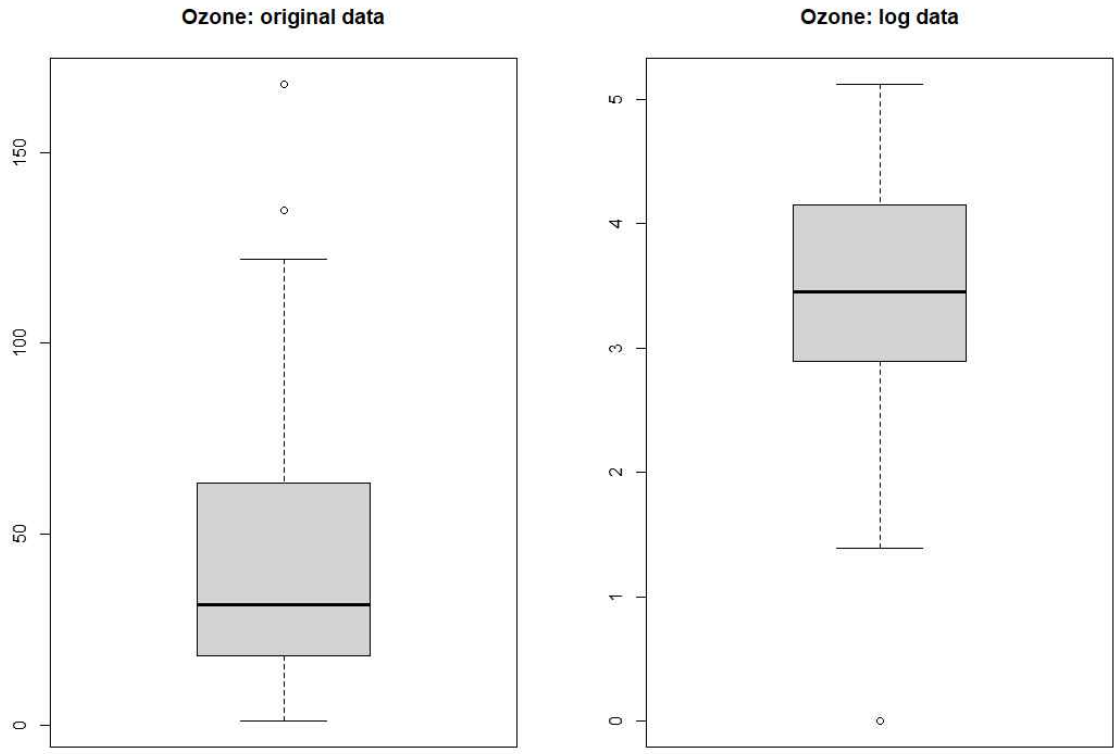
EDA 5장 숙제

2017122007 정헌영

(1)

자료를 대칭화되도록 변환:

```
Oz <- airquality$Ozone
#X를 re-expression 하기 위한 p 계산 (p: power of X)
findp <- function(x){
  Hl <- fivenum(x)[2]
  M <- fivenum(x)[3]
  Hu <- fivenum(x)[4]
  p=1-2*M*(Hu-M+Hl-M)/((Hl-M)^2+(Hu-M)^2)
  p
}
findp(Oz)
#p=0.03378238
#p가 0에 가까우므로 logarithms를 사용한다.

boxplot(Oz, main="islands: original data")
boxplot(log(Oz), main="islands: log data")
```



그래프 분석: 계산된 p가 0에 가까우므로 Ozone 데이터에 log를 취해 자료를 재표현한다. 원래 Ozone 데이터와 재표현한 log(Ozone)데이터의 boxplot를 그려 비교한다. log(Ozone)의 boxplot은 median이 상자의 가운데에 더 가깝고 상자의 양 끝 줄도 길이가 비슷해 대칭화되었음이 확인된다.

skewness 계산:

```
skew <- function(x) {
  H_L = quantile(x, 0.25)
  H_U = quantile(x, 0.75)
  Med = median(x)
  print(((H_U-Med)-(Med-H_L)) / ((H_U-Med)+(Med-H_L)))
}
Oz2<-na.omit(Oz)
skewO <- skew(Oz2)
skewlogO <- skew(log(Oz2))
matrix(c(skewO,skewlogO),dimnames = list(c("Original", "log"), "skewness"))
```

```
#          skewness
Original 0.4065934
log        0.1123698
```

[Ozone의 변환 전(Original)과 변환 후(log)의 skewness]

(2)

