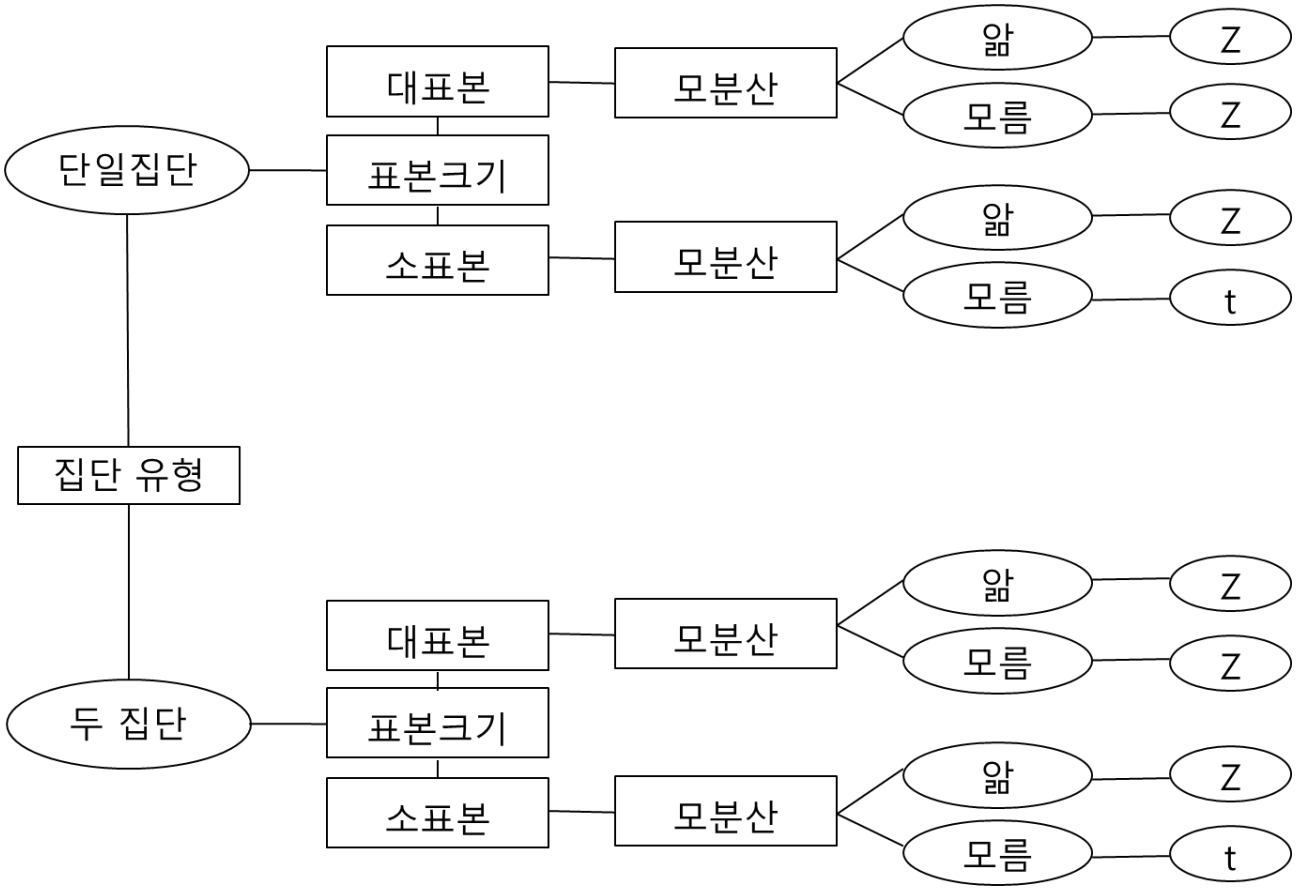
R 가설검정

I. 단일집단 가설검정

II. 두 집단 가설검정

1. 의사결정 트리

- 모평균에 대한 가설검정은 단일집단이든 두 집단이든, 모분산을 알든 모르든 관계없이
  - 표본의 크기가 30 이상의 대표본이면 Z-통계량을 이용
  - 표본의 크기가 30 미만의 소표본이면 t-통계량을 이용



