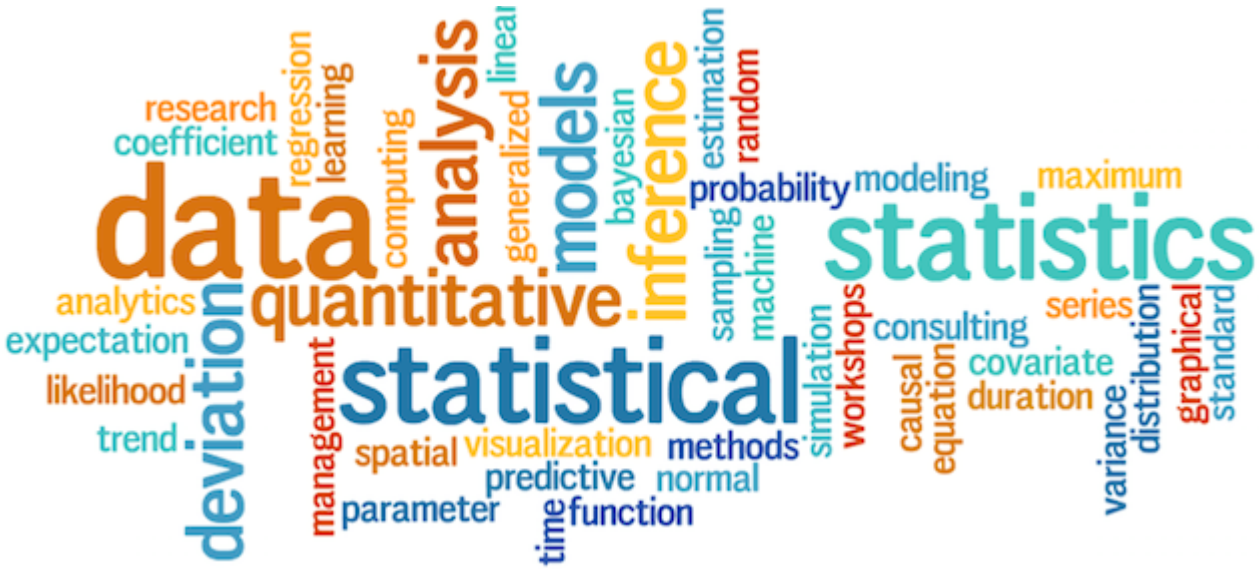


통계기반 데이터 분석 (Ch 6)

데이터사이언스 & A.I 정화민 교수

평균비교

모집단이 두 개인 경우

- 두 모집단의 종류
- 서로 독립인 두 집단에서의 평균 차이 검정
- 서로 대응인 두 집단에서의 평균 차이 검정

모집단이 세 개 이상

- 분석방법 : 일원분산분석
- 분산분석표

독립표본 t검정

모집단이 두 개인 경우

- **모집단이 두 개인 경우는 '서로 독립인 두 집단'과 '대응을 이루는 두 집단'**
- **서로 독립인 두 집단 : 독립표본**

- 각 집단을 변수에 의해 두 개로 구분할 때 서로 영향을 끼치지 않는 집단입니다.

예) 성별 변수에 의해 나눠진 남자 아이와 여자 아이의 몸무게

여아	3837	3334	2208	1745	2576	3208	3746	3523	3430	3480
	3116	3428	2184	2383	3500	3866	3542	3278		
남아	3554	3838	3625	2846	3166	3520	3380	3294	3521	2902
	2635	3920	3690	3783	3345	3034	3300	3428	4162	3630
	3406	3402	3736	3370	2121	3150				

