R을 이용한 통계 기초와 데이터 분석

Lecture 5

남현진

한성대학교

2020

T test

두 집단의 평균의 차이 유무를 비교하는 테스트이다. Z test의 경우 모집단의 표준편차를 알고 있어야 하지만 현실에서는 모집단에 대한 정확한 평균이나 분산에 대한 통계량을 알지 못하는 경우가 대부분이다. 따라서 Z 검정 보다는 조금 더 일반적인 경우에도 이용할 수 있는 T 검정을 많이 사용한다.

T test 가정

Student T 검정을 위해서는 두가지 가정이 만족되어야한다. T test에서도 여러가지 종류가 있는데 만약 등분산성 가정이 만족되면 Student T test를 사용하고 등분산성이 만족되지 않는다면 Welch T-Test를 이용해야 한다.

Assumptions

- 정규성 가정 (Normality Assumption)
- 등분산성 가정 (Homogeneity of variance) (2 표본일 경우)

정규성 가정

일반적으로 표본의 수가 30개 이상이면 중심극한 정리에 의해 검정 없이도 정규성을 가정할 수 있다. 하지만 표본의 개수가 30개 이하라면 Shapiro–Wilk test,Pearson's chi-squared test 등의 방법을 이용하여 표본이 정규분포를 따름을 증명해야한다. 알파를 0.05라고 두면 p-value가 0.05보다 크면 자료가 정규성을 만족한다고 말할 수 있다.

Assumptions

- 귀무가설 H0: 자료는 정규분포를 따른다.
- 대립가설 H1: 자료는 정규분포를 따르지 않는다.

등분산성 가정

등분산성은 분석하는 두 표본의 분산이 같다는 뜻이다. 만약 두 표본의 분산이 같다면 Student T test를 이용하고 두 표본의 분산이 같지 않다면 Welch T test를 이용해야 한다. R 에서 Student T test는 t.test 함수의 var.equl = F를 통해 사용할 수 있고 Student T Test는 var.equl = T를 통해 사용할 수 있다. 알파를 0.05라고 두면 p-value가 0.05보다 크면 자료가 등분산성을 만족한다고 말할 수 있다.

Assumptions

- 귀무가설 H0: 두 자료의 분산은 같다.
- 대립가설 H1: 두 자료의 분산은 같지 않다.

일표본 T 검정

어떤 전구의 지속 시간이 39000분으로 알려져 있다고 한다. 이 때 75개의 전구 표본을 뽑아 지속시간을 알아보니 평균 36500분과 표준편차 2000분으로 나타났다고 한다. 이 때 이 전구는 평균 지속시간이 39000분이라 말할 수 있을까?

- 귀무가설 H0: 전구의 지속시간은 39000분이다.
- 대립가설 H1: 전구의 지속시간은 39000분이 아니다.

일표본 T 검정

우선 평균이 36500 이고 표준편차가 2000인 75개의 난수를 생성해보자. 이 때 생성된 난수 집단의 평균이 39000과 같은지 T 검정을 해보려 한다.

