootstrap") %>%  
 calculate(stat = "t") %>%  
 print()

## Response: 차이 (numeric)  
## Null Hypothesis: point  
## # A tibble: 1,000 x 2  
## replicate stat  
## <int> <dbl>  
## 1 1 1.16   
## 2 2 -1.42   
## 3 3 0.287  
## 4 4 2.00   
## 5 5 3.60   
## 6 6 -0.223  
## 7 7 1.51   
## 8 8 -0.870  
## 9 9 0.114  
## 10 10 0.299  
## # ... with 990 more rows

### 신뢰구간 생성

null\_dist\_ci <- null\_dist\_t %>%  
 get\_ci(level = 0.95,   
 type = "percentile") %>%  
 print()

## # A tibble: 1 x 2  
## lower\_ci upper\_ci  
## <dbl> <dbl>  
## 1 -2.06 2.35

null\_dist\_t %>%  
 visualize(method = "both") + #method = "both": 이론분포+boot분포  
 shade\_p\_value(obs\_stat = t\_cal,  
 direction = "two-sided") +  
 shade\_confidence\_interval(endpoints = null\_dist\_ci)

## Warning: Check to make sure the conditions have been met for the theoretical  
## method. {infer} currently does not check these for you.