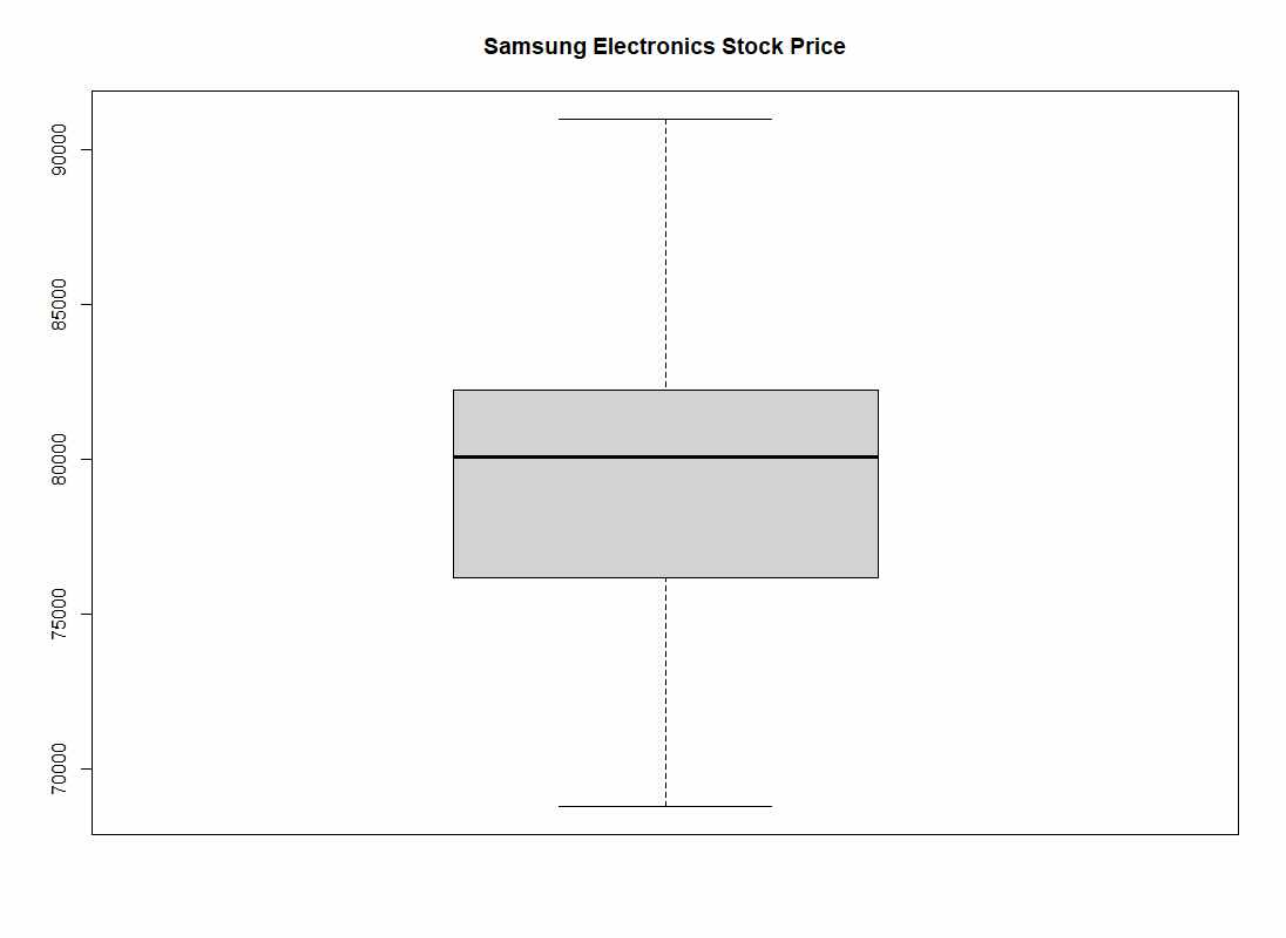
S: 삼성전자 2021년 일별 주식 가격

줄기와 잎 그림:

```
stem(S)
  68 | 8
  69 | 04899
  70 | 112222223445666677
  71 | 123334556
  72 | 2337
  73 | 12379
  74 | 1123446689
  75 | 3336678
  76 | 0113333667889
  77 | 0001233444678
  78 | 01235555588
  79 | 00023344556667778889999
  80 | 000111112234555566788999
  81 | 00001222244455556667788899999
  82 | 000000111122223344556667888899999
  83 | 000222355667999
  84 | 0001246789
  85 | 03466
  86 | 078
  87 | 02
  88 | 018
  89 | 477
  90 | 6
  91 | 0
```

상자그림:

boxplot(S, main="Samsung Electronics")



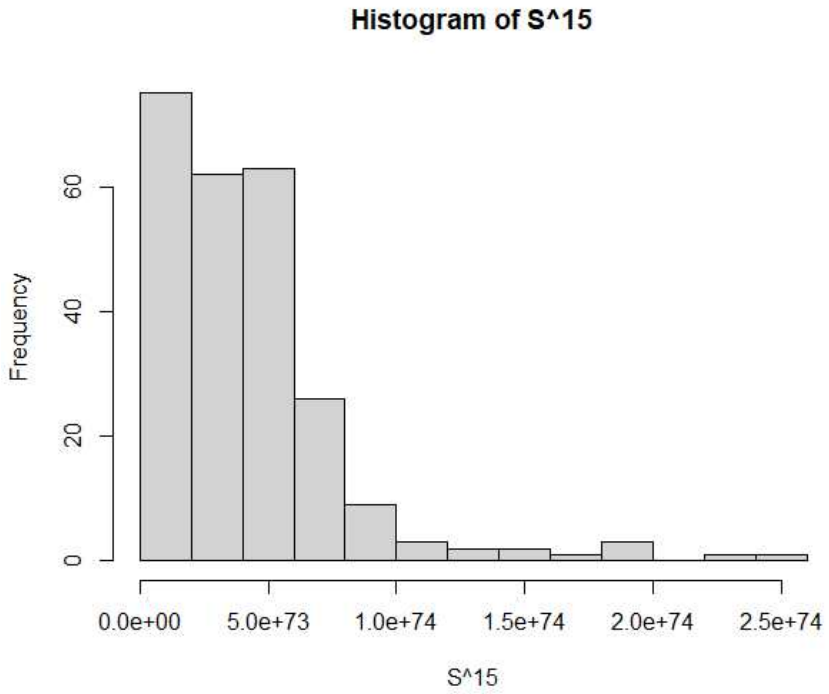
분석: 줄기와 잎 그림은 전체적으로 대칭성이 있으나 높은 주가쪽에 자료가 군집되어 낮은 값 방향으로 skewed 모습을 다소 보인다. 이는 상자그림에서도 확인 가능한데 전반적으로 양 끝 줄이 비슷하고 중앙값도 상자의 중간쯤에 있으나 조금 더 위쪽에 위치해 낮은 값으로 다소 skewed되어있음을 보여준다. 즉 2021년의 삼성전자 주가는 낮은 값 방향으로 skewed 되어있다.

