EDA 6장(2) 과제

주의사항을 숙지하였고 모든 책임을 지겠습니다.

2019122041 송유진

---

1. Shortly after metric units of length were officially introduced in Australia, each of a group of 44 students was asked to guess, to the nearest metre, the width of the lecture hall in which they were sitting. Another group of 69 students in the same room was asked to guess the width in feet, to the nearest foot. The true width of the hall was 13.1 metres (43.0 feet).

Guesses in metres:

8	9	10	10	10	10	10	10	11	11	11	11	12	12	13
13	13	14	14	14	15	15	15	15	15	15	15	15	16	16
16	17	17	17	17	18	18	20	22	25	27	35	38	40	

Guesses in feet:

24	25	27	30	30	30	30	30	30	32	32	33	34	34	34
35	35	36	36	36	37	37	40	40	40	40	40	40	40	40
40	41	41	42	42	42	42	43	43	44	44	44	45	45	45
45	45	45	46	46	47	48	48	50	50	50	51	54	54	54
55	55	60	60	63	70	75	80	94						

(1) Check separately if the guesses in metres and guesses in feet follow normal distributions.

학생들이 metres 그리고 feet로 강의실 폭을 예측한 거리 자료를 각각 metres와 feet에 할댕해주었다.

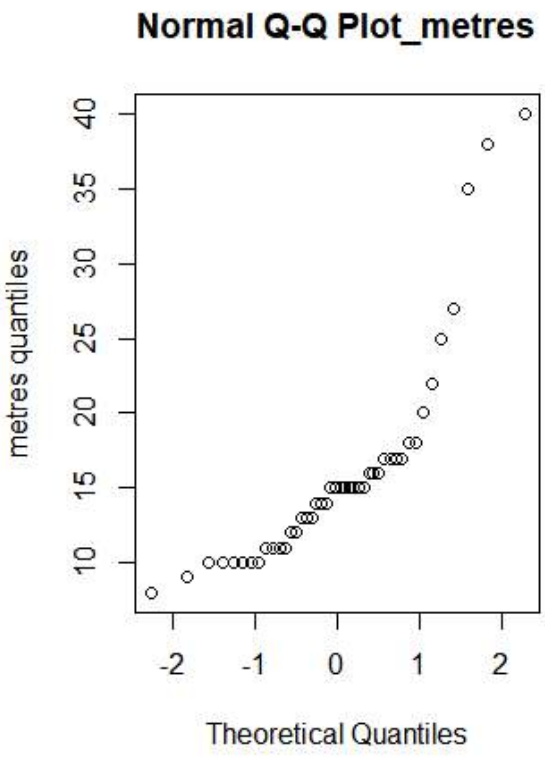
```
> metres <- c(8,9,10,10,10,10,10,10,11, 11,11,11,12,12,13,
+             13, 13,14,14,14,15,15,15,15,15,15,15,15,16,16,
+             16,17,17,17,17,18,18,20,22,25,27,35,38,40
+             )
>
> length(metres)
[1] 44
>
> feet <- c(24,25,27,30,30,30,30,30,30,32,32,33,34,34,34,
+           35,35,36,36,36,37,37,40,40,40,40,40,40,40,40,
+           40,41,41,42,42,42,42,43,43,44,44,44,45,45,45,
+           45,45,45,46,46,47,48,48,50,50,50,51,54,54,54,
+           55,55,60,60,63,70,75,80,94)
>
> length(feet)
[1] 69
```

두 자료의 대략적인 분포를 파악하기 위해 stem and leaf display를 그려주었다.

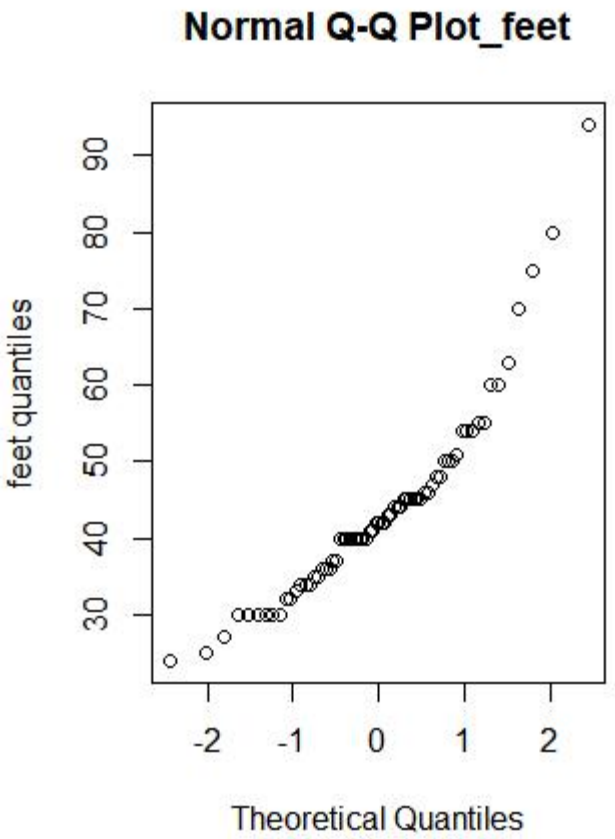
```
> stem(metres)

  The decimal point is 1 digit(s) to the right of the |

  0 | 89
  1 | 000000111122333444
  1 | 55555555666777788
  2 | 02
  2 | 57
  3 |
  3 | 58
  4 | 0

>
> stem(feet)

  The decimal point is 1 digit(s) to the right of the |

  2 | 457
  3 | 0000002234445566677
  4 | 00000000011222233444555555566788
  5 | 000144455
  6 | 003
  7 | 05
  8 | 0
  9 | 4
```