2. 모평균에 대한 가설검정

- (연습 1) 어떤 아이스크림 회사의 영업부 사원은 체인점의 여름 판매량보다 겨울 판매량이 평균 34.5%감소한다고 한다. 전국 체인점 중 15개를 표본추출하여 판매 감소량을 조사해 보니 다음과 같이 나타났다. 모집단이 정규분포한다고 가정하고 판매량의 감소가 평균 34.5%라는 귀무가설을 10% 유의수준에서 양측검정하라.

33.46	33.38	32.73	32.15	33.99	34.10	33.97	34.34
22.95	33.85	34.23	34.05	34.13	34.45	34.19	

b1-ch7-2-rev.R

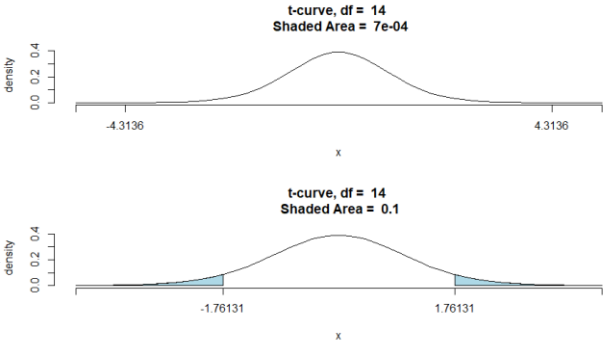
```
library(foreign)
library(tigerstats)
time<-read.dta(file="http://kanggc.iptime.org/book/data/chap11-2-1.dta")
time$var1
(n<-length(time$var1))
(m<-mean(time$var1))
(s<-sd(time$var1))
(tc<-(m-34.5)/(s/sqrt(n)))
t.test(time$var1, mu=34.5, conf.level=0.9)
(t14<-qt(0.05, df=14, lower.tail=T))
(t14<-qt(0.05, df=14, lower.tail=F))
par(mfrow=c(2,1))
ptGC(c(-4.3136,4.3136),region="outside",df=14, graph=T)
ptGC(c(-1.76131,1.76131),region="outside",df=14, graph=T)
```

```
> (n<-length(time$var1))
[1] 15

> (m<-mean(time$var1))
[1] 33.798

> (s<-sd(time$var1))
[1] 0.6302968

> (tc<-(m-34.5)/(s/sqrt(n)))
[1] -4.313578
```




```
> t.test(time$var1, mu=34.5, conf.level=0.9)

        One Sample t-test

data:  time$var1
t = -4.3136, df = 14, p-value = 0.0007145
alternative hypothesis: true mean is not equal to 34.5
90 percent confidence interval:
 33.51136 34.08464
sample estimates:
mean of x
   33.798
```

```
> (t14<-qt(0.05, df=14, lower.tail=T))
[1] -1.76131

> (t14<-qt(0.05, df=14, lower.tail=F))
[1] 1.76131
```

- 계산된 검정통계량의 값은 -4.3136이고 10% 유의수준 하에서 자유도가 14인 t-분포의 임계치는 -1.7631과 1.7631이므로 10% 유의수준 하에서 모평균이 34.5라는 귀무가설을 기각

3. 모분산에 대한 가설검정

- (연습 2) 한 기술연구소에서 휴대전화 배터리 무게의 분산이 62g이라는 주장에 대한 양측검정을 하려고 한다. 휴대전화 7개를 무작위 선정하여 조사한 무게가 다음과 같으며, 휴대전화 배터리 무게는 정규분포한다고 하자. 5% 유의수준에서 휴대전화 배터리 무게의 분산이 62g이라는 귀무가설을 양측검정하라.

