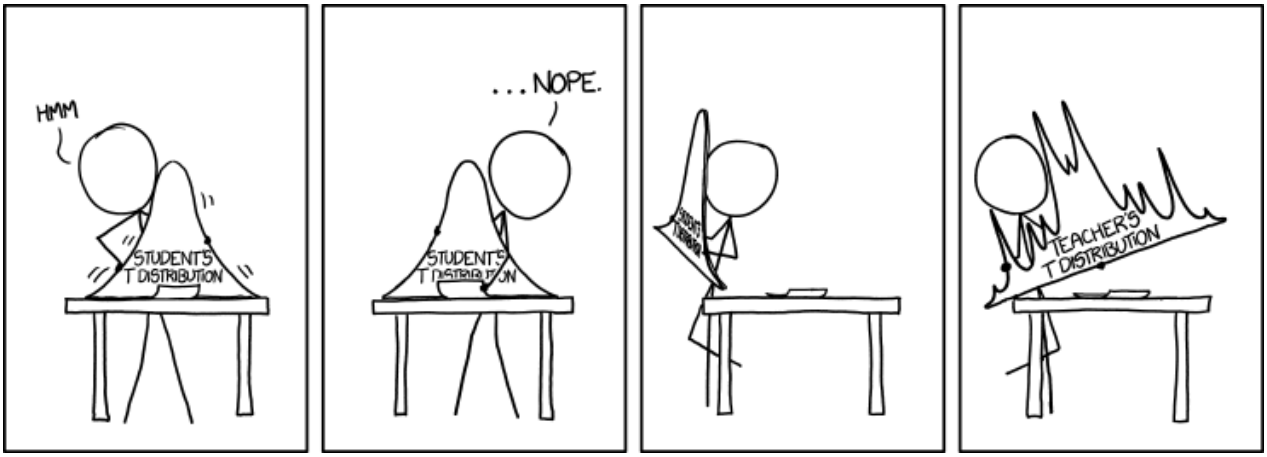
DODOMIRA

데이터 분석하는 게임회사 직원의 이야기

R을 사용한 t-test – 두 그룹 간 평균 차이가 유의미 한 지를 비교해 보자.

예전 포스팅을 통해 3개 이상의 집단이 있을 때 집단 별 차이가 의미가 있는지를 확인해 보는 일원배치 분산분석(ANOVA)에 대해 알아보았습니다.

이번 포스팅에서는 여러 집단 별 차이가 아니라 두 개의 집단 간 차이가 의미가 있는지를 확인해 보는 t-test 를 R에서 실행하는 방법을 정리해 보겠습니다.



t-test에서 사용하는 Student's T-distribution, 소스: xkcd.com

t-test의 유형

t-test는 비교의 대상이 되는 두개의 그룹이 무엇인지에 따라 크게 세가지 유형으로 구분할 수 있습니다.

A. 독립 표본 t-test: 서로 다른 두개의 그룹 간의 평균 비교
(예: 남자와 여자 간 소득의 차이 비교)

B. 대응 표본 t-test : 하나의 집단에 대한 비교

(예: 과외를 하기 전과 후의 반 학생들의 성적 변화)

C. 단일 표본 t-test : 특정 집단의 평균이 어떤 숫자와 같은지 다른지를 비교

t-test의 조건

두개의 집단에 대한 t-test를 실시하기 위해서는 등분산성, 정규성이 만족되어야 합니다.

정규성의 경우 일반적으로 관측 갯수가 30개 이상일때 만족한다고 판단할 수 있습니다.

등분산성을 확인하기 위해서는 var.test라는 함수를 사용하면 됩니다. 한번 실습해 볼까요?

```
a = c(175, 168, 168, 190, 156, 181, 182, 175, 174, 179)
b = c(185, 169, 173, 173, 188, 186, 175, 174, 179, 180)

var.test(a,b)

F test to compare two variances

data: a and b
F = 2.1028, num df = 9, denom df = 9, p-value = 0.2834
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.5223017 8.4657950
sample estimates:
ratio of variances
 2.102784
```

p-value가 0.05보다 작은 경우 두 집단의 분산은 유의미하게 다르다고 볼 수 있습니다. 위 예제에서는 p-value가 0.2834로 0.05보다 크네요. 귀무가설 기각에 실패했으므로 두 집단의 분산은 같다고 볼 수 있겠네요.

독립표본 t-test (independent two sample t-test)

서로 다른 두개의 그룹 간 평균의 차이가 유의미 한지 여부를 판단하기 위한 t-test는 독립표본 t-test라고 말씀드렸죠?

