STA106 - Project1

Shih-Chi Chen

2/8/2021

I. Introduction

This project aims at analyzing whether there are any differences of average length of patient's hospital stay among 4 regional hospital (NC, NE, S, W). The cell means model ($Y_{ij} = \mu_i + \varepsilon_{ij}$) is applied. By looking at some plots of this data, the normality, equal variance assumption and outliers are analyzed at first. Then, the ANOVA table is used to analyze the p-value. Based on the p-value, if the average length of patient's hospital stay for these four regions hospitals are not the same, the confidence interval is used to analyze which region hospitals are different or not.

II. Summary of data

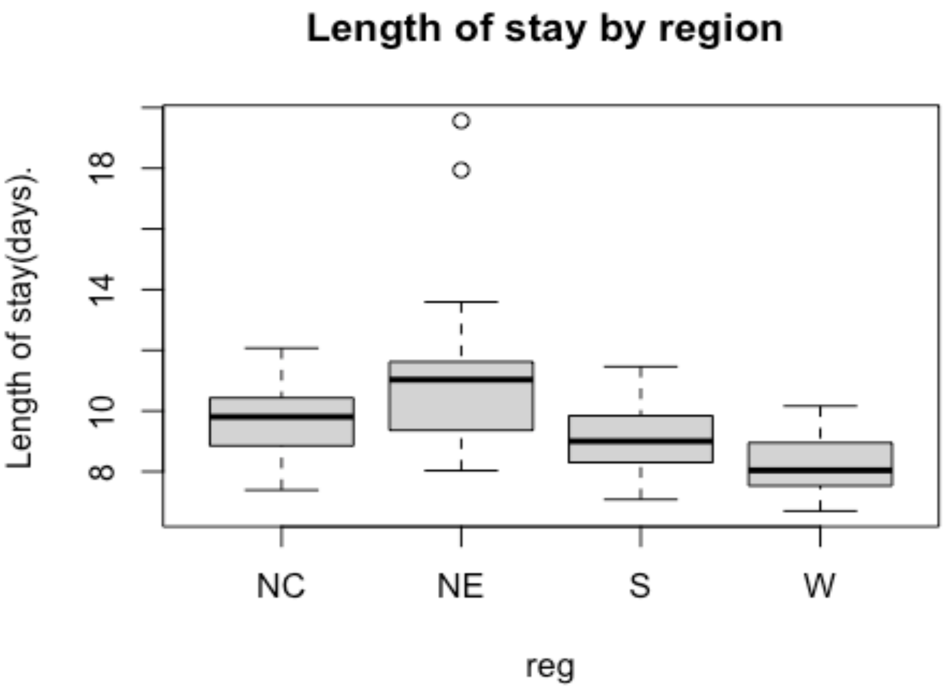
- Sample mean and Standard deviation

The null hypothesis should be rejected because some group's mean values are quite different, such as NE and W. The mean of NE is three standard deviations away from the mean of W; it means that NE and W are much different.

	NC	NE	S	W
Mean	9.683438	11.19429	9.20303	8.218571
Standard Deviations	1.192938	2.936654	1.269529	1.027999

