R통계 과제4 대륙별 코로나 징후 분산분석

조성우_201823869

2020년 11월 2일

분산분석

분산 분석(analysis of variance, ANOVA, 변량 분석)은 통계학에서 두 개 이상 다수의 집단을 비교하고자 할 때 집단 내의 분산, 총평균과 각 집단의 평균의 차이에 의해 생긴 집단 간 분산의 비교를 통해 만들어진 F분포를 이용하여 변수들의 모평균에 대한 가설검정을 하는 방법이다.

*1.*개요

목적:

Worldometers 웹사이트에 게시되는 전세계 코로나 확진 현황표를 스크래핑한 자료를 데이터셋으로, 전처리부터 기술통계, 조건 검증부터 분산분석 및 사후검증까지의 구조적인 과정을 통해 대륙별 COVID-19 발생률 및 사망률에 차이가 있는지 통계적으로 면밀히 규명해보고자 한다.

가설 설정:

발생률

귀무가설*(H0) :* 대륙 집단간 코로나 발생률 평균의 분산은 같을것이다.
연구가설*(H1) :* 대륙 집단간 코로나 발생률 평균의 분산은 같지 않을것이다.

사망률

귀무가설*(H0) :* 대륙 집단간 코로나 사망률 평균의 분산은 같을것이다.
연구가설*(H1) :* 대륙 집단간 코로나 사망률 평균의 분산은 같지 않을것이다.

- **독립변수 *: [대륙]* 유럽,북아메리카,아시아,남아메리카,아프리카,오세아니아**
  -> 비연속척도*(명목척도/서열척도)*
- **종속변수 *:* 발생률과 사망률**
  -> 연속척도*(명목척도/서열척도)*

활용 데이터셋:

---

https://www.worldometers.info/coronavirus/ (https://www.worldometers.info/coronavirus/) 에 게시되는 전세계 코로나 현황표를 스크래핑한 데이터셋

상세:

국가,대륙,총 확진자,신규확진자,사망자 등의 변수를 포함한 데이터셋이다.

해당 보고서의 분석을 위해
[Country:*국가*, Continent:*대륙*, Totalcases_per_1M:*발생률*, Deaths_per_1M:*사망률 *]을 이용한다.

2.본문

2.1 데이터 수집 및 전처리

2.1.1 Html source 스크래핑 및 table data 추출

```
#관련패키지 불러오기

#<<스크래핑>>
url = "https://www.worldometers.info/coronavirus/"
library(XML)
library(httr)


html_source = GET(url) # 해당 html의 전체 소스를 받아옴

tabs = readHTMLTable(rawToChar(html_source$content),stringsAsFactord=F) #html의 콘텐츠 중에서 테이블만 추출

world_table = tabs$main_table_countries_yesterday  # 11월 1일자 보고서를 추출하여 사용하도록 한다.

setwd("C:WWUsersWWJSWWWDesktopWW아주대학교WW1_학기WW3학년 2학기_2020_2WW강의자료WW1_R통계WWR통계 과제WW과제4_분산분석")
write.csv(world_table,"Rstat_ass4_covid19_1.csv",row.names=F)

world_table = read.csv("Rstat_ass4_covid19_1.csv")
```

```
str(world_table)
```

```
## 'data.frame':    226 obs. of  19 variables:
##  $ X.                 : int  NA NA NA NA NA NA NA NA 1 2 ...
##  $ Country.Other      : chr  "Asia" "North America" "South America" "Europe" ...
##  $ TotalCases         : chr  "13,941,168" "11,554,713" "9,777,852" "10,704,180" ...
##  $ NewCases           : chr  "+91,921" "+105,870" "+36,075" "+242,228" ...
##  $ TotalDeaths        : chr  "247,394" "356,158" "297,449" "273,962" ...
##  $ NewDeaths          : chr  "+1,512" "+1,545" "+970" "+3,849" ...
##  $ TotalRecovered     : chr  "12,433,058" "7,629,712" "8,769,829" "3,991,137" ...
##  $ NewRecovered       : chr  "+84,530" "+59,243" "+71,477" "+78,307" ...
##  $ ActiveCases        : chr  "1,260,716" "3,568,843" "710,574" "6,439,081" ...
##  $ Serious.Critical   : chr  "22,427" "21,501" "17,975" "23,802" ...
##  $ Tot훆Cases.1M.pop  : chr  "" "" "" "" ...
##  $ Deaths.1M.pop      : chr  "" "" "" "" ...
##  $ TotalTests         : chr  "" "" "" "" ...
##  $ Tests.1M.pop       : chr  "" "" "" "" ...
##  $ Population         : chr  "" "" "" "" ...
##  $ Continent          : chr  "Asia" "North America" "South America" "Europe" ...
##  $ X1.Caseevery.X.ppl : chr  "" "" "" "" ...
##  $ X1.Deathevery.X.ppl: chr  "" "" "" "" ...
##  $ X1.Testevery.X.ppl : chr  "" "" "" "" ...
```

스크래핑한 데이터 구조를 요약출력한 모습

데이터 특이사항 확인

- 일부 변수명 인코딩 오류

2.1.2 데이터 전처리

전처리 내용

- 변수명 변경
- 불필요한 변수를 제외한 데이터프레임 재구성
- 세계 및 대륙 총계 record(행) 제거
- Character type로 저장된 수치변수값을 numeric type으로 바꾸기

원데이터 확인

```
head(world_table,3)
```

```
##   X. Country.Other TotalCases NewCases TotalDeaths NewDeaths TotalRecovered
## 1 NA          Asia 13,941,168  +91,921     247,394    +1,512     12,433,058
## 2 NA North America 11,554,713 +105,870     356,158    +1,545      7,629,712
## 3 NA South America  9,777,852  +36,075     297,449      +970      8,769,829
##   NewRecovered ActiveCases Serious.Critical Tot휷Cases.1M.pop Deaths.1M.pop
## 1      +84,530   1,260,716           22,427
## 2      +59,243   3,568,843           21,501
## 3      +71,477     710,574           17,975
##   TotalTests Tests.1M.pop Population     Continent X1.Caseevery.X.ppl
## 1                                           Asia
## 2                                  North America
## 3                                  South America
##   X1.Deathevery.X.ppl X1.Testevery.X.ppl
## 1
## 2
## 3
```

스크래핑한 원데이터를 *3*행만 출력한 모습

변수명 변경

변수값이 존재하지 않거나 불필요한 변수를 논리형을 활용해 일괄적으로 제외시킨 데이터프레임을 만들기 위해 의미없는 변수의 변수명을 일괄적으로 a로 바꿔준다.

```
#<<데이터 전처리>>

colnames(world_table) = c('a',"Country","Total_case","New_case","Total_death","New_death","Total_r
ecovered",'a',"Active_case","Serious","Totalcases_per_1M","Deaths_per_1M","a","a",'a',"Continent",
'a','a','a')
 #(1)변수명 변경
```

불필요한 변수를 제외하고 데이터프레임 재구성

```
 #(2)불필요한 변수를 제하고 데이터프레임 재구성
world_table = world_table[,colnames(world_table)!='a']  #world_table2의 이름으로 전처리 데이터셋
 분기복제(추후 분석에서의 나머지 변수들의 필요성을 고려 첫번째 분기로 world_table을 저장해두고, wo
rld_table2로 이후 전처리 마저 실시)

write.csv(world_table,"Rstat_ass4_covid19_2.csv",row.names=F)


# 본문에 최우선적으로 필요한 변수만 뽑아내기 위한 2차 변수제외 실시
colnames(world_table) = c("Country","Total_case","New_case","Total_death","New_death","a","a","a",
"Totalcases_per_1M","Deaths_per_1M","Continent")

world_table2 = world_table[,colnames(world_table)!='a']
str(world_table2)
```

```
## 'data.frame':    226 obs. of  8 variables:
##  $ Country          : chr  "Asia" "North America" "South America" "Europe" ...
##  $ Total_case       : chr  "13,941,168" "11,554,713" "9,777,852" "10,704,180" ...
##  $ New_case         : chr  "+91,921" "+105,870" "+36,075" "+242,228" ...
##  $ Total_death      : chr  "247,394" "356,158" "297,449" "273,962" ...
##  $ New_death        : chr  "+1,512" "+1,545" "+970" "+3,849" ...
##  $ Totalcases_per_1M: chr  "" "" "" "" ...
##  $ Deaths_per_1M    : chr  "" "" "" "" ...
##  $ Continent        : chr  "Asia" "North America" "South America" "Europe" ...
```

```
write.csv(world_table2,"Rstat_ass4_covid19_3.csv",row.names=F)
```

필요한 변수로만 재구성된 데이터프레임의 모습

관측값에 문제가 있거나 불필요한 record(행) 제거

```
#세계 및 대륙총계 record 행 제거
world_table3 = world_table2[-(1:8),]

world_table3 = world_table3[world_table3$Country!="MS Zaandam",]
world_table3 <- world_table3[world_table3$Country!="Diamond Princess",] #대륙 factor값이 할당되지
 않은 유사국가 제외
rownames(world_table3) = 1:nrow(world_table3) # 행이름 재설정
```

변수 특성에 따라 본래의 모습에 맞는 데이터 타입으로 형변환

- 반복문의 변수로 데이터프레임의 인덱스를 지정하여 gsub과 as.numeric을 적용시켜 일괄적으로 처리한다.
- Continent변수의 관측값들은 factor형으로 변환

```
 #(4) 문자열로 저장된 수치형의 변수값을 수치형으로 바꾸기
  #--> 모든 수치형 record에 단위구분자로 삽입된 쉼표제거 & numeric으로 변형 실시
library(dplyr)
attach(world_table3)

for (i in 2:7){
world_table3[,i] <- as.numeric(gsub(",","",world_table3[,i]))
}

world_table3$Continent= factor(world_table3$Continent) # 대륙 변수값들은 factor로
levels(world_table3$Continent)
```

```
## [1] "Africa"           "Asia"             "Australia/Oceania"
## [4] "Europe"           "North America"    "South America"
```

```
write.csv(world_table3,"Rstat_Covid19_4.csv",row.names = FALSE)
```

*factor*로 변환된 대륙변수의 레벨 확인

```
str(world_table3)
```

```
## 'data.frame':    216 obs. of  8 variables:
##  $ Country          : chr  "China" "USA" "India" "Brazil" ...
##  $ Total_case       : num  86070 9692528 8312947 5567126 1673686 ...
##  $ New_case         : num  49 94463 46033 12920 18648 ...
##  $ Total_death      : num  4634 238641 123650 160548 28828 ...
##  $ New_death        : num  NA 1199 511 276 355 ...
##  $ Totalcases_per_1M: num  60 29223 6004 26127 11467 ...
##  $ Deaths_per_1M    : num  3 720 89 753 198 586 780 707 624 695 ...
##  $ Continent        : Factor w/ 6 levels "Africa","Asia",..: 2 5 2 6 4 4 4 6 6 4 ...
```

모든 전처리가 완료된 데이터셋의 구조

2.2 탐색적 기술통계

전처리한 데이터를 기술하며 해당 데이터가 가지는 기본 통계량을 탐색적으로 분석합니다.

