Lesson7

Sung Won Kang

2016년 12월 일

모집단이 2개

독립표본, 대응표본

- 독립표본 : 서로 독립인 2개의 모집단에서 추출한 2개의 표본
- 대응표본 : 같은 대상을 실험(Treatment) 이전과 이후에 측정하여 수집한 표본

(독립표본) 모평균 차이 검정(Two sample test)

자료: 여아 신생아 몸무게(X1), 남아 신생아 몸무게(X2)

```
#data=read.table("http://www.amstat.org/publications/jse/datasets/babyboom.dat.txt",header=F)
#names(data)=c("time","gender","weight","minutes")#시간.분, 성별, 체중, 분#
#chapter7=data[,c(2,3)]
#save(chapter7,file="ch7.Rdata")
load("ch7.Rdata")
X1=chapter7[chapter7$gender==1,]#girls weight
X2=chapter7[chapter7$gender==2,]#boys' weight
```

기본가정 : 독립표본, 정규분포, 분산이 같음(등분산성)

$$X_1 \sim N(\mu_1, \sigma_1^2) \qquad X_2 \sim N(\mu_2, \sigma_2^2) \sigma_1^2 = \sigma_2^2$$

1단계 : 등분산성 검정 - 유의수준 5% 양측검정

```
1.가설 수립 :  남아 신생아 체중 분산 = 여아 신생아 체중 분산
2.검정 통계량 계산:
3.가설 채택 기준 설정: 유의 수준 5%
4.채택/기각 결정
```

1. 가설 수립

$$H_o : \frac{\sigma_1^2}{\sigma_2^2} = 1 \qquad .\text{vs} \qquad H_1 : \frac{\sigma_1^2}{\sigma_2^2} \neq 1$$

2. 검정통계량 계산

$$F = \frac{\frac{(n-1)S_1^2}{\sigma_1^2(n-1)}}{\frac{(m-1)S_2^2}{\sigma_2^2(m-1)}} = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n-1, m-1)$$

귀무가설이 성립한다면

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{S_1^2/\sigma^2}{S_2^2/\sigma^2} = \frac{S_1^2}{S_2^2} \sim F(n-1, m-1) = 2.177$$

3. 가설 채택 기준 설정: 양측검정, F 분포

$$\text{임계치} : P(F > c_u) = 0.025, P(F < C_l) = 0.025 \Rightarrow C_l = 0.392, C_u = 2.360$$

$$\text{기각역} : \{F < 0.392\} \cup \{F > 2.36\}$$

$$\text{P value} : 2 \times [1 - P(F > 2.177)] = 0.076$$

4. 가설 기각/채택

$$F = 2.177 < 2.36 \rightarrow \text{ not reject } H_o$$

$$P = 2 \times [1 - P(F > 2.177)] = 0.076 > 0.05 \rightarrow \text{ not reject } H_o$$

```
# 검정통계량
S1=var(X1$weight)
S2=var(X2$weight)
(Ftest=S1/S2)
```

```
## [1] 2.177104
```

```
# 임계치, 기각역
alpha=0.05
(n=length(X1$weight))
```

```
## [1] 18
```

```r
(m=length(X2$weight))
```

```
## [1] 26
```

```r
(c.l=qf(alpha/2,(n-1),(m-1)))
```

```
## [1] 0.3924002
```

```r
(c.u=qf(1-(alpha/2),(n-1),(m-1)))
```

```
## [1] 2.359863
```

```r
# p value


(Pv=2*(1-pf(Ftest,17,25)))
```

```
## [1] 0.07526262
```

```r
# test critical value
(Ftest > c.u)|(Ftest<c.l)
```

```
## [1] FALSE
```

```r
#test p value
Pv < alpha
```

```
## [1] FALSE
```

```r
var.test(chapter7$weight~chapter7$gender)
```

```
##
##  F test to compare two variances
##
## data:  chapter7$weight by chapter7$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
```