36	37	38	39	39	44	47

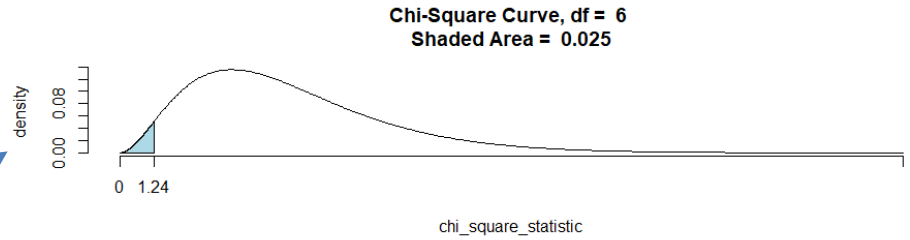
b1-ch7-3-rev.R

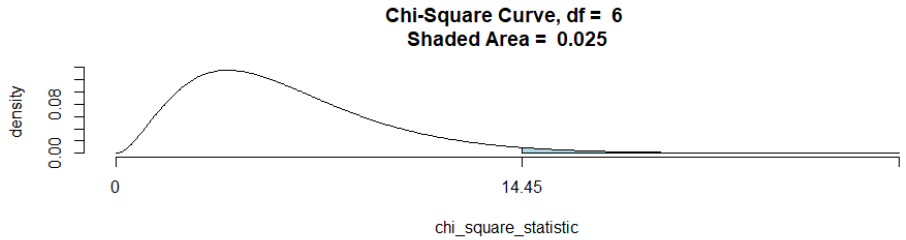
```
library(tigerstats)
x<-c(36,37,38,39,39,44,47)
xbar<-mean(x)
xbar
s.sq<-var(x)
s.sq
df=6
q=length(x)-1
chi<-(df*s.sq)/62
chi
pchisq(chi, df=q,lower.tail=F)
UCV<-qchisq(0.025, df=q, lower.tail=F)
LCV<-qchisq(0.975, df=q, lower.tail=F)
UCV;LCV
par(mfrow=c(2,1))
pchisqGC(1.237,region="below",df=6, gr
aph=T)
pchisqGC(14.45,region="above",df=6, gr
aph=T)
```

```
> chi
[1] 1.548387

> pchisq(chi, df=q,lower.tail=F)
[1] 0.9562158

> UCV<-qchisq(0.025, df=q, lower.tail=F)
> LCV<-qchisq(0.975, df=q, lower.tail=F)
> UCV;LCV
[1] 14.44938
[1] 1.237344
```





- 계산된 검정통계량의 값은 1.548387이고 5% 유의수준 하에서 자유도가 6인 χ^2-분포의 임계치는 1.237344와 14.44938이므로 5% 유의수준 하에서 모분산이 62g이라는 귀무가설을 허용

1. 모평균에 대한 가설검정
(1) 두 집단의 분산을 모르지만 같다고 가정할 경우

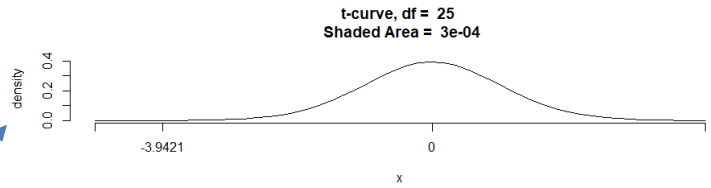
- (연습 3) 대기업과 중소기업에 근무하는 근로자의 직장 만족도가 다음과 같이 조사되었다. 두 모집단의 분산은 동일하다고 가정하고, 중소기업 근로자들의 평균 직장 만족도가 대기업 근로자보다 높다는 귀무가설($\mu_1 \geq \mu_2$) 및 대립가설($\mu_1 < \mu_2$)을 설정하고 1% 유의수준에서 단측검정하라..