두개의 표본이 “독립”적 이기 위해서는 아래 조건을 만족해야 합니다.

A. 두개의 표본이 서로 관계 없는 모집단에서 추출 되었을 것
B. 표본 간에는 아무런 관계가 없을 것

그럼 mtcars 데이터셋을 사용해서 독립표본 t-test를 해 보겠습니다. mtcars 데이터셋은 1974년 미국에서 자동차 별 가스 마일리지를 측정한 데이터 입니다. 보다 자세한 설명은 위 mtcars링크를 클릭하시거나 R 콘솔에서 ?mtcars 를 입력하면 됩니다.

str과 head 함수를 사용해서 mtcars 데이터셋이 어떻게 생겼는지를 확인해 보죠.

(참고로, 데이터 셋을 R에 로드할 때 str과 head 함수를 사용해서 로드가 잘 되었는지, 데이터 셋이 어떤 모양인지를 확인하는 습관을 들이면 좋습니다.)

```
 str(mtcars)
'data.frame':   32 obs. of  11 variables:
 $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
 $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
 $ disp: num  160 160 108 258 360 ...
 $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
 $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
 $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
 $ qsec: num  16.5 17 18.6 19.4 17 ...
 $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
 $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
 $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
 $ carb: num  4 4 1 1 2 1 4 2 2 4 ...


head(mtcars)
                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

11개의 변수가 있는 32개의 관측 데이터네요. 오늘 사용할 변수는 mtcars$am 변수입니다. 이 변수는 자동차 기어가 오토인지 수동인지를 기록한 변수 입니다. (0=오토, 1=수동)

자동차 기어 종류(오토/수동)에 따른 mpg의 차이가 통계적으로 유의한지 t-test를 통해 확인해 보겠습니다.

우선 두 표본이 등분산성을 만족하는지 확인해 보아야 겠죠?

```
var.test(mtcars[mtcars$am==1,1 ], mtcars[mtcars$am==0, 1])

    F test to compare two variances

data:  mtcars[mtcars$am == 1, 1] and mtcars[mtcars$am == 0, 1]
F = 2.5869, num df = 12, denom df = 18, p-value = 0.06691
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.934280 8.040391
```

```
sample estimates:
ratio of variances
          2.586911
```

p-value가 0.06691로 0.05보다 크므로 등분산성을 만족합니다. 그럼 다음 단계인 t-test단계로 넘어가 보겠습니다.

R에서 독립표본 t-test를 하는 방법은 두가지가 있습니다. 하나는 분석을 원하는 두 집단의 평균을 각각 별개의 벡터 객체로 만들어 입력하는 방법 입니다.

유형 1 문법: t.test(group 1의 관측치, group2의 관측치, t-test 유형, 신뢰범위)

다른 방법은 하나의 데이터 프레임에서 집단을 구분하고자 하는 기준을 입력하는 것입니다.

유형 2 문법: t.test(관측치~집단 구분 기준, 데이터프레임, t-test 유형, 신뢰범위)

mtcars 데이터셋으로 돌아가서, 한번 분석을 실시해 보겠습니다. 독립표본 t-test의 경우 t-test 유형을 var.equal을 TRUE로 지정하면 됩니다. 신뢰범위는 default로 0.95로 지정되어 있으므로 별도로 지정할 필요는 없습니다.

```
t.test(mtcars[mtcars$am==0,1 ], mtcars[mtcars$am==1, 1],  paired =
+            FALSE, var.equal = TRUE, conf.level = 0.95)

    Two Sample t-test

data:  mtcars[mtcars$am == 0, 1] and mtcars[mtcars$am == 1, 1]
t = -4.1061, df = 30, p-value = 0.000285
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -10.84837  -3.64151
sample estimates:
mean of x mean of y
 17.14737  24.39231
```

```
t.test(mpg ~ am, data=mtcars, var.equal=TRUE, conf.level = 0.95)