2단계 : 모평균 차이 검정

1. 가설 수립 : 여아 신생아 체중 평균 < 남아 신생아 체중 평균
2. 검정 통계량 계산:

3.가설 채택 기준 설정: 유의 수준 5% 단측검정(왼쪽)
4.채택/기각 결정

1. 가설 수립

$$H_o : \mu_1 - \mu_2 = 0 \qquad .\text{vs} \qquad H_1 : \mu_1 - \mu_2 < 0$$

2. 검정통계량 계산

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{1/n + 1/m}} \sim t(n+m-2)$$

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{1/n + 1/m}} \sim N(0,1)$$

$$\frac{(n-1)S_1^2 + (m-1)S_2^2}{\sigma^2} \sim \chi^2(n+m-2)$$

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{1/n + 1/m}} / \sqrt{\frac{(n-1)S_1^2 + (m-1)S_2^2}{\sigma^2(n+m-2)}} \sim t(n+m-2)$$

$$s_p = \sqrt{\frac{(n-1)S_1^2 + (m-1)S_2^2}{(n+m-2)}}$$

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{1/n + 1/m}} \sim t(n+m-2)$$

귀무가설 하에서는 $\mu_1 - \mu_2 = 0$ 이므로 검정통계량은

$$T = \frac{3132.444 - 3375.308}{520.1151 \times \sqrt{1/18 + 1/26}} = -1.5229$$

3. 가설 채택 기준 설정: 단측검정, T 분포

$$\text{임계치} : P(T < c_l) = 0.05 \Rightarrow C_l = -1.6820$$

$$\text{기각역} : T < -1.6820$$

$$\text{P value} : P(T < -1.5229) = 0.0676$$

4. 가설 기각/채택

$$T = -1.5229 > -1.6820 \rightarrow \text{ not reject } H_o$$

$$P(T < -1.5229) = 0.0676 > 0.05 \rightarrow \text{ not reject } H_o$$

```
# 검정통계량
(barx1=mean(X1$weight))
```

```
## [1] 3132.444
```

```
(barx2=mean(X2$weight))
```

```
## [1] 3375.308
```

```
(n=length(X1$weight))
```

```
## [1] 18
```

```
(m=length(X2$weight))
```

```
## [1] 26
```

```
S1=var(X1$weight)
S2=var(X2$weight)
sp=sqrt(((n-1)*S1+(m-1)*S2)/(n+m-2))
(T=(barx1-barx2)/(sp*sqrt(1/n+1/m)))
```

```
## [1] -1.522856
```

```
# 임계치, 기각역
alpha=0.05
(c.l=qt(alpha,df=n+m-2))
```

```
## [1] -1.681952
```

```
# p value
(Pv=pt(T,df=n+m-2))
```

```
## [1] 0.06764459
```

```
# test critical value
(T < c.l)
```

```
## [1] FALSE
```

```
#test p value
Pv < alpha
```

```
## [1] FALSE
```

```r
t.test(chapter7$weight~chapter7$gender,mu=0,alternative="less",var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  chapter7$weight by chapter7$gender
## t = -1.4211, df = 27.631, p-value = 0.08324
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf 47.99869
## sample estimates:
## mean in group 1 mean in group 2
##        3132.444        3375.308
```

분산이 다를 경우

검정통계량

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{s_1^2/n + s_2^2/m}} \sim t(v)$$

$$v = \frac{(s_1^2/n)^2 + (s_2^2/m)^2}{s_1^4/n^2(n-1) + s_2^4/m^2(m-1)}$$


```r
t.test(chapter7$weight~chapter7$gender,mu=0,alternative="less",var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  chapter7$weight by chapter7$gender
## t = -1.4211, df = 27.631, p-value = 0.08324
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf 47.99869
## sample estimates:
## mean in group 1 mean in group 2
##        3132.444        3375.308
```

(대응표본) 모평균 차이 검정(Paired Data)

