Independent Sample\_t-test

Youngjin\_Lee

2022 2 13

Table of Contents

[독립표본(Independent Sample) t-test 2](#_Toc95668623)

[1.기본 package 설정, library 로드 2](#_Toc95668624)

[2.데이터 불러오기 2](#_Toc95668625)

[3.기술통계분석 3](#_Toc95668626)

[4.그래프 그리기(박스그래프,히스토그램) 4](#_Toc95668627)

[5.이상치 제거(이상치 확인,이상치 제거) 6](#_Toc95668628)

[6.정규분포 검정 7](#_Toc95668629)

[7.등분산 검정 7](#_Toc95668630)

[8.등분산 t-검정 8](#_Toc95668631)

[9.추론(infer)을 이용한 가설검정 및 그래프 9](#_Toc95668632)

[9.1 표본평균(x)을 이용한 검정그래프 9](#_Toc95668633)

[표본평균(x) 계산 9](#_Toc95668634)

[Bootstrapping을 이용한 귀무가설 분포 생성 9](#_Toc95668635)

[신뢰구간 생성 10](#_Toc95668636)

[그래프 그리기 10](#_Toc95668637)

[9.2 t값을 이용한 검정그래프 11](#_Toc95668638)

[t\_cal 계산 11](#_Toc95668639)

[Bootstrapping을 이용한 귀무가설 분포 생성 11](#_Toc95668640)

[신뢰구간 생성 12](#_Toc95668641)

# 독립표본(Independent Sample) t-test

통계분석 차이검정 : 독립표본(Independent Sample) t-test 문제 이교수는 이번에 자동차 타이어를 교체하려고 하는데 수명이 긴 타이어로 교체하려고 한다. 시중에는 A회사의 타이어와 B회사의 타이어가 있는데, 이 교수는 이 중에서 어느 타이어를 골라야 하는가 ?

# 1.기본 package 설정, library 로드

# 2.데이터 불러오기

ist\_tb <- read\_csv('data\\IST.csv',   
 col\_names = TRUE,  
 locale=locale('ko', encoding='euc-kr'), # 한글  
 na=".") %>%  
 round(2) %>% # 소수점 2자리로 반올림  
 mutate\_if(is.character, as.factor) %>%  
 mutate(자동차회사 = factor(자동차회사,  
 levels = c("1","2"),  
 labels = c("A자동차",  
 "B자동차")))

## Rows: 60 Columns: 2  
## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## dbl (2): 자동차회사, 타이어수명  
##   
## 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(ist\_tb)

## spec\_tbl\_df [60 x 2] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ 자동차회사: Factor w/ 2 levels "A자동차","B자동차": 1 1 1 1 1 1 1 1 1 1 ...  
## $ 타이어수명: num [1:60] 48187 47245 51020 50732 52416 ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. 자동차회사 = col\_double(),  
## .. 타이어수명 = col\_double()  
## .. )  
## - attr(\*, "problems")=<externalptr>

ist\_tb

## # A tibble: 60 x 2  
## 자동차회사 타이어수명  
## <fct> <dbl>  
## 1 A자동차 48187  
## 2 A자동차 47245  
## 3 A자동차 51020  
## 4 A자동차 50732  
## 5 A자동차 52416  
## 6 A자동차 49278  
## 7 A자동차 38214  
## 8 A자동차 46742  
## 9 A자동차 48706  
## 10 A자동차 54280  
## # ... with 50 more rows

# 3.기술통계분석

skim(ist\_tb)

Data summary

|  |  |
| --- | --- |
| Name | ist\_tb |
| Number of rows | 60 |
| Number of columns | 2 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| factor | 1 |
| numeric | 1 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: factor**

| skim\_variable | n\_missing | complete\_rate | ordered | n\_unique | top\_counts |
| --- | --- | --- | --- | --- | --- |
| 자동차회사 | 0 | 1 | FALSE | 2 | A자동: 30, B자동: 30 |

**Variable type: numeric**

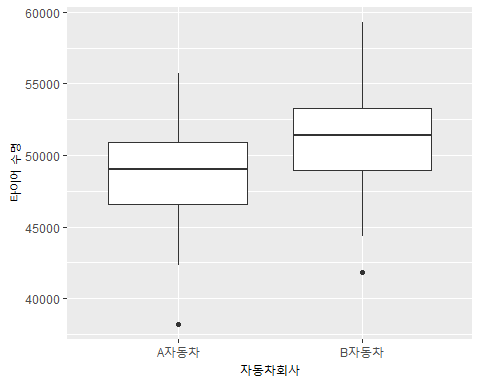
| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 타이어수명 | 0 | 1 | 50024.08 | 4113.28 | 38214 | 47167.5 | 50452 | 52199.75 | 59299 | ▁▃▇▆▃ |

ist\_tb %>%   
 group\_by(자동차회사) %>%  
 get\_summary\_stats(타이어수명)

