Statistics Lab 033.020
Seoul National University

Assignment #5

2016-19516, Sangjun Son

Example 1

(textbooks.txt) 주어진 자료는 UCLA 내의 서점과 Amazon.com 에서 판매되는 교재들의 가격에 대한 자료이다. 2010년 봄학기에 개설된 UCLA의 강의 중에서 73개를 선택하여 각 강의에 쓰이는 교재의 온라인 판매 가격 (amazNew) 과 오프라인의 판매 가격 (uclaNew)을 조사하였다. 교재의 판매가격은 판매 장소 (온라인 또는 오프라인)에 따라 차이가 난다고 볼 수 있는가? 적절한 가설을 세우고 유의수준 5%에서 이를 검정하시오.

```
textbooks = read.table("textbooks.txt", header=T)
# 대응 비교: 기각
t.test(textbooks$amazNew, textbooks$uclaNew, paired=T)
```

```
----------------------------------------
        Paired t-test

data:  textbooks$amazNew and textbooks$uclaNew
t = -7.6488, df = 72, p-value = 6.928e-11
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -16.087652  -9.435636
sample estimates:
mean of the differences
              -12.76164
----------------------------------------
```

Explanation: 대응비교는 실험단위를 동질적인 쌍으로 묶고, 각 쌍에 두 처리를 랜덤하게 적용한 다음, 각 쌍에서 얻은 관측값의 차로 두 모평균의 차인 $\mu_1 - \mu_2 = \delta$에 대한 추론 문제를 다루는 방법이다. 이때 귀무가설 $H_0 : \delta = 0$와 대립가설 $H_1 : \delta \neq 0$을 설정할 수 있다. textbook 데이터의 amazNew와 uclaNew에 대해 대응비교 검정 (paired=T)을 한 결과 유의확률은 6.928e-11로 유의수준 5% 내에서 귀무가설을 기각할 수 있다. 즉, 교재의 판매가격은 판매 장소 (온라인 또는 오프라인)에 따라 차이가 난다.

Example 2

(run10samp.txt) 주어진 자료는 2012년 Washington, DC에서 열렸던 Cherry Blossom 10 mile run 경기에서 완주를 한 선수 100명의 자료이다. 주요 변수에 대한 설명은 다음과 같다.

변수	설명
time	10 마일 달리기 완주 기록 (분)
age	선수 나이
gender	성별 (M=남성, F=여성)
state	출신 지역 (또는 국가)

성별에 따른 완주시간은 차이가 있는가? 적절한 가설을 세우고 유의수준 5%에서 이를 검정하시오.

```
run10samp = read.table("run10samp.txt", header=T)
tapply(run10samp$time, run10samp$gender, mean)

# 등분산 여부 검정: 기각X (등분산)
var.test(time ~ gender, data=run10samp)
# 등분산 가정에서의 독립 이표본 평균 검정: 기각
t.test(time ~ gender, data=run10samp, var.equal=T, alternative="greater")
```

```
        F         M
102.13491  87.64533
```

```
----------------------------------------
        F test to compare two variances

data:  time by gender
F = 1.4781, num df = 54, denom df = 44, p-value =
0.1833
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.8292103 2.5909362
sample estimates:
ratio of variances
           1.47806
----------------------------------------
        Two Sample t-test

data:  time by gender
t = 5.1173, df = 98, p-value = 7.719e-07
alternative hypothesis: true difference in means between group F and group M is greater than 0
95 percent confidence interval:
 9.787749      Inf
sample estimates:
mean in group F mean in group M
      102.13491        87.64533
----------------------------------------
```

Explanation: 이표본에 의한 모평균의 비교는 두 모집단의 분포의 등분산 가정 여부에 따라 검정이 달라진다. 독립 이표본 검정에서 등분산 가정 여부는 두 모분산에 대한 검정 결과를 이용할 수 있다. 먼저 성별에 따른 완주시간의 평균을 비교하기 위해 tapply() 함수를 사용하였고 여성의 완주시간 평균보다 남성의 경우 더 낮은 값을 확인하였다.
따라서 여성, 남성의 평균 μ_1, μ_2에 대한 귀무가설 $H_0 : \mu_1 - \mu_2 = 0$와 대립가설 $H_1 : \mu_1 - \mu_2 > 0$을 세울 수 있다. 먼저 등분산 여부를 위한 검정 var.test()을 진행하고 유의확률 0.1833으로 유의수준 5%에서 기각을 하지 못하여 등분산을 가정할 수 있었다. 등분산 가정에서의 독립 이표본 평균 검정 결과, 귀무가설을 기각, 즉 마일 달리기 완주 기록은 남성이 여성의 기록보다 짧다고 할 수 있다.

Example 3

서울대입구의 한 안과에 양쪽 눈의 시력이 다른 문제를 호소하며 병원을 찾는 환자들이 유독 늘었다고 한다. 병원에서는 이런 환자들의 양쪽 시력이 유의수준 5%에서 정말로 다른지가 궁금하여, 때마침 병원을 찾은 학생 A와 학생 B에게 검정을 의뢰하였다. 환자들 중 열 명의 샘플을 추출하였고 시력은 다음과 같다. `left`는 왼쪽 눈, `right`는 오른쪽 눈의 시력을 측정한 결과이다. 자료는 정규분포 가정을 만족한다고 한다.

`left`	1.8	0.2	1.8	1.8	0.9	1.0	0.4	0.2	0.5	2.9
`right`	2.4	0.5	2.1	2.2	1.4	1.5	1.0	0.5	1.1	3.2

(a) 학생 A는 대응 비교를, 학생 B는 등분산 가정에서의 독립 이표본 평균 검정을 실행하였다. 두 학생이 각각 1) 설정한 가설, 2) 검정을 진행한 결과, 3) 최종적으로 내렸을 결론을 서술하시 오.