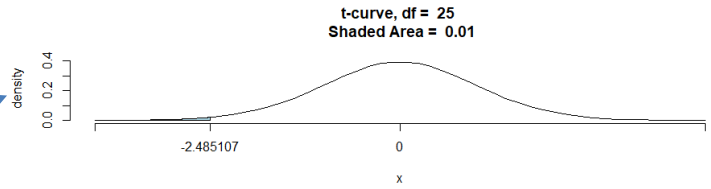
b1-ch7-4-rev.R

```
library(tigerstats)
small<-c(41,45,42,62,68,54,52,55,44,60)
large<-c(74,74,70,52,76,91,71,78,76,78,83,50,
52,66,65,53,72)
ns<-length(small);nl<-length(large)
ms<-mean(small);ml<-mean(large)
vs<-var(small);vl<-var(large)
ss<-sd(small);sl<-sd(large)
ns;nl;ms;ml;vs;vl;ss;sl
sp.sq<-((ns-1)*vs+(nl-1)*vl)/(ns+nl-2)
sp<-sqrt(sp.sq)
tc<-(ms-ml)/(sp*sqrt(1/ns+1/nl))
sp.sq;sp;tc
t.test(small, large, alternative="less", var.equal=
T, conf.level=0.99)
t25<-qt(0.01, df=25, lower.tail=T);t25
par(mfrow=c(2,1))
ptGC(-3.9421,region="below",df=25, graph=T)
ptGC(-2.485107,region="below",df=25,  graph=
T)
```

표본 1(중소기업 근로자, $n_1 = 10$)									
41	45	42	62	68	54	52	55	44	60
표본 1(대기업 근로자, $n_2 = 17$)									
74	74	70	52	76	91	71	78	76	78
83	50	52	66	65	53	72			

```
> ns;nl;ms;ml;vs;vl;ss;sl
[1] 10
[1] 17
[1] 52.3
[1] 69.47059
[1] 85.12222
[1] 138.7647
[1] 9.226171
[1] 11.77984
> sp.sq;sp;tc
[1] 119.4534
[1] 10.92947
[1] -3.942108
```






```
        Two Sample t-test

data:  small and large
t = -3.9421, df = 25, p-value = 0.0002874
alternative hypothesis: true difference in means is less than 0
99 percent confidence interval:
      -Inf -6.346238
sample estimates:
mean of x mean of y
 52.30000  69.47059
```

```
> t25
[1] -2.485107
```

- 계산된 검정통계량의 값은 -3.9421이고 1% 유의수준 하에서 자유도가 25인 t-분포의 임계치는 -2.485107이므로 1% 유의수준 하에서 귀무가설을 기각

(2) 두 집단의 분산을 모르지만 다르다고 가정할 경우

- **(연습 4) 통계학을 수강하는 경제학과 학생과 경영학과 학생들의 중간고사 성적이 각각 다음과 같았다. 두 모집단의 분산은 다르다고 가정하고, 경제학과 학생과 경영학과 학생의 모평균이 동일하다는 귀무가설을 5% 유의수준에서 양측검정하라.**

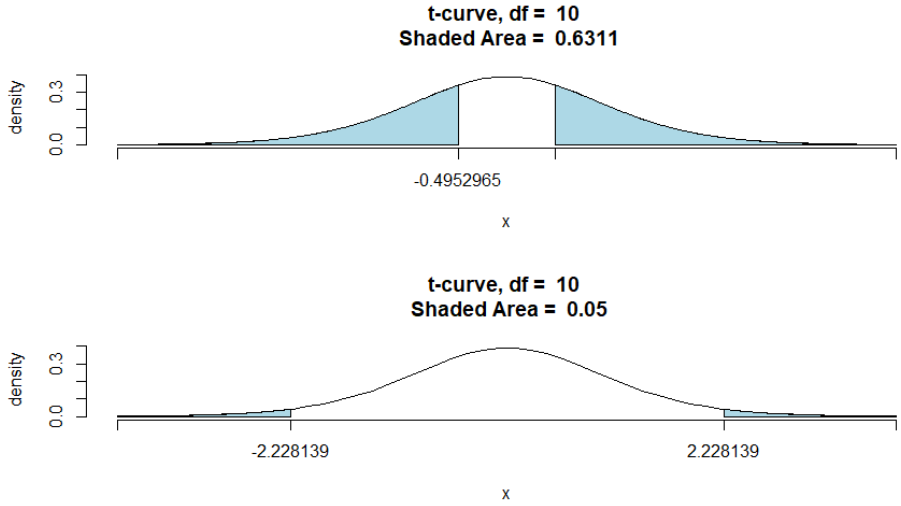
	학생 1	학생 2	학생 3	학생 4	학생 5	학생 6
경제학과	75	71	52	46	70	83
경영학과	82	73	59	48	68	93