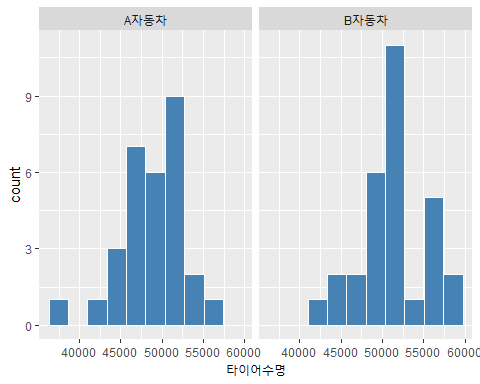
## # A tibble: 2 x 14  
## 자동차회사 variable n min max median q1 q3 iqr mad mean  
## <fct> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 A자동차 타이어수~ 30 38214 55750 49047 46590. 50948 4358. 3274. 48671.  
## 2 B자동차 타이어수~ 30 41852 59299 51396. 48928. 53310. 4381. 3641. 51378.  
## # ... with 3 more variables: sd <dbl>, se <dbl>, ci <dbl>

# 4.그래프 그리기(박스그래프,히스토그램)

ist\_tb %>%   
 ggplot(mapping = aes(x = 자동차회사,  
 y = 타이어수명)) +  
 geom\_boxplot() +  
 labs(y = "타이어 수명")



ist\_tb %>%   
 ggplot(mapping = aes(x = 타이어수명)) +  
 geom\_histogram(bins = 10, # binwidth=1,000: 값차이  
 color = "white",   
 fill = "steelblue") +  
 facet\_wrap(~ 자동차회사) # 그룹별 분리

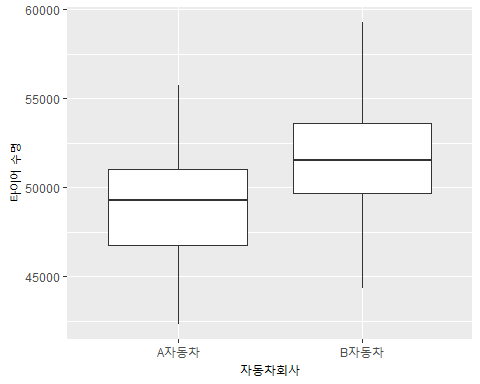


# 5.이상치 제거(이상치 확인,이상치 제거)

ist\_tb %>%  
 group\_by(자동차회사) %>%  
 identify\_outliers(타이어수명)

## # A tibble: 2 x 4  
## 자동차회사 타이어수명 is.outlier is.extreme  
## <fct> <dbl> <lgl> <lgl>   
## 1 A자동차 38214 TRUE FALSE   
## 2 B자동차 41852 TRUE FALSE

ist\_tb <- ist\_tb %>%  
 filter(!(자동차회사 == "A자동차" & 타이어수명 <= 38214 )) %>%  
 filter(!(자동차회사 == "B자동차" & 타이어수명 <= 41852 ))   
  
ist\_tb %>%   
 ggplot(mapping = aes(x = 자동차회사,  
 y = 타이어수명)) +  
 geom\_boxplot() +  
 labs(y = "타이어 수명")



# 6.정규분포 검정

ist\_tb %>%  
 group\_by(자동차회사) %>%  
 shapiro\_test(타이어수명)

## # A tibble: 2 x 4  
## 자동차회사 variable statistic p  
## <fct> <chr> <dbl> <dbl>  
## 1 A자동차 타이어수명 0.992 0.997  
## 2 B자동차 타이어수명 0.965 0.445

# 7.등분산 검정

ist\_tb %>%  
 levene\_test(타이어수명 ~ 자동차회사)

## # A tibble: 1 x 4  
## df1 df2 statistic p  
## <int> <int> <dbl> <dbl>  
## 1 1 56 0.754 0.389

# 8.등분산 t-검정

two-sided test: alternative = c(“two.sided”) right-sided test: alternative = c(“greater”) left-sided test: alternative = c(“less”) 등분산이면 var.equal=TRUE, 이분산이면 var.equal=FALSE

ist\_tb %>%   
 t\_test(formula = 타이어수명 ~ 자동차회사,  
 comparisons = c("A자동차", "B자동차"),  
 paired = FALSE,  
 var.equal = TRUE,  
 alternative = "two.sided",  
 detailed = TRUE)

## # A tibble: 1 x 15  
## estimate estimate1 estimate2 .y. group1 group2 n1 n2 statistic p  
## \* <dbl> <dbl> <dbl> <chr> <chr> <chr> <int> <int> <dbl> <dbl>  
## 1 -2675. 49031. 51706. 타이~ A자동~ B자동~ 29 29 -2.92 0.00503  
## # ... with 5 more variables: df <dbl>, conf.low <dbl>, 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)

ist\_tb %>%   
 cohens\_d(formula = 타이어수명 ~ 자동차회사,   
 var.equal = TRUE)

## # A tibble: 1 x 7  
## .y. group1 group2 effsize n1 n2 magnitude  
## \* <chr> <chr> <chr> <dbl> <int> <int> <ord>   
## 1 타이어수명 A자동차 B자동차 -0.767 29 29 moderate