대칭화 변환 시행착오:

히스토그램을 이용해 대칭성을 판단하고자 한다.

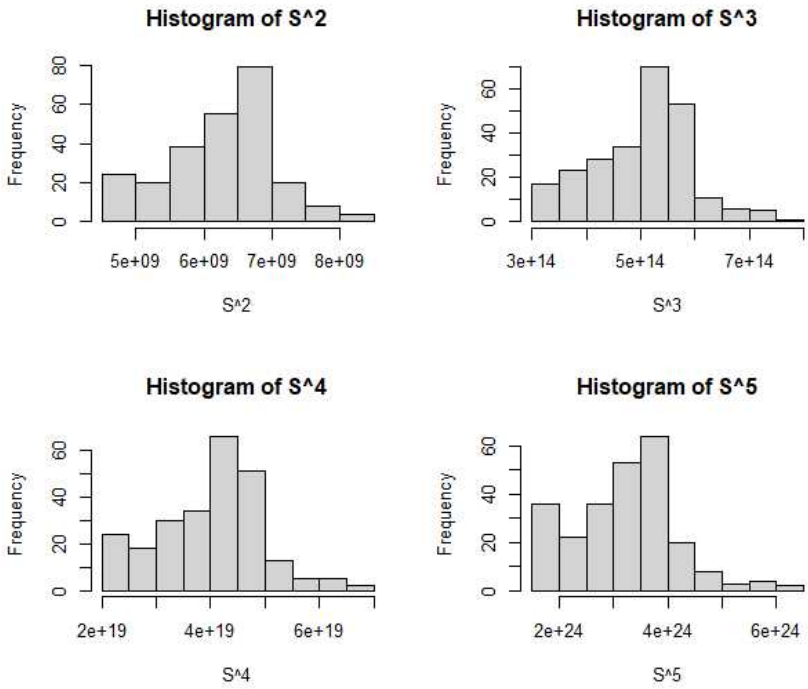
우선 앞에서 지정한 함수로 p를 찾은 다음 x^p를 그려본다.

```
findp(S)
#p=15.13589
hist(S^15)
```



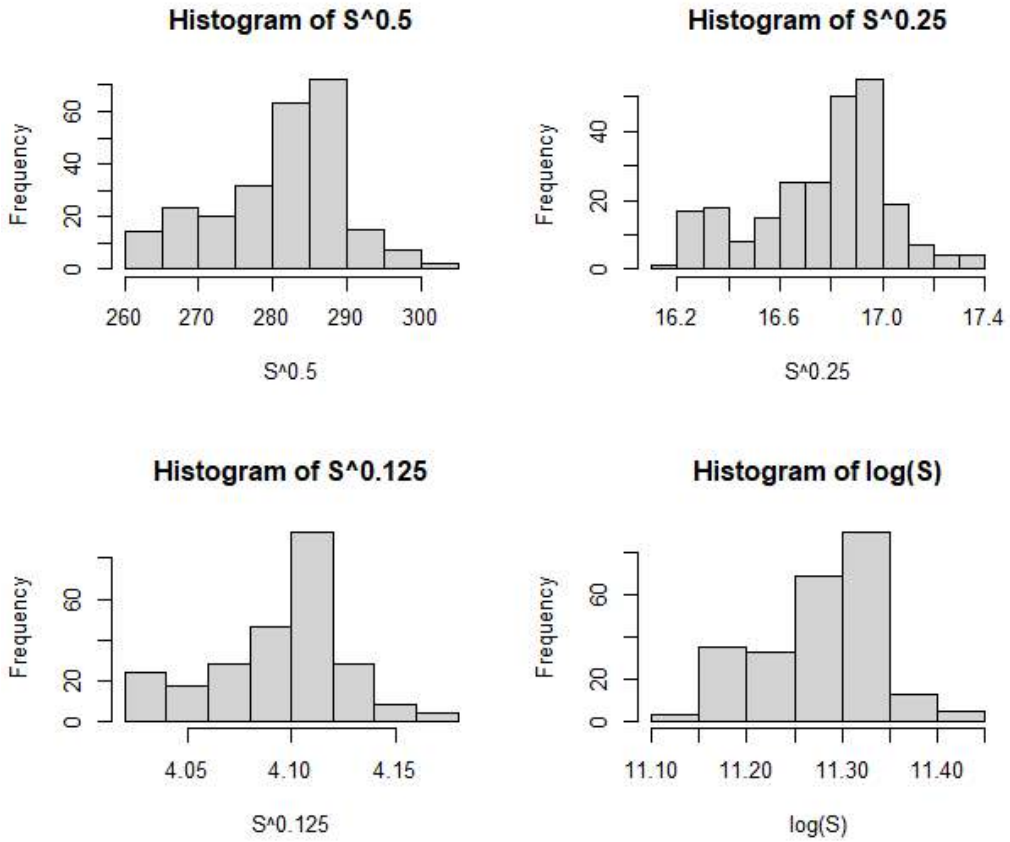
자료가 과도하게 right skewed되어 power값을 낮추면서 확인해 본다.