Source : 제대로 알고 쓰는 통계분석

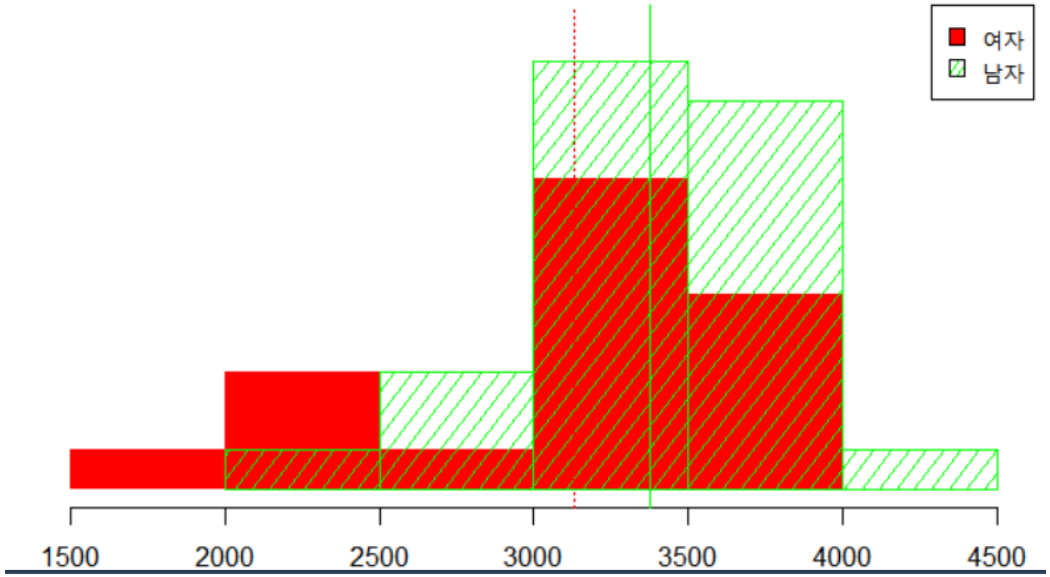
독립표본 t검정

모집단이 두 개인 경우

히스토그램코딩

```
data <- read.table("D:/Line_삼성공유폴더/서강대학교 정보통신대학원/2020년_통계기반데이터분석/age3.txt", header=T)

par(mar=c(2, 1, 1, 1))
hist(data$weight[data$gender==1], xlim=c(1500, 4500), ylim=c(0, 12), col="red", border=NA, main="", xlab="", ylab="", axes=F)
hist(data$weight[data$gender==2], density=10, angle=45, add=TRUE, col="green")
axis(1)
abline(v = mean(data$weight[data$gender==1]), lty=3, lwd=1.5, col="red")
abline(v = mean(data$weight[data$gender==2]), lty=1, lwd=1.5, col="green")
legends = c("여자", "남자")
legend("topright", legend=legends, fill=c("red", "green"), density=c(NA, 20))
```



Source : 제대로 알고 쓰는 통계분석 수정사용

독립표본 t검정

모집단이 두 개인 경우

- 서로 독립인 두 모집단은 정규분포를 이룬다.

'정규성'이라 하며, 이를 만족하는지 검정해야 합니다. 본 책에서는 '정규성'은 만족하는 것으로 가정.

- 두 집단의 분산은 서로 동일하다.

'등분산성'이라 하며, R의 분산 비교 검정함수를 이용하여 분산이 서로 동일한지 검정

⇒ 독립인 두 모집단의 분산의 동질성 검증을 실시한다.

엑셀로 먼저 확인 후 R 코딩 (1반수학점수와 2반 수학점수간의 평균차이 검정)

검정 순서 : 두 집단의 분산의 동질성 검정 (p값 기준 0.05 미만 : 비 등분산 , 0.05 이상 : 등분산)