b1-ch7-5-rev.R

```
library(tigerstats)
econ<-c(75,71,52,46,70,83)
mgt<-c(82,73,59,48,68,93)
me<-mean(econ)
mm<-mean(mgt)
ve<-var(econ)
vm<-var(mgt)
me;mm;ve;vm
tc<-(me-mm)/(sqrt((ve/6)+(vm/6)))
ptc1<-pt(tc, df=10, lower.tail=T)
ptc2<-pt(-tc, df=10, lower.tail=F)
t1<-qt(0.025, df=10, lower.tail=T)
t2<-qt(0.025, df=10, lower.tail=F)
tc;ptc1;ptc2;t1;t2
par(mfrow=c(2,1))
ptGC(c(-0.4952965,0.4952965),region="o
utside",df=10, graph=T)
ptGC(c(-2.228139,2.228139),region="outs
ide",df=10, graph=T)
#t.test(econ,  mgt,onf.level=0.95)  #Welch
Two Sample t-test
```

```
> me;mm;ve;vm
[1] 66.16667
[1] 70.5
[1] 201.3667
[1] 257.9

> tc;ptc1;ptc2;t1;t2
[1] -0.4952965
[1] 0.3155466
[1] 0.3155466
[1] -2.228139
[1] 2.228139
```



- **계산된 검정통계량의 값은 -0.4952965이고 5% 유의수준 하에서 자유도가 10 인 t-분포의 임계치는 -2.228139와 2.229139이므로 5% 유의수준 하에서 귀무가설을 허용**

(3) 독립이 아닌 짝진 표본

- (연습 5) 통계학을 수강하는 경영학부 학생들을 대상으로 보충수업이 학생들에게 도움이 되는지 알아보기 위해 6명을 임의로 선정하였다. 보충수업 전에 시험을 보게 하고 보충수업을 수강한 후 다시 시험을 보게 하였는데 보충수업이 학생들의 성적 향상에 도움이 되는지 5% 유의수준에서 단측검정하라.

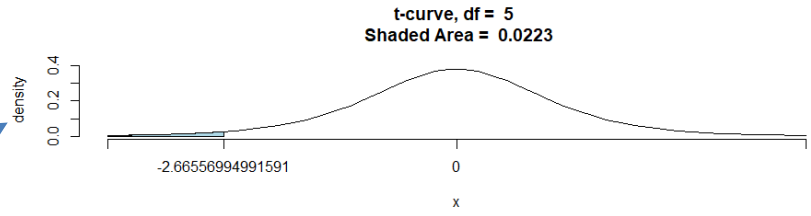
학생	보충수업 전(X_1)	보충수업 후(X_2)	점수 차이($d = X_1 - X_2$)
1	75	82	−7
2	71	73	−2
3	52	59	−7
4	46	48	−2
5	70	69	1
6	83	93	−10

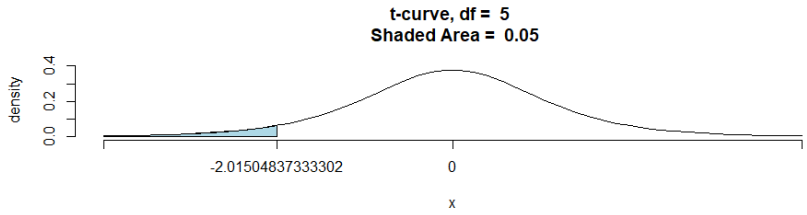
b1-ch7-6-rev.R

```
library(foreign)
library(tigerstats)
data<-read.dta(file="http://kanggc.iptime.
org/book/data/chap11-3-2.dta")
data$var1
data$var2
(d<-data$var1-data$var2)
(md<-mean(d))
(n<-length(d))
(dsd<-sd(d))
tc<-md/(dsd/sqrt(n))
t5<-qt(0.05, df=5, lower.tail=T)
tc;t5
t.test(d, mu=0, alternative="less", conf.lev
el=0.95)
par(mfrow=c(2,1))
ptGC(tc,region="below",df=5, graph=T)
ptGC(t5,region="below",df=5, graph=T)
```

```
> tc;t5
[1] -2.66557
[1] -2.015048
```

```
data:  d
t = -2.6656, df = 5, p-value = 0.02229
alternative hypothesis: true mean is less than 0
95 percent confidence interval:
      -Inf -1.098207
sample estimates:
mean of x
     -4.5
```





