임상시험자료분석 2

1ST ASSIGNMENT

조현선 | 162STG26 | 2016/09/26

1. 3 장의 TRIAL 자료를 SAS 로 분석한 내용을 그대로 R 로 재현하시오.

(단, UNIVARIATE 절차 부분은 제외.)

Data setting in SAS & R

```
DATA TRIAL;
INFILE 'C:\Users\user\Desktop\trial3.csv' DELIMITER=','
FIRSTOBS=2;
INPUT TRT $ CENTER PAT SEX $ AGE SCORE @@;
RESP = (SCORE GT 0) ;
IF (SCORE=0) THEN SEV=0;
IF (1 LE SCORE LE 30) THEN SEV=1;
IF (31 LE SCORE LE 69) THEN SEV=2;
IF (SCORE GE 70) THEN SEV=3;
RUN;
```

```
data3.1<-read.table("trial3.csv",sep=",",header = T)
attach(data3.1)
RESP <- NULL
for (i in 1:nrow(data3.1)){
  if (SCORE[i]==0) { RESP[i]<-0} else  RESP[i]<-1}

SEV <- NULL
for (i in 1:nrow(data3.1)){if (SCORE[i]==0){SEV[i]<-0}
   else if (1<=SCORE[i] & SCORE[i]<=30) {SEV[i]<-1}
   else if (31<SCORE[i] & SCORE[i]<=69) {SEV[i]<-2}
   else  SEV[i]<-3 }

data3.1$RESP<-RESP
data3.1$SEV<-SEV
```

1

```
PROC SORT DATA = TRIAL ;
        BY PAT;

PROC PRINT DATA = TRIAL;
        VAR PAT TRT CENTER SEX AGE RESP SEV SCORE;
RUN;
```

2

```
PROC SORT DATA = TRIAL;
        BY TRT;
PROC MEANS MEAN STD N MIN MAX DATA = TRIAL;
        BY TRT;
        VAR SCORE AGE;
RUN;
```

3

```
PROC CHART DATA = TRIAL;
        VBAR SCORE / MIDPOINTS = 10 30 50 70 90 GROUP=TRT;
RUN;
```

4

```
PROC FORMAT;
        VALUE RSPFMT 0 = '0=Abs.' 1='1=Pres';
RUN;
PROC FREQ DATA = TRIAL;
        TABLES TRT*RESP /NOCOL NOPCT;
        FORMAT RESP RSPFMT.;
RUN;
```

5

```
PROC FORMAT;
        VALUE SEVFMT 0 = '0=None'
        1 = '1=Mild'
        2 = '2=Mod.'
        3 = '3=Sev.'
RUN;
PROC FREQ DATA = TRIAL;
        TABLES TRT*SEV /NOCOL NOPCT;
        FORMAT SEV SEVFMT.;
RUN;
```

6

```
PROC FREQ DATA = TRIAL;
        TABLES SEX*TRT*SEV / NOCOL NOPCT;
        FORMAT SEV SEVFMT.;
RUN;
```

7

```
PROC CHART DATA = TRIAL;
        VBAR SEV / MIDPOINTS = 0 1 2 3 GROUP = TRT;
RUN;
```

1.

SAS 코드

```
PROC SORT DATA = TRIAL ;
        BY PAT;

PROC PRINT DATA = TRIAL;
        VAR PAT TRT CENTER SEX AGE RESP SEV SCORE;
RUN;
```

R 코드

```
trial.1<-data3.1[order(PAT),]
trial.1<-subset(trial.1,select = c(3,1,2,4,5,6,8,7))
head(trial.10)
```

```
> head(trial.1,10)
    PAT TRT CENTER SEX AGE SCORE SEV RESP
1   101   A      1   M  55     5   1    1
67  102   B      1   M  19    68   2    1
85  103   B      1   F  51    10   1    1
16  104   A      1   F  27     0   0    0
49  105   B      1   M  45    20   1    1
31  106   A      1   M  31    35   2    1
2   107   A      1   F  44    21   1    1
68  108   B      1   F  44    65   2    1
17  109   A      1   M  47    15   1    1
86  110   B      1   M  32    25   1    1
```

2.

