# 통계학 입문

**2021-05-16**

**데이터과학융합스쿨**

**손 대 순 교수**

앞서 배웠던 t-test에 대해서 R로 구현하는 방법을 해 봅시다. 이는 한 개나 두 개의 서로 다른 군 간에 평균을 비교하는 검정 방법이예요. A학교 학생 100명, B 학교 학생 100명을 표본으로 시험 성적으로 조사하여 데이터를 확보하고, 이를 이용해서 A, B 학교의 시험 성적 평균에 차이가 있는지를 통계적 논리로 테스트하는 작업을 t-test라고 합니다. 이 과정은 통계학에서 중요한 개념이예요. 모든 통계적 검정이라는 것이 같은 개념을 사용하고 있으므로, 이 t-test를 예로 확실히 이해해 두면, 다른 방법들로 확장하기 수월합니다.

t-test는 크게 3가지로 나눌 수 있습니다.

1) one-sample t-test

   이는 모집단이 1개인 경우를 말합니다. 그리고, 내가 알고 있는 모평균이 있는 경우죠. 예를 들어 우리나라 사람의 키 평균이 165라고 알고 있는데, 1000명의 표본을 뽑아 키를 측정하였더니 표본 평균이 170이었다. 모평균이 165가 맞나? 하는 생각이 들죠. 이를 검정하는 방법이 one-sample t-test입니다.

2) paired t-test

   이는 두 개의 모집단인 것처럼 보이지만 사실상 하나의 모집단을 테스트하는 방법이예요.^^ 말이 이상하죠? 예를 들어보면 이해가 쉬울겁니다. 어떤 약을 먹거나, 어떤 치료를 받은 환자를 대상으로 하는 경우가 많아요. 혈압강하제라고 해 보죠. 1000명의 고혈압 환자를 대상으로 혈압강하제를 투약하여 투약 전과 후의 혈압을 측정하였다고 해 봅시다. 이 혈압강하제는 혈압을 떨어뜨리는 효과가 있을까? 하는 것이 관심사겠죠. 투약 전 혈압 데이터가 있고, 투약 후의 혈압 데이터가 있으니 마치 두 개의 모집단이 있는 것처럼 보입니다. 하지만, 이는 동일 환자에서 (투약 후 혈압)-(투약 전 혈압) 데이터를 가지고 분석하는 것이죠. 이 "혈압 차이"가 0인가? 에 대한 통계적 검정이 됩니다. 따라서, one-sample t-test과 같은 형태가 되지요. 데이터의 형태가 전/후로 나뉘어서 수집되는 특징 때문에 paired t-test라는 이름이 붙었습니다.

3) 2-sample t-test

모집단이 두 개인 경우죠. 앞서 예를 든 대로 A학교와 B학교의 성적 비교가 이에 해당합니다.

오늘은 위의 3가지 검정법을 R에서 어떻게 수행하는지를 배워 보겠습니다. 여러분들이 충실히 실습을 하면서 익히기 바랍니다.