- 계산된 검정통계량의 값은 -0.4521418이고 5% 유의수준 하에서 자유도가 8 인 t-분포의 임계치는 -2.306004와 2.306004이므로 5% 유의수준 하에서 귀무가설을 기각

2. 모분산에 대한 가설검정

- (연습 6) 대기업과 중소기업에 근무하는 근로자의 직장 만족도가 다음과 같이 조사되었다. 두 모집단의 모분산은 동일성 여부를 5% 유의수준에서 단측검정하라

표본 1(중소기업 근로자, $n_1 = 10$)									
41	45	42	62	68	54	52	55	44	60
표본 1(대기업 근로자, $n_2 = 17$)									
74	74	70	52	76	91	71	78	76	78
83	50	52	66	65	53	72			

b1-ch7-7-rev.R

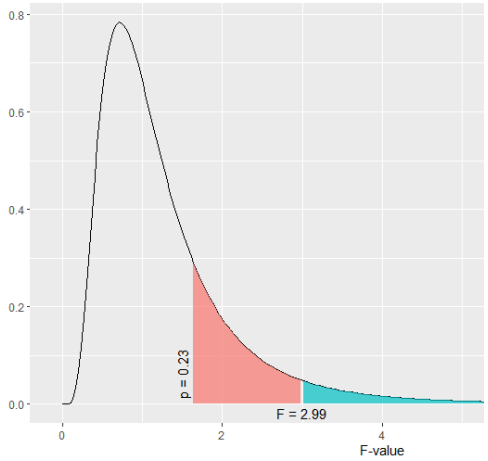
```
small<-c(41,45,42,62,68,54,52,55,44,60)
large<-c(74,74,70,52,76,91,71,78,76,78,8
3,50,52,66,65,53,72)
(mean(small))
(mean(large))
(vs<-var(small))
(vl<-var(large))
(fc<-vl/vs)
(pfc<-pf(fc, df1=16, df2=9,lower.tail=F))
var.test(large, small, alternative="greater",
conf.level=0.95)
(f<-qf(0.05, df1=16, df2=9, lower.tail=F))
library(sjPlot)
par(mfrow=c(2,1))
dist_f(p=0.05, deg.f1 = 16, deg.f2 = 9, xm
ax=8)
dist_f(f=1.630182, deg.f1 = 16, deg.f2 = 9
, xmax=8)
```

```
> (vs<-var(small))
[1] 85.12222

> (vl<-var(large))
[1] 138.7647

> (fc<-vl/vs)
[1] 1.630182

> (pfc<-pf(fc, df1=16, df2=9,lower.tail=F))
[1] 0.2310163
```




```
        F test to compare two variances

data:  large and small
F = 1.6302, num df = 16, denom df = 9, p-value = 0.231
alternative hypothesis: true ratio of variances is greater than 1
95 percent confidence interval:
 0.5454    Inf
sample estimates:
ratio of variances
          1.630182
```

```
> (f<-qf(0.05, df1=16, df2=9, lower.tail=F))
[1] 2.988966
```

- 계산된 검정통계량의 값은 1.6302이고 5% 유의수준 하에서 분자의 자유도가 16이고, 분모의 자유도가 9인 F-분포의 임계치는 2.988966이므로 5% 유의수준 하에서 귀무가설을 허용