자료: 식욕부진증 치료요법 이전의 체중과 이후의 체중

```
A.data=read.csv("./data/01.anorexia.csv",header=T)
str(A.data)
```

```
## 'data.frame':    17 obs. of  2 variables:
##  $ Prior: num  83.8 83.3 86 82.5 86.7 79.6 76.9 94.2 73.4 80.5 ...
##  $ Post : num  95.2 94.3 91.5 91.9 100.3 ...
```

기본가정 : 체중의 차이는 정규분포를 한다.

$$D = X_{pre} - X_{post} \sim N(\mu_D, \sigma_D)$$

가설검정: 식욕부진증 치료요법 이후에 체중이 증가하는가?

```
1.가설 수립 :  치료요법 이전 체중 <치료요법 이후 체중
2.검정 통계량 계산:
3.가설 채택 기준 설정: 유의 수준 5%
4.채택/기각 결정
```

1. 가설 수립

$$H_0 : \mu_D = 0 \qquad .\text{vs} \qquad H_1 : \mu_D < 0$$

2. 검정통계량

$$T = \frac{\bar{D} - \mu_D}{s_D/\sqrt{N}} \sim t(N-1)$$

귀무가설 하에서는 $\mu_D = 0$ 이므로 검정통계량은

$$T = \frac{7.2647}{51.2287/\sqrt{16}} = -4.1849$$

3. 가설 채택 기준 설정

$$\text{임계치, 기각역} : P(T < C_l) = 0.05 \rightarrow C_l = -1.7458 \qquad \{T < -1.7458\}$$

$$\text{P value} : P(T < -4.1849) = 0.0003$$

4. 가설 기각/채택

$$T = -4.1849 < C_l = -1.7458 \Rightarrow \text{ Reject } H_0$$

$$P(T < -4.1849) = 0.0003 < 0.05 \Rightarrow \text{ Reject } H_0$$

```r
D=A.data$Prior-A.data$Post
# 검정통계량
(bard=mean(D))
```

```
## [1] -7.264706
```

```r
ssd=var(D)
(T=bard/sqrt(ssd/length(D)))
```

```
## [1] -4.184908
```

```r
# 임계점, 기각역, pvalue
(c.l=qt(alpha,df=(length(D)-1)))
```

```
## [1] -1.745884
```

```r
(pvalue=pt(T,df=(length(D)-1)))
```

```
## [1] 0.0003501266
```

```r
# 가설 기각/채택


T<c.l
```

```
## [1] TRUE
```

```r
pvalue<alpha
```

```
## [1] TRUE
```

```r
t.test(A.data$Prior,A.data$Post,paired=T,alternative="less")
```

```
##
##  Paired t-test
##
## data:  A.data$Prior and A.data$Post
## t = -4.1849, df = 16, p-value = 0.0003501
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -4.233975
## sample estimates:
```

```
## mean of the differences
##               -7.264706
```

모집단이 3개 이상 (ANOVA)

- 예제: 대도시, 중소도시, 읍면지역의 서비스 만족도 차이가 있는가, 없는가?
- 가설: 차이가 없다. vs. 셋 중 하나는 다르다.

기본가정 : 정규분포,독립표본, 등분산성

$$y_{i,j} \sim N(\mu + \alpha_i, \sigma^2) \qquad j \in \{1, 2, \cdots, N\}(\text{ individual}), \quad i \in \{1, 2, \cdots, k\}(\text{ group })$$

$$y_{1,j} \sim N(\mu + \alpha_1, \sigma^2), j = 1, 2, \cdots, n_1$$

$$y_{2,j} \sim N(\mu + \alpha_2, \sigma^2), j = 1, 2, \cdots, n_2$$

$$\vdots$$

$$y_{k,j} \sim N(\mu + \alpha_k, \sigma^2), j = 1, 2, \cdots, n_k$$

- 특징

$$\sum_i^k \sum_j^n (y_{ij} - \bar{y})^2 = \sum_i^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 + \sum_i^k \sum_{j=1}^{n_i} (\bar{y}_i - \bar{y})^2$$

$$SST = SSE + SSt$$

```
+ SST(Total Sum of Squres): 개별 data - 전체평균
+ SSE(Error Sum of Squres): 개별 data - Group 평균 (within difference)
+ SSt(Treatment Sum of Squres): Group 평균 - 전체평균 (between difference)
```

