

## HW#1 - solution

1. SCORE 가 20 보다 작으면 treatment 의 효과가 있다고 알려져 있다고 가정하자. 이 때 treatment A 가 효과가 있는지를 알고자 한다.

① 귀무가설과 대립가설을 쓰시오.

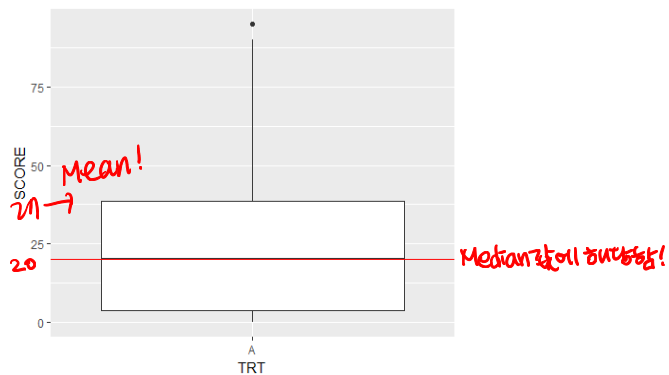
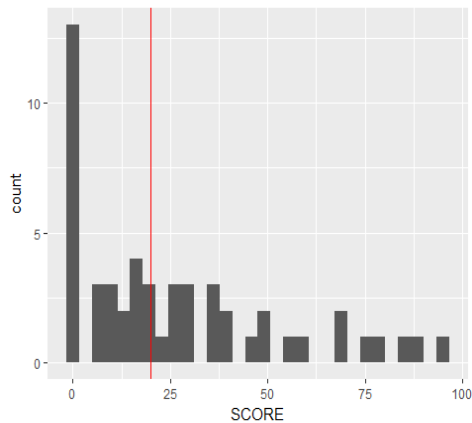
```
trial<-read.csv("ex3-1-new.csv")
```

$H_0: \mu(A) = 20$

$H_1: \mu(A) < 20$  *효과가 있다!*

② 위의 가설을 그림으로 확인하고자 한다. 해당하는 그림을 그리시오

```
library(ggplot2)
trialA<-subset(trial, TRT=="A")
ggplot(trialA, aes(SCORE))+geom_histogram()+geom_vline(xintercept=20, col="Red")
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
ggplot(trialA, aes(TRT, SCORE))+geom_boxplot()+geom_hline(yintercept=20, col="Red")
```



③ ① 의 가설 검정을 위하여 어떠한 방법을 써야 하는가? 기술하시오.

==> one sample t-test (*one-sided test*)

④ 위에서 기술한 방법을 SAS 와 R 을 각각 이용하여 결과를 얻은 후 이 두 결과를 비교하여 결론을 내리시오.

un-sided test

```
t.test(trialA$SCORE,mu=20,alternative="less")

##
## One Sample t-test
##
## data: trialA$SCORE
## t = 1.7855, df = 51, p-value = 0.9599
## alternative hypothesis: true mean is less than 20
## 95 percent confidence interval:
##      -Inf 32.93428
## sample estimates:
## mean of x
## 26.67308
```

분산 커서 가장 중요!

==> treatment A 는 효과가 있다고 할 수 없다. (H0 채택)

```
DATA trial;
  INFILE 'ex3-1-new.csv' DELIMITER=',' FIRSTOBS=2;
  INPUT TRT$ CENTER PAT SEX$ AGE SCORE @@;
RUN;
```

```
DATA trialA;
  set trial;
  IF (TRT="A") THEN OUTPUT;
RUN;
```

```
PROC TTEST DATA=trialA H0=20 sides=1;
  VAR SCORE;
RUN;
```

#### The TTEST Procedure

Variable: SCORE

N	Mean	Std Dev	Std Err	Minimum	Maximum
52	26.6731	26.9507	3.7374	0	95.0000

Mean	95% CL Mean	Std Dev	95% CL Std Dev
26.6731	-Infy 32.9343	26.9507	22.5860 33.4223

DF	t Value	Pr < t
51	1.79	0.9599

★ TRT A, B 2개지만, one-way ANOVA  
 사용가능함!

## 2. SCORE 변수를 이용하여

① treatment group 간에 차이가 있는지를 one-way ANOVA 를 이용하여 분석하시오.

```
anova(lm(SCORE~TRT,data = trial))
```

```
## Analysis of Variance Table
```

```
##
```

```
## Response: SCORE
```

```
##          Df Sum Sq Mean Sq F value Pr(>F)
```

```
## TRT          1    3394   3393.6    4.2375 0.0422 *
```

```
## Residuals  98  78484    800.9
```

```
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A와 B는  
차이 있다.

```
PROC GLM DATA = trial;
```

```
  CLASS TRT CENTER; TRT; 필요없는거 그냥 생략
```

```
  MODEL SCORE = TRT;
```

```
RUN;
```

### The GLM Procedure

Dependent Variable: SCORE

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	3393.60103	3393.60103	4.24	0.0422
Error	98	78484.10897	800.85825		
Corrected Total	99	81877.71000			

R-Square	Coeff Var	Root MSE	SCORE Mean
0.041447	87.69581	28.29944	32.27000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
TRT	1	3393.601026	3393.601026	4.24	0.0422

Source	DF	Type III SS	Mean Square	F Value	Pr > F
TRT	1	3393.601026	3393.601026	4.24	0.0422

⇒ 차이 있음. SAS 와 R 의 결과 일치

② treatment group 과 center 간에 교호작용이 있는지를 알고자 한다. ANOVA 를 이용하여 결론을 내리시오.

