```
### 인천 3년 평균 세대수 EDA ###
## set working directory
setwd('C:/Users/wndy4/Desktop/Project_DEMA')
## csv file load
dat_pop = read.csv('정동호/인천광역시세대수/인천광역시 월별 구별 세대수(new).csv',header=T,stringsAsFactors=F)

# 기초정보
dim(dat_pop)
```

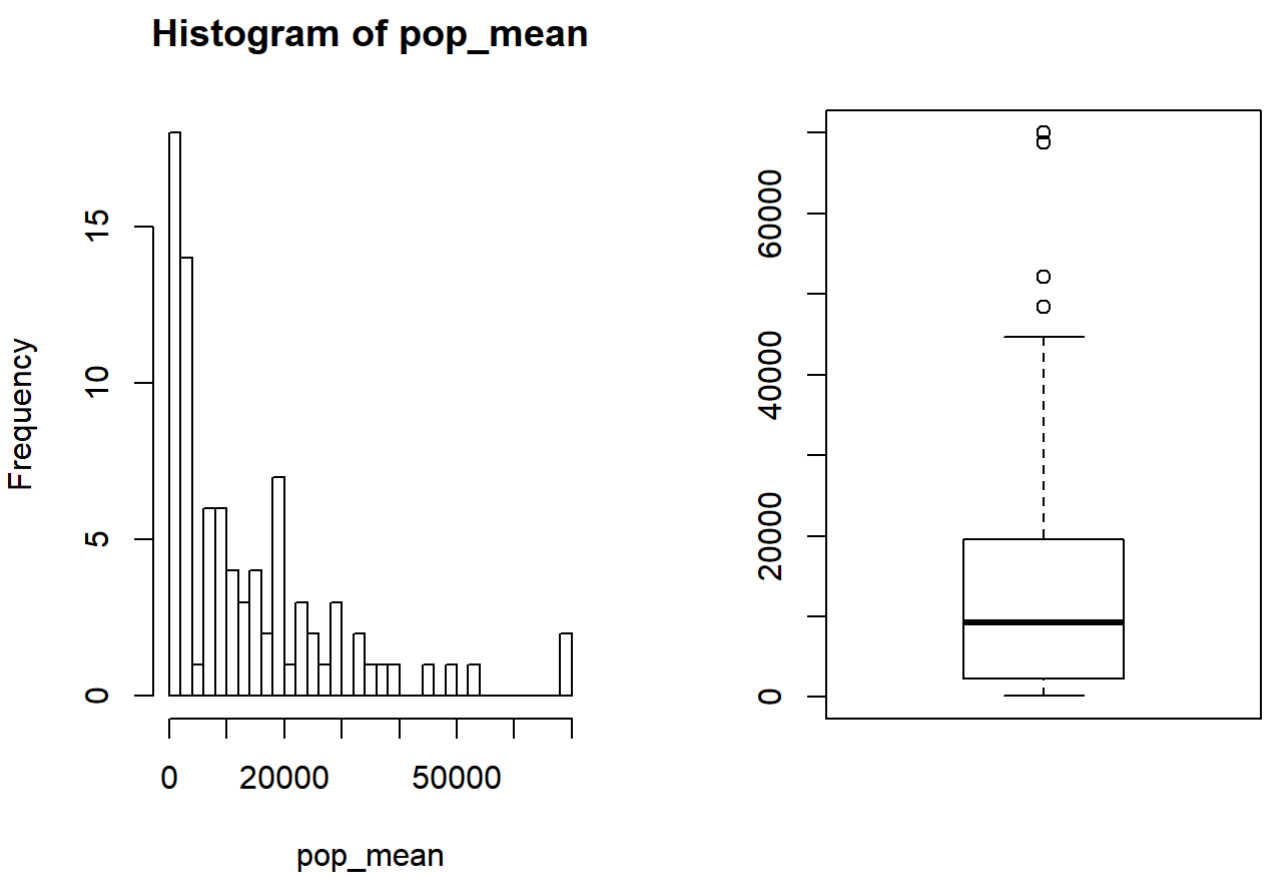
```
## [1] 95 39
```

```