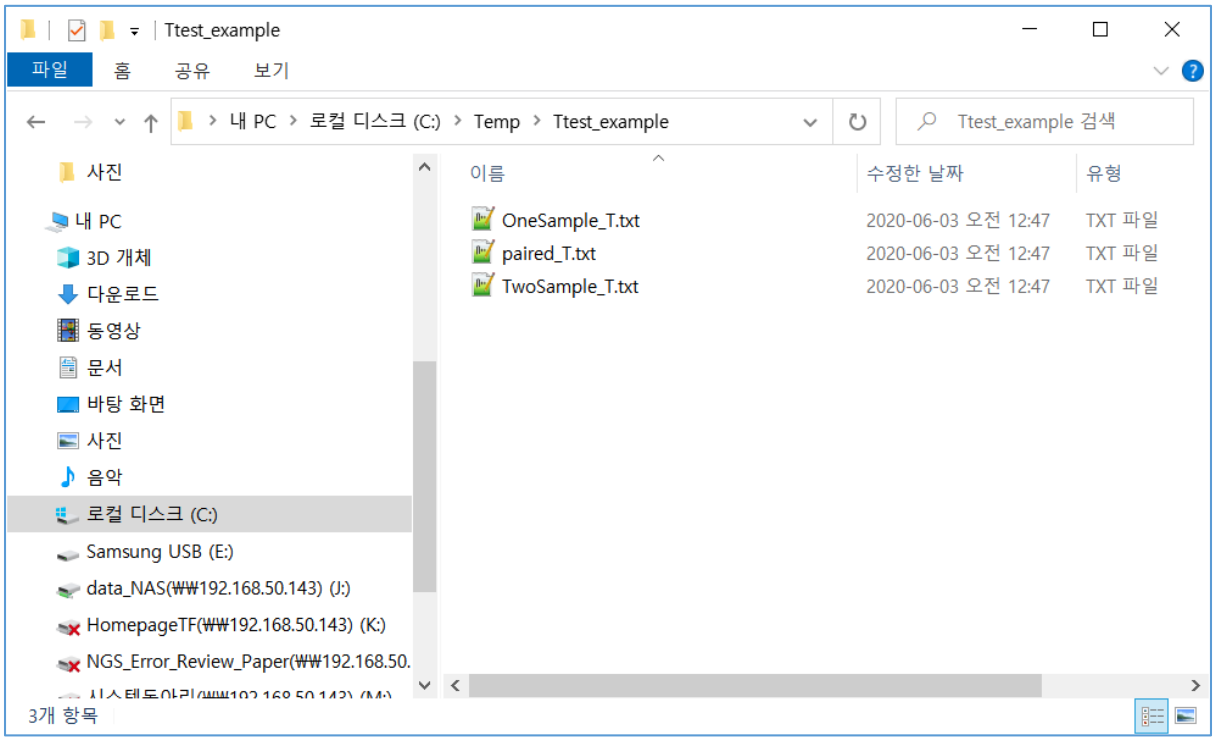
강의교안과 함께 첨부한 압축 파일을 받아서, 적절한 폴더에 풀어주세요. 이 폴더를 WorkingDirectory라고 합시다. 그리고, R에서 다음을 실행시켜 봅시다.

```
dat <- read.delim("WorkingDirectory/OneSample_T.txt", header=T)
View(dat)
t.test(dat$mmhg, mu=140)
```

위의 내용이 one-sample t-test입니다.

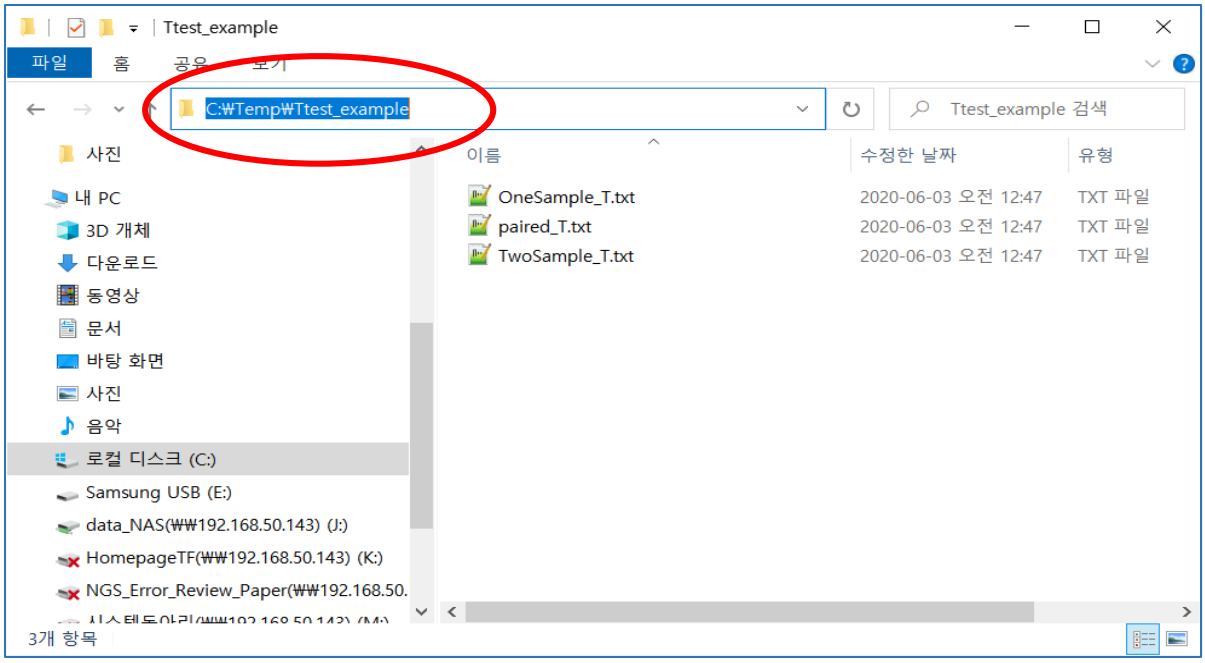
첫번째 줄은 데이터를 불러오는 과정이예요. read.delim 이라는 명령어를 사용하면 R 외부에 있는 데이터를 불러올 수 있습니다. "Working directory"는 데이터를 넣어 놓은 폴더의 경로를 말합니다. 여기에서 주의할 부분이 있어요.