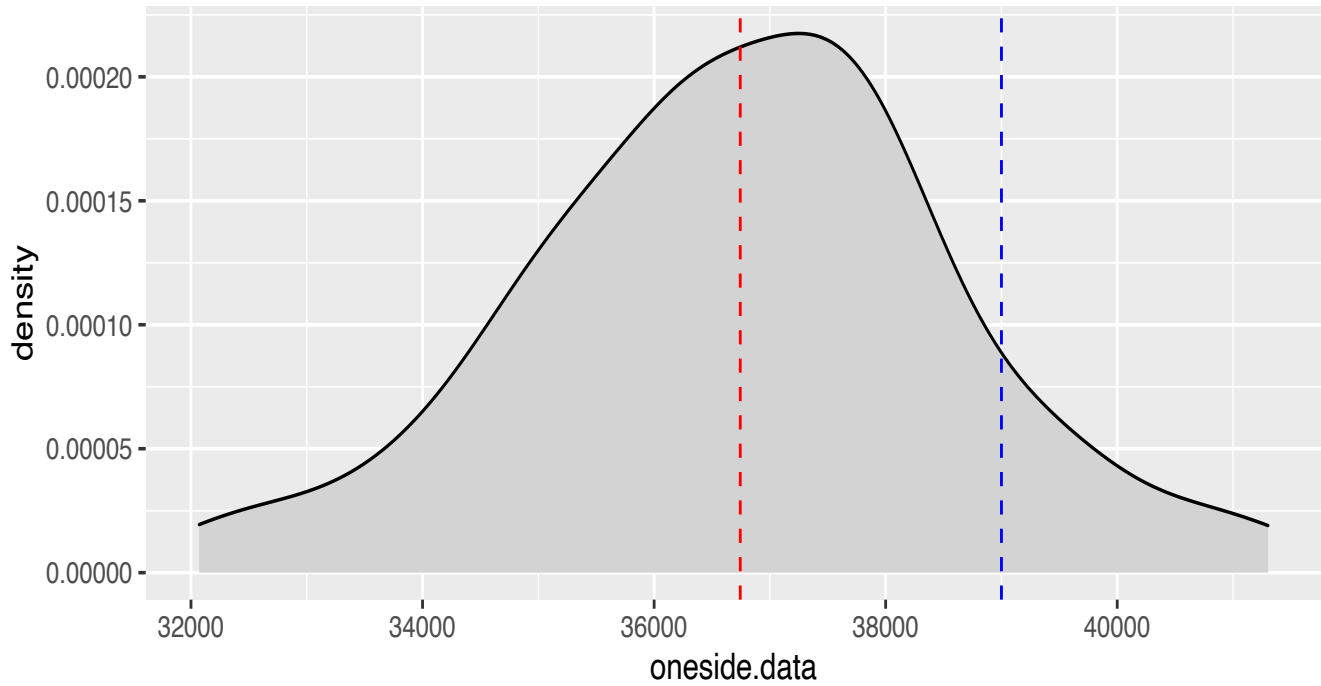
```
set.seed(1)
oneside.data <- c(rnorm(75, mean = 36500, sd = 2000))
head(oneside.data)
```

```
## [1] 35247.09 36867.29 34828.74 39690.56 37159.02 34859.06
```

일표본 T 검정

생성한 난수의 density plot은 다음과 같다. 빨간색 선은 난수 집단의 평균을 뜻하고 파란색 선은 비교대상인 39000이다.

```
ggplot() +
  geom_density(mapping = aes(x = oneside.data, y = ..density..), fill = "lightgray") +
  geom_vline(xintercept = mean(oneside.data), color = "red", linetype = "dashed") +
  geom_vline(xintercept = 39000, color = "blue", linetype = "dashed")
```



일표본 T 검정

```r
t.test(oneside.data, mu = 39000)
```

```
##
##  One Sample t-test
##
## data:  oneside.data
## t = -10.565, df = 74, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 39000
## 95 percent confidence interval:
##  36318.76 37169.64
## sample estimates:
## mean of x
##   36744.2
```

일표본 T 검정

Result

- t-test의 검정통계량 t = -10.565이다.
- 자유도 df는 74이다.
- 유의확률 p-value는< 2.2e-16 이다.
- 전구의 지속 시간의 95 % 신뢰구간은 [36318.76 37169.64]이다.
- 표본의 평균은 36744.2이다.

p value는 < 2.2e-16로 0.05보다 작다. 따라서 전구의 지속시간은 39000분이라는 귀무가설을 기각한다.

독립표본 T 검정

9명의 여성 집단과 9명의 남성 집단이 있다. 모든 사람들의 몸무게를 측정했을 때 남성 집단의 몸무게의 평균과 여성 집단의 몸무게 평균을 비교해보려 한다. 이 때 남성 집단의 몸무게가 여성 집단의 몸무게와 같다고 말할 수 있을까?

- 귀무가설 H0: 남성과 여성의 몸무게는 같다.
- 대립가설 H1: 남성 집단과 여성 집단의 몸무게는 다르다.

독립표본 T 검정

남성과 여성의 몸무게가 다음과 같다고 가정하고 데이터를 만들어 보자.