(b) 만약 내가 안과 의사라면 두 학생 중 누구의 결론을 신뢰할 것인가? 그 이유는 무엇인가?

```r
left  = c(1.8, 0.2, 1.8, 1.8, 0.9, 1.0, 0.4, 0.2, 0.5, 2.9)
right = c(2.4, 0.5, 2.1, 2.2, 1.4, 1.5, 1.0, 0.5, 1.1, 3.2)

```

```r
# 대응 비교: 기각
t.test(left, right, paired=T)
# 등분산 가정에서의 독립 이표본 평균 검정: 기각X
t.test(left, right, var.equal=T)
```

```
----------------------------------------
        Paired t-test

data:  left and right
t = -10.307, df = 9, p-value = 2.779e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.5365658 -0.3434342
sample estimates:
mean of the differences
                 -0.44

----------------------------------------
        Two Sample t-test

data:  left and right
t = -1.1119, df = 18, p-value = 0.2808
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.271392  0.391392
sample estimates:
mean of x mean of y
     1.15      1.59

----------------------------------------
```

Explanation: ***(a)*** *학생 A는 대응 비교를 하기 위해 설정한 통계량 $D = L - R$ (L: 왼쪽 눈의 시력, R: 오른쪽 눈의 시력)에 대해 1) 귀무가설 $H_0 : \delta = 0$와 대립가설 $H_1 : \delta \neq 0$의 검정을 하였고 2) 유의 확률 2.779e-06로 유의수준 5% 기각할 수 있다. 즉, 3) 환자들의 양쪽 시력이 다르다고 할 수 있다. 학생 B는 등분산 가정에서의 독립 이표본 평균 검정을 하기 위해 1) 귀무가설 $H_0 : \mu_L = \mu_R$와 대립가설 $H_1 : \mu_L \neq \mu_R$을 설정하고 검정 결과 2) 유의 확률 0.2808로 유의수준 5% 기각할 수 없다. 즉, 3) 환자들의 양쪽 시력이 다르다고 할 수 없다.*
(b) *대응비교를 한 A의 결론을 신뢰한다. left와 right의 시력은 각각의 사람마다 추출한 값으로 각 쌍의 차이를 비교하는데 필요한 검정은 대응비교이다. 학생 B의 결론을 채택하기 위해서는 위의 표가 10명에 대한 양안 시력을 측정한 값이 아닌 20명에 대해 랜덤으로 왼쪽/오른쪽 눈을 선택하고 한쪽 눈의 시력 대한 데이터에 적합할 것이다.*

Example 4

3장 과제에서 사용하였던 `iris` 데이터셋을 분석하려고 한다. `data('iris')`의 코드를 실행하여 데이터셋을 불러올 수 있으며, 붓꽃의 세 가지 품종 (Species) 중 setosa, virginica의 두 종류만을 분석에 사용할 것이다. setosa의 꽃받침 길이(Sepal.Length)가 virginica의 꽃받침 길이보다 짧다고 할 수 있을까? 적절한 가설을 세우고, 유의수준 5%에서 이를 검정하시오.

```r
data(iris)
iris = iris[(iris$Species=="setosa" | iris$Species=="virginica"),]
tapply(iris$Sepal.Length, iris$Species, mean)

# 등분산 여부 검정: 기각 (이분산)
var.test(Sepal.Length ~ Species, data=iris)
# 이분산 가정에서의 독립 이표본 평균 검정: 기각
```

Assignment #5

2016-19516, Sangjun Son

```
8  t.test(Sepal.Length ~ Species, data=iris, var.equal=F, alternative="less")

     setosa versicolor  virginica
      5.006         NA      6.588
-----------------------------------------
        F test to compare two variances

data:  Sepal.Length by Species
F = 0.30729, num df = 49, denom df = 49, p-value =
6.366e-05
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1743776 0.5414962
sample estimates:
ratio of variances
         0.3072862
-----------------------------------------
        Welch Two Sample t-test

data:  Sepal.Length by Species
t = -15.386, df = 76.516, p-value < 2.2e-16
alternative hypothesis: true difference in means between group setosa and group virginica is less than 0
95 percent confidence interval:
      -Inf -1.410804
sample estimates:
   mean in group setosa mean in group virginica
                  5.006                  6.588
-----------------------------------------
```

Explanation: iris 데이터를 로드하고 이 중 Species가 setosa 또는 virginica에 대한 데이터 행들을 추출하여 다시 데이터셋을 구성 (iris)한다. 먼저 품종 (Species)에 따른 꽃받침 길이(Sepal.Length) 평균을 비교하기 위해 tapply() 함수를 사용하였고 이를 통해 setosa의 꽃받침 길이가 virginica의 꽃받침 길이보다 짧다는 사실을 유추하였다. setosa와 virginica의 꽃받침 길이 각각의 평균 μ_s, μ_v에 대한 귀무가설 $H_0 : \mu_s - \mu_v = 0$와 대립가설 $H_1 : \mu_s - \mu_v < 0$을 세울 수 있다. 이표본에 의한 모평균의 비교는 두 모집단의 분포의 등분산 가정 여부에 따라 검정이 달라지므로 등분산 검정 var.test()을 진행하였고 유의확률 6.366e-05으로 유의수준 5%에서 기각을 하여 등분산이 아닌 이분산을 가정하였다. 이분산 가정에서의 독립 이표본 평균 좌측 검정 t.test(var.equal=F, alternative="less") 결과, 유의확률 2.2e-16, 유의수준 5%에서 귀무가설을 기각, 즉 setosa의 꽃받침 길이(Sepal.Length)가 virginica의 꽃받침 길이보다 짧다고 할 수 있다.