기술통계 목차

2.2.0 covid_table에 관한 요약통계량 출력

```
covid_table = read.csv("Rstat_Covid19_4.csv")
covid_table$Continent = factor(covid_table$Continent)


######### 2.2.0 요약통계량 출력
summary(covid_table)
```

```
##    Country            Total_case         New_case      Total_death
##  Length:216         Min.   :      1   Min.   :    1   Min.   :     1
##  Class :character   1st Qu.:   1494   1st Qu.:   29   1st Qu.:    60
##  Mode  :character   Median :  11660   Median :  311   Median :   298
##                     Mean   : 221491   Mean   : 2992   Mean   :  6386
##                     3rd Qu.:  93519   3rd Qu.: 1470   3rd Qu.:  1809
##                     Max.   :9692528   Max.   :94463   Max.   :238641
##                                       NA's   :53      NA's   :25
##    New_death       Totalcases_per_1M Deaths_per_1M                Continent
##  Min.   :   1.00   Min.   :    3.0   Min.   :   0.08   Africa           :57
##  1st Qu.:   4.00   1st Qu.:  661.8   1st Qu.:  16.00   Asia             :49
##  Median :  13.00   Median : 4215.5   Median :  72.00   Australia/Oceania: 9
##  Mean   :  77.28   Mean   : 8603.9   Mean   : 170.96   Europe           :48
##  3rd Qu.:  66.00   3rd Qu.:12302.8   3rd Qu.: 205.50   North America    :39
##  Max.   :1199.00   Max.   :63512.0   Max.   :1237.00   South America    :14
##  NA's   :110                         NA's   :25
```

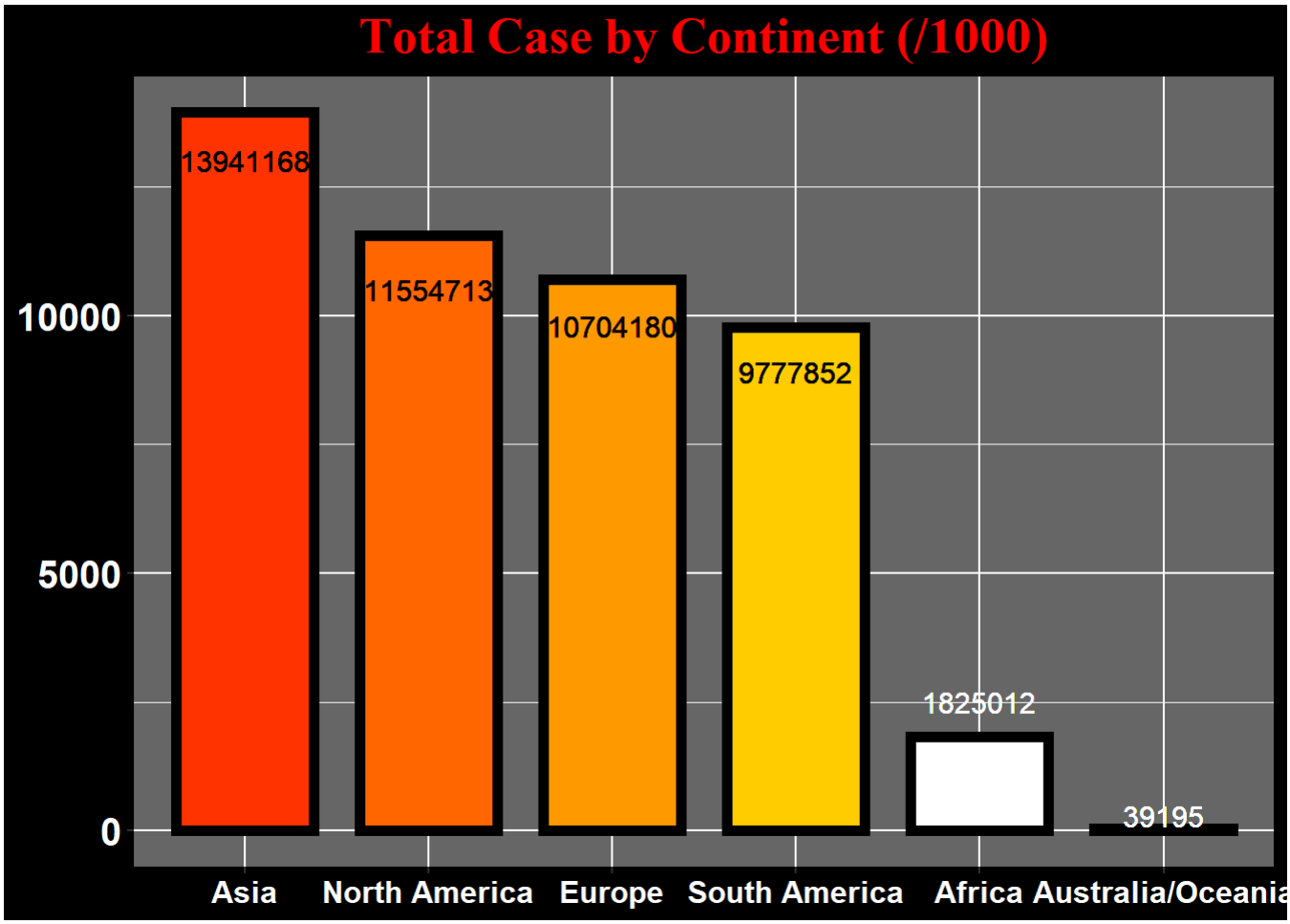
***covid_table*의 기본통계량을 *summary* 함수로 알아본 모습**

2.2.1 대륙별 발병자수 순위

```
library(ggplot2)

#########2.2.1 대륙별 발병자수 순위

Totalcase_byContinent = aggregate(Total_case~Continent,covid_table,FUN = sum) # aggregate함수를 사용하여 '대륙별 확진자수 총계 출력'
Totalcase_byContinent = Totalcase_byContinent[order(-Totalcase_byContinent$Total_case),] #확진자수별 순위정렬
rownames(Totalcase_byContinent) = NULL #행이름 재설정

print(Totalcase_byContinent) #표출력
```

```
##           Continent Total_case
## 1              Asia   13941168
## 2     North America   11554713
## 3            Europe   10704180
## 4     South America    9777852
## 5            Africa    1825012
## 6 Australia/Oceania      39195
```

```
      #바차트 출력

ggplot(Totalcase_byContinent,aes(reorder(Continent,-Total_case/1000),Total_case/1000,fill=Continent))+ geom_bar(width=0.75,stat = 'identity',colour='black',size=2)+ xlab("")+ylab("")+ggtitle('Total Case by Continent (/1000)') + theme(plot.title= element_text(family='serif',face="bold",hjust=0.5,size=20,color='RED'),legend.position='none') +labs(x=NULL,y=NULL) + theme(plot.subtitle =element_text(vjust=1),plot.caption=element_text(vjust=1),axis.text.x=element_text(angle=,color='white',size=12,face='bold'),axis.text.y=element_text(color='white',size=15,face='bold'))+
  geom_text(aes(x=1,y=13000,label=Totalcase_byContinent$Total_case[1]),color='black',size=4)+
  geom_text(aes(x=2,y=10500,label=Totalcase_byContinent$Total_case[2]),color='black',size=4)+
  geom_text(aes(x=3,y=9800,label=Totalcase_byContinent$Total_case[3]),color='black',size=4)+
  geom_text(aes(x=4,y=8900,label=Totalcase_byContinent$Total_case[4]),color='black',size=4)+
  geom_text(aes(x=5,y=2500,label=Totalcase_byContinent$Total_case[5]),color='white',size=4)+
  geom_text(aes(x=6,y=300,label=Totalcase_byContinent$Total_case[6]),color='white',size=4)+
  theme(panel.background = element_rect(fill='grey40'),plot.background = element_rect(fill='black'))+
 scale_fill_manual(values = c( "#FFFFFF", "#FF3300", "#FFFFFF",
                               "#FF9900", "#FF6600","#FFCC00"))
```



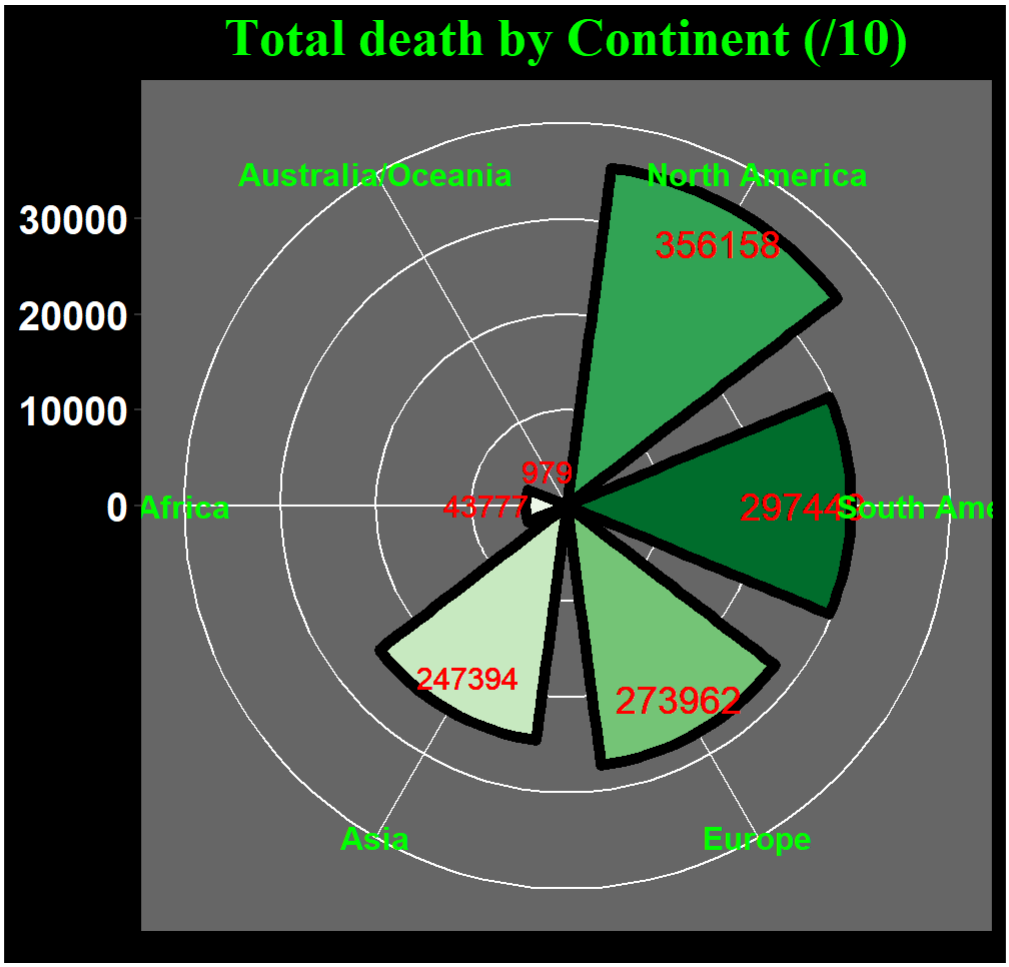
발병자수는 큰곳부터 아시아,북아메리카,유럽,남아메리카,아프리카,오세아니아 순으로 나타난다