데이터의 경로는 보통 탐색기의 상단 주소를 복사해서 가져오면 됩니다.




데이터가 들어있는 경로를 탐색기로 들어가면 위와 같은 화면이겠지요. 여기에서 주소 부분을 클

릭하면 복사할 수 있습니다.




만일 이 경로를 사용한다면 위의 명령어는 아래와 같이 입력해야 합니다.

dat <- read.delim("C:/Temp/Ttest_example/OneSample_T.txt", header=T)

윈도우에서 폴더 표시를 '₩'로 하는데 R에서는 '/'로 입력해야 합니다. 간혹, 쌍따옴표("")가 폰트에 따라 R에서 에러가 나는 경우도 있어요. 특히, 외부에서 명령어를 복사해 오면 그런 문제가 많이 발생하니, 가급적 직접 키보드로 입력하는 것이 좋습니다. 그리고 또 한가지, 경로에는 한글이 포함되지 않도록 하는 것이 필요합니다. 한글 경로도 일반적으로는 사용가능하지만, 늘 문제가 될 때에는 한글 때문인 경우가 태반입니다. 따라서, 경로명에 한글이 포함되지 않도록 하는 습관을 가지는 것이 좋습니다.

명령어의 두번째 줄 View(dat)는 dat라는 데이터셋을 보여달라는 명령입니다. 이 명령을 실행하면 데이터셋의 형태로 테이블이 보일겁니다. mmhg라는 변수명으로 55개의 데이터가 있는 구조라는 것을 확인할 수 있을 겁니다.

마지막 줄의 명령어가 사실상 one-sample t-test의 명령어 입니다. t.test라는 함수로 이를 수행할 수 있고, dat$mmhg라는 것은 dat 데이터셋에 mmhg라는 변수를 의미합니다. 데이터셋 안에 있는 특정 변수를 지칭하는 방법이예요. 이를 이용해서 히스토그램이나 박스그림을 그릴 수 있습니다. 한 번 해 보세요. hist(dat$mmhg), boxplot(dat$mmhg)와 같이 사용하면 됩니다.

mu=140이라는 것은 모평균이 140인지를 검정하라는 말입니다.

one-sample t-test를 하면 다음과 같은 결과가 나옵니다.

```
        One Sample t-test

data:  bp$mmhg
t = -3.8693, df = 54, p-value = 0.0002961
alternative hypothesis: true mean is not equal to 140
95 percent confidence interval:
 124.8185 135.1815
sample estimates:
mean of x
      130
```

t값과 자유도, p-value, 그리고, 지난 시간에 배웠던 95% 신뢰구간이 결과로 나오는 군요. t값과 p-value도 알아서 계산해 줍니다. 참 편리하죠? 데이터의 개수가 55개이니, 자유도는 54로 적혀 있네요. 데이터를 가지고 모평균이 어디에 있을지 95% 신뢰도로 신뢰구간을 계산해 본 결과도 제시되어 있습니다. 결론적으로 p-value가 5%보다 작다면 대립가설(alternative hypothesis)을 채택합니다 (reject H0). 대립가설이 무엇일까요? 위에 적혀 있습니다. "true mean is not equal to 140". 모집단의 평균은 140이 아니라는 것이죠. 명령어를 줄 때, mu=140이라는 옵션을 통해서 "모집단의 평균이 140이라고 할 수 있는가?"하고 물은 것이고, 이에 대한 답변으로 p-value가 0.05보다 작으므로 "유의수준 5%에서 모집단의 평균은 140이라고 할 수 없다"는 답을 얻은 것이죠. 이것이 검정입니다. 명령어를 통해서 질문을 하고, 답을 받는 것이라 할 수 있겠네요.

paired t-test를 한 번 해 볼까요?

```
dat2 <- read.delim("WorkingDirectory/paired_T.txt", header=T)
View(dat2)
t.test(dat2$before, dat2$after, paired=T)
```

데이터를 불러들이고 나면, 반드시 데이터의 구조를 잘 살펴보아야 합니다. View 명령으로 살펴보면, before, after, diff 세 개의 변수가 있군요. 이 분석에서는 before, after 두 개의 변수만 사용합니다.