hist(S^2)
hist(S^3)
hist(S^4)
hist(S^5)



power값을 더할수록 right skewed되는 경향이 강해지므로 1보다 작은 power를 적용해 본다.

hist(S^0.5)
hist(S^0.25)
hist(S^0.125)
hist(log(S)

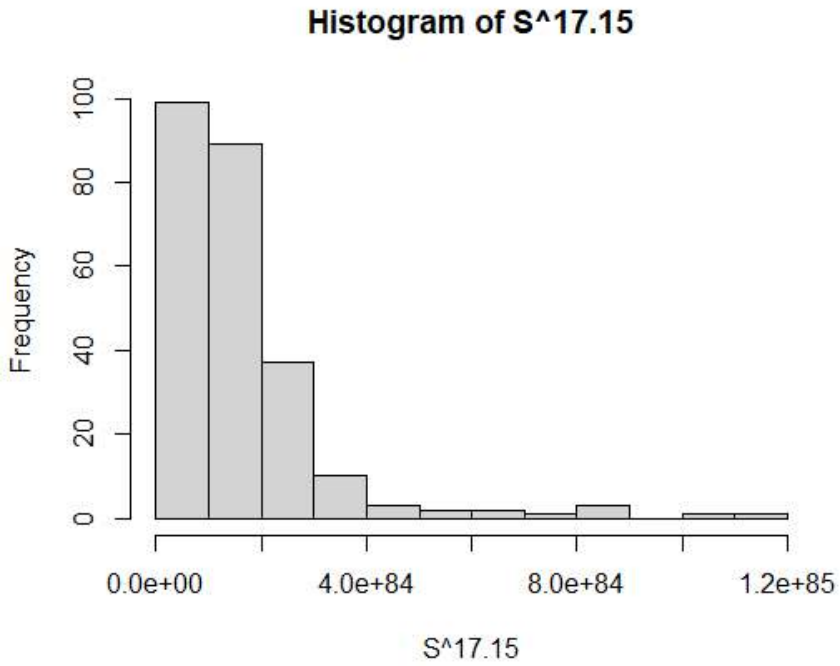


작은 power로 바꿔가며 히스토그램을 확인해 보았지만 left skewed된 경향을 뚜렷하게 상쇄시킨 변환은 찾아지지 않았다.
p에 (0.01, 0.02, ... , 30)를 넣어가며 skewness를 가장 적게 하는 p를 찾아본다.

```
x = (1:3000)/100
y = rep(0,3000)
for(i in 1:3000){
  y[i] <- skew(S^x[i])

}

which(y==min(abs(y)))/100
#[1] 17.15
```

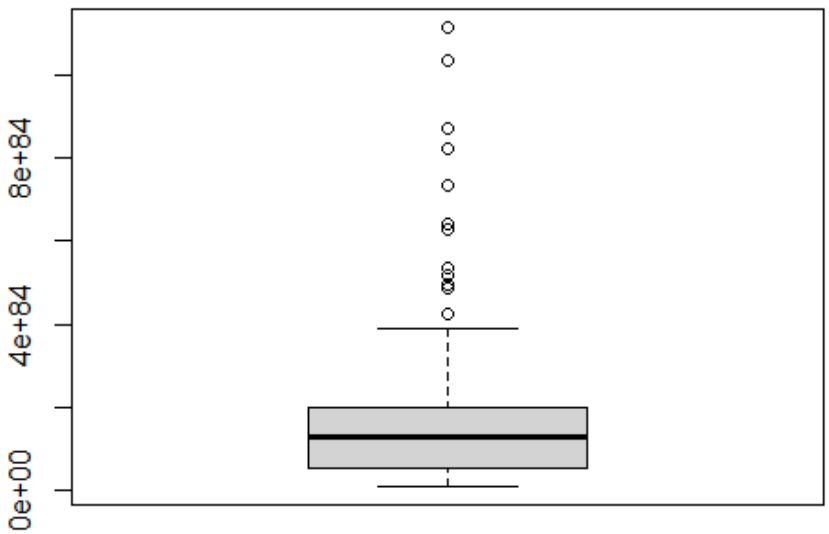
p는 17.15가 찾아진다. 이를 histogram으로 그려본다.

```
hist(S^17.15)
```