    Two Sample t-test

data:  mpg by am
t = -4.1061, df = 30, p-value = 0.000285
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -10.84837  -3.64151
sample estimates:
mean in group 0 mean in group 1
       17.14737        24.39231
```

위의 코드 박스는 각각 유형 1, 유형2의 문법을 사용한 t-test입니다.

결과를 해석해 보겠습니다. 우선 가장 아래쪽 집단 별 mpg의평균을 보면 오토는 17.14, 수동은 24.39로 차이가 나는 것 같네요. 이러한 차이가 유의한지를 판단하기 위해서는 p-value를 확인하면 됩니다.

p-value를 확인해 보면 0.001374로 오토와 수동 자동차의 mpg차이는 유의하네요.

대응표본 t-test (Paired sample t-test)

대응표본 t-test는 동일한 집단의 전-후 차이를 비교하기 위해 사용됩니다.

예를 들어 초콜렛을 하루 30g씩 섭취하는 것이 수면 시간에 영향을 미치는지 여부나, 과외를 받는 것이 학교 성적에 영향을 미치는지 등등 특정 변인의 영향을 측정하기 위해 주로 사용되죠.

주의할 점은 대응 표본은 실험 전-후를 비교하는 것이기 때문에 입력하는 관측치의 수가 반드시 같아야 합니다.

중간고사 이후 과외를 받은 10명의 학생의 중간고사 – 기말고사 점수 데이터를 가상으로 만들어서 비교해 보겠습니다.

```
mid = c(16, 20, 21, 22, 23, 22, 27, 25, 27, 28)
final = c(19, 22, 24, 24, 25, 25, 26, 26, 28, 32)
t.test(mid,final, paired=TRUE)

    Paired t-test

data:  mid and final
t = -4.4721, df = 9, p-value = 0.00155
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.0116674 -0.9883326
sample estimates:
mean of the differences
                     -2
```

p-value가 0.00155로 과외를 받은 전과 이후의 평균 성적 차이는 통계적으로 유의미하다고 말할 수 있겠네요.

단일표본 t-test (One sample t-test)

단일 표본 t-test는 하나의 집단의 평균이 특정 기준보다 유의미하게 다른지 혹은 큰지/작은지를 알아보는 분석 방법입니다. 아래와 같은 문법을 사용합니다.

문법: t.test(관측치, alternative = 판별 방향, mu=특정기준, conf.level = 신뢰수준)

alternative에는 “greater”, “less”, “two.sided”가 있습니다. 각각 큰지/작은지/같은지를 구분하라는 명령입니다.

위의 중간-기말고사 점수 데이터를 계속 사용해 보겠습니다. 기말고사의 평균 점수는 25.1점이네요. 이 학생들의 기말고사 점수가 24점보다 유의하게 높은지를 확인해 볼까요?

```
mean(final)
[1] 25.1
```

```
t.test(prem, alternative="greater", mu=24, conf.level = .95)

    One Sample t-test

data:  prem
t = 1.0093, df = 9, p-value = 0.1696
alternative hypothesis: true mean is greater than 24
95 percent confidence interval:
 23.10218      Inf
sample estimates:
mean of x
     25.1
```

p-value가 0.1696으로 0.05보다 크므로 귀무가설을 기각할 수 없습니다. 95% 신뢰 수준에서 학생들의 기말고사 성적은 24점보다 높다고 말할 수 없겠네요.

23점 기준으로 확인해 볼까요? p-value가 0.05보다 낮으므로 학생들의 기말고사 성적은 23점보다는 유의하게 높다고 말할 수 있겠습니다.

```
t.test(prem, alternative="greater", mu=23, conf.level = .95)

    One Sample t-test

data:  prem
t = 1.9269, df = 9, p-value = 0.04305
alternative hypothesis: true mean is greater than 23
95 percent confidence interval:
 23.10218      Inf
sample estimates:
mean of x
     25.1
```

지금까지 두개 집단의 평균 차이가 유의한지를 분석하는 t-test를 알아보았습니다.

세개 이상의 집단의 평균을 비교하는 ANOVA 분석과 함께 t-test를 알아두면 평균의 차이

가 날 때 이 차이가 유의미한지 아닌지를 자신있게 판단할 수 있겠죠?

참고 자료

1) Cookbook for R, t-test

2) R-tutorial, t.test

3) R-bloggers, two sample Student's t-test #1

4) Statistical Data Analysis R, 집단간 평균 분석

관련

R을 사용한 일원배치 분산분석 (ANOVA in R) - 집단 별 차이가 통계적으로 의미가 있는지 검증해 보자.
2월 24, 2016
"Data Analysis"에서

R을 사용한 다중회귀분석 (Multiple regression in R)
1월 31, 2016
"Data Analysis"에서

데이터 분석의 유형 6가지 - 목적에 따라 달라지는 분석 방법
1월 12, 2016
"Data Analysis"에서

4월 2, 2016 Mira Jang Data Analysis, R ANOVA, mtcars, R, t-test, 단일표본 t-test, 대응표본 t-test, 데이터분석, 독립표본 t-test, 등분산, 통계, 평균 비교