동일하게 t.test라는 함수를 쓰고, 변수는 2개를 모두 써 줍니다. 그리고, paired=T라는 옵션을 줌으로 paired t-test를 수행하라는 명령을 주는 것이죠. 결과 역시 간단합니다.

```
        Paired t-test

data:  after and before
t = -1.8628, df = 14, p-value = 0.08361
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -9.7529438  0.6862772
sample estimates:
mean of the differences
              -4.533333
```

이는 앞서 언급한 대로, before와 after의 차이가 0이라고 할 수 있는가?라는 질문에 대한 답입니다. 역시 t값과 p-value가 계산되고, 신뢰구간도 나오는 군요.

데이터를 살펴 보면 diff라는 변수가 있었지요. 이는 (after-before) 값을 계산해 놓은 것입니다. paired t-test는 사실상 one sample t-test라고 했으니, 정말 그런지 한 번 확인해 봅시다.

```
t.test(dat2$diff, mu=0)
```

이는 앞서 해 보았던 one-sample t-test의 명령어입니다. 다만 사용하는 변수가 dat2라는 이름의 데이터셋의 diff 변수를 이용하는 것이고, 모평균이 0인지를 검정하라는 의미가 되겠지요. 결과를 한 번 살펴 보세요.

```
        One Sample t-test

data:  dat2$diff
t = -1.8628, df = 14, p-value = 0.08361
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -9.7529438  0.6862772
sample estimates:
mean of x
-4.533333
```

paired t-test와 동일한 결과지요? paired t-test는 2개의 변수를 사용하니, 2개의 집단인듯 보이지만 결국 '차이'를 이용하는 one-sample t-test와 같은 것이죠. 다만, paired 데이터를 가지고 있을 때, one-sample t-test를 이용하기 위해서는 'diff'값을 계산해서 변수로 만들어 주는 과정이 필요하겠지만, paired t-test 명령어를 이용하면 R이 알아서 계산해 주는 장점이 있다는 것이죠. 위의 paired t-test 명령어에서는 diff 변수를 사용하지 않았으니까요.

마지막으로 two-sample t-test를 한 번 해 보겠습니다. 방법이 크게 다르지 않아요.

```
dat3 <- read.delim("WorkingDirectory/TwoSample_T.txt", header=T)
View(dat3)
t.test(dat3$smoker, dat3$nonsmoker)
```

모집단이 두 개인 경우도 위와 같이 수행하면 아래와 같은 결과를 얻을 수 있습니다. 흡연자군과 비흡연자군, 두 개의 군이 있는 데이터이고, 각 군에서의 측정값들이 기록되어 있는 데이터입니다. 두 군의 모집단에 측정값의 평균차이가 있다고 할 수 있는지, two-sample t-test의 과정이예요.

```
        Welch Two Sample t-test

data:  dat$smoker and dat$nonsmoker
t = 3.1702, df = 26.479, p-value = 0.003826
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.772020 8.291329
sample estimates:
mean of x mean of y
 17.34706  12.31538
```

자, 수행 방법이 간단하죠? 역시, 검정통계량과 p-value, 신뢰구간을 계산해 주고, 각 군의 평균이 얼마인지도 아래쪽에 계산이 되어 있습니다. 간단하게 계산되는 방법론이지만, 중요한 통계적 개념들이 들어 있습니다. 이를 이해하고 있어야, 제대로 해석할 수 있는 것이죠.

t-test를 통해서, 통계적 검정방법의 원리를 설명하고, 그에 따른 용어들도 정리했어요. R로 어떻게 구현할 수 있는지도 간략하게 살펴보았습니다.

어렵지만, 조금 더 이해할 수 있는 시간이 되면 좋겠네요.^^

오늘도 수고 많았습니다.