상단의 줄기잎그림을 통해 두 분포가 정규분포를 하고 있다고 보기에는 어려움이 있었다. 정규분포는 중간을 기준으로 종 모양으로 퍼져 있는 형태를 보이고 있는데 상단의 두 분포는 오른쪽으로 skewed된 모양을 확인할 수 있다. 다음으로, qqnorm()을 사용하여 각각의 Normal QQplot을 그려보았다.

```
qqnorm(metres, ylab="metres quantiles", main="Normal Q-Q Plot_metres")
```



```
qqnorm(feet, ylab="feet quantiles", main= "Normal Q-Q Plot_feet")
```



우선, metres 데이터의 normal qqplot은 대체적으로 convex모양으로 분포하고 있고, feet자료 또한 마찬가지로 점들이 convex 모양으로 분포하고 있다. 만약 두 자료가 정규분포를 따른다면 직선의 형태로 점들이 분포해야 하기 때문에 두 데이터가 정규분포를 따르지 않는다는 것을 다시 한 번 확인해볼 수 있었다.
metres 데이터의 경우, 점들이 10~20에 집중되어 있으나, theoretical 정규분포에서는 점들이 -1~1 사이에 넓게 분포하고 있다. 즉, metres 데이터는 오른쪽으로 꼬리가 길게 늘어진 분포임을 확인할 수 있다.
feet 데이터의 경우, 30~50(왼쪽) 구간에 자료들이 촘촘하게 모여있지만, theoretical 정규분포는 -1~1에 분포하고 있다. 따라서 feet 데이터들도 마찬가지로 오른쪽으로 skewed된 분포임을 알 수 있다.

```
> lsum(metres)
  letter depth lower  mid upper spread
1      M  22.5    15 15.0    15      0
2      H  11.5    11 14.0    17      6
3      E   6.0    10 16.0    22     12
4      D   3.5    10 20.5    31     21
5      C   2.0     9 23.5    38     29
> lsum(feet)
  letter depth lower   mid upper spread
1      M  35.0    42 42.00  42.0    0.0
2      H  18.0    36 42.00  48.0   12.0
3      E   9.5    31 42.75  54.5   23.5
4      D   5.0    30 46.50  63.0   33.0
5      C   3.0    27 51.00  75.0   48.0
```

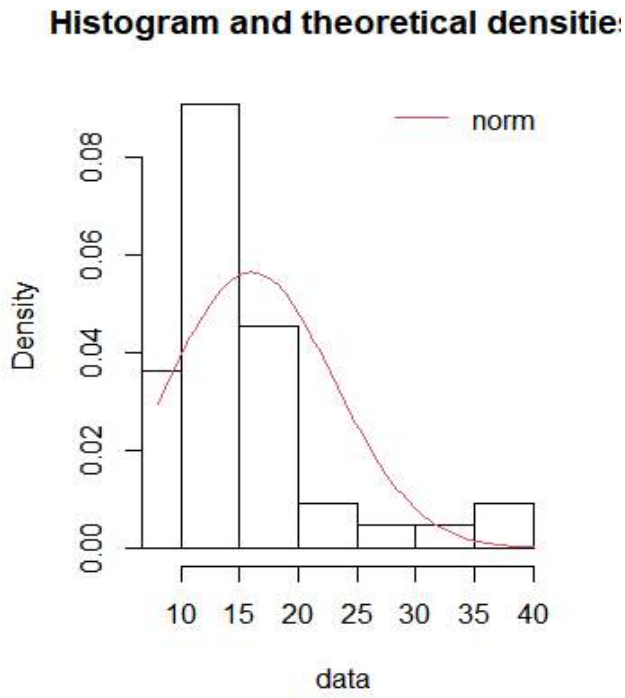
해당 분석 내용은 lsum을 통해서도 살펴볼 수 있다. 두 분포에서 모두 mid 값이 분위수가 증가함에 따라 증가한다는 것을 알 수 있다. 이는 분포가 오른쪽으로 skewed되어있다는 것을 뜻하기 때문에 두 데이터

모두 정규분포를 따르지 않는다는 결론을 내릴 수 있다.