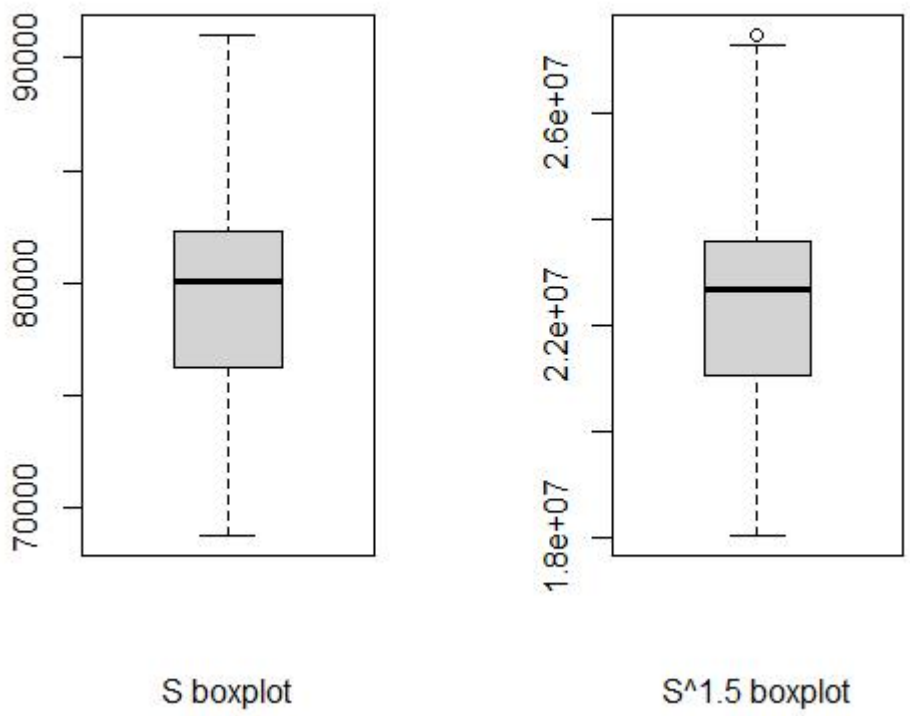
역시 p=15를 넣었을 때 처럼 symmetric하다고 보기 힘든 right skewed된 히스토그램이 나타난다, 그 이유를 boxplot으로 그려 이해할 수 있었다.

boxplot(S^17.15)



boxplot을 보면 상자 안에서 median만을 보면 상당히 symmetric하다. 그러나 상자를 벗어난, 즉 1분위수와 3분위수를 벗어난 구간에서는 symmetric함을 유지하지 못하게 된다. 즉 skewness에 의존하는 방법은 1분위수와 3분위수 구간을 벗어난 구간의 대칭성을 극단적으로 무시하게 된다.

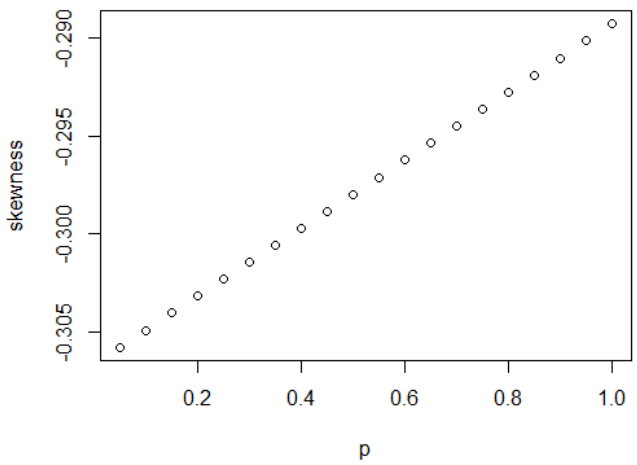
```
boxplot(S, xlab = "S boxplot")
boxplot(S^1.5, xlab = "S^1.5 boxplot")
```



p를 1.5수준으로만 올려도 이상값이 나타나기 시작한다. 즉 1분위수와 3분위수 구간을 벗어난 구간의 대칭성 훼손이 시작되는 경향성은 p가 1보다 커질 때 나타난다.

따라서 반대로 p에 (0.05, 0.10, ... 1)을 대입하여 skewness를 확인해본다.

```
x <- (1:20)/20
y <- rep(0,20)
for(i in 1:20){
  y[i] <- skew(S^x[i])

}
plot(x,y,xlab="p",ylab="skewness")
```



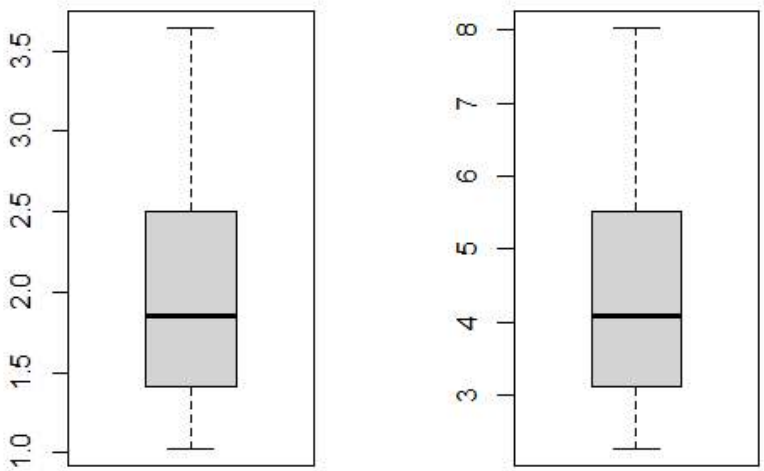
얻어진 그래프를 확인해보면
반대로 1 이하로 p가 작아질수록
1분위수와 3분위수 사이의 범위에서
대칭성이 훼손되기 시작한다.

즉 삼성전자 주가 자료를 대칭적이게 변환하기 위해 p를 조정하면 p>1일 경우 1분위수와 3분위수를 벗어난 구간에서 대칭성이 훼손되고, p<1일경우 1분위수와 3분위수 사이의 구간에서 대칭성이 훼손된다.
결론적으로 p=1로 설정해 오히려 자료를 변환하지 않는것이 대칭성을 유지하는데에 적절하다고 생각된다.