2.2.2 대륙별 사망자수 순위

```
######### 2.2.2 대륙별 사망자수 순위

Totaldeath_byContinent = aggregate(Total_death~Continent,covid_table,FUN = sum) # aggregate함수를
 사용하여 '대륙별 사망자수 총계 출력'

Totaldeath_byContinent = Totaldeath_byContinent[order(-Totaldeath_byContinent$Total_death),]
rownames(Totaldeath_byContinent) = NULL #행이름 재설정

print(Totaldeath_byContinent) #표출력
```

```
##           Continent Total_death
## 1     North America      356158
## 2     South America      297449
## 3            Europe      273962
## 4              Asia      247394
## 5            Africa       43777
## 6 Australia/Oceania         979
```

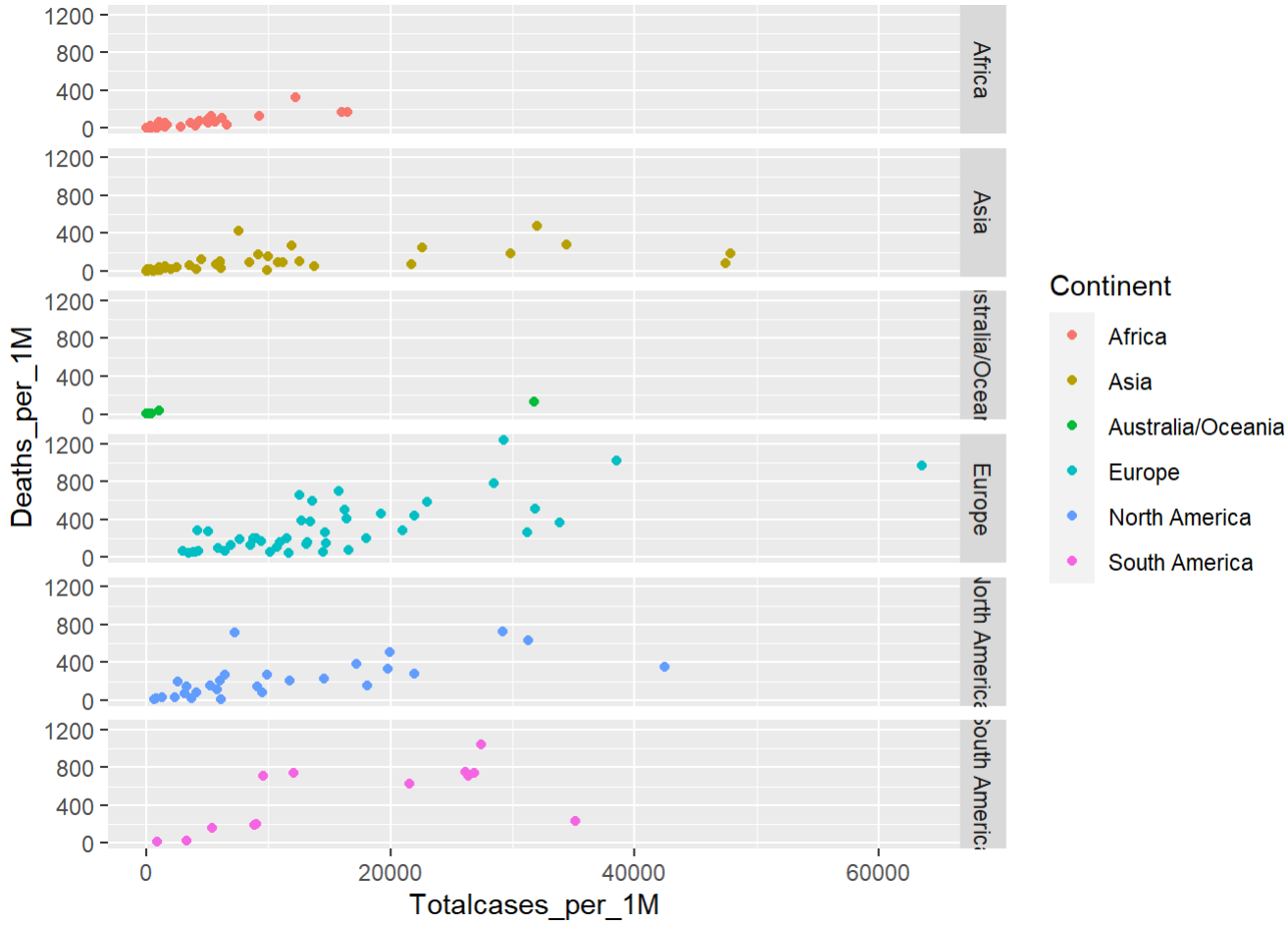
```
#레이더 차트 출력

ggplot(Totaldeath_byContinent,aes(reorder(Continent,-Total_death/10),Total_death/10,fill=Continen
t))+ geom_bar(width=0.75,stat = 'identity',colour='black',size=2)+ xlab("")+ylab("")+ggtitle('Tota
l death by Continent (/10)') + theme(plot.title= element_text(family='serif',face="bold",hjust=0.5
,size=20,color='green '),legend.position='none') +labs(x=NULL,y=NULL) + theme(plot.subtitle =eleme
nt_text(vjust=1),plot.caption=element_text(vjust=1),axis.text.x=element_text(angle=,color='green',
size=12,face='bold'),axis.text.y=element_text(color='white',size=15,face='bold'))+
  geom_text(aes(x=1,y=Totaldeath_byContinent$Total_death[1]/10 - 4000,label=Totaldeath_byContinent
$Total_death[1]),color='red',size=5)+
  geom_text(aes(x=2,y=Totaldeath_byContinent$Total_death[2]/10 - 5000,label=Totaldeath_byContinent
$Total_death[2]),color='red',size=5)+
  geom_text(aes(x=3,y=Totaldeath_byContinent$Total_death[3]/10 - 4000,label=Totaldeath_byContinent
$Total_death[3]),color='red',size=5)+
  geom_text(aes(x=4,y=Totaldeath_byContinent$Total_death[4]/10 - 4000,label=Totaldeath_byContinent
$Total_death[4]),color='red',size=4)+
  geom_text(aes(x=5,y=Totaldeath_byContinent$Total_death[5]/10 + 4000,label=Totaldeath_byContinent
$Total_death[5]),color='red',size=4)+
  geom_text(aes(x=6,y=Totaldeath_byContinent$Total_death[6]/10 + 4000,label=Totaldeath_byContinent
$Total_death[6]),color='red',size=4)+
  theme(panel.background = element_rect(fill='grey40'),plot.background = element_rect(fill='black'
))+
  scale_fill_brewer(palette = "green")+
  coord_polar()
```



사망자수는 큰곳부터 북아메리카,남아메리카,유럽,아시아,아프리카,오세아니아 순으로 나타난다

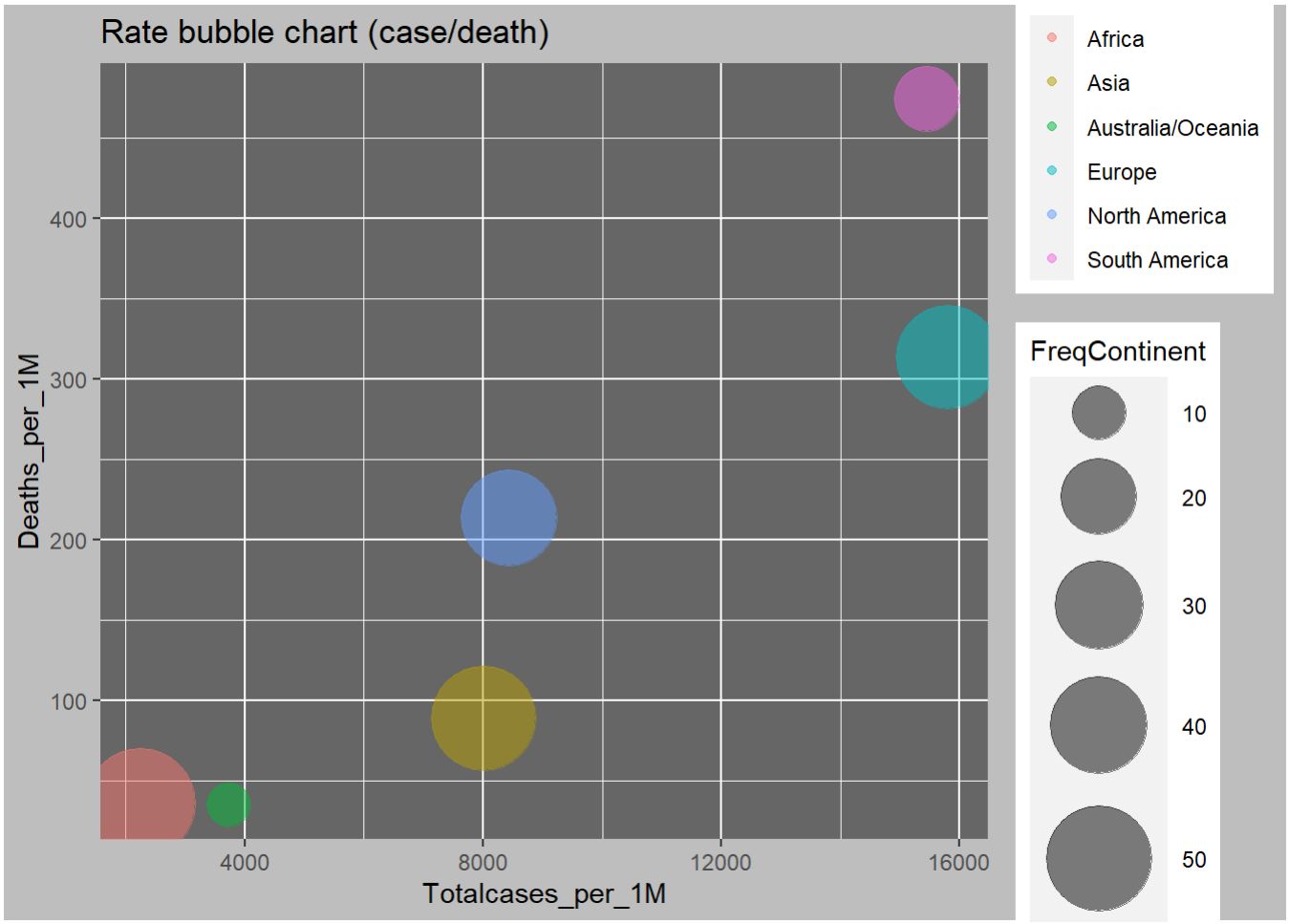
2.2.3 발병률 및 사망률

```
T_case = aggregate(Totalcases_per_1M~Continent,covid_table,FUN=mean)
T_D = aggregate(Deaths_per_1M~Continent,covid_table,FUN=mean)

T_case_D = merge(T_case,T_D,by="Continent")
T_case_D$FreqContinent = c(57,49,9,48,39,14)



library(viridis)

    ##대륙내 국가별 발병률 및 사망률을 scatter diagram으로 표현
ggplot(data = covid_table, aes(x=Totalcases_per_1M, y=Deaths_per_1M, color = Continent, group = Co
ntinent)) +
  geom_point() +
  facet_grid(Continent ~ .)
```