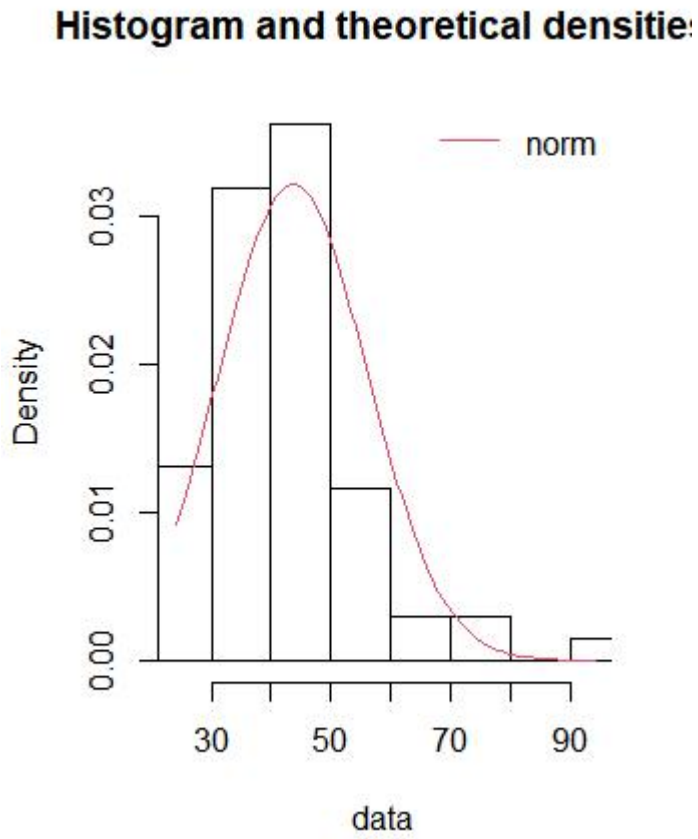
```
> shapiro.test(metres)

        Shapiro-Wilk normality test

data:  metres
W = 0.76568, p-value = 5.658e-07

> shapiro.test(feet)

        Shapiro-Wilk normality test

data:  feet
W = 0.88616, p-value = 1.306e-05
```

다음으로, shaprio-wilks test를 통해 정규성을 확인해보았다. 두 데이터 모두 p-value인 0.05보다 훨씬 작은 값이 산출되었기 때문에 정규분포를 따른다는 귀무가설을 기각해야만 한다.

```
library(fitdistrplus)
fnormal_metres <- fitdist(metres, "norm")
denscomp(list(fnormal_metres))
```



```
fnormal_feet <- fitdist(feet, "norm")
denscomp(list(fnormal_feet))
```



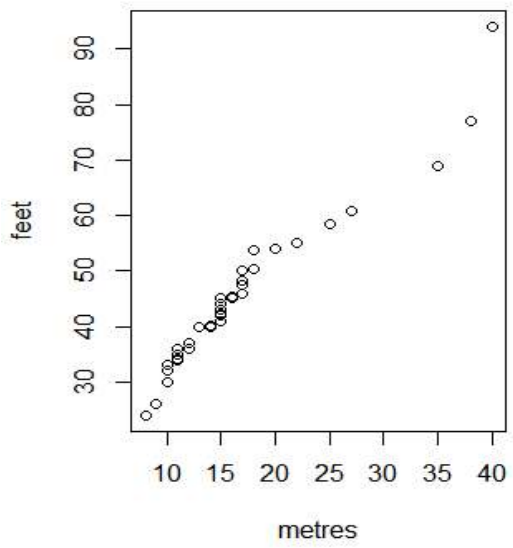
다음으로, histogram과 normal 분포를 그려보았다. metres와 feet 모두 정규분포를 벗어남을 확인할 수 있었다.