(3)

가. 전체 자료, 파운드 변환된 자료의 상자그림과 줄기그림:

```
D <- read.table('C:/Users/dhkd0/Desktop/강의/DISTRESS_rmstar.DAT')
D <- as.numeric(unlist(D))
par(mfrow=c(1,2))
boxplot(D,xlab = "D boxplot")
D_p <- D*2.20462
boxplot(D_p,xlab = "D_pound boxplot")
```




```
stem(D,2)
  1 | 0111222233334
  1 | 566677778888999
  2 | 001223344
  2 | 56666778
  3 | 0024
  3 | 6
```

[원래 자료의 줄기그림]

```
stem(D_p,2)
  2 | 334566778999
  3 | 134557888999
  4 | 023345699
  5 | 003456677
  6 | 00256
  7 | 05
  8 | 0
```

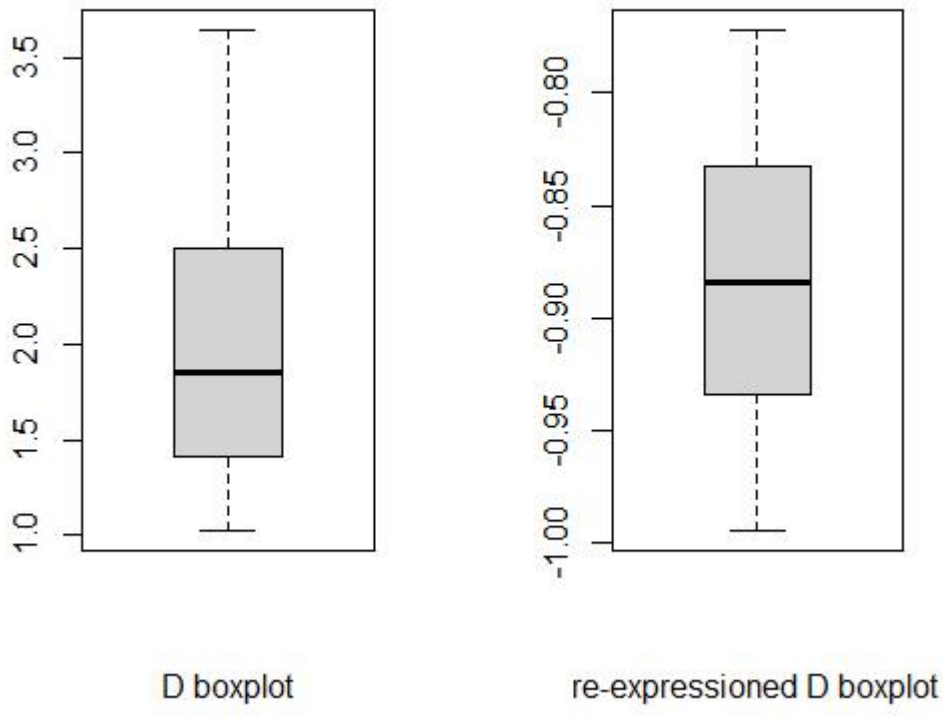
[파운드화된 자료의 줄기그림]

줄기 그림은 대체로 비슷하지만 차이가 나는 이유는 소수점이 배수화 되면서 정수 구간 자료수에 차이가 생겼기 때문이다. 구간을 특정하지 않은 분포는 상자그림에서 확인되는데 완전히 동일함을 알 수 있다.

나. 대칭화 후 비교

위에서 설정한 findp 함수로 대칭화에 적절한 p를 찾아낸 후 이를 통해 대칭화한다.

```
findp(D)
#p=-0.2083707
boxplot(D, xlab = "D boxplot")
boxplot(-D^(-.2), xlab = "re-expressioned D boxplot")
```



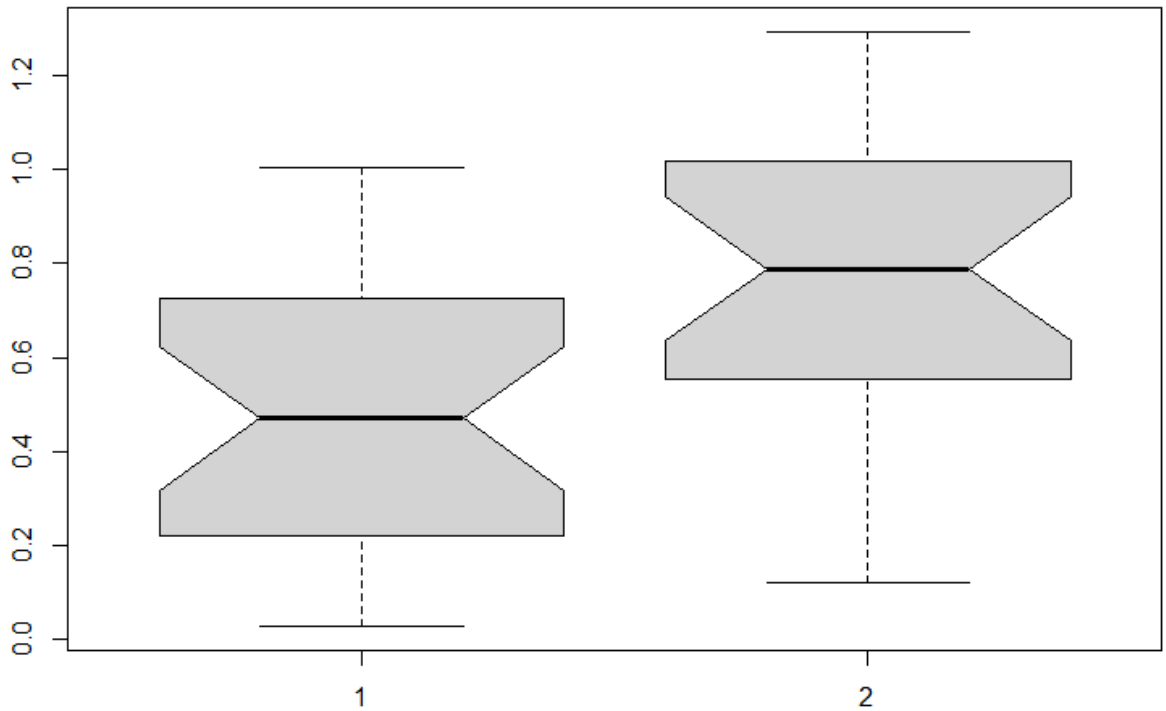
비교: 상자 그림에서 median이 상자의 가운데로 오며 양끝줄의 길이가 비슷해진 점으로 보아 분포가 대칭화되었음을 확인할 수 있다.