```
anova(lm(SCORE ~ TRT*factor(CENTER), data = trial))
```

```
## Analysis of Variance Table
```

```
##
```

```
## Response: SCORE
```

```
##
```

```
##          Df Sum Sq Mean Sq F value Pr(>F)
```

```
## TRT          1    3394   3393.6    4.2047 0.0431 *
```

```
## factor(CENTER)  2    1145    572.5    0.7093 0.4946
```

```
## TRT:factor(CENTER) 2    1472    736.1    0.9121 0.4052
```

```
## Residuals      94   75867    807.1
```

```
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

자유의 확인해서 잘 들어갔나 보기! 만약 center가 numeric으로 들어간다면 DF가 이상한 것임

교호작용 없다!

```
PROC GLM DATA = trial;
```

```
CLASS TRT CENTER;
```

```
MODEL SCORE = TRT CENTER TRT*CENTER;
```

```
RUN;
```

#### The GLM Procedure

Dependent Variable: SCORE

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	6010.89531	1202.17906	1.49	0.2006
Error	94	75866.81469	807.09377		
Corrected Total	99	81877.71000			

R-Square	Coeff Var	Root MSE	SCORE Mean
0.073413	88.03655	28.40940	32.27000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
TRT	1	3393.601026	3393.601026	4.20	0.0431
CENTER	2	1145.014504	572.507252	0.71	0.4946
TRT*CENTER	2	1472.279781	736.139891	0.91	0.4052

Source	DF	Type III SS	Mean Square	F Value	Pr > F
TRT	1	2706.907813	2706.907813	3.35	0.0702
CENTER	2	999.961195	499.980597	0.62	0.5404
TRT*CENTER	2	1472.279781	736.139891	0.91	0.4052

⇒ 교호작용 없음 R 과 SAS 의 Type I SS 결과와 일치.

↓  
R은 type I default!

③ treatment group 과 center 간에 교호작용이 없다고 가정하고 treatment 와 center 간 차이를 알고자 한다. ANOVA 를 이용하여 결론을 내리시오.

```
anova(lm(SCORE~factor(CENTER)+TRT,data = trial))

## Analysis of Variance Table
##
## Response: SCORE
##          Df Sum Sq Mean Sq F value    Pr(>F)
## factor(CENTER)  2     966    482.8    0.5993 0.55121
## TRT             1    3573   3572.9    4.4350 0.03782 *
## Residuals      96   77339    805.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
PROC GLM DATA = trial;
  CLASS TRT CENTER;
  MODEL SCORE = CENTER TRT;
RUN;
```

Dependent Variable: SCORE

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	4538.61553	1512.87184	1.88	0.1385
Error	96	77339.09447	805.61557		
Corrected Total	99	81877.71000			

R-Square	Coeff Var	Root MSE	SCORE Mean
0.055432	87.95590	28.38337	32.27000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
CENTER	2	965.689286	482.844643	0.60	0.5512
TRT	1	3572.926244	3572.926244	4.44	0.0378

Source	DF	Type III SS	Mean Square	F Value	Pr > F
CENTER	2	1145.014504	572.507252	0.71	0.4939
TRT	1	3572.926244	3572.926244	4.44	0.0378

⇒ Center 차이 없음, TRT 차이 있음.

숨겨진 변수,  
type III 결과는 같은데, type I 결과는 달라진다.  
따라서 type III 결과는 읽는게 더 편하다!

```
anova(lm(SCORE~TRT+factor(CENTER),data = trial))

## Analysis of Variance Table
##
## Response: SCORE
##              Df Sum Sq Mean Sq F value    Pr(>F)
## TRT           1   3394   3393.6    4.2124 0.04285 *
## factor(CENTER) 2   1145    572.5    0.7106 0.49389
## Residuals     96  77339    805.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
PROC GLM DATA = trial;
  CLASS TRT CENTER;
  MODEL SCORE = TRT CENTER;

RUN;
```

Dependent Variable: SCORE

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	4538.61553	1512.87184	1.88	0.1385
Error	96	77339.09447	805.61557		
Corrected Total	99	81877.71000			

R-Square	Coeff Var	Root MSE	SCORE Mean
0.055432	87.95590	28.38337	32.27000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
TRT	1	3393.601026	3393.601026	4.21	0.0429
CENTER	2	1145.014504	572.507252	0.71	0.4939

Source	DF	Type III SS	Mean Square	F Value	Pr > F
TRT	1	3572.926244	3572.926244	4.44	0.0378
CENTER	2	1145.014504	572.507252	0.71	0.4939

⇒ model 에서 CENTER, TRT 순서를 바꾸면 Type I SS 는 달라짐. Type III SS 는 그대로

④ ③에서 treatment 또는 center 간 차이가 있다는 결론이 나왔을 경우 어떻게 해야 하는가? ==> 사후분석

CENTER 간 차이 없다는 결론 나온 → 데이터할 필요 X  
하나는 2개 이므로 데이터의 사후분석 필요 X

↓  
하아 A,B 중 어느것이 더 큰지  
확인해야 해봐주세요!