a = which(dat_pop$행정구역 == '합계')
# 합계 제외한 3년 평균 인구수 데이터
pop_mean = dat_pop[-a,39]

# 평균, 분산
mean(pop_mean);sd(pop_mean)
```

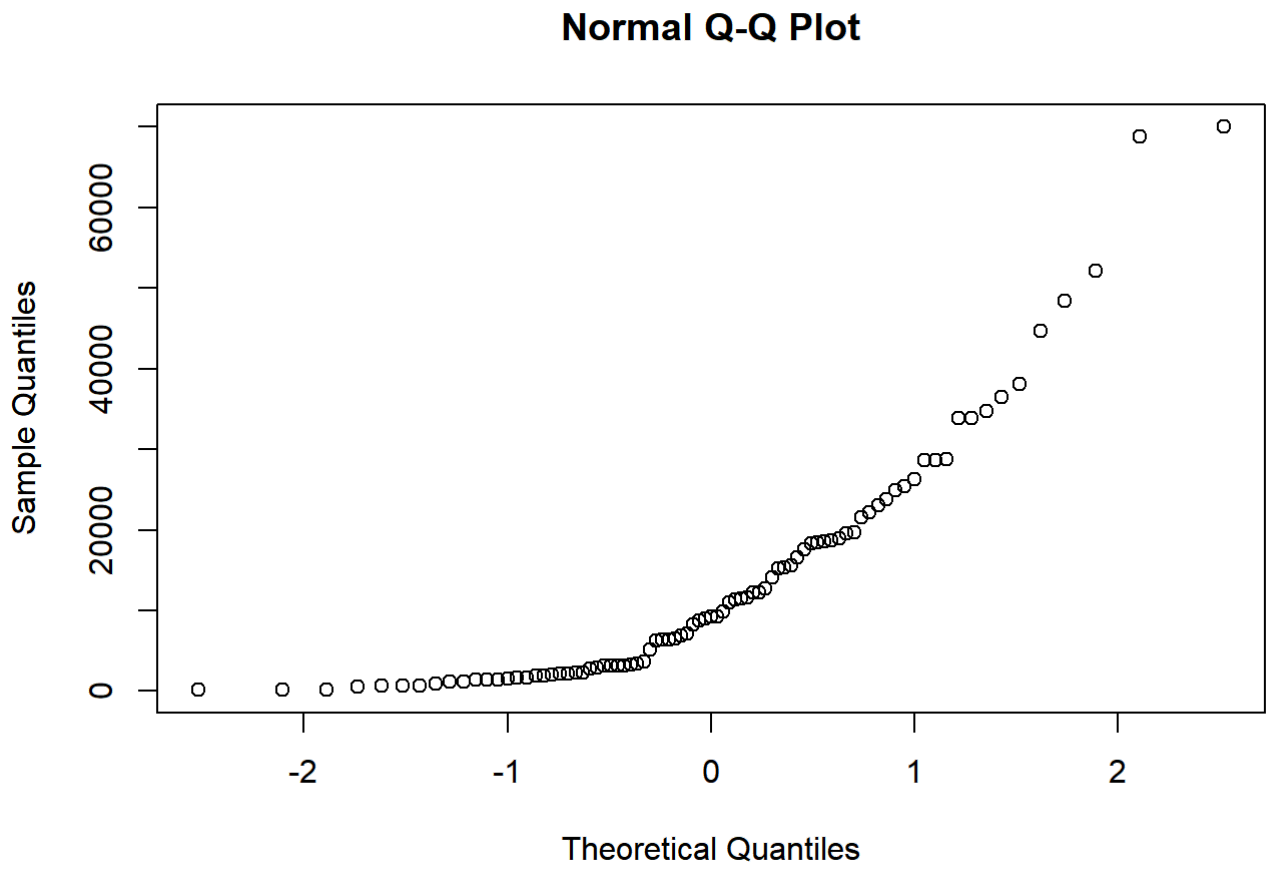
```
## [1] 13905.48
```

```
## [1] 15002.02
```

```
graphics.off()
# 그래프
windows()
par(mfrow=c(1,2))
hist(pop_mean,breaks = 40)
boxplot(pop_mean)
```