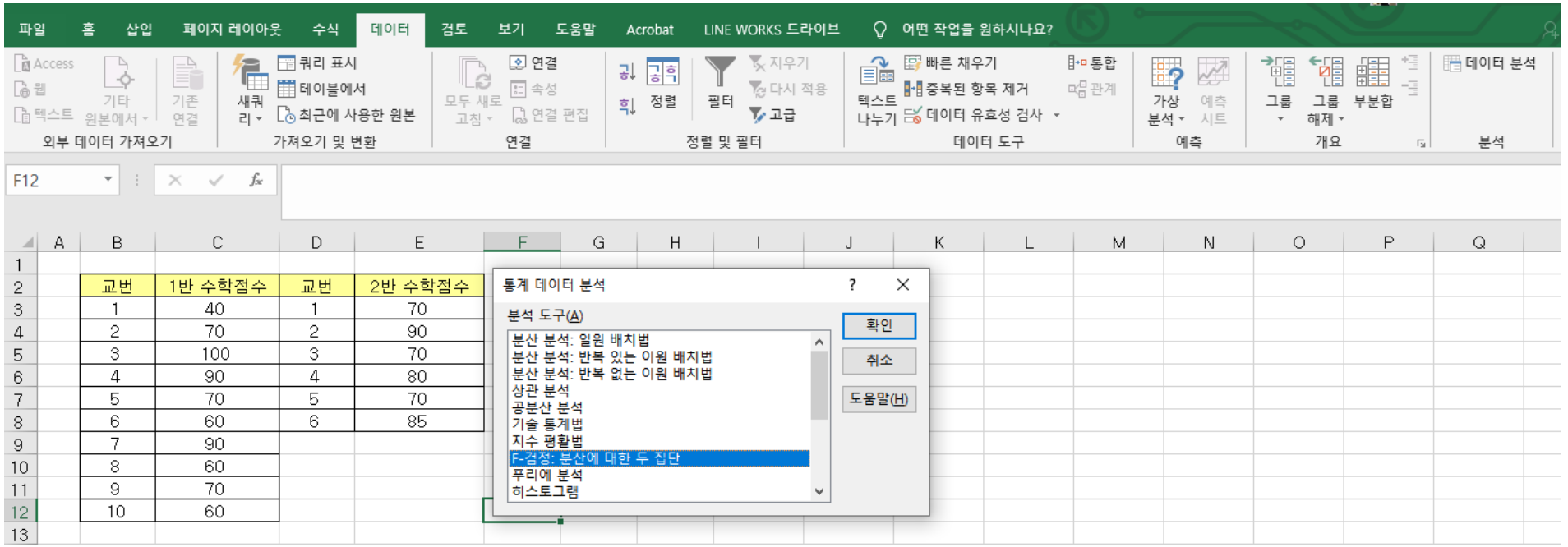
가설 검정 : 독립된 두 집단으로 두 집간 간의 인원수도 다르고 수학 평균도 다르다 .

통계적으로 평균의 차이가 있는 지를 검증)

가설설정 : 연구가설은 ?

독립표본 t검정(등분산검정)