대륙내 국가별 발병률 및 사망률을 *scatter diagram*으로 표현한 모습

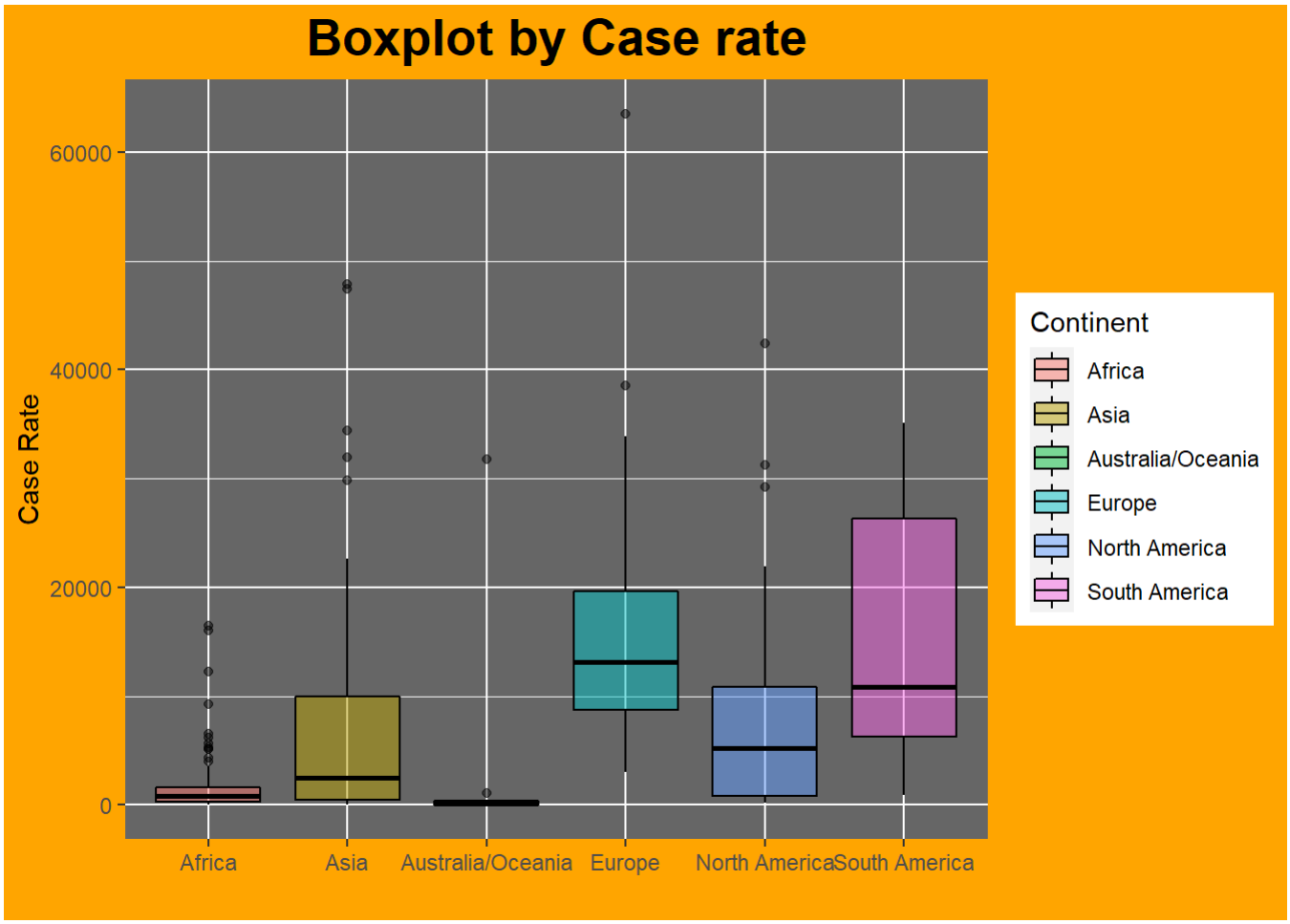
```
   ##대륙별 발병률 및 사망률을  버블차트로 표현
T_case_D %>%
  ggplot(aes(x=Totalcases_per_1M,y=Deaths_per_1M,color=Continent,size=FreqContinent))+
  geom_point(alpha=0.5)+
  scale_size(range=c(8,20))+
  ggtitle('Rate bubble chart (case/death)')+
  theme(panel.background = element_rect(fill='grey40'),plot.background = element_rect(fill='grey'
))
```



대륙별 발병률 및 사망률을 버블차트로 표현한 모습

2.2.4 BOX PLOT for Case Rate

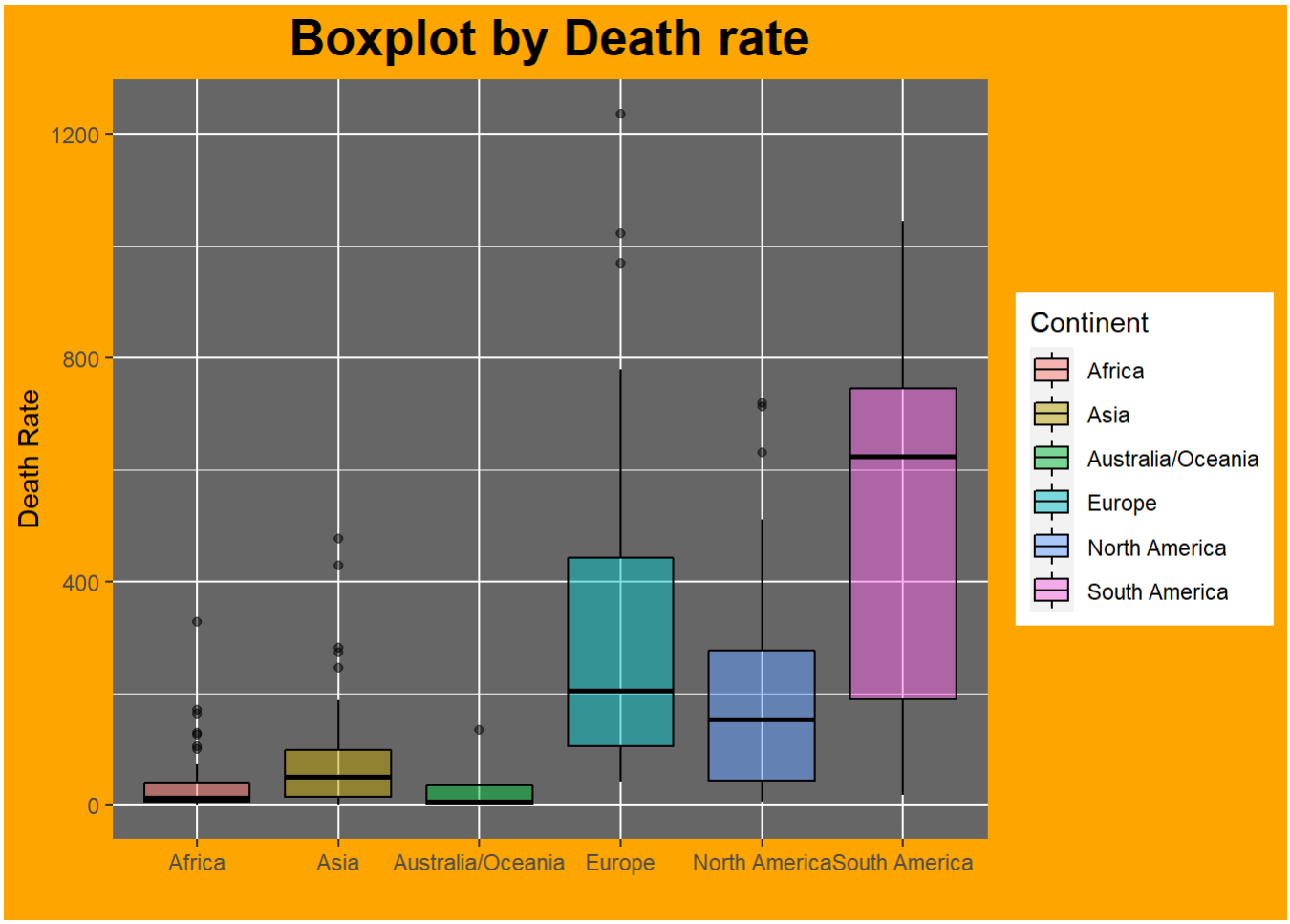
```
# 2.2.4 BOXPLOT for Case Rate
ggplot(covid_table, aes(x = Continent, y = Totalcases_per_1M,fill=Continent)) +
  geom_boxplot( colour = "Black",alpha=0.5) +
  scale_x_discrete() + xlab("") +
  ylab("Case Rate")+
  ggtitle('Boxplot by Case rate')+
  theme(plot.title= element_text(face="bold",hjust=0.5,size=20,color='black'),panel.background = e
lement_rect(fill='grey40'),plot.background = element_rect(fill='orange'))
```



대륙별 확진률을 박스플롯으로 나타낸 모습

2.2.5 BOX PLOT for Death Rate

```
# 2.2.5 BOXPLOT for Deaths Rate
ggplot(covid_table, aes(x = Continent, y = Deaths_per_1M,fill=Continent)) +
  geom_boxplot( colour = "black",alpha=0.5) +
  scale_x_discrete() + xlab("") +
  ylab("Death Rate")+
  ggtitle('Boxplot by Death rate')+
  theme(plot.title= element_text(face="bold",hjust=0.5,size=20,color='black'),panel.background = e
lement_rect(fill='grey40'),plot.background = element_rect(fill='orange'))
```



대륙별 사망률을 박스플롯으로 나타낸 모습

2.3 일원분산분석 실시

비연속척도의 독립변수인 대륙변수에 따라 연속척도의 종속변수인 확진률과 사망률을 개별 종속변수로 독립변수가 집단인 하나의 변수만 사용하는 분석이기 때문에 일원배치 분산분석을 실시한다.