```
# 왼쪽으로 쏠려 있는 모습이며, outlier가 있는 것으로 보인다.

# 정규확률그림
windows()
qqnorm(pop_mean)
```




```
# 정규확률그림이 아래로 볼록한 형태이므로 데이터의 분포가 왼쪽으로 쏠려있는 형태라는 것을 알 수 있다.

# 기술통계량
summary(pop_mean)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     161    2255    9249   13905   19596   69983
```

```
# 평균이 중앙값보다 크므로 왼쪽으로 쏠려 있는 분포라는 것을 알 수 있다.

# 왜도 및 첨도
# 왜도, 첨도
library(moments)
```

```
## Warning: package 'moments' was built under R version 3.4.4
```

```
skewness(pop_mean)
```

```
## [1] 1.684005
```

```
kurtosis(pop_mean) # 정규분포와 비슷한 꼬리를 갖는다고 하기 힘들다
```

```
## [1] 6.051515
```

```
# Agostino skewness test
# 왜도가 0인지 검정
agostino.test(pop_mean)
```

```
##
##  D'Agostino skewness test
##
## data:  pop_mean
## skew = 1.6840, z = 5.0823, p-value = 3.728e-07
## alternative hypothesis: data have a skewness
```

```
# p-value가 매우 작으므로 유의수준 0.05에서 왜도가 0이라고 하기 힘들다

# 첨도가 3인지 검정
anscombe.test(pop_mean)
```

```
##
##  Anscombe-Glynn kurtosis test
##
## data:  pop_mean
## kurt = 6.0515, z = 3.3500, p-value = 0.0008082
## alternative hypothesis: kurtosis is not equal to 3
```

```
# p-value가 작으므로 유의수준 0.05에서 첨도가 0이라고 하기 힘들다

# 정규성 검정
shapiro.test(pop_mean)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  pop_mean
## W = 0.81433, p-value = 5.813e-09
```

```
# p-value가 매우 작으므로 유의수준 0.05에서 귀무가설 기각(귀무가설 : 정규분포를 따른다)
# 정규분포를 따른다고 하기 힘들다

jarque.test(pop_mean)
```

```
##
##  Jarque-Bera Normality Test
##
## data:  pop_mean
## JB = 73.154, p-value < 2.2e-16
## alternative hypothesis: greater
```

```
library(nortest)

# kolmogorov-smirnov test
lillie.test(pop_mean)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  pop_mean
## D = 0.17979, p-value = 3.563e-07
```

```
ad.test(pop_mean)
```

```
##
##  Anderson-Darling normality test
##
## data:  pop_mean
## A = 4.3491, p-value = 6.853e-11
```