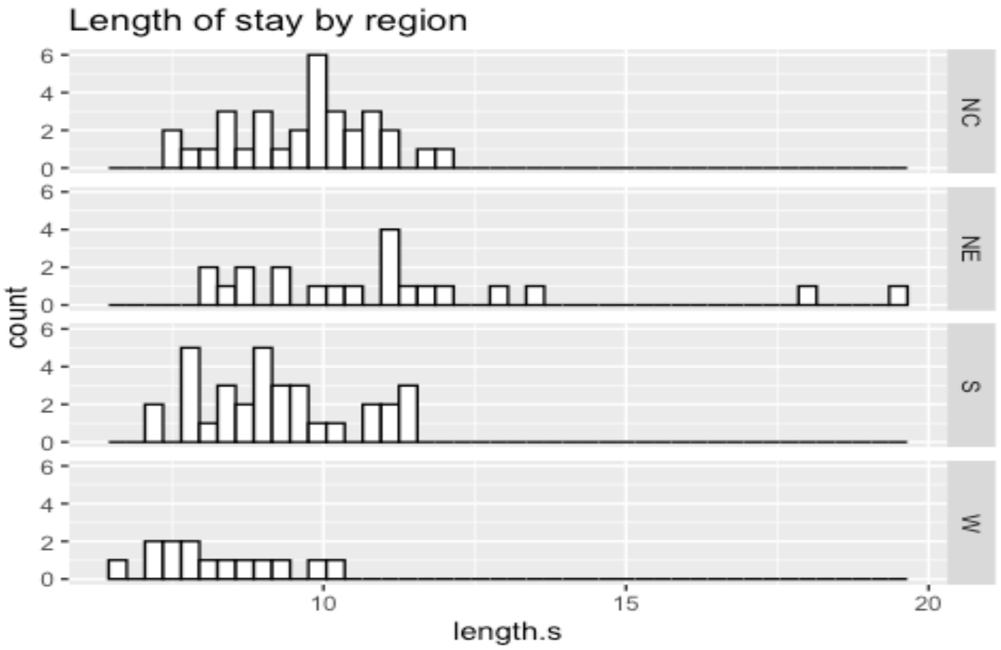
- Boxplots

The NE group is found to have the significant differences on median of long stay, which is larger than other groups. Also, the NE group has the largest IQR (interquartile range), which means that data of this group are more scattered.



- Histogram

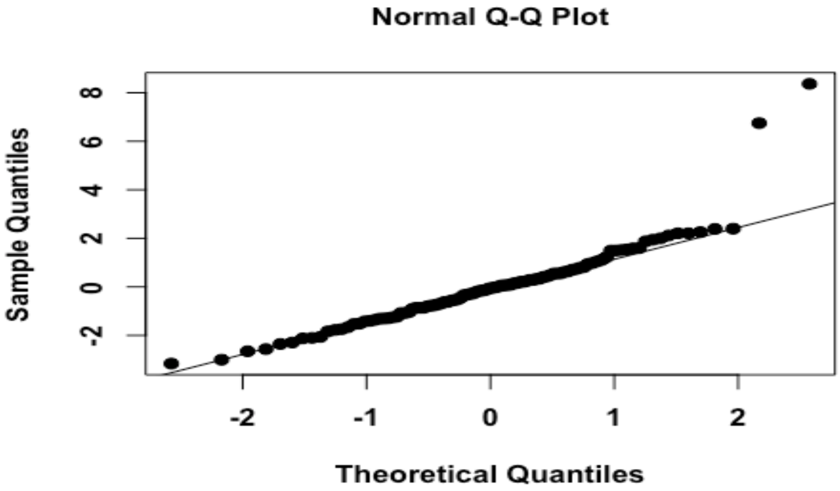
It is found that the smallest length of stay day occurs in the W group while the largest length of stay day occurs in the NE group. Therefore, data of the NE group are more scattered than other groups. Besides, there are some outliers in the NE group.



III. Diagnostics

- Check Normality: Q-Q Plot

From the Q-Q Plot, the data seem to be approximately normal because most of the dots are close to the line or on the line although there are some outliers exist.



- Check Normality: Shaprio-Wilks test

Since plots are subjective, the Shaprio-Wilks test is used to check.