다. 사망 집단과 생존 집단을 각각 분석

```
D_death <- 사망한 신생아의 몸무게
D_live <- 생존한 신생아의 몸무게
#spread 함수 생성
spread <- function(x) {
  H_L = quantile(x, 0.25)
  H_U = quantile(x, 0.75)
  H_U-H_L

}
```

spread에 일치되도록 하는 적절한 power값인 p를 찾기 위해
-1부터 0.5까지 값을 가지는 x를 생성하고 이를 p에 대입해 spread차이를 0에 가깝게 만드는 x를 찾는다.

```
x<-(-100:50)/100
y<-rep(0,length(x))

for(i in 1:length(x)){
  y[i] <- spread(D_death^x[i])-spread(D_live^x[i])

}
plot(x,y)
n=which(y==-min(abs(y)))
x[n]
#x=0 , p가 0으로 계산되었으므로 log변환을 사용한다.
par(mfrow=c(1,1))
boxplot(log(D_death), log(D_live), notch=T)
```



[1=death, 2=live]

notch가 겹치는지 확인해 보면 사망한 신생아의 notch와 생존한 신생아의 notch가 겹치지 않아 몸무게의 차이는 생존에 유의미하다고 볼 수 있다.

사망한 신생아 몸무게의 notch 구간을 계산해보면 다음과 같다.

```
c(median(D_death)+1.58*(fivenum(D_death)[4]-fivenum(D_death)[2])/length(D_death)^0.5,
median(D_death)-1.58*(fivenum(D_death)[4]-fivenum(D_death)[2])/length(D_death)^0.5)
```

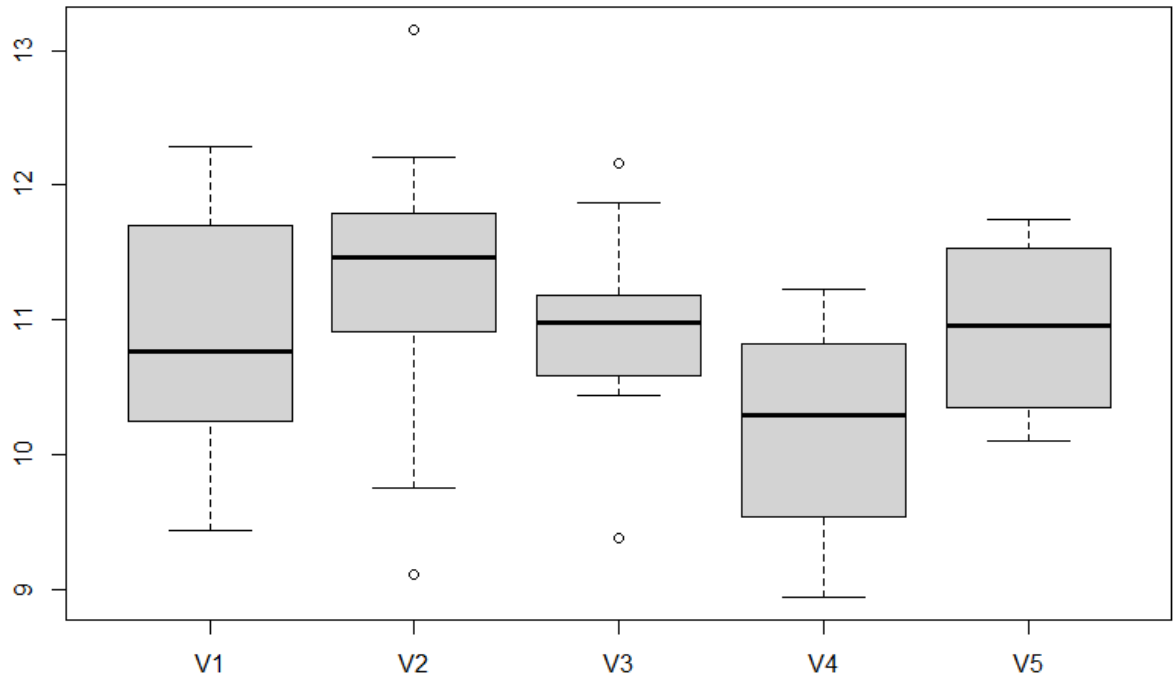
```
#[1] 1.850555 1.349445
```

즉, 몸무게가 1.85 킬로그램 미만일 경우 사망할 가능성이 크다고 볼 수 있다.

(4)

원 자료의 boxplot을 그려 확인하면 다음과 같다.