# 9.추론(infer)을 이용한 가설검정 및 그래프

## 9.1 표본평균(x)을 이용한 검정그래프

### 표본평균(x) 계산

x\_bar <- ist\_tb %>%  
 specify(formula = 타이어수명 ~ 자동차회사) %>% # hypothesize 없음  
 calculate(stat = "diff in means", # diff in means 변경  
 order= c("A자동차", "B자동차")) %>%  
 print()

## Response: 타이어수명 (numeric)  
## Explanatory: 자동차회사 (factor)  
## # A tibble: 1 x 1  
## stat  
## <dbl>  
## 1 -2675.

### Bootstrapping을 이용한 귀무가설 분포 생성

set.seed(123)   
null\_dist\_x <- ist\_tb %>%  
 specify(formula = 타이어수명 ~ 자동차회사) %>%  
 hypothesize(null = "independence") %>%  
 generate(reps = 1000, type = "permute") %>%  
 calculate(stat = "diff in means",   
 order= c("A자동차", "B자동차")) %>%  
 print()

## Response: 타이어수명 (numeric)  
## Explanatory: 자동차회사 (factor)  
## Null Hypothesis: independence  
## # A tibble: 1,000 x 2  
## replicate stat  
## <int> <dbl>  
## 1 1 -587.   
## 2 2 44.3  
## 3 3 -368.   
## 4 4 439.   
## 5 5 626.   
## 6 6 246.   
## 7 7 -971.   
## 8 8 911.   
## 9 9 -183.   
## 10 10 -405.   
## # ... with 990 more rows

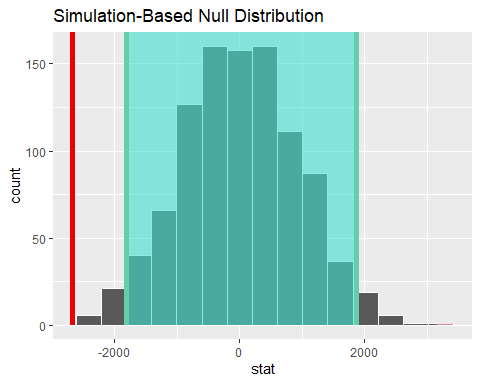
### 신뢰구간 생성

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 -1814. 1859.

### 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

## 9.2 t값을 이용한 검정그래프

### t\_cal 계산

t\_cal <- ist\_tb %>%  
 specify(formula = 타이어수명 ~ 자동차회사) %>% # hypothesize 없음  
 calculate(stat = "t",   
 order= c("A자동차", "B자동차")) %>%  
 print()

## Response: 타이어수명 (numeric)  
## Explanatory: 자동차회사 (factor)  
## # A tibble: 1 x 1  
## stat  
## <dbl>  
## 1 -2.92

### Bootstrapping을 이용한 귀무가설 분포 생성

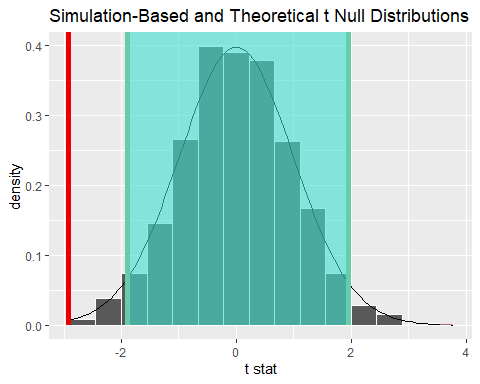
set.seed(123)   
null\_dist\_t <- ist\_tb %>%  
 specify(formula = 타이어수명 ~ 자동차회사) %>%  
 hypothesize(null = "independence") %>%  
 generate(reps = 1000, type = "permute") %>%  
 calculate(stat = "t",   
 order= c("A자동차", "B자동차")) %>%  
 print()

## Response: 타이어수명 (numeric)  
## Explanatory: 자동차회사 (factor)  
## Null Hypothesis: independence  
## # A tibble: 1,000 x 2  
## replicate stat  
## <int> <dbl>  
## 1 1 -0.599   
## 2 2 0.0451  
## 3 3 -0.374   
## 4 4 0.447   
## 5 5 0.639   
## 6 6 0.250   
## 7 7 -0.997   
## 8 8 0.934   
## 9 9 -0.186   
## 10 10 -0.413   
## # ... 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 -1.90 1.95



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.

부록. 비모수통계분석 정규분포검정 p < 0.05일 때 비모수 wilcox.test로 분석

ist\_tb %>%   
 wilcox\_test(formula = 타이어수명 ~ 자동차회사,  
 alternative = "two.sided")

## # A tibble: 1 x 7  
## .y. group1 group2 n1 n2 statistic p  
## \* <chr> <chr> <chr> <int> <int> <dbl> <dbl>  
## 1 타이어수명 A자동차 B자동차 29 29 250 0.00751