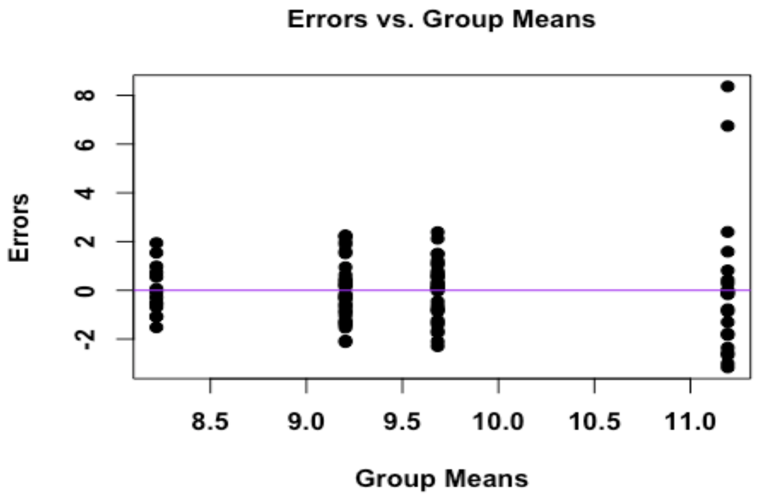
$H_0 =$ Data are normally distributed.

$H_a =$ Data are not normally distributed.

Since p-value is very small (2.200273e-07), the H_0 is rejected, and it is concluded that the data are not normally distributed. Therefore, the normality assumption is violated.

- Check Equal Variance: Error vs. Group Means Plot

From this plot, it is found that the vertical spreads of variance are not equal. Especially, the variance of NE group (rightmost on the plot) is larger than that of others. Therefore, the equal variance assumption is violated.



- Check Equal Variance: Brown-Forsythe test

The Brown-Forsythe test is used to check the equal variance.

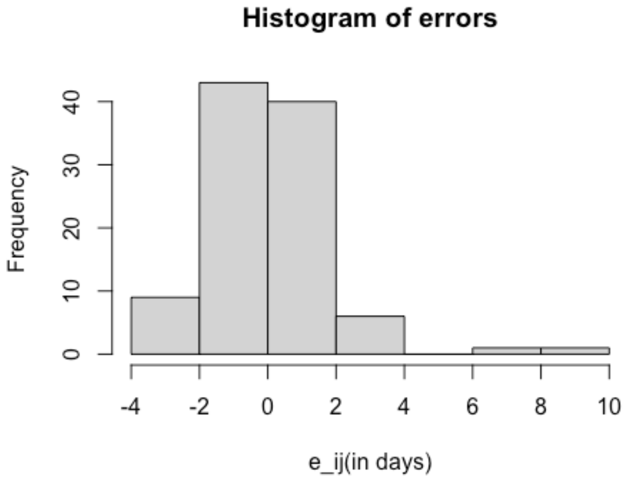
$H_0 = \sigma_{NC}^2 = \sigma_{NE}^2 = \sigma_S^2 = \sigma_W^2$

$H_a =$ at least one σ_i^2 is not equal.

It is found that the p-value (0.0102322) is smaller than α = 0.05 but a little bit larger than α = 0.01. Therefore, the conclusion depends on what values of α used. However, the equal variance is still a problem because the plot shows that the variances are significant different.

- Check Outliers: Histogram of Errors

Based on the Histogram, it is found that the residuals fall more than 4 standard deviations away from 0, which means that there are some outliers existing.

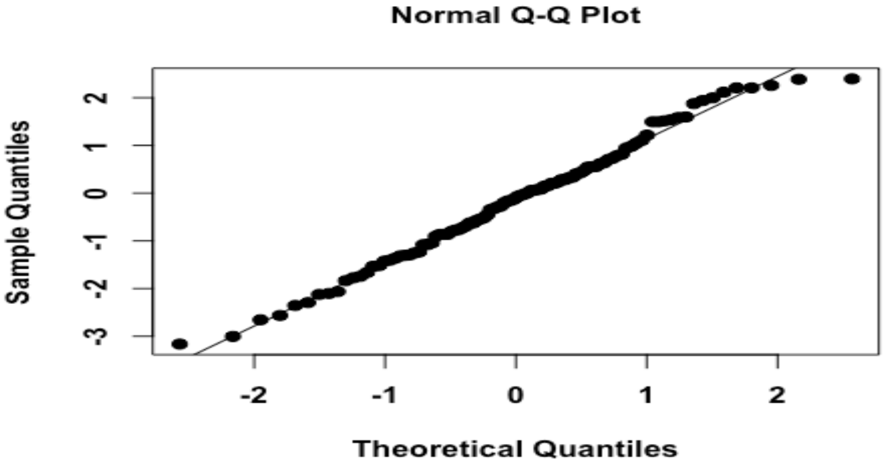


Remove outliers

Any outliers can be removed according to Standardized Residuals, with a cutoff of $t_{1-0.01;nt-a}$