가설

확진률

- **귀무가설*(H0)* : 대륙 집단간 코로나 발생률 평균의 분산은 같을것이다.**
- **연구가설*(H1)* : 대륙 집단간 코로나 발생률 평균의 분산은 같지 않을것이다.**

사망률

- **귀무가설*(H0) :* 대륙 집단간 코로나 사망률 평균의 분산은 같을것이다.**
- **연구가설*(H1) :* 대륙 집단간 코로나 사망률 평균의 분산은 같지 않을것이다.**

일원배치 분산분석(one-way ANOVA)의 조건

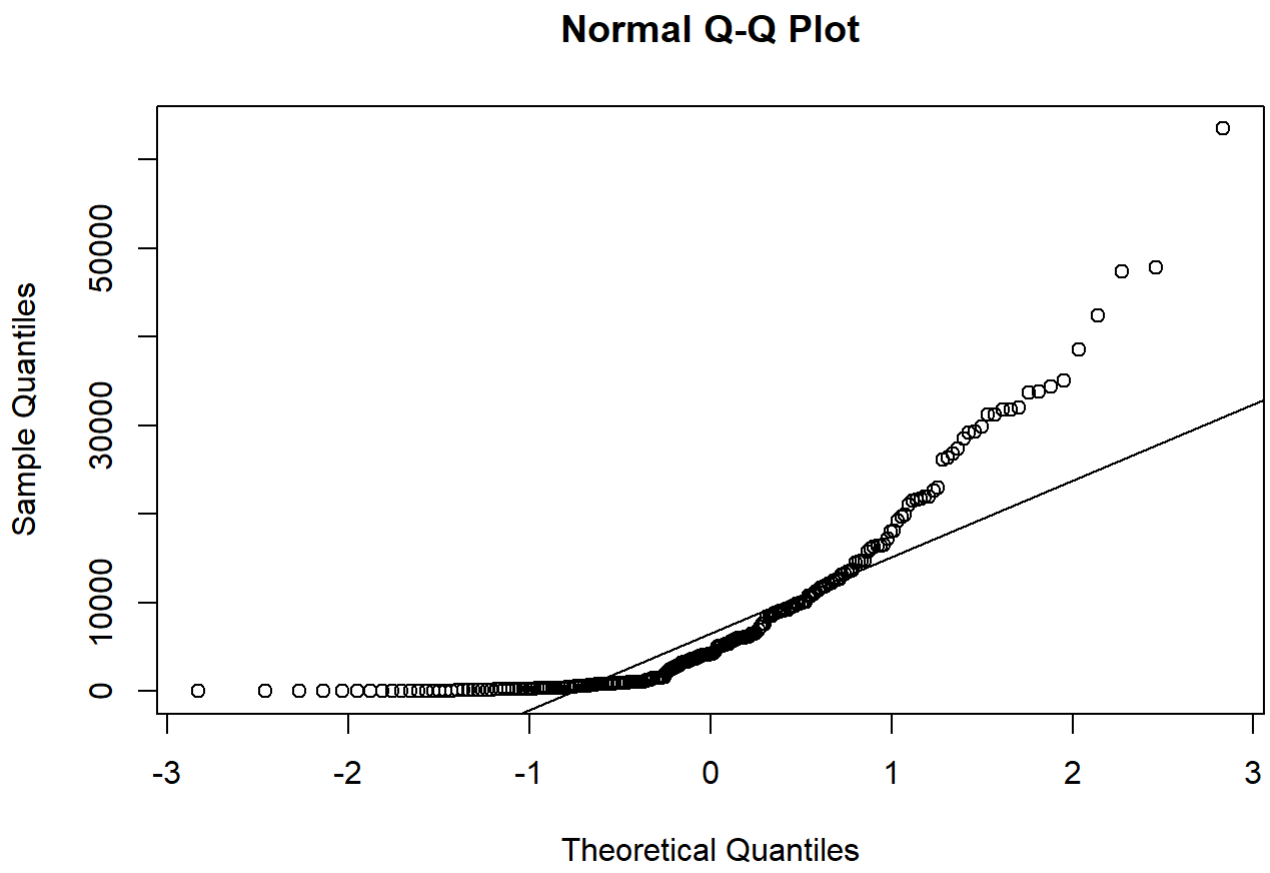
- **독립성** : 독립변수의 그룹군은 상호 독립적이여야 한다.
- **정규성** : 독립변수에 대한 종속변수는 정규분포를 만족해야한다.
- **등분산성** : 독립변수에 대한 종속변수의 분산은 각 군마다 유사해야함

=> 일원배치 분산분석의 조건충족여부를 알아보고 조건 충족여부에 맞춘 분석법을 사용하기 위해 **[정규성 검정->등분산 검정->일원배치분산분석]**의 순서로 검정결과에 따라 분기하는 구조로 분석을 진행하고자 한다.

2.3.0 정규성 검증 실시

table()함수 등 내장함수를 활용하여 빈도표 및 교차표를 출력합니다.

```
 # 확진률
         #qqnorm
qqnorm(covid_table$Totalcases_per_1M)
qqline(covid_table$Totalcases_per_1M)
```

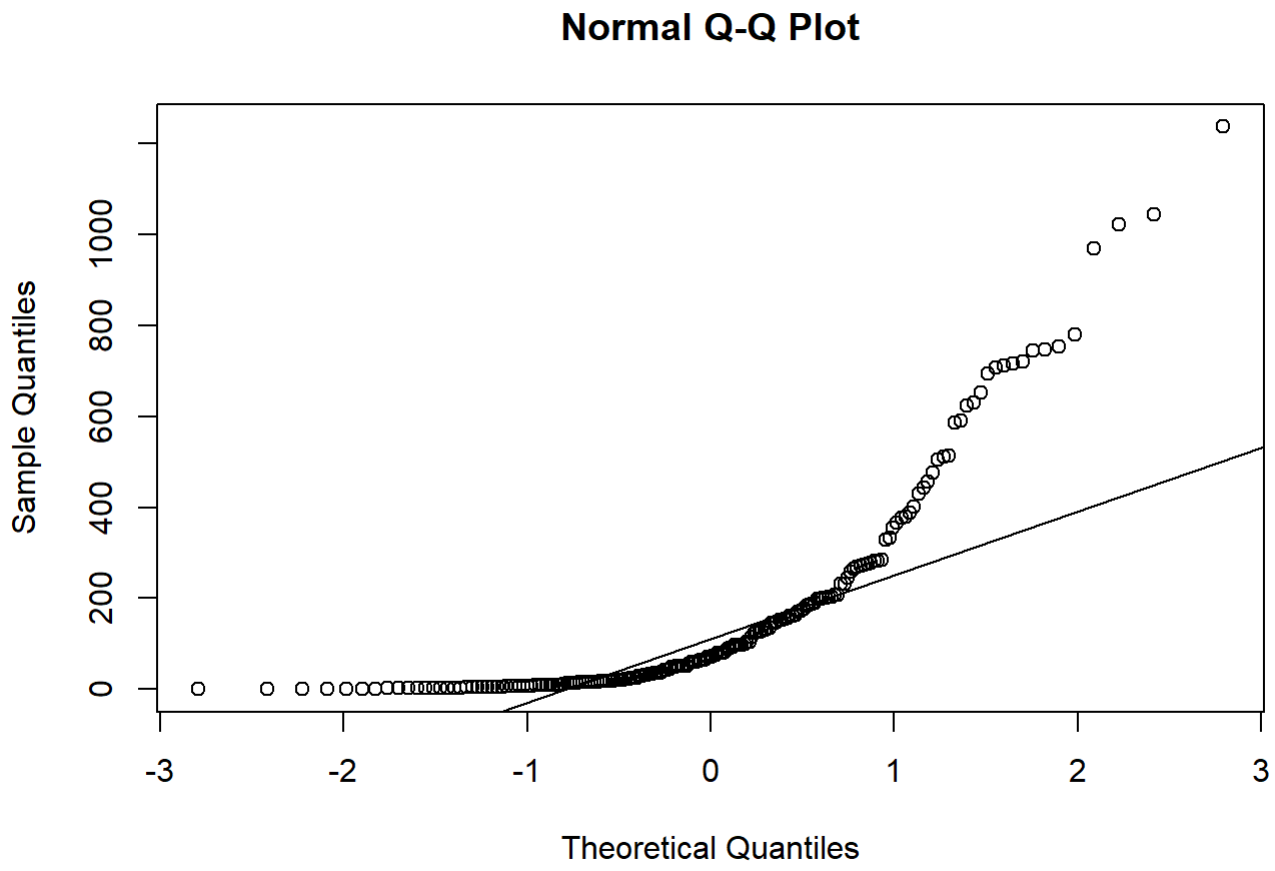



```
         #shapiro test
shapiro.test(covid_table$Totalcases_per_1M)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  covid_table$Totalcases_per_1M
## W = 0.77245, p-value < 2.2e-16
```

```
           #해석 : 정규성 검사결과 p-value가 소수점 15자리까지 0으로, 유의수준 5%에서 H0을 기각하기 때문에 확진률 데이터는 정규성을 갖는다고 하기 어렵습니다.
```

```
     # 사망률
         #qqnorm
qqnorm(covid_table$Deaths_per_1M)
qqline(covid_table$Deaths_per_1M)
```




```
          #shapiro test
shapiro.test(covid_table$Deaths_per_1M)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  covid_table$Deaths_per_1M
## W = 0.72354, p-value < 2.2e-16
```

정규성 해석

*1.*확진률

정규성 검사결과 p-value가 소수점 15자리까지 0으로, 유의수준 5%에서 H0을 기각하기 때문에 확진률 데이터는 정규성을 갖는다고 하기 어렵다

*2.*사망률

정규성 검사결과 p-value가 소수점 15자리까지 0으로, 유의수준 5%에서 H0을 기각하기 때문에 확진률 데이터는 정

규성을 갖는다고 하기 어렵다.