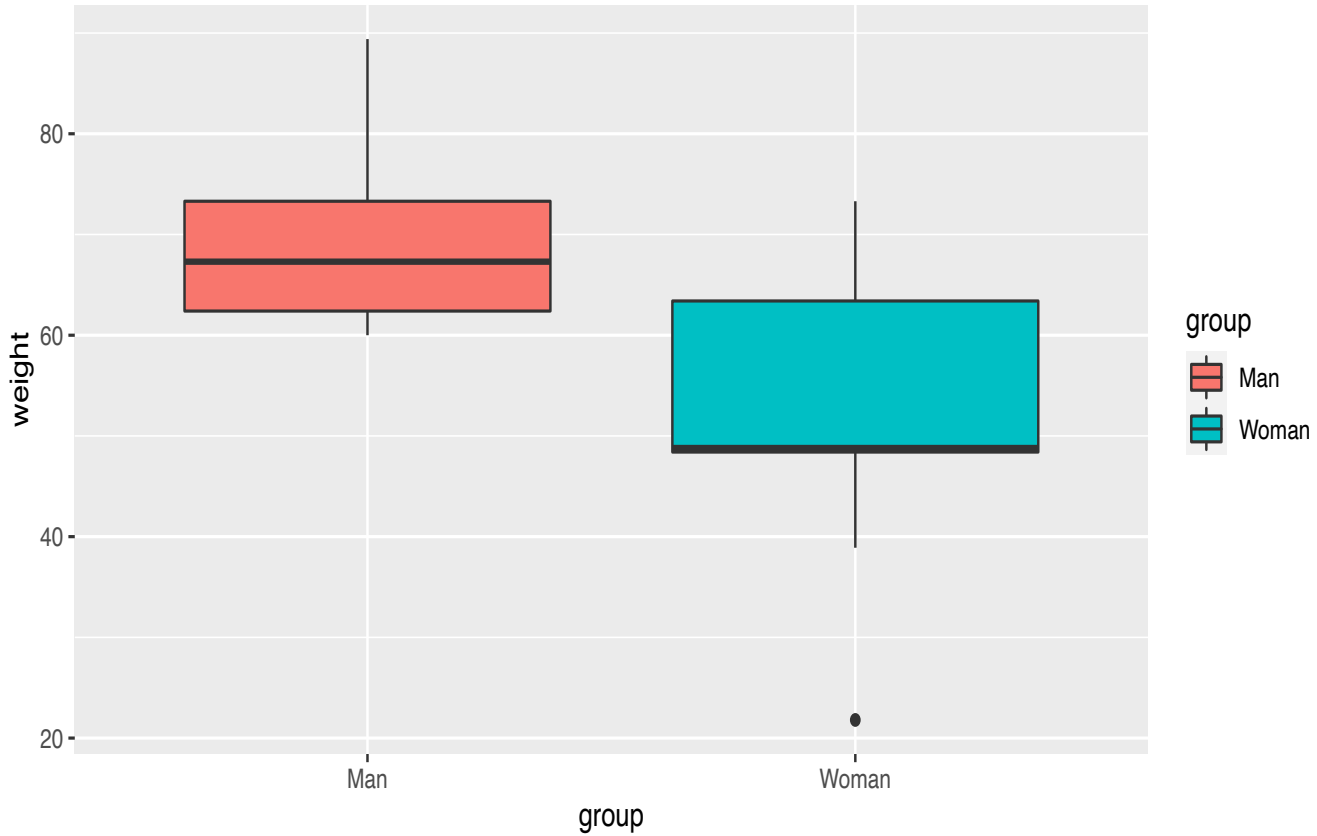
```
# Data in two numeric vectors
women.weight <- c(38.9, 61.2, 73.3, 21.8, 63.4, 64.6, 48.4, 48.8, 48.5)
men.weight <- c(67.8, 60, 63.4, 76, 89.4, 73.3, 67.3, 61.3, 62.4)
# Create a data frame
weight.data <- data.frame(
                group = rep(c("Man","Woman"), each = 9),
                weight = c(men.weight,  women.weight)
                )
str(weight.data)
```

```
## 'data.frame':    18 obs. of  2 variables:
##  $ group : chr  "Man" "Man" "Man" "Man" ...
##  $ weight: num  67.8 60 63.4 76 89.4 73.3 67.3 61.3 62.4 38.9 ...
```