- Check Normality: Q-Q Plot

  After the outliers are removed, all the dots seem to be close to the line or on the line.

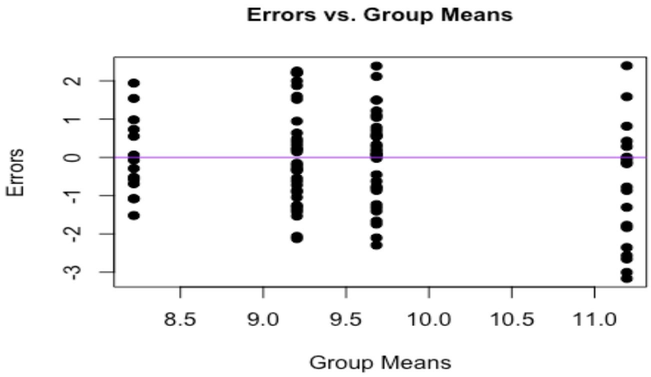


- Check Normality: Shaprio-Wilks test

The p-value (0.5054769) in Shaprio-Wilks test becomes larger than before. Therefore, It could be concluded that the data are normally distributed after the outliers are removed.

- Check Equal Variance: Error vs. Group Means Plot

After the outliers are removed, it is found that the vertical spreads are less significantly different than before.

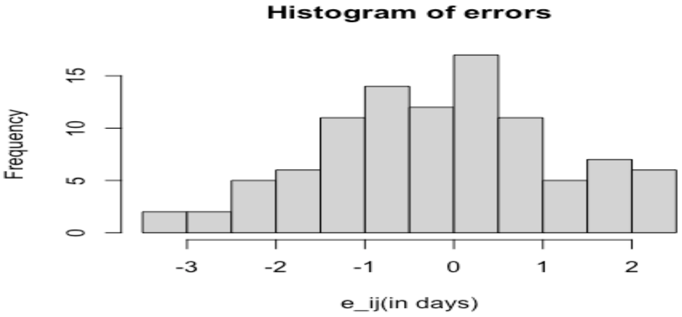


- Check Equal Variance: Brown-Forsythe test

The p-value (0.2668604) in Brown-Forsythe test becomes larger than before. Therefore, it can be concluded that the equal variance assumption is hold after the outliers are removed.

- Check Outliers: Histogram of Errors

Based on the Histogram, it is found that the residuals don't fall more than 4 standard deviations away from 0.



IV. Analysis

- ANOVA

State the null and alternative for testing

H_0: $\mu_{NC} = \mu_{NE} = \mu_S = \mu_W$

H_a: At least one μ_i is not equal

In the ANOVA table, the P-value = 1.198e-05. Since p-value is very small, H_0 will be rejected by more evidences.

	Df	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
reg	3	85.379	28.4597	9.6916	1.198e-05
Residuals	96	281.908	2.9365		

- Power Calculations

From the function code, it is found that Power = 0.9969844 when $\alpha = 0.05$ and Power = 0.9822976 when $\alpha = 0.01$. Since the power is large, H_0 will be rejected.

- Confidence Intervals (95% confidence intervals)

Confidence interval for each μ_i :

It is found that the NE group has longer days (11.93656) of upper bound and the W group has shorter days (7.309473) of lower bound.

Confidence interval for $\mu_{i-} \mu_i{}'$:

Since H_0 is rejected, every possibility of each two groups is compared to find which groups are different or not.