SAS 코드

```
PROC SORT DATA = TRIAL;
        BY TRT;
PROC MEANS MEAN STD N MIN MAX DATA = TRIAL;
        BY TRT;
        VAR SCORE AGE;
RUN;
```

R 코드

```
trial.2<-data3.1[order(TRT),]

trial.2.A<-subset(trial.2,TRT=="A",select=c("SCORE","AGE"))
sd.A<-round(apply(trial.2.A,2,sd),2)
summary.A<-rbind(summary(trial.2.A),paste0("sd.    :",sd.A))

trial.2.B<-subset(trial.2,TRT=="B",select=c("SCORE","AGE"))
sd.B<-round(apply(trial.2.B,2,sd),2)
summary.B<-rbind(as.matrix(summary(trial.2.B)),paste0("sd.    :",sd.B))
```

```
> summary.A
     SCORE              AGE
 "Min.   : 0.00  " "Min.   : 0.00  "
 "1st Qu.: 3.75  " "1st Qu.:29.50  "
 "Median :20.00  " "Median :41.50  "
 "Mean   :26.67  " "Mean   :39.75  "
 "3rd Qu.:38.50  " "3rd Qu.:51.75  "
 "Max.   :95.00  " "Max.   :80.00  "
 "sd.    :26.95" "sd.    :18.46"
```

```
> summary.B
     SCORE              AGE
 "Min.   : 0.00  " "Min.   :19.00  "
 "1st Qu.:11.50  " "1st Qu.:32.75  "
 "Median :39.00  " "Median :41.50  "
 "Mean   :38.33  " "Mean   :41.33  "
 "3rd Qu.:65.00  " "3rd Qu.:49.50  "
 "Max.   :95.00  " "Max.   :63.00  "
 "sd.    :29.69" "sd.    :11.79"
```

3.

SAS 코드

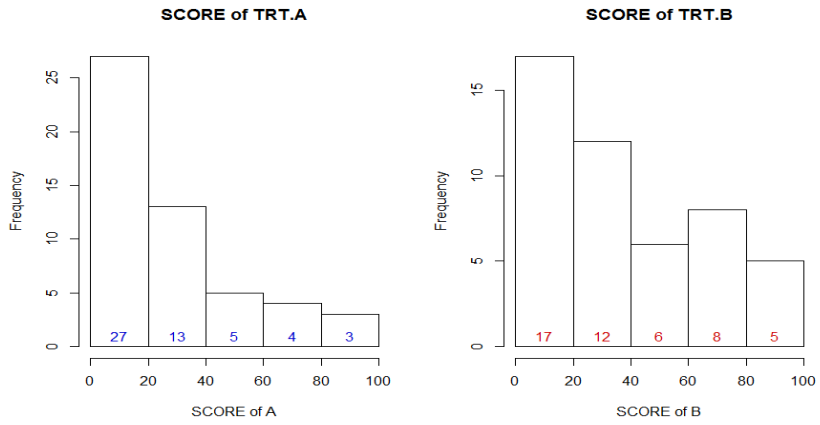
```
PROC CHART DATA = TRIAL;
        VBAR SCORE / MIDPOINTS = 10 30 50 70 90 GROUP=TRT;
RUN;
```

R 코드

```
par(mfrow=c(1,2))

hist.a<-hist(trial.2.A$SCORE, breaks=6,freq=TRUE,main="SCORE of TRT.A",xlab="SCORE of A")
text(hist.a$mids,hist.a$density,hist.a$counts,adj=c(0.5,-0.5),col="blue3")

hist.b<-hist(trial.2.B$SCORE, breaks=6,freq=TRUE,main="SCORE of TRT.B",xlab="SCORE of B")
text(hist.b$mids,hist.b$density,hist.b$counts,adj=c(0.5,-0.5),col="red3")
```



4.