독립표본 T 검정

weight.data 데이터의 상자그림은 다음과 같다.

```
library(ggplot2)
ggplot(data = weight.data) +
  geom_boxplot(mapping = aes(x = group, y = weight, fill = group ))
```



독립표본 T 검정

우선 두 그룹의 표본이 정규성을 따르는지를 확인해야한다. 정규성 검사는 shapiro.test 함수를 이용해 Shapiro-Wilk test를 실행한다. 귀무가설은 자료가 정규분포를 따른다는 것이다.남성 집단의 경우 p-value가 0.1066으로 0,05를 넘기 때문에 정규분포를 따른다고 말할 수 있다. 즉 남성 집단의 분포는 정규성을 만족한다.

```
shapiro.test(men.weight)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  men.weight
## W = 0.86425, p-value = 0.1066
```

독립표본 T 검정

여성 집단 또한 동일하게 정규성 검사를 실행한다. 여성 집단의 경우 또한 p-value가 0.6101으로 0,05를 넘기 때문에 정규분포를 따른다고 말할 수 있다. 즉 여성 집단의 분포는 정규성을 만족한다.

```
shapiro.test(women.weight)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  women.weight
## W = 0.94266, p-value = 0.6101
```

독립표본 T 검정

남성 집단과 여성 집단의 등분산성이 만족되는지 확인한다. p-value는 0.17 는 0.05 수준보다 높다. 따라서 두 분산의 차이가 통계적으로 유의한 차이가 없다고 말할 수 있다. 따라서 두 집단은 등분산성을 만족한다.

```
var.test(women.weight, men.weight)
```

```
##
##  F test to compare two variances
##
## data:  women.weight and men.weight
## F = 2.7675, num df = 8, denom df = 8, p-value = 0.1714
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##   0.6242536 12.2689506
## sample estimates:
## ratio of variances
##           2.767478
```

독립표본 T 검정

```
t.test(women.weight, men.weight, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  women.weight and men.weight
## t = -2.7842, df = 16, p-value = 0.01327
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -29.748019  -4.029759
## sample estimates:
## mean of x mean of y
##  52.10000  68.98889
```

독립표본 T 검정

Result

- t-test의 검정통계량 t = -2.784이다.
- 자유도 df는 16이다다.
- 유의확률 p-value는 0.01327이다.
- 두 집단의 평균 차이의 95% 의 신뢰구간은 [-29.748019,-4.029759]이다.
- 두 집단의 평균은 다음과 같다. $\bar{x} = 52.1$, $\bar{y}$ =68.99

p value는 0.01327로 0.05보다 작다. 따라서 남성의 평균 몸무게는 여성의 평균 몸무게와 같다라는 귀무가설을 기각하고 두 집단의 몸무게가 다르다라고 결론내릴 수 있다.

대응표본 T 검정

어떤 학교의 1000명의 학생들의 중간고사 수학 성적의 평균은 64점 표준편차는 10이라고 알려져 있다. 이 때 이 학교에 새로운 교육 방식이 적용되었고 기말고사를 치르게 되었다. 이 때 기말고사 수학 성적의 평균은 68점 표준편차는 20점이라고 했을 때 새로운 교육 방식 이후의 학생들의 점수가 통계적으로 유의미하게 올랐다고 말할 수 있을까?

- 귀무가설 H0: 새로운 교육 방식의 전과 후의 학생들의 성적은 같다.
- 대립가설 H1: 새로운 교육 방식의 적용 후 학생들의 성적은 상향되었다.

대응표본 T 검정

우선 시험 전과 시험 후의 학생들의 데이터를 만들어 보자.