The 95% confidence interval for $\mu_{NC-} \mu_{NE}$ is: (-2.4661225, -0.5555739)

The 95% confidence interval for $\mu_{NC-} \mu_S$ is: (-0.3635093, 1.3243237)

The 95% confidence interval for $\mu_{NC-} \mu_W$ is: (0.3748957, 2.5548364)

The 95% confidence interval for $\mu_{NE-} \mu_S$ is: (1.041733, 2.940777)

The 95% confidence interval for $\mu_{NE-} \mu_W$ is: (1.802074, 4.149355)

The 95% confidence interval for $\mu_{S-} \mu_W$ is: (-0.1004736, 2.0693914)

From these confidence intervals, it is found that NC group and S group are not different. In addition, S group and W group are not different. The rest of 4 confidence intervals show that the true average lengths of stay days are different. Especially, the largest difference is 4.149355 days, which occurs between NE group and W group.

V. Interpretation

From the F test-statistic, very small p-value is obtained, which means that H_0 should be rejected so that the true average lengths of patients stay at hospital for at least one region are not the same. From the power calculation, the large power value is found, which can also support this inference. Since larger power = larger ϕ, it means more differences existing in the μ_i

After at least one group is found to be different, the differences among groups can be traced. Therefore, confidence intervals can be used to check.

From the CI for $\mu_i - \mu_i{}'$, it is found that:

The 95% confidence interval for $\mu_{NC} - \mu_{NE}$ is: (-2.4661225, -0.5555739)

It means that we are 95% confident that the true average length of stay days of NC group is less than that of NE group by between 0.5555739 days and 2.4661225 days.

The 95% confidence interval for $\mu_{NC} - \mu_S$ is: (-0.3635093, 1.3243237)

It means that the true average lengths of stay days of NC group and S group are not significant different because the confidence interval contains zero.

The 95% confidence interval for $\mu_{NC} - \mu_{W}$ is: (0.3748957, 2.5548364)

It means that we are 95% confident that the true average length of stay days of NC group is more than W group by between 0.3748957 days and 2.5548364 days.

The 95% confidence interval for $\mu_{NE} - \mu_{S}$ is: (1.041733, 2.940777)

It means that we are 95% confident that the true average length of stay days of NE group is more than S group by between 1.041733 days and 2.940777 days.

The 95% confidence interval for $\mu_{NE} - \mu_{W}$ is: (1.802074, 4.149355)

It means that we are 95% confident that the true average length of stay days of NE group is more than W group by between 1.802074 days and 4.149355 days.

The 95% confidence interval for $\mu_{S} - \mu_{W}$ is: (-0.1004736, 2.0693914)

It means that the true average lengths of stay days of S group and W group are not significant different because the confidence interval contains zero.

From the CI for μ_i only, it is found that:

The 95% confidence interval for μ_{NC} is: (9.082126, 10.284749)

It means that we are 95% confident that the true average length of stay days of NC group is between 9.082126 days and 10.284749 days.

The 95% confidence interval for μ_{NE} is:(10.45201, 11.93656)

It means that we are 95% confident that the true average length of stay days of NE group is between 10.45201 days and 11.93656 days.

The 95% confidence interval for μ_{S} is:(8.610899, 9.795161)

It means that we are 95% confident that the true average length of stay days of S group is between 8.610899 days and 9.795161 days.

The 95% confidence interval for μ_{W} is:(7.309473, 9.127670)

It means that we are 95% confident that the true average length of stay days of W group is between 7.309473 days and 9.127670 days.

VI. Conclusion

Firstly, the normality of this data look acceptable from the Q-Q plot while the equal variance could be the problem because significant differences are found in the Error vs. group means plot. After some outliers are removed, the Q-Q plot looks better than before because all dots are close to the line or on the line. Furthermore, the equal variance assumption in Error vs. group means plot looks acceptable, too. Therefore, the Shaprio-Wilks test and the Brown-Forsy test are applied to double check the assumption. As a result, the p-values are obviously larger than before; it is indicates that the normality and equal variance are hold after the outliers are removed.