모집단이 두 개인 경우



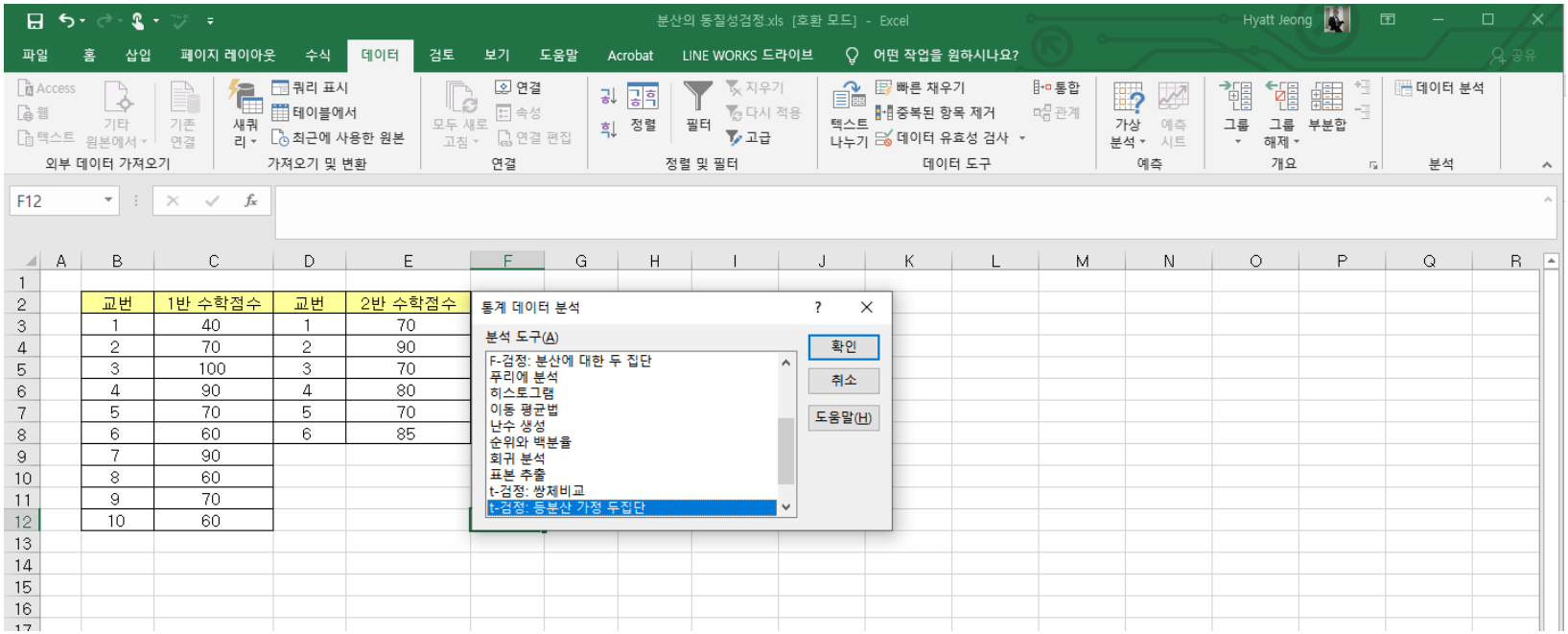
교번	1반 수학점수	교번	2반 수학점수
1	40	1	70
2	70	2	90
3	100	3	70
4	90	4	80
5	70	5	70
6	60	6	85
7	90		
8	60		
9	70		
10	60		

F-검정: 분산에 대한 두 집단

	1반 수학점수	2반 수학점수
평균	71	77.5
분산	321.1111111	77.5
관측수	10	6
자유도	9	5
F 비	4.143369176	
P(F<=f) 단측 검정	0.065958468	0.131916935
F 기각치: 단측 검정	4.772465613	

독립표본 t검정(가설검정)

모집단이 두 개인 경우



t-검정: 등분산 가정 두 집단

	1반 수학점수	2반 수학점수
평균	71	77.5
분산	321.1111111	77.5
관측수	10	6
공동(Pooled) 분산	234.1071429	
가설 평균차	0	
자유도	14	
t 통계량	-0.822662419	
P(T<=t) 단측 검정	0.212245526	
t 기각치 단측 검정	1.761310136	
P(T<=t) 양측 검정	0.424491051	
t 기각치 양측 검정	2.144786688	

독립표본 t검정(R 분석)

모집단이 두 개인 경우

```r
data <- read.table("D:/Line_삼성공유폴더/서강대학교 정보통신대학원/2020년_통계기반데이터분석/age3.txt", header=T)
var.test(data$weight ~ data$gender)
# 등분산
t.test(data$weight ~ data$gender, mu=0, alternative="less", var.equal=TRUE)
#비등분산
t.test(data$weight ~ data$gender, mu=0, alternative="less", var.equal=FALSE)
```

① 성별로 몸무게가 결정됨을 나타내는 수식 '몸무게 ~ 성별'을 전달.

② Mu는 모집단의 평균을 나타냄 (연구가설)

③ alternative는 대안가설에 따라 결정됩니다. (양쪽검정, 왼쪽 한쪽 검정, 오른쪽 한쪽 검정)

④ var.equal을 통해 분산의 동일성 여부를 전달

- TRUE이면 동일한 분산(등분산), FALSE이면 서로 다른 분산(비등분산 = 이분산).