SAS 코드

```
PROC FORMAT;
        VALUE RSPFMT 0 = '0=Abs.' 1='1=Pres';
RUN;
PROC FREQ DATA = TRIAL;
        TABLES TRT*RESP /NOCOL NOPCT;
        FORMAT RESP RSPFMT.;
RUN;
```

R 코드

```
trial.4<-subset(trial.1,select=c(2,8))
tab.4<-table(trial.4)
attr(tab.4,"dimnames")$RESP<-c("0=Abs.","1=Pres")
table.4.freq<-ftable(addmargins(tab.4))
table.4.rel.freq<-ftable(prop.table(tab.4,margin = 1))*100
```

```
> table.4.freq
    RESP 0=Abs. 1=Pres Sum
TRT
A           13      39  52
B            9      39  48
Sum         22      78 100
```

```
> table.4.rel.freq
    RESP 0=Abs. 1=Pres
TRT
A         25.00  75.00
B         18.75  81.25
```

5.

SAS 코드

```
PROC FORMAT;
            VALUE SEVFMT 0 = '0=None'
            1 = '1=Mild'
            2 = '2=Mod.'
            3 = '3=Sev.'
RUN;
PROC FREQ DATA = TRIAL;
            TABLES TRT*SEV /NOCOL NOPCT;
            FORMAT SEV SEVFMT.;
RUN;
```

R 코드

```
trial.5<-subset(trial.1,select=c(2,7))
tab.5<-table(trial.5)
attr(tab.5,"dimnames")$SEV<-c("0=None.","1=Mild","2=Mod","3=Sev.RUN")
table.5.freq<-ftable(addmargins(tab.5))
table.5.rel.freq<-round(ftable(prop.table(tab.5,margin = 1))*100,2)
```

```
> table.5.freq
    SEV 0=None. 1=Mild 2=Mod 3=Sev.RUN Sum
TRT
A            13      22    11         6  52
B             9      12    17        10  48
Sum          22      34    28        16 100
> table.5.rel.freq
    SEV 0=None. 1=Mild 2=Mod 3=Sev.RUN
TRT
A         25.00  42.31 21.15     11.54
B         18.75  25.00 35.42     20.83
```

6.

SAS 코드

```
PROC FREQ DATA = TRIAL;
            TABLES SEX*TRT*SEV / NOCOL NOPCT;
            FORMAT SEV SEVFMT.;
RUN;
```

R 코드

```
trial.6<-subset(trial.1,select=c(2,7,4))
tab.6<-table(trial.6)
attr(tab.6,"dimnames")$SEV<-c("0=None","1=Mild","2=Mod","3=Sev.RUN")
table.6.freq<-addmargins(tab.6,c(1,2))

table.6.rel.freq<-tab.6
table.6.rel.freq[1:8]<-matrix((round(tab.6*100/c(31,25),2)[1:8]),2,4)
table.6.rel.freq[9:16]<-matrix((round(tab.6*100/c(21,23),2)[9:16]),2,4)
```

```
> table.6.freq
, , SEX = F

     SEV
TRT   0=None 1=Mild 2=Mod 3=Sev.RUN Sum
  A       10      12     5         4  31
  B        7       4    11         3  25
  Sum     17      16    16         7  56

, , SEX = M

     SEV
TRT   0=None 1=Mild 2=Mod 3=Sev.RUN Sum
  A        3      10     6         2  21
  B        2       8     6         7  23
  Sum      5      18    12         9  44
```

```
> table.6.rel.freq
, , SEX = F

    SEV
TRT 0=None 1=Mild 2=Mod 3=Sev.RUN
  A  32.26  38.71 16.13      12.90
  B  28.00  16.00 44.00      12.00

, , SEX = M

    SEV
TRT 0=None 1=Mild 2=Mod 3=Sev.RUN
  A  14.29  47.62 28.57       9.52
  B   8.70  34.78 26.09      30.43
```

7.