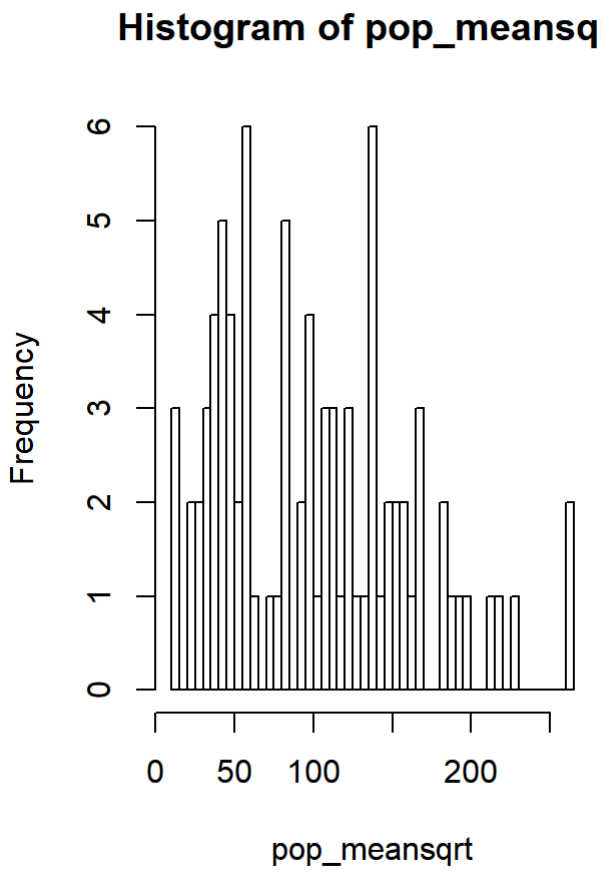
```
# 모든 검정에서의 p-value가 0.05보다 작으므로 데이터가
# 정규분포로부터 나왔다는 귀무가설을 채택하기 힘들다.
# 즉, 데이터가 정규분포로부터 나왔다고 하기 힘들다

# 변수 변환
# 변수변환전 범위 확인
min(pop_mean)
```

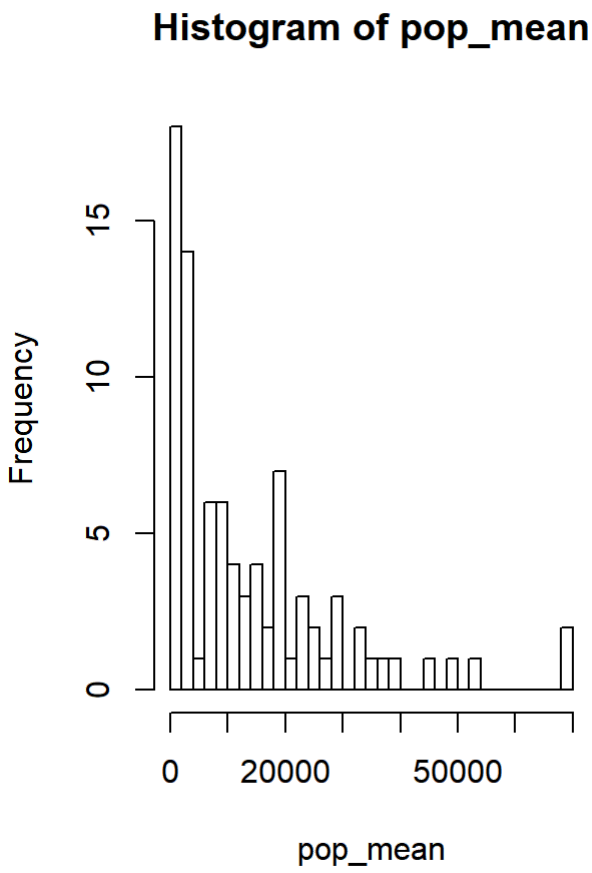
```
## [1] 161
```

```
max(pop_mean)
```

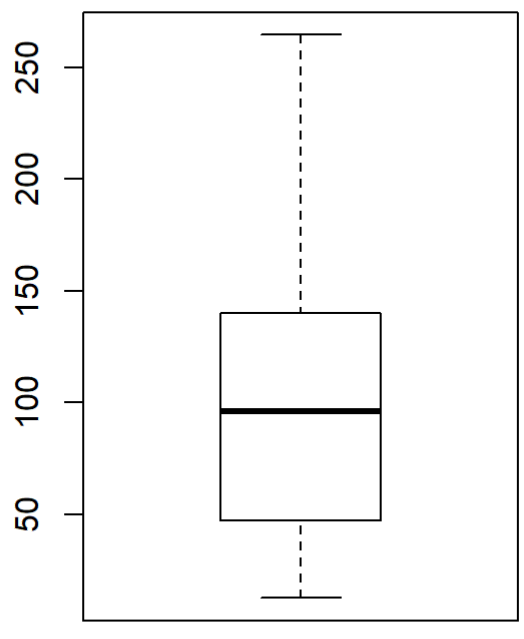
```
## [1] 69983
```

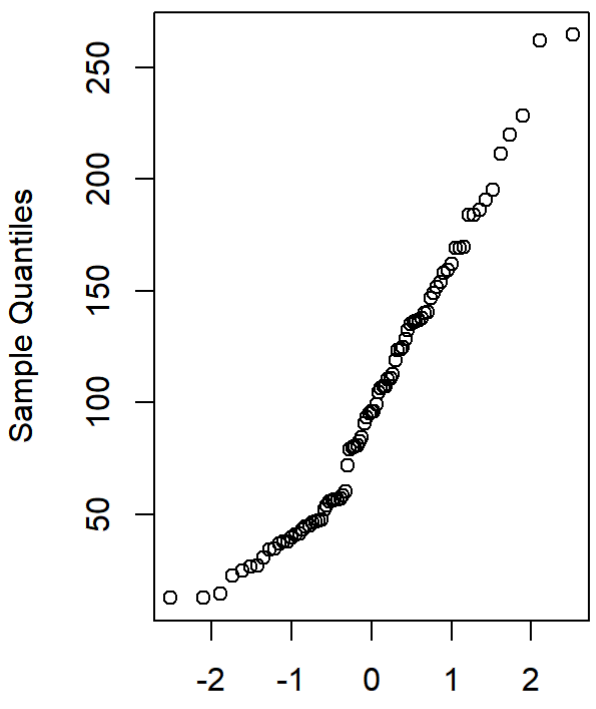
```
# 변수변환 : sqrt, log

pop_meansqrt = sqrt(pop_mean)
pop_meanlog = log(pop_mean)

# sqrt 분석
windows()
par(mfrow=c(1,2))
hist(pop_meansqrt,breaks = 40)
hist(pop_mean,breaks = 40)
```






```
windows()
par(mfrow=c(1,2))
boxplot(pop_meansqrt)
qqnorm(pop_meansqrt)
```






```
skewness(pop_meansqrt)
```

```
## [1] 0.5907933
```