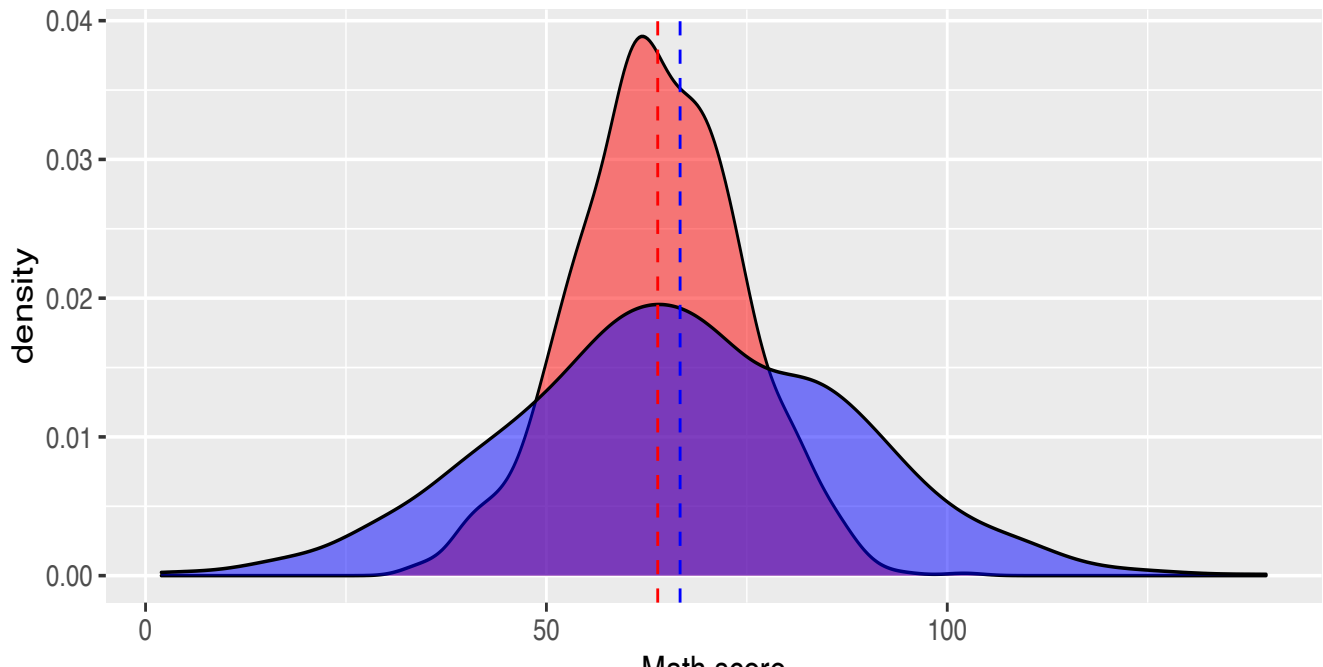
```
set.seed(1)
pre.edu <- c(rnorm(1000, mean = 64, sd = 10))
post.edu <- c(rnorm(1000, mean = 67, sd = 20))

math.data <- data.frame(pre.edu, post.edu)
```

대응표본 T 검정

math.data 데이터의 상자그림은 다음과 같다.

```
ggplot() +
  geom_density(mapping = aes(x = pre.edu, y = ..density..), fill = "red", alpha = 0.
  geom_density(mapping = aes(x = post.edu, y = ..density..), fill = "blue", alpha =
  geom_vline(xintercept = mean(pre.edu), color = "red", linetype = "dashed") +
  geom_vline(xintercept = mean(post.edu), color = "blue", linetype = "dashed") +
  labs(x = 'Math score')
```



대응표본 T 검정

```
t.test(pre.edu, post.edu, paired = TRUE, var.equal = FALSE, alternative = "less")
```

```
##
##  Paired t-test
##
## data:  pre.edu and post.edu
## t = -3.8091, df = 999, p-value = 7.4e-05
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##     -Inf -1.5848
## sample estimates:
## mean of the differences
##               -2.791243
```

대응표본 T 검정

Result

- t-test의 검정통계량 t = -3.8091이다.
- 자유도 df는 999이다.
- 유의확률 p-value는 7.4e-05이다.
- 시험 전과 후의 점수 차이의 95% 의 신뢰구간은 [-Inf,-1.5848]이다.
- 두 집단의 평균의 차이는 -2.791243 이다.

p value는 7.4e-05로 0.05보다 작다. 따라서 새로운 교육 방식의 전과 후의 학생들의 성적은 같다라는 귀무가설을 기각하고 새로운 교육 방식의 적용 후 학생들의 성적은 상향되었다고 결론내릴 수 있다.