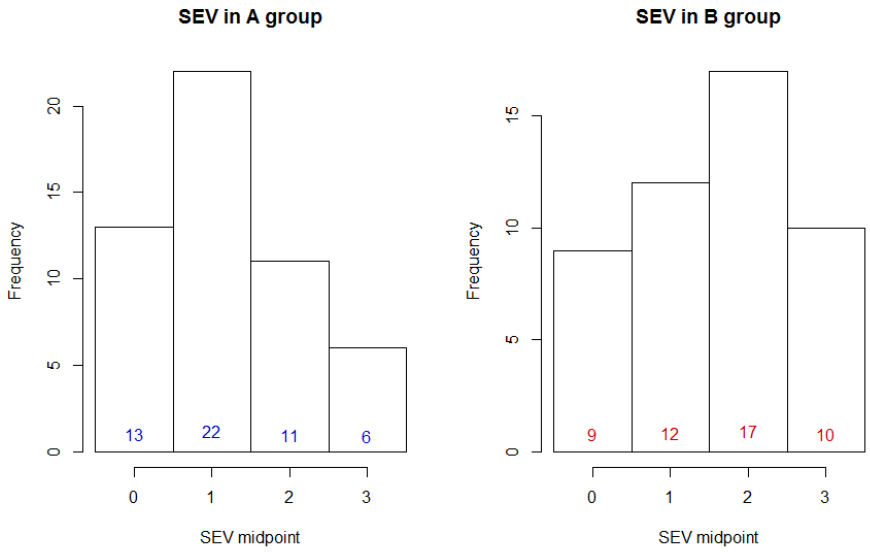
SAS 코드

```
PROC CHART DATA = TRIAL;
        VBAR SEV / MIDPOINTS = 0 1 2 3 GROUP = TRT;
RUN;
```

R 코드

```
trial.6<-subset(trial.1,select=c(2,7))
trial.6.A<-subset(trial.6,TRT=="A")
trial.6.B<-subset(trial.6,TRT=="B")
hist.a<-hist(trial.6.A$SEV, freq=TRUE,main="SEV in A group",xlab="SEV midpoint",mids<-c(-0.5,0.5,1.5,2.5,3.5))
text(hist.a$mids,hist.a$density,hist.a$counts,adj=c(0.5,-0.5),col="blue3")

hist.b<-hist(trial.6.B$SEV, freq=TRUE,main="SEV in B group",xlab="SEV midpoint",mids<-c(-0.5,0.5,1.5,2.5,3.5))
text(hist.b$mids,hist.b$density,hist.b$counts,adj=c(0.5,-0.5),col="red3")
```

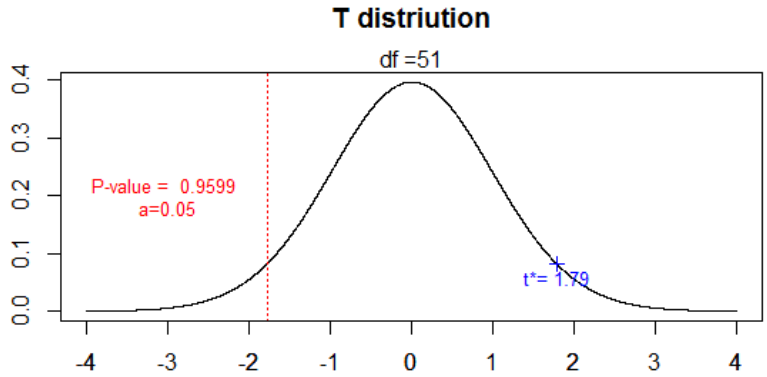


2. SCORE 가 20 보다 작으면 treatment 의 효과가 있다고 알려져 있다고 가정하자. 이 때 treatment A 가 효과가 있는지를 알고자 한다.

1) 귀무가설과 대립가설을 쓰시오.

$H_0: \mu_A = 20$ vs $H_1: \mu_A < 20$

2) 위의 가설을 그림으로 확인하고자 한다. 해당하는 그림을 그리시오.



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

한 변수의 평균이 특정값과 같은지 아닌지를 알아보기 위한 검정으로, 단일 평균치를 분석하는 가장 간단한 One sample t-test 를 수행한다. 대립가설이 '표본의 평균이 20 보다 작다.'이므로 양측 검정이 아닌, 단측 검정을 하는 것이 옳다. 현재 모집단의 분포형태는 모르지만, sample 의 크기가 100 으로 큰 크기의 표본이 추출된 경우이므로 t 검정을 수행할 수 있다.