```
agostino.test(pop_meansqrt)
```

```
##
##  D'Agostino skewness test
##
## data:  pop_meansqrt
## skew = 0.59079, z = 2.24950, p-value = 0.02448
## alternative hypothesis: data have a skewness
```

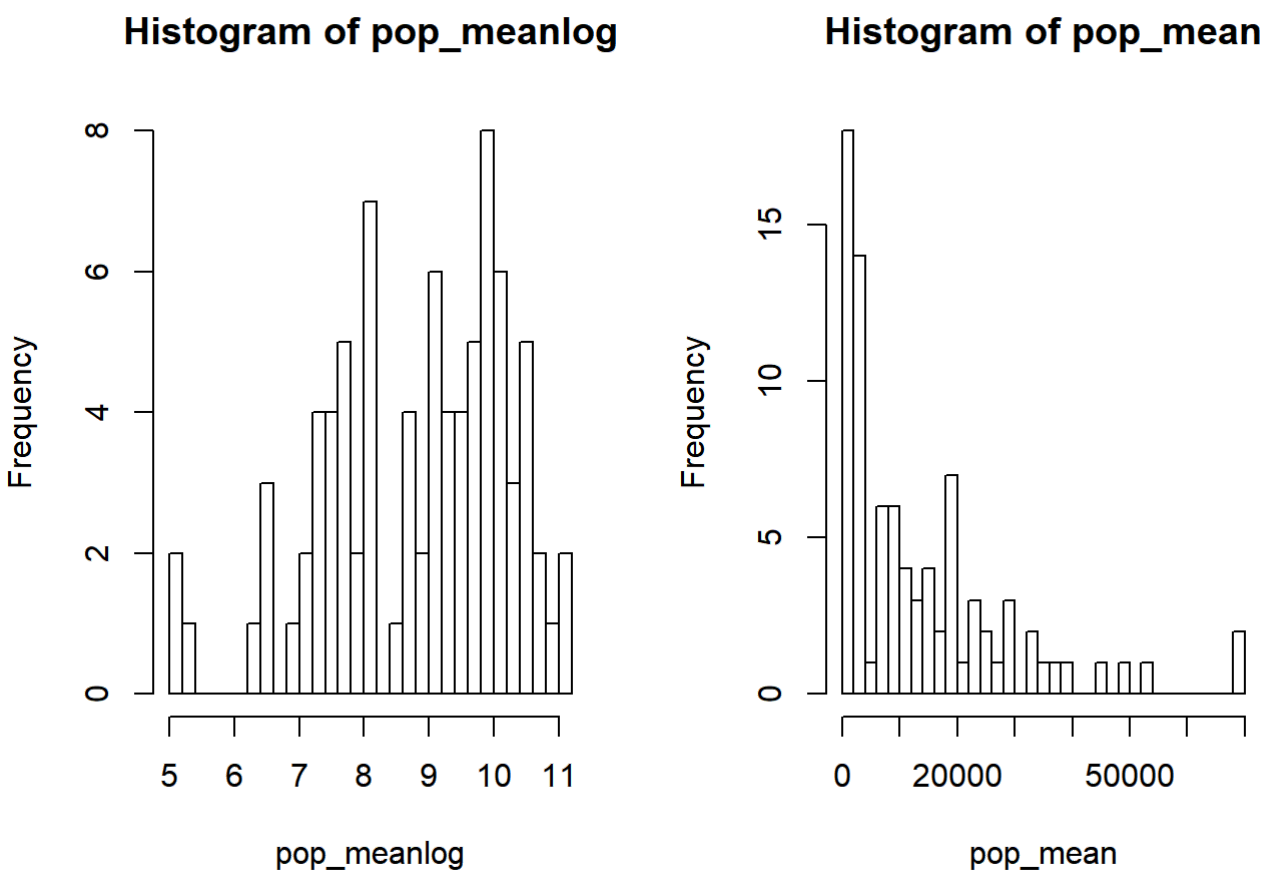
```
ad.test(pop_meansqrt)
```

```
##
##  Anderson-Darling normality test
##
## data:  pop_meansqrt
## A = 1.1588, p-value = 0.004715
```

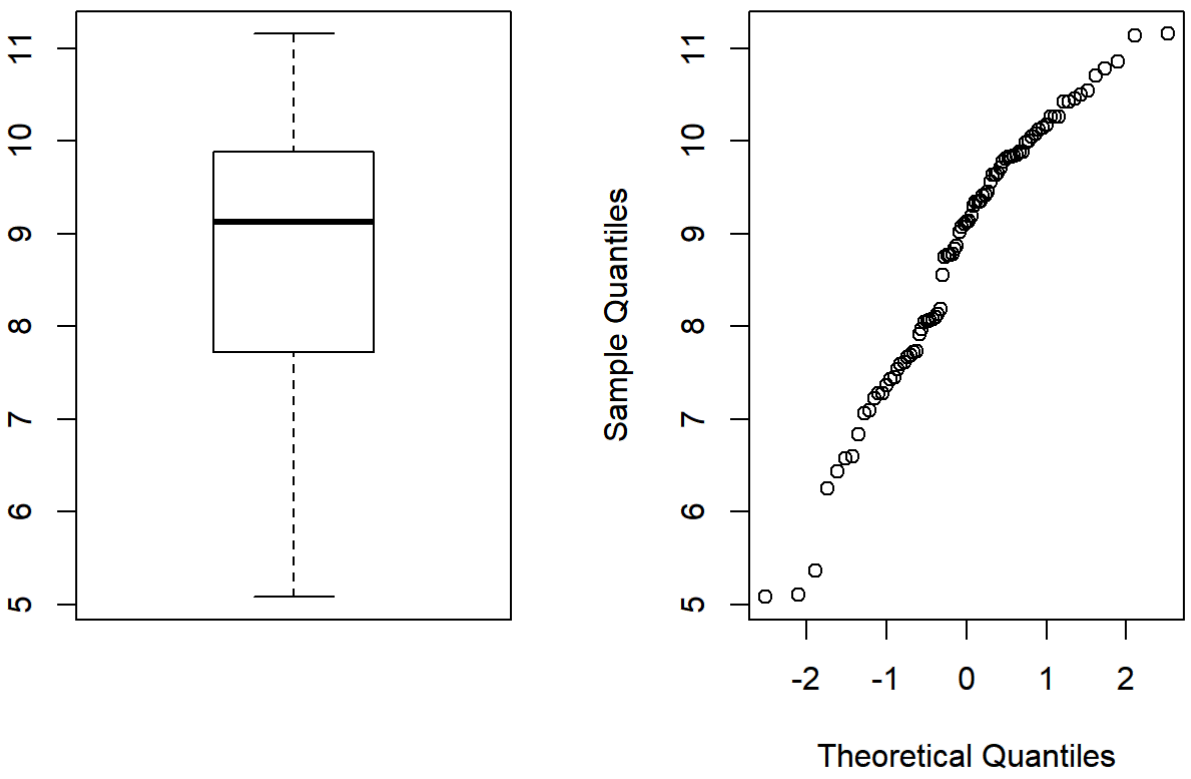
```
lillie.test(pop_meansqrt)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  pop_meansqrt
## D = 0.13009, p-value = 0.001156
```

```
# 히스토그램과 상자그림에서 변환 전보다 쏠린 정도가 완화되었지만
# 아직 왼쪽으로 쏠린 분포를 보이고 있다.
# 왜도도 변환전보다 0에 가까운 값을 갖지만,왜도가 0인지에대한 검정에서 p-value가
# 0.05보다 작으므로 왜도가 0이라고 할 수 없다.
# 정규성 검정에서 역시 p-value가 0.05보다 작으므로
# 데이터가 정규분포로부터 나왔다고 하기 힘들다.

# log 분석
windows()