(2) Examine if the two data sets follow the same distribution. If so, then analyze their variances and means from the qq-plot and the sample estimates.

```
> qq.x <- qqplot(metres, feet)$x
> qq.y <- qqplot(metres, feet)$y
> length(qq.x)
[1] 44
> length(qq.y)
[1] 44
```

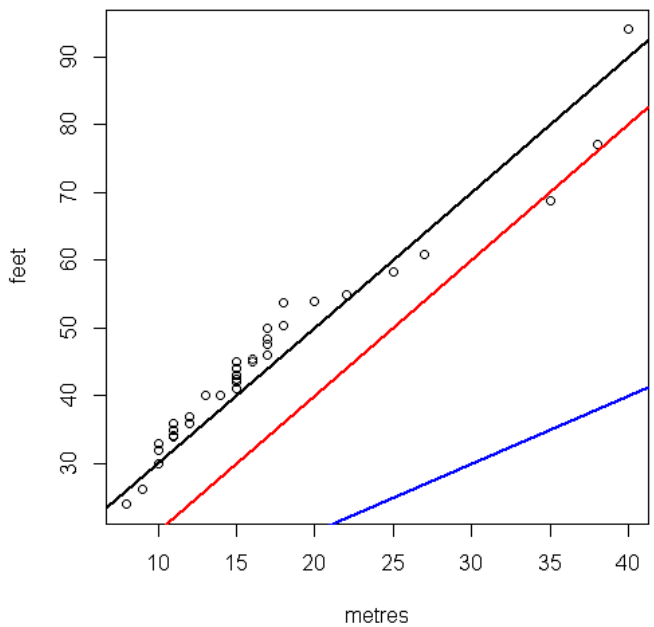
(1)에서 확인한 바와 같이 metres와 feet은 개수가 다르다. 그러나 qqplot은 자동으로 interpolation을 해주기 때문에 상단의 코드를 실행하고 아래 qqplot을 그려보았다.

```
qqplot(metres, feet)
```



x축은 metres를, y축은 feet을 나타낸다. 만약 두 분포가 같은 분포라면, qqplot은 직선 형태로 나오게 된다. 상단의 qqplot은 어느정도의 선형을 확인할 수 있다.
만약, 두 분포가 평균도 같고 분산도 같은 분포라면 qqplot은 기울기가 1이고 절편이 0인 직선의 형태로 나올 것이다. 만약 직선의 형태지만 기울기가 1이 아니고 절편이 0이 아니라면, 이는 분포는 같지만 평균과 분산이 다른 분포임을 알 수 있다.

```
qqplot(metres, feet)
line(qqplot(metres, feet))
abline(0,1, lwd=2, col="blue")
abline(0,2, lwd=2, col="red")
abline(10,2, lwd=2, col="black")
```



상단의 코드를 실행하여 기울기가 2이고 절편이 10인 검정선, 기울기가 2이고 절편이 0인 빨간선 그리고 기울기가 1이고 절편이 0인 파란선을 그려보았다.

Metres와 feet의 점들은 검정선에 따라 어느정도의 선형의 형태를 보이고 있음을 확인할 수 있었다. 따라서 metres와 feet는 같은 분포에서 나왔다고 판단할 수 있었다.

If X and Y are standardized, $\frac{X-\mu_X}{\sigma_X} = \frac{Y-\mu_Y}{\sigma_Y}$.

Substituting $X = \frac{Y-\mu}{\sigma}$, $\mu = \mu_Y - \frac{\sigma_Y}{\sigma_X}\mu_X$ and $\sigma = \frac{\sigma_Y}{\sigma_X}$.

$$F_2^{-1}(p) = \mu + \sigma F_1^{-1}(p)$$

상단의 수식을 이용하여 metres, feet 각각의 평균과 분산을 추정해보았다. slope=2로 추정하였기 때문에 sigma_y(feet)와 sigma_x(slope)의 비율은 2이다. 따라서 feet의 standard deviation이 metres의 standard deviation보다 약 2배 정도 크다. metres와 feet 표본자료들의 표준편차를 계산하여 이를 확인해보았다.

```
> mean(metres)
[1] 16.02273
> sd(metres)
[1] 7.144647
>
> mean(feet)
[1] 43.69565
> sd(feet)
[1] 12.49742
```

Feet의 표준편차가 12.5, metres의 표분편차가 7.1로 대략 2배 정도 차이가 나는 것을 확인하였다.
μ intercept는 10이라고 추정하였는데, 이를 통해 값을 다음과 같이 구할 수 있다.
Feet의 평균은 43, metres의 평균은 16으로, 2*16(x-metres값)≈43.6(y-feet값)-10=33.6, 각각 32, 33.6으로 거의 비슷하다는 것을 알 수 있다.

따라서 metres와 feet 데이터는 qqplot을 그렸을 때 기울기가 2, 절편이 10인 직선의 형태를 지니고 있기 때문에 같은 분포임을 알 수 있었다. 또한, 표준편차는 feet 데이터가 metres에 비해 2배 정도 크고, 평균은 대략 30(43.6-16=27.6) 차이나는 분포임을 알 수 있다.

qqplot 값과 계산해서 구한 값들을 고려하여 최종적으로 추정한 metres의 평균과 표준편차 각각 15, 7이며, feet의 평균과 표