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

SAS 시스템

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	-Infty	32.9343	26.9507	22.5860	33.4223

DF	t Value	Pr < t
51	1.79	0.9599

R result

```
        One Sample t-test

data:  data3.1[TRT == "A", ]$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
```

SAS 와 R 에서 동일한 결과를 줌을 알 수 있었고, 1-pvalue<0.05 이므로 귀무가설을 유의수준 0.05 에서는 기각하지 못하므로 treatment A 의 효과가 있다고 결론 내릴 수 없다..

3. SCORE 변수를 이용하여

1) TREATMENT GROUP 간에 차이가 있는지를 ONE-WAY ANOVA 를 이용하여 분석하시오.

SAS 시스템

The GLM Procedure

E

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

```
> Anova(lm(SCORE ~ TRT),type="II")
Anova Table (Type II tests)

Response: SCORE
          Sum Sq Df F value Pr(>F)
TRT         3394  1  4.2375 0.0422 *
Residuals  78484 98
```

```
> Anova(lm(SCORE ~ TRT),type="III")
Anova Table (Type III tests)

Response: SCORE
             Sum Sq Df F value    Pr(>F)
(Intercept)  36996  1 46.1949 8.417e-10 ***
TRT           3394  1  4.2375    0.0422 *
Residuals    78484 98
```

SAS에서의 결과와 R에서의 결과가 같음을 볼 수 있다. SAS에서는 ANOVA test를 수행하는 프로시져도 있지만, PROC GLM을 써도 동일한 결과가 나오기 때문에 이를 사용하였다. TRT에 대한 효과의 pvalue가 0.0422이므로 유의수준 0.05에서 귀무가설 $\boldsymbol{H_0}: \boldsymbol{\mu_A} = \boldsymbol{\mu_B}$ 을 기각하고 대립가설 $\boldsymbol{H_1}: \boldsymbol{\mu_A} \neq \boldsymbol{\mu_B}$ 을 채택한다. 다시 말해, oneway anova 분석결과 treatment group간에 통계적으로 유의한 차이가 있다고 결론 내릴 수 있다.

2) TREATMENT GROUP 과 CENTER 간에 교호작용이 있는지를 알고자 한다.

ANOVA 를 이용하여 결론을 내리시오.

SAS 시스템

The GLM Procedure

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

```
> Anova(lm(SCORE ~ TRT+CENTER+TRT*CENTER),type="II")
Anova Table (Type II tests)

Response: SCORE
           Sum Sq Df F value  Pr(>F)
TRT          3567  1  4.5080 0.03631 *
CENTER       1130  1  1.4285 0.23496
TRT:CENTER   1396  1  1.7644 0.18723
Residuals   75958 96
```

```
> Anova(lm(SCORE ~ TRT+CENTER+TRT*CENTER),type="III")
Anova Table (Type III tests)

Response: SCORE
             Sum Sq Df F value  Pr(>F)
(Intercept)    846  1  1.0692 0.30373
TRT           3349  1  4.2321 0.04238 *
CENTER        2508  1  3.1700 0.07816 .
TRT:CENTER    1396  1  1.7644 0.18723
Residuals    75958 96
```

TRT*CENTER 의 p-value 가 0.18723 으로 귀무가설인 $\boldsymbol{H_0}: (\boldsymbol{\alpha\beta})_{ij} = \mathbf{0}$ 를 기각하지 못한다.

따라서 TREATMENT GROUP 과 CENTER 간에 교호작용은 통계적으로 유의하지 않다고 본다.

3) TREATMENT GROUP 과 CENTER 간에 교호작용이 없다고 가정하고 TREATMENT 와 CENTER 간 차이를 알고자 한다. ANOVA 를 이용하여 결론을 내리시오.

SAS 시스템

The GLM Procedure

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

```
> Anova(lm(SCORE ~ TRT+CENTER),type="II")
Anova Table (Type II tests)

Response: SCORE
          Sum Sq Df F value Pr(>F)
TRT         3567  1  4.4728 0.0370 *
CENTER      1130  1  1.4173 0.2368
Residuals  77354 97
```

```
> Anova(lm(SCORE ~ TRT+CENTER),type="III")
Anova Table (Type III tests)

Response: SCORE
             Sum Sq Df F value  Pr(>F)
(Intercept)    4588  1  5.7528 0.01837 *
TRT            3567  1  4.4728 0.03700 *
CENTER         1130  1  1.4173 0.23675
Residuals     77354 97
```

SAS output의 Type III을 보면,. TRT의 P-value는α=0.05보다 작아 $H_0: \beta_{TRT} = 0$를 기각한다. 그러나 Center의 P-value는 0.4939로 α=0.05보다 크므로 $H_0: \beta_{CENTER} = 0$를 기각하지 못하므로 Treatment group에 따라 SCORE값은 차이가 나지만 Center간의 차이는 통계적으로 유의하지 않다.