가설검정: Group 별 격차가 있는가?

```
1.가설 수립 :  각 Group 별 격차는 없다 vs. 한 Group에서라도 격차는 존재한다
2.검정 통계량 계산:
3.가설 채택 기준 설정: 유의 수준 5%
4.채택/기각 결정
```

가설 수립

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$$

$$H_1 : \exists i \qquad s.t. \qquad \alpha_i \neq 0$$

검정통계량 계산(귀무가설)

$$F = \frac{SSt/(k-1)}{SSE/(n-k)} \sim F(k-1, n-k)$$

$$F = \frac{150.0933/(3-1)}{30139.38/(150-3)} = 0.366$$

```
ad=read.csv(file="./data/age.data.csv",header=T)
ad$scale=factor(ad$scale)
y1=ad$age[ad$scale==1]
y2=ad$age[ad$scale==2]
y3=ad$age[ad$scale==3]

y1.mean=mean(y1)
y2.mean=mean(y2)
y3.mean=mean(y3)
y=mean(ad$age)

sse.1=sum((y1-y1.mean)^2)
sse.2=sum((y2-y2.mean)^2)
sse.3=sum((y3-y3.mean)^2)

(sse=sse.1+sse.2+sse.3)
```

```
## [1] 30139.38
```

```
(dfe=length(y1)-1+length(y2)-1+length(y3)-1)
```

```
## [1] 147
```

```
sst.1=length(y1)*sum(y1.mean-y)^2
sst.2=length(y2)*sum(y2.mean-y)^2
sst.3=length(y3)*sum(y3.mean-y)^2
```

```r
(sst=sst.1+sst.2+sst.3)
```

```
## [1] 150.0933
```

```r
(dft=length(levels(ad$scale))-1)
```

```
## [1] 2
```

```r
# check the decomposition
(SST=sum((ad$age-y)^2))
```

```
## [1] 30289.47
```

```r
(SS=sst+sse)
```

```
## [1] 30289.47
```

```r
# 검정통계량
(F=(sst/dft)/(sse/dfe))
```

```
## [1] 0.3660281
```

가설 채택 기준 설정: 유의 수준 5%

$$\text{임계치, 기각역} : P(F > C.u) = 0.05 \rightarrow C.u = 3.057621 \qquad \{F > 3.0576\}$$

$$\text{P value} : P(F > 0.3660) = 0.6941$$

채택/기각 결정

$$F = 0.3660 < 3.0576 \rightarrow \text{ cannot reject } H_0$$

$$P(F > 0.3660) = 0.6941 > 0.05 \rightarrow \text{ cannot reject } H_0$$

```r
#임계치
(c.u=qf(1-alpha,dft,dfe))
```

```
## [1] 3.057621
```

```r
# p value
(p.v=1-pf(F,dft,dfe))
```

```
## [1] 0.6941136
```

```r
F>c.u
```

```
## [1] FALSE
```