par(mfrow=c(1,2))
hist(pop_meanlog,breaks = 40)
hist(pop_mean,breaks = 40)
```




```
# 오히려 로그 변환은 기존보다 오른쪽으로 쏠리는 모양을 보인다.
windows()
par(mfrow=c(1,2))
boxplot(pop_meanlog)
qqnorm(pop_meanlog)
```




```
# 정규확률그림이 오목한 모양으로 보아 오른쪽으로 쏠린 모양을 유추할 수 있다.
skewness(pop_meanlog)
```

```
## [1] -0.6055561
```

```
agostino.test(pop_meanlog)
```

```
##
##  D'Agostino skewness test
##
## data:  pop_meanlog
## skew = -0.60556, z = -2.29960, p-value = 0.02147
## alternative hypothesis: data have a skewness
```

```
# 왜도가 -0.60이고 왜도가 0인지 검정하기 위한 가설검정에서도 p-value가 0.021으로 0.05보다 작으므로
# 귀무가설을 기각할 수 있다. 즉 왜도가 0이라고 할 수 없다.

kurtosis(pop_meanlog)
```

```
## [1] 2.759352
```

```
anscombe.test(pop_meanlog)
```

```
##
##  Anscombe-Glynn kurtosis test
##
## data:  pop_meanlog
## kurt = 2.75940, z = -0.20612, p-value = 0.8367
## alternative hypothesis: kurtosis is not equal to 3
```

```
# 첨도가 약 2.7로 3보다 작지만, 첨도가 3인지에 대한 검정에서
#p-value가 0.83로 매우 크므로 첨도가 3이라고 할 수 있다.

ad.test(pop_meanlog)
```

```
##
##  Anderson-Darling normality test
##
## data:  pop_meanlog
## A = 1.1957, p-value = 0.003819
```

```
lillie.test(pop_meanlog)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  pop_meanlog
## D = 0.10515, p-value = 0.02126
```

```
# p-value가 0.05보다 작으므로 귀무가설 기각
# 즉 데이터의 분포가 정규분포라 하기 힘들다
```

NA