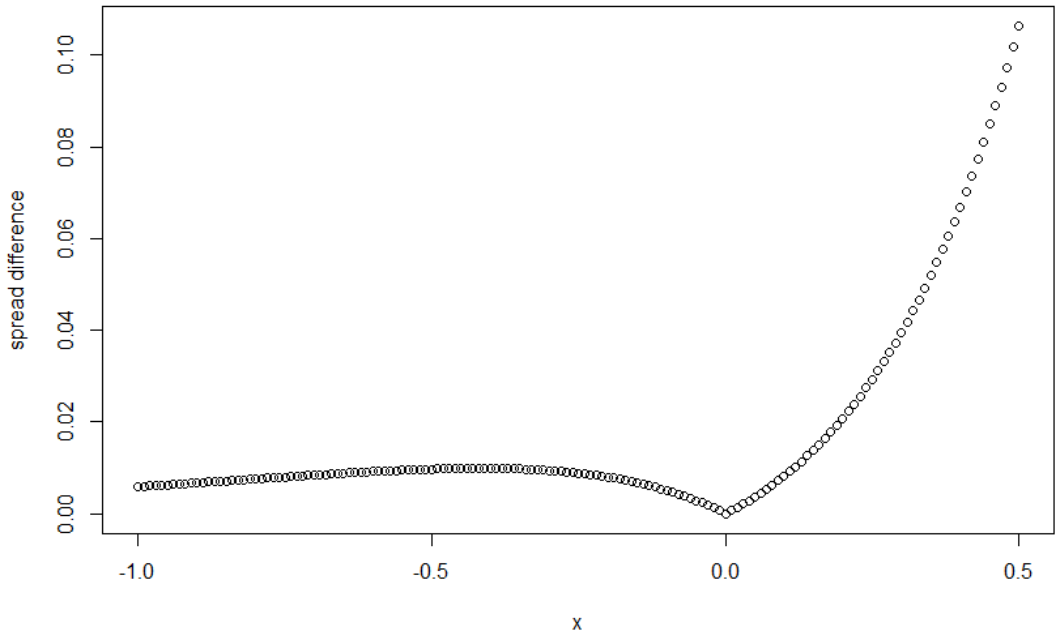
```
P <- PLASMA.dat
boxplot(P)
```



각 시간대(V1~V5)간의 상자 길이가 조금씩 달라 재표현이 필요하다고 생각된다. 특히 V1이 가장 크고 V3가 가장 작은 길이를 가지고 있음을 확인할 수 있다. 따라서 V1과 V3의 spread차이를 최소화하는 p를 찾아 re-expression하기 위해 x에 (-1, -0.99 ... , 0,5)를 넣어가며 V1과 V3간의 spread차이를 가장 적게 하는 p를 찾는다.

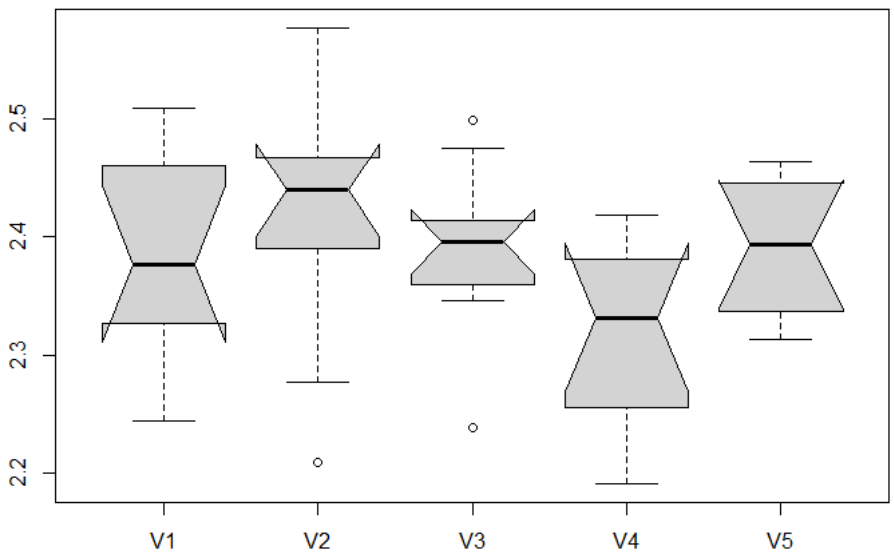
```
x<-(-100:50)/100
y<-rep(0,length(x))
for(i in 1:length(x)){
  y[i] <- spread(P$V1^x[i])-spread(P$V3^x[i])
```

}
plot(x,y)



x가 0일때 spread 차이가 가장 적으므로, 적절한 p값은 0에 가까워 로그변환을 하는것이 가장 적절하다고 생각된다.

boxplot(log(P), notch = T)



분석: V1부터 V5까지는 순서대로 오전 8시, 11시, 오후 2시, 5시, 8시를 나타낸다. 오후8시와 11시는 notch가 대부분 겹치기 때문에 혈장 농도에 큰 차이는 없다고 볼 수 있다. 오전 11시부터는 점점 오후2시, 오후5시로 될수록 혈장 농도가 줄어듦을 학인할 수 있다. 다시 오후8시붜는 혈장농도가 다소 증가하나 notch가 조금 겹치기 때문에 명확한 증가라고는 볼 수 없다고 생각된다. 결론적으로 혈장 농도는 다른 시간대에는 차이가 있다고 보기 힘들지만 오후 11시를 지나면서 오후5시가 되기까지 줄어드는 경향이 강하.