Secondly, very small p-values are found in the ANOVA table. It means that more evidences are available to reject H_0 and the outcome that the true average length of stay days for at least one group is not equal can be concluded. Accordingly, it can also be concluded the true average average lengths of stay days for these four regions are not all the same. Such outcomes can also be observed from the boxplot. In the boxplot, it clearly shows that the NE region is significantly higher than other regions. Therefore, p-value and boxplot have the same conclusions.

Thirdly, in each region confidence interval, some information about regional data can be observed. For example, it is found that the highest bound is in the NE region, which means that most of the patients stay longer days in NE region hospital than other region hospitals. On the other hand, the lowest bound is in the W region, which means that most of patients stay short days in W region hospital than other region hospitals. Therefore, W region hospital seems to be more acceptable.

Finally, a meaningful observation is to trace out the degree of difference between each region pair. Thus, the confidence interval is used to compare each region pair. It is found that NC region and S region are not significant different, and S region and W region are not significant different, either. Therefore, it can be concluded that the true average length of stay days between NC region and NE region, NC region and W region, NE region and S region, NE region and W region are significant different. Among them, the largest difference is found between NE region and W region. Accordingly, the widest confidence interval is also found in these two regions.

VI. Appendix

```r
#sample mean and sd
senic=read.csv("~/Downloads/senic.csv")
length.s=senic$Length
reg=senic$Region
group.means = by(length.s,reg,mean)
group.means
group.sd = by(length.s,reg,sd)
group.sd

#boxplot
boxplot(length.s ~ reg, main = "Length of stay by region ",ylab = "Length of
stay(days).")

#histogram
library(ggplot2)
ggplot(senic, aes(x = length.s)) + geom_histogram(binwidth = 0.3,color = "bla
ck",fill = "white") +
facet_grid(senic$Region ~.) +ggtitle("Length of stay by region")

#QQ-plot
the.model = lm(length.s ~ reg, data = senic)
senic$ei = the.model$residuals
qqnorm(the.model$residuals,pch = 19,font = 2,font.lab =2,cex =1,cex.lab=1, ce
x.axis=1, cex.main=1, cex.sub=1)
qqline(the.model$residuals)

#Shaprio-Wilks test
sp.pval = shapiro.test(the.model$residuals)$p.val
sp.pval

#Error vs group means plot
plot(the.model$fitted.values, the.model$residuals, main = "Errors vs. Group M
eans",xlab = "Group Means",ylab = "Errors",pch = 19,font = 2,font.lab =2,cex
=1,cex.lab=1, cex.axis=1, cex.main=1, cex.sub=1)
abline(h = 0,col = "purple")

#Brown-Forsythe test
the.BFtest = leveneTest(the.model$residuals~reg , data=senic, center=median)
bf.p.val = the.BFtest[[3]][1]
bf.p.val

#Histogram of errors(outlier)
hist(senic$ei,main = "Histogram of errors",xlab = "e_ij(in days)")

#QQ plot (Remove outliers)
alpha = 0.01
a = length(unique(senic$Region))
nt = nrow(senic)
```

```