4) 3)에서 TREATMENT 또는 CENTER 간 차이가 있다는 결론이 나왔을 경우 어떻게 해야 하는가?

만약 TREATMENT 간 차이가 있고 CENTER 간에는 차이가 없다면, CENTER 에 대한 효과는 모형에서 빼고, 약의 효과가 어느 정도인지에 대해 자세히 추가적으로 분석을 해야 한다.

만약 CENTER 간 차이가 있다는 결론이 나온다면, 우리가 알고자 하는 실험의 목적은 결국, TREATMENT A 의 효과이므로, 모형에 CENTER 변수를 넣어, 그 효과를 제어한 후에 A 의 효과에 대해 검증해야 한다.

APPENDIX

R code

```
###############1:: 앞 페이지 table 첨부####################

###############  2-1  ######################

(t.A<-t.test(data3.1[TRT=='A',]$SCORE,mu=20, alternative = c("less")))

###############  2-2  ######################

par(mfrow=c(1,1))

x<-seq(-4,4,by=0.001)

plot(x,dt(x,df=nrow(data3.1[TRT=='A',])),type="l",xlab=" ",ylab=" ",

    main="T distriution")

points(t.A$statistic,dt(t.A$statistic,51),pch=3,col="blue")

abline(v=qt(1-t.A$p.value,51),col="red",lty=3)  # 기준

text(t.A$statistic, 0.06,paste("t*=",round(t.A$statistic,2)),cex=0.8,col="blue")

text(-3, 0.2,paste("P-value = ",round(t.A$p.value,4),"\n a=0.05 "),cex=0.8,col="red")

result <- paste("df =51")

mtext(result,3)

axis(1, at=seq(-4, 4, 1))

###############  3-1  ######################

library(car)

attach(data3.1)

Anova(lm(SCORE ~ TRT),type="II")

Anova(lm(SCORE ~ TRT),type="III")

###############  3-2  ######################

Anova(lm(SCORE ~ TRT+CENTER+TRT*CENTER),type="II")

Anova(lm(SCORE ~ TRT+CENTER+TRT*CENTER),type="III")

###############  3-3  ######################

Anova(lm(SCORE ~ TRT+CENTER),type="II")

Anova(lm(SCORE ~ TRT+CENTER),type="III")
```

SAS code

```
DATA TRIAL;
INFILE "C:\Users\user\Desktop\trial3.csv"  DELIMITER="," FIRSTOBS = 2;
INPUT TRT $ CENTER PAT SEX $ AGE SCORE @@;
RUN;

/* 2_2 */
DATA PROBLEM2_2;
SET TRIAL (KEEP = TRT SCORE);
RUN;
DATA A_TRT;
SET PROBLEM2_2;
IF TRT = "A";
RUN;
/* TRT_A :: T-TEST */
PROC TTEST H0 = 20  SIDES=L DATA = A_TRT;
VAR SCORE;
RUN;

/* 3_1*/
DATA PROBLEM3;
SET TRIAL (KEEP = TRT CENTER SCORE);
RUN;

/* ONE WAY ANOVA FOR TRT */
PROC ANOVA DATA = PROBLEM3;
CLASS TRT;
MODEL SCORE = TRT;
RUN;

PROC GLM DATA = PROBLEM3;
CLASS TRT;
MODEL SCORE = TRT;
RUN;

/* 3_2 */
PROC GLM DATA = PROBLEM3;
CLASS TRT CENTER;
MODEL SCORE = TRT CENTER TRT*CENTER;
RUN;

/* 3_3 */
PROC GLM DATA = PROBLEM3;
CLASS TRT CENTER;
MODEL SCORE = TRT CENTER;
RUN;
```