독립표본 t검정(R 분석)

모집단이 두 개인 경우

```
        F test to compare two variances

data:  data$weight by data$gender
F = 2.1771, num df = 17, denom df = 25, p-value = 0.07526
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.9225552 5.5481739
sample estimates:
ratio of variances
          2.177104

> # 등분산
> t.test(data$weight ~ data$gender, mu=0, alternative="less", var.equal=TRUE)

        Two Sample t-test

data:  data$weight by data$gender
t = -1.5229, df = 42, p-value = 0.06764
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
     -Inf 25.37242
sample estimates:
mean in group 1 mean in group 2
       3132.444        3375.308
```

대응표본 t검정(R 분석)

모집단이 두 개인 경우

[3] 검정방법에 적합한 검정 통계량 결정과 p-값을 산출한다.

주어진 데이터를 가지고 검정 통계량 계산에 요구되는 표본 평균 혹은 표본 분산 등을 R을 포함한 다양한 통계 패키지를 활용하여 계산하고, 검정 통계량의 분포를 확인하여 p-값을 계산한다.

1. 데이터 분석 예시에 대한 p-값 산출

(1) 1-1절에서 살펴보았던 데이터 분석 예시 데이터 불러오기

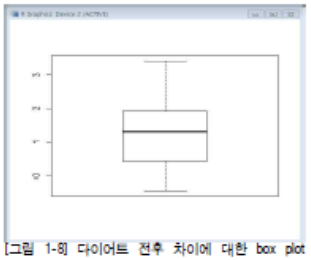
예를 들어 예시 데이터가 "diet.csv" 파일형식으로 [그림 1-1]과 같이 저장되어 있다고 가정하자.

	A	B	C
1	Subject	Before	After
2	1	57.9	58.31
3	2	64.68	61.27
4	3	66.3	66.43
5	4	59.97	58.39
6	5	74.12	73.71
7	6	72.71	71.72
8	7	72.5	70.59
9	8	68.36	66.83
10	9	78.86	76.84
11	10	49.24	48.17

[그림 1-7] Diet.csv 파일 저장내용 예시

```
> # "diet.csv"파일에 있는 내용을 가져온다.
> data <- read.csv("diet.csv",header=T)
> attach(data)
> names(data)
[1] "Subject" "Before"  "After"
```

```
> # "diet.csv" 파일에 있는 내용을 가져온다.
> diff <- Before - After
> diff
 [1] -0.41  3.41 -0.13  1.58  0.41  0.99  1.91  1.53  2.02  1.07
> # "diet.csv" 파일에 있는 내용을 가져온다.
> boxplot(diff)
```



[그림 1-8] 다이어트 전후 차이에 대한 box plot (boxplot()함수 수행결과)

```
> # T 통계량 계산
> mean_diff <- mean(diff)
> mean_diff
[1] 1.238
> sd_diff <- sd(diff)
> sd_diff
[1] 1.122772
> t_stat <- mean_diff/(sd_diff/sqrt(length(diff)))
> t_stat
[1] 3.486815
```

만약 양측검증을 수행하는 경우 다음과 같은 명령어를 수행한다.

```
>    t.test(Before,    After,    alternative=c("two.sided"),    paired=TRUE,
conf.level=0.95)

        Paired t-test

data:  Before and After
```

Source : NCS 통계기반 데이터분석

대응표본 t검정(R 분석)

모집단이 두 개인 경우

```r

data <-read.csv("diet.csv", header =T)
attach(data)
names(data)
summary(data)
#Before와 after의 차이확이 확인. 가설을 설정해 보자 귀무가설, 대립가설
diff <- Before -After
diff
#diet.csv 파일에 있는 내용을 boxplot로 그린다. 다이어트전후 차이. 최저값 최저값 중앙값, 3사분위 표시
boxplot(diff)
# 대응표본t test를 해보자
t.test(Before, After, alternative=c("two.sided"), paired=TRUE,
       conf.level=0.95)

```

Source : NCS 통계기반 데이터분석