정규성 결론

확진률/사망률은 둘 다 정규성을 만족한다고 보기 힘들기 때문에,Bartlett's test를 실시하는 대신 정규분포를 벗어나도 사용할 수 있는 등분산성 검증법인 **Levene's test를 사용한다.**

2.3.1 등분산성 검증 실시

정규성 검증 결과에 따라 Levene's test로 등분산성 검증을 실시합니다.

Levene's test 결과에 따른 분기 명시

- ***H0 :* 등분산성 만족 *-> ANOVA test***
- ***H1 :* 등분산성 만족안함 = 유의수준 *5%*에서, *P-VALUE*가 *0.05*보다 클경우 *-> Welch test***

Levene's Test - 확진률

```
library(car)

       #확진률 LEVENE TEST
leveneTest(Totalcases_per_1M ~ Continent,data=covid_table)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value    Pr(>F)
## group   5  4.3796 0.0008181 ***
##       210
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
           #해석: p-값이 0.0008181로 유의수준인 0.05보다 작으므로 "등분산성을 만족"한다는 귀무가설이 채택된다.
            #분기 : '등분산성 만족' -> 'ANOVA TEST'를 실시한다.
```

Levene's Test - 사망률

```
 #사망률 LEVENE TEST
leveneTest(Deaths_per_1M ~ Continent, data =covid_table )
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value    Pr(>F)
## group   5  12.547 1.695e-10 ***
##       185
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
          #해석: p-값이 소숫점 9자리까지 0이므로 유의수준인 0.05보다 작으므로 "등분산성을 만족"한다는 귀무가설이 채택된다.
            #분기 : '등분산성 만족' -> 'ANOVA TEST'를 실시한다.
```

등분산성 검증 결론

확진률과 사망률 둘 다 등분산성을 만족하므로 **ANOVA TEST를 실시한다.**

2.3.2 ANOVA TEST 실시

등분산성 만족에 따라 대륙별 확진률과 사망률에 대해 'ANOVA TEST'를 실시한다.

ANOVA TEST - 확진률

```
library(car)

ano_caseRate = aov( Totalcases_per_1M ~ Continent, data= covid_table)
anova(ano_caseRate)
```

```
## Analysis of Variance Table
##
## Response: Totalcases_per_1M
##            Df     Sum Sq    Mean Sq F value    Pr(>F)
## Continent   5 5.6654e+09 1133072337  11.898 3.794e-10 ***
## Residuals 210 1.9999e+10   95231031
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

확진률 분석결과 해석

- 1. 대륙집단 간 제곱합 (BSS)은 5665361686이고, 자유도는 5 이므로 집단 간 평균 제곱합(BMS)은 1133072337
- 2. 대륙집단 내 제곱합 (WSS) 은 19998516479이고 , 자유도는 210이므로 집단내 평균제곱합은 95231031
- 3. F 값은 대륙집단 간 평균 제곱합을 집단 내 평균 제곱합으로 나눈 11.89814가 되어, 집단 간 제곱합의 자유도와 집단 내 제곱합의 자유도를 감안한 유의도는 소수점 9 번째 자리까지 0 이므로 영가설의 채택 혹은 기각을 판단할 수 있는 유의수준 0.05 보다 낮게 나타나므로 "대륙 집단에 따른 확진률의 차이가 없다는 영가설은 기각하고,연구가설을 채택

===> *H1* 채택 *: "대륙 집단에 따른 확진률의 차이가 존재한다""*

ANOVA TEST - 사망률

```
      #사망률 ANOVA TEST
ano_deathRate = aov(Deaths_per_1M ~ Continent , data = covid_table)
anova(ano_deathRate)
```

```
## Analysis of Variance Table
##
## Response: Deaths_per_1M
##            Df  Sum Sq Mean Sq F value   Pr(>F)
## Continent   5 3560796  712159  19.073 2.78e-15 ***
## Residuals 185 6907543   37338
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

사망률 분석결과 해석

- 1. 집단 간 제곱합 (BSS)은 3560796.2이고, 자유도는 5 이므로 집단 간 평균 제곱합(BMS)은 712159.24
- 2. 집단 내 제곱합 (WSS) 은 6907542.6이고 , 자유도는 185이므로 집단내 평균제곱합은 37338.07
- 3. F 값은 집단 간 평균 제곱합을 집단 내 평균 제곱합으로 나눈 19.07328가 되어, 집단 간 제곱합의 자유도와 집단 내 제곱합의 자유도를 감안한 유의도는 소수점 14 번째 자리까지 0 이므로 영가설의 채택 혹은 기각을 판단할 수 있는 유의수준 0.05 보다 낮게 나타나므로 "대륙 집단에 따른 사망률의 차이가 없다는 영가설은 기각하고,연구가설을 채택

===> *H1* 채택 *: "대륙 집단에 따른 사망률의 차이가 존재한다""*

2.3.3 describeBy로 집단별 평균값 구하기

```
library(psych)

  #확진률
describeBy(covid_table$Totalcases_per_1M, covid_table$Continent)
```

```
##
##  Descriptive statistics by group
## group: Africa
##    vars  n    mean      sd median trimmed    mad min   max range skew kurtosis
## X1    1 57 2254.28 3651.76    841 1441.87 858.43   8 16449 16441 2.46     5.91
##        se
## X1 483.69
## ------------------------------------------------------------
## group: Asia
##    vars  n    mean       sd median trimmed     mad min   max range skew
## X1    1 49 8018.53 11917.66   2497  5639.1 3667.95   3 47854 47851 1.96
##    kurtosis      se
## X1      3.2 1702.52
## ------------------------------------------------------------
## group: Australia/Oceania
##    vars n    mean       sd median trimmed   mad min   max range skew kurtosis
## X1    1 9 3734.44 10527.85     90 3734.44 83.03  19 31794 31775 2.07     2.62
##         se
## X1 3509.28
## ------------------------------------------------------------
## group: Europe
##    vars  n     mean       sd  median  trimmed     mad  min   max range skew
## X1    1 48 15792.69 11435.71 13150.5 14353.12 7685.06 3001 63512 60511 1.79
##    kurtosis     se
## X1     4.34 1650.6
## ------------------------------------------------------------
## group: North America
##    vars  n    mean      sd median trimmed     mad min   max range skew kurtosis
## X1    1 39 8439.33 9924.52   5235    6832 6588.67 199 42442 42243 1.59      2.1
##         se
## X1 1589.19
## ------------------------------------------------------------
## group: South America
##    vars  n     mean       sd  median  trimmed      mad min   max range skew
## X1    1 14 15446.14 11294.26 10844.5 15018.25 12967.56 919 35108 34189 0.26
##    kurtosis      se
## X1    -1.61 3018.52
```

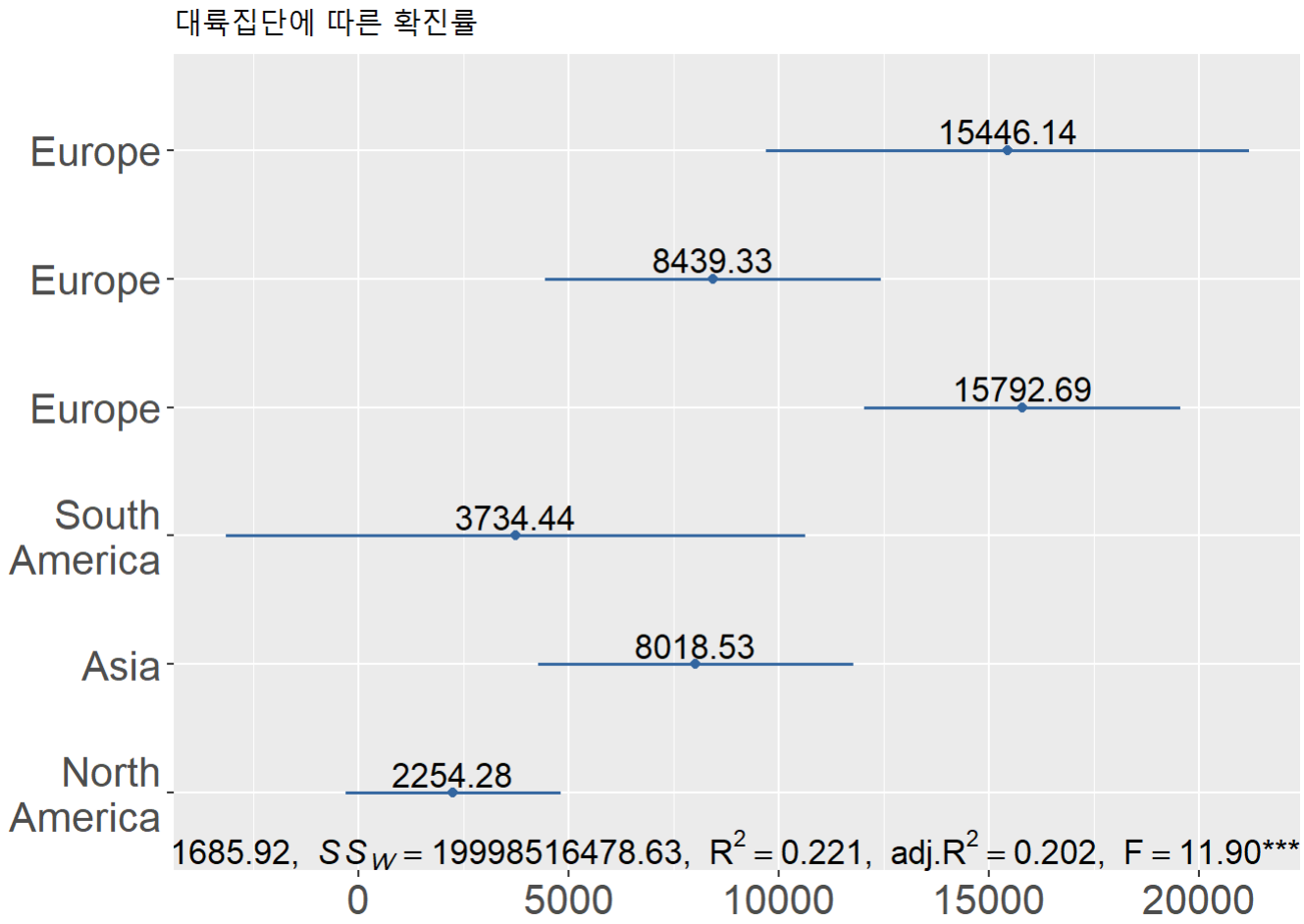
```
  #사망률
describeBy(covid_table$Deaths_per_1M, covid_table$Continent)
```

```
##
##  Descriptive statistics by group
## group: Africa
##    vars  n  mean    sd median trimmed   mad  min max  range skew kurtosis   se
## X1    1 55 35.68 57.32     13   23.09 11.86 0.08 328 327.92 2.99    10.65 7.73
## ------------------------------------------------------------
## group: Asia
##    vars  n  mean    sd median trimmed   mad min max range skew kurtosis    se
## X1    1 43 88.64 110.9     51   67.09 63.75 0.3 477 476.7 1.86      3.2 16.91
## ------------------------------------------------------------
## group: Australia/Oceania
##    vars n  mean    sd median trimmed  mad min max range skew kurtosis    se
## X1    1 5 35.56 57.35      5   35.56 6.23 0.8 135 134.2 0.93    -1.13 25.65
## ------------------------------------------------------------
## group: Europe
##    vars  n   mean     sd median trimmed    mad min  max range skew kurtosis
## X1    1 45 313.87 285.03    204  268.35 209.05  42 1237  1195 1.39     1.43
##       se
## X1 42.49
## ------------------------------------------------------------
## group: North America
##    vars  n   mean     sd median trimmed    mad min max range skew kurtosis
## X1    1 30 213.67 205.17    153  179.83 179.39   6 720   714 1.13     0.37
##       se
## X1 37.46
## ------------------------------------------------------------
## group: South America
##    vars  n   mean     sd median trimmed    mad min  max range skew kurtosis
## X1    1 13 474.46 342.15    624  464.09 581.18  18 1045  1027 0.01    -1.67
##      se
## X1 94.9
```