rij = rstandard(the.model)
t.cutoff= qt(1-alpha, nt-a)
CO.rij = abs(rij) > t.cutoff
qqnorm(the.model$residuals[!CO.rij],pch = 19,font = 2,font.lab =2,cex =1,cex.
lab=1, cex.axis=1, cex.main=1, cex.sub=1)
qqline(the.model$residuals)

#Shaprio-Wilks test (Remove outliers)
sp.pval.2 = shapiro.test(the.model$residuals[!CO.rij])$p.val
sp.pval.2

#Errors vs. Group Means plot (Remove outliers)
plot(the.model$fitted.values[!CO.rij], the.model$residuals[!CO.rij], main = "
Errors vs. Group Means",xlab = "Group Means",ylab = "Errors",pch = 19,font =
1,font.lab =1,cex =1,cex.lab=1, cex.axis=1, cex.main=1, cex.sub=1)
abline(h = 0,col = "purple")

#Brown-Forsythe test (Remove outliers)
the.BFtest2 = leveneTest(the.model$residuals[!CO.rij]~reg[!CO.rij] , data=sen
ic, center=median)
bf.p.val.2 = the.BFtest2[[3]][1]
bf.p.val.2

#Histogram of errors (Remove outliers)
hist(the.model$residuals[!CO.rij],main = "Histogram of errors",xlab = "e_ij(i
n days)")

#ANOVA
anova.table = anova(the.model)
anova.table

#Power Calculations
give.me.power = function(ybar,ni,MSE,alpha){
a = length(ybar) # Finds a
nt = sum(ni) #Finds the overall sample size
overall.mean = sum(ni*ybar)/nt # Finds the overall mean
phi = (1/sqrt(MSE))*sqrt( sum(ni*(ybar - overall.mean)^2)/a) #Finds the books
 value of phi
phi.star = a *phi^2 #Finds the value of phi we will use for R
Fc = qf(1-alpha,a-1,nt-a) #The critical value of F, use in R’s function
power = 1 - pf(Fc, a-1, nt-a, phi.star)# The power, calculated using a non-ce
ntral F
return(power)
}
group.means = by(senic$Length,reg,mean)
group.nis = by(senic$Length,senic$Region,length)
MSE = anova.table[2,3]
the.power.5 = give.me.power(group.means,group.nis,MSE,0.05)
the.power.1 = give.me.power(group.means,group.nis,MSE,0.01)
the.power.5
the.power.1

#CI only ui
give.me.CI = function(ybar,ni,ci,MSE,multiplier){
if(sum(ci) != 0 & sum(ci !=0 ) != 1){
return("Error - you did not input a valid contrast")
} else if(length(ci) != length(ni)){
return("Error - not enough contrasts given")
}
else{
estimate = sum(ybar*ci)
SE = sqrt(MSE*sum(ci^2/ni))
CI = estimate + c(-1,1)*multiplier*SE
result = c(estimate,CI)
names(result) = c("Estimate","Lower Bound","Upper Bound")
return(result)
}
}
t.value = qt(1-0.05/2, sum(group.nis) - length(group.nis))
ci.nc=c(1,0,0,0)
ci.ne=c(0,1,0,0)
ci.s=c(0,0,1,0)
ci.w=c(0,0,0,1)
CI.nc = give.me.CI(group.means,group.nis,ci.nc,MSE,t.value)
CI.ne = give.me.CI(group.means,group.nis,ci.ne,MSE,t.value)
CI.s = give.me.CI(group.means,group.nis,ci.s,MSE,t.value)
CI.w = give.me.CI(group.means,group.nis,ci.w,MSE,t.value)
CI.nc
CI.ne
CI.s
CI.w
#CI for ui-ui'
ci.1 = c(1,-1,0,0)
ci.2 = c(1,0,-1,0)
ci.3 = c(1,0,0,-1)
ci.4 = c(0,1,-1,0)
ci.5 = c(0,1,0,-1)
ci.6 = c(0,0,1,-1)
CI.ncne = give.me.CI(group.means,group.nis,ci.1,MSE,t.value)
CI.ncs = give.me.CI(group.means,group.nis,ci.2,MSE,t.value)
CI.ncw = give.me.CI(group.means,group.nis,ci.3,MSE,t.value)
CI.nes = give.me.CI(group.means,group.nis,ci.4,MSE,t.value)
CI.new = give.me.CI(group.means,group.nis,ci.5,MSE,t.value)
CI.sw = give.me.CI(group.means,group.nis,ci.6,MSE,t.value)
CI.ncne
CI.ncs
CI.ncw
CI.nes
CI.new
CI.sw
```