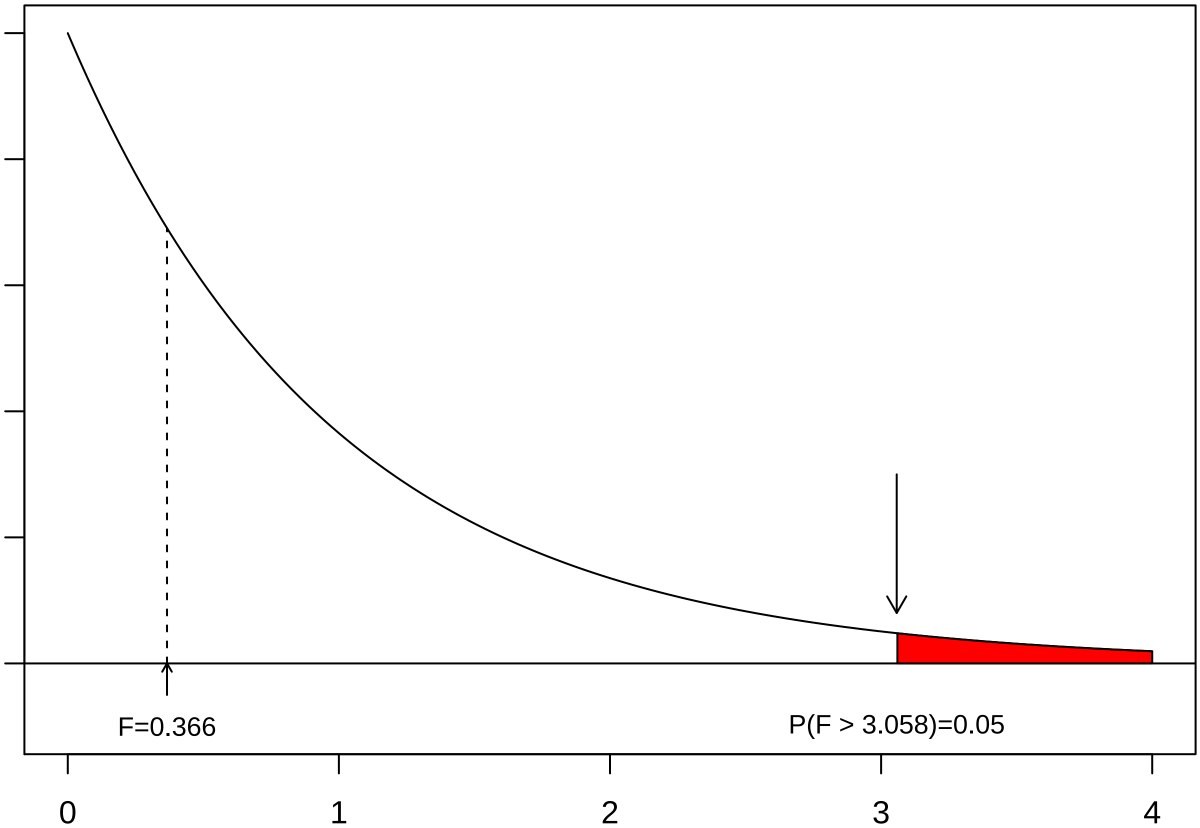
```r
p.v<alpha
```

```
## [1] FALSE
```

```r
ow=lm(age~scale,data=ad)
anova(ow)
```

```
## Analysis of Variance Table
##
## Response: age
##            Df  Sum Sq Mean Sq F value Pr(>F)
## scale       2   150.1  75.047   0.366 0.6941
## Residuals 147 30139.4 205.030
```

```r
x <- seq(0, 4, by=0.01)
yf <- df(x, 2, 147)
par(mar=c(2, 1, 1, 1))
plot(x, yf, type="l", ylim=c(-0.1, 1), xlab="", ylab="", axes=F)
abline(h=0)
cu.r <- round(c.u, 2)
polygon(c(cu.r, x[x>=cu.r], 4), c(0, yf[x>=cu.r], 0), col="red")
arrows(c.u, 0.3, c.u, 0.08, length=0.1)
#abline(v=c.u)
text(c.u, -0.1, paste("P(F > ", round(c.u, 3),")=0.05", sep=""), cex=0.8)
lines(c(F, F), c(0,df(F, 2, 147)), lty=2)
arrows(F, -0.05, F, 0, length=0.05)
#abline(v=F)
text(F, -0.1, paste("F=", round(F, 3),sep=""), cex=0.8)
```


F=0.366
P(F > 3.058)=0.05
0
1
2
3
4