2.3.4 sjplot 분산분석 그래프 출력

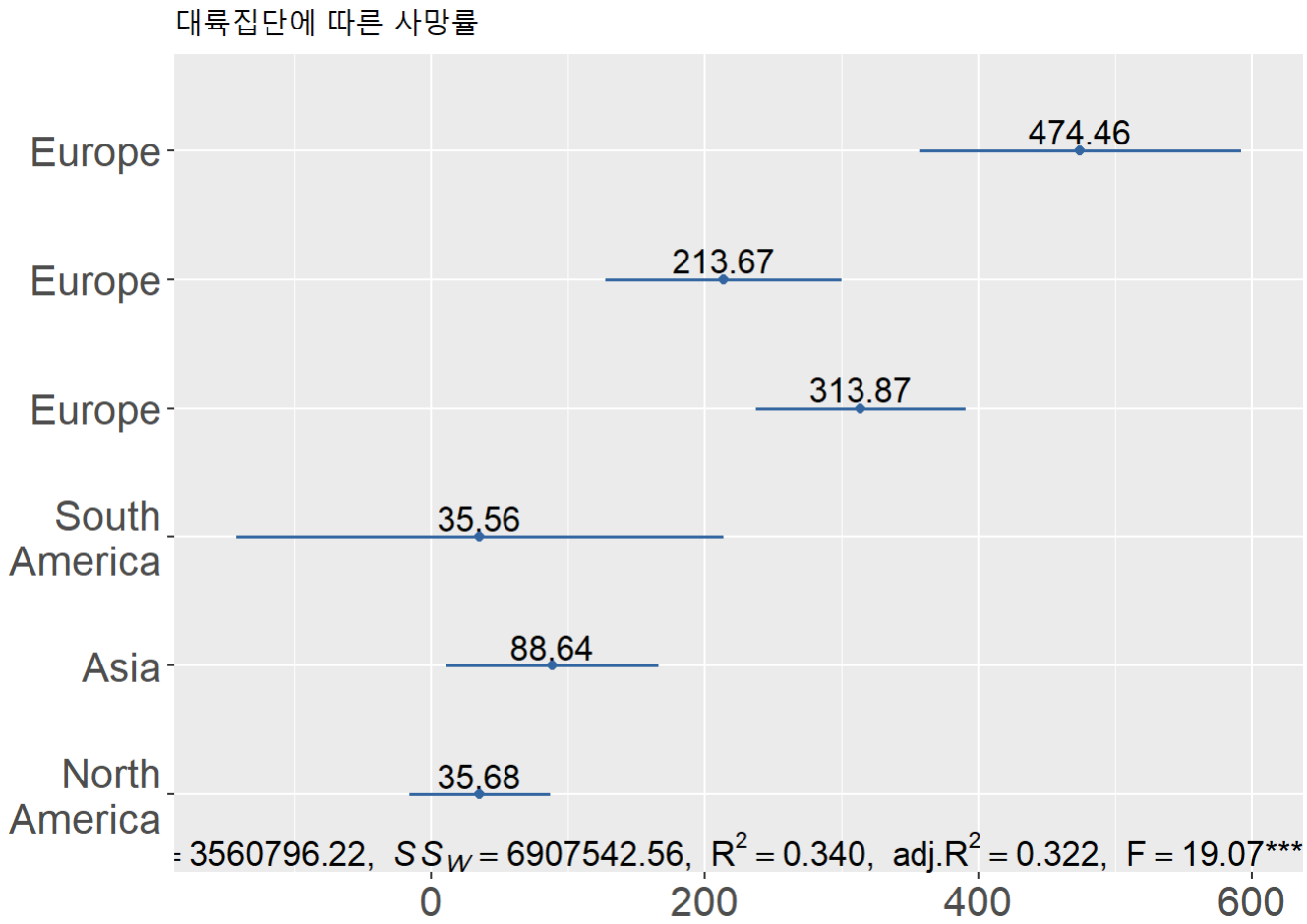
확진률 그래프

```
library(sjPlot)
   #확진률
set_theme(axis.textsize=1.2, geom.label.size=4.5)
sjp.aov1(covid_table$Totalcases_per_1M, covid_table$Continent,
         geom.size = 1.5, wrap.labels = 7,
         axis.lim = , meansums = T,
         show.summary = T, show.p = F,
         title = "대륙집단에 따른 확진률",
         axis.labels = covid_table$Continent)
```



사망률 그래프

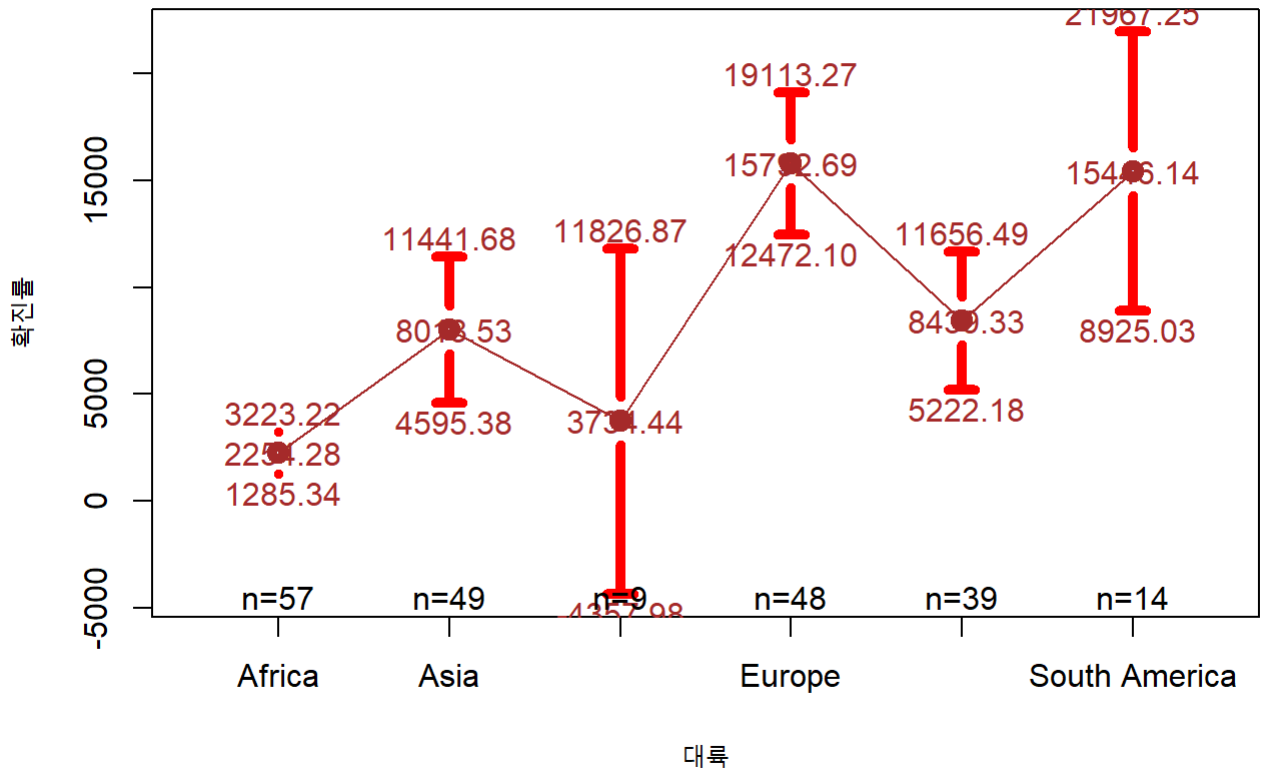
```
    #사망률
set_theme(axis.textsize=1.2, geom.label.size=4.5)
sjp.aov1(covid_table$Deaths_per_1M, covid_table$Continent,
         geom.size = 1.5, wrap.labels = 7,
         axis.lim = , meansums = T,
         show.summary = T, show.p = F,
         title = "대륙집단에 따른 사망률",
         axis.labels = covid_table$Continent)
```



2.3.5 gplots 분산분석 그래프 출력

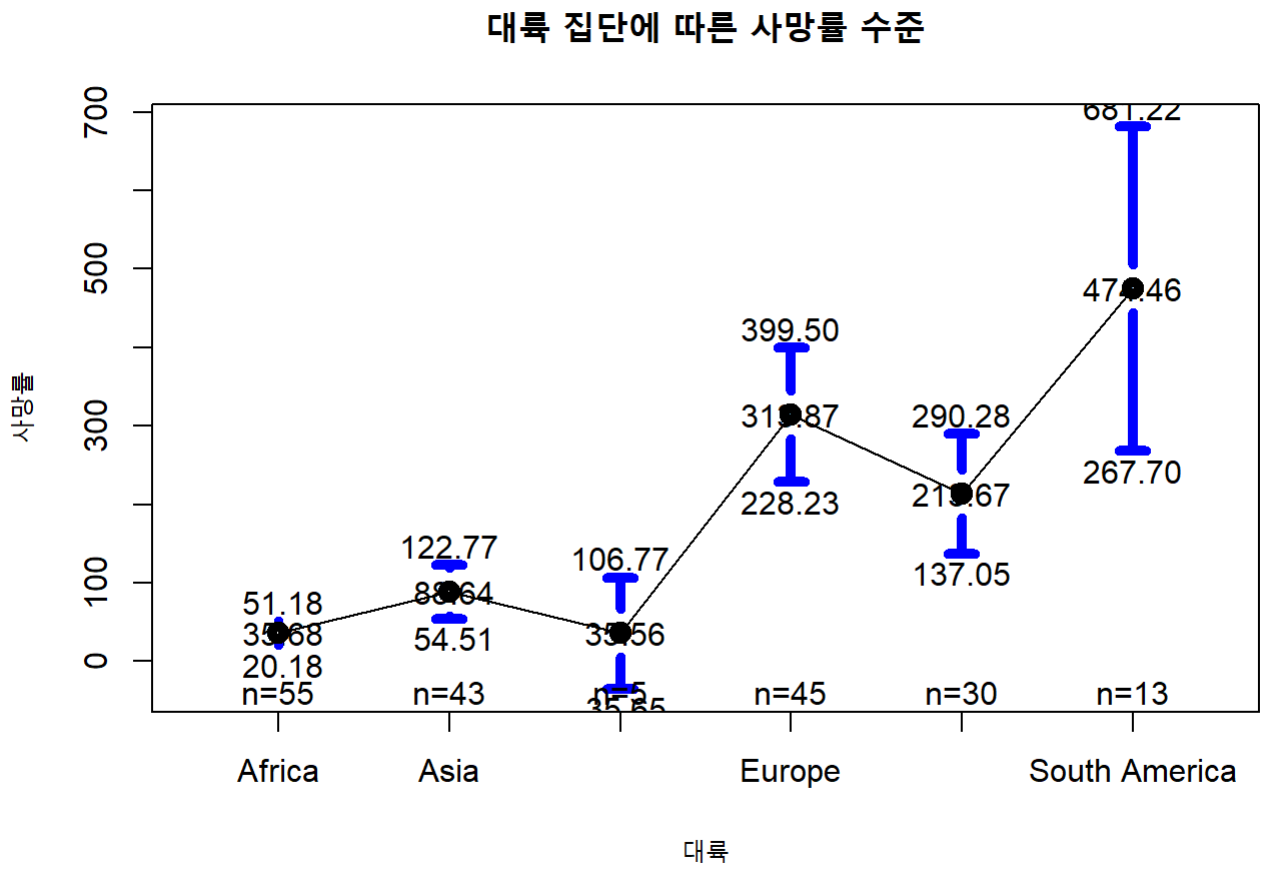
확진률 그래프

```
 #확진률
library(gplots)
plotmeans(Totalcases_per_1M ~ Continent, data=covid_table,
          xlab="대륙", ylab="확진률", ci.label=T,
          mean.label=T, barwidth=5, digits=2,
          col="brown", pch=1, barcol="red",
          legends = levels(covid_table$Continent),
          main="대륙 집단에 따른 확진률 수준")
```



사망률 그래프

```
      #사망률
plotmeans(Deaths_per_1M ~ Continent, data=covid_table,
          xlab="대륙", ylab="사망률", ci.label=T,
          mean.label=T, barwidth=5, digits=2,
          col="black", pch=1, barcol="blue",
          legends = levels(covid_table$Continent),
          main="대륙 집단에 따른 사망률 수준")
```



2.3.6 사후검증 scheffe

확진률 사후검증

```
# scheffe 사후검증
library(agricolae)

      #확진률
scheffe.test(ano_caseRate, "Continent", alpha=0.05, console=TRUE)
```

```
##
## Study: ano_caseRate ~ "Continent"
##
## Scheffe Test for Totalcases_per_1M
##
## Mean Square Error  : 95231031
##
## Continent,  means
##
##                   Totalcases_per_1M       std  r  Min   Max
## Africa                     2254.281  3651.757 57    8 16449
## Asia                       8018.531 11917.662 49    3 47854
## Australia/Oceania          3734.444 10527.852  9   19 31794
## Europe                    15792.688 11435.712 48 3001 63512
## North America              8439.333  9924.515 39  199 42442
## South America             15446.143 11294.257 14  919 35108
##
## Alpha: 0.05 ; DF Error: 210
## Critical Value of F: 2.257066
##
## Groups according to probability of means differences and alpha level( 0.05 )
##
## Means with the same letter are not significantly different.
##
##                   Totalcases_per_1M groups
## Europe                    15792.688      a
## South America             15446.143     ab
## North America              8439.333     bc
## Asia                       8018.531     bc
## Australia/Oceania          3734.444     bc
## Africa                     2254.281      c
```

```
#해석 : 집단간 차이를 검증하기 위한 기준인 유의도 (alpha: 0.05)와 자유도 210이 출력됨
                #---> 집단 내 제곱합의 자유도가 210에서  유의도가 0.05수준인 경우의 F값이 2.26
 이다
              #집단간의 평균이 유의한 차이가 있는가? => 사후검증 결과, 집단(Groups)은 a,ab,bc 그
리고 c로 나누어졌다. 즉 독립변수의 6집단의 평균중, Europe과 sOUTH America 그리고 North America,Asi
a ,Oceania 그리고 Africa 간의 유의한 차이가 존재한다는 것이라고 해석된다.
```

확진률 분석결과 해석

- 1. 집단간 차이를 검증하기 위한 기준인 유의도 (alpha: 0.05)와 자유도 210이 출력됨
- 2. 집단 내 제곱합의 자유도가 210에서 유의도가 0.05수준인 경우의 F값이 2.26 이다

- **(3) 집단간의 평균이 유의한 차이가 있는가? => 사후검증 결과, 집단(Groups)은 a,ab,bc 그리고 c로 나누어졌다. 즉 독립변수의 6집단의 평균중, Europe과 sOUTH America 그리고 North America,Asia ,Oceania 그리고 Africa 간의 유의한 차이가 존재한다는 것이라고 해석된다.**

사망률 사후검증

```
#사망률
scheffe.test(ano_deathRate, "Continent", alpha=0.05, console=TRUE)
```

```
##
## Study: ano_deathRate ~ "Continent"
##
## Scheffe Test for Deaths_per_1M
##
## Mean Square Error  : 37338.07
##
## Continent,  means
##
##                   Deaths_per_1M       std  r   Min  Max
## Africa                 35.67964  57.32382 55  0.08  328
## Asia                   88.63953 110.90486 43  0.30  477
## Australia/Oceania      35.56000  57.35214  5  0.80  135
## Europe                313.86667 285.03473 45 42.00 1237
## North America         213.66667 205.17013 30  6.00  720
## South America         474.46154 342.15314 13 18.00 1045
##
## Alpha: 0.05 ; DF Error: 185
## Critical Value of F: 2.262937
##
## Groups according to probability of means differences and alpha level( 0.05 )
##
## Means with the same letter are not significantly different.
##
##                   Deaths_per_1M groups
## South America         474.46154      a
## Europe                313.86667     ab
## North America         213.66667     bc
## Asia                   88.63953     cd
## Africa                 35.67964      d
## Australia/Oceania      35.56000      d
```

```
            #해석 : 집단간 차이를 검증하기 위한 기준인 유의도 (alpha: 0.05)와 자유도 185이 출력됨
                  #---> 집단 내 제곱합의 자유도가 185에서  유의도가 0.05수준인 경우의 F값이 2.26이다
               #집단간의 평균이 유의한 차이가 있는가? => 사후검증 결과, 집단(Groups)은 a,ab,bc,cd 그리고 d로 나누어졌다. 즉 독립변수의 6집단의 평균중, South America가 단독으로 a집단 Europe이 ab집단 그리고 North America가 bc집단, Asia가 cd집단 그리고 Oceania와 Africa 가 d집단으로 아프리카와 오세아니아의 유사점만 제외하고 사실상 모든 집단간의 사망률 차이가 존재한다는 것이라고 해석된다.
```

사망률 분석결과 해석

- 1. 집단간 차이를 검증하기 위한 기준인 유의도 (alpha: 0.05)와 자유도 185이 출력
- 2. 집단 내 제곱합의 자유도가 185에서 유의도가 0.05수준인 경우의 F값이 2.26 이다

- **(3) 집단간의 평균이 유의한 차이가 있는가? => 사후검증 결과, 집단(Groups)은 a,ab,bc,cd 그리고 d로 나누어졌다. 즉 독립변수의 6집단의 평균중, South America가 단독으로 a집단 Europe이 ab집단 그리고 North America가 bc집단, Asia가 cd집단 그리고 Oceania와 Africa 가 d집단으로 아프리카와 오세아니아의 유사점만 제외하고 사실상 모든 집단간의 사망률 차이가 존재한다는 것이라고 해석된다.**

*3.*결론

해당 보고서에선 Worldometers에서 공유되는 세계 코로나 통계현황 데이터를 데이터로, 탐색적 기술통계부터 시작해서 “대륙 집단별 확진률/사망률 에 유의한 차이가 있을것인가”에 대한 대립가설을 연구과제로 분산분석을 실시하기까지의 일련의 구조적인 절차에 따라 연구를 시행했다.

필요한 분석파트마다 해석을 명시해놨으며, 분산분석을 통해궁극적으로 “비연속 척도의 독립변수인 대륙집단과 연속척도인 확진률/사망률”사이의 유의미한 관계를 확인할 수 있었고 추가적으로 사후검증을 통해 집단간에 어느 수준의 차이를 보이는지 대략적으로 알아볼 수 있었다

연구가설과 결론 정리

확진률

귀무가설*(H0) :* 대륙 집단간 코로나 확진률 평균의 분산은 같을것이다*.*

연구가설*(H1) :* 대륙 집단간 코로나 확진률 평균의 분산은 같지 않을것이다*.*

확진률 가설 결론 : 연구가설(H1)을 채택

집단간의 평균이 유의한 차이가 있는가? => 사후검증 결과, 집단(Groups)은 a,ab,bc 그리고 c로 나누어졌다. 즉 독립변수의 6집단의 평균중, Europe과 sOUTH America 그리고 North America,Asia ,Oceania 그리고 Africa 간의 유의한 차이가 존재한다는 것이라고 해석된다.

사망률

귀무가설*(H0) :* 대륙 집단간 코로나 사망률 평균의 분산은 같을것이다.

연구가설*(H1) :* 대륙 집단간 코로나 사망률 평균의 분산은 같지 않을것이다.

사망률 가설 결론 : 연구가설(H1)을 채택

집단간의 평균이 유의한 차이가 있는가? => 사후검증 결과, 집단(Groups)은 a,ab,bc,cd 그리고 d로 나누어졌다. 즉 독립변수의 6집단의 평균중, South America가 단독으로 a집단 Europe이 ab집단 그리고 North America가 bc집단, Asia가 cd집단 그리고 Oceania와 Africa 가 d집단으로 아프리카와 오세아니아의 유사점만 제외하고 사실상 모든 집단간의 사망률 차이가 존재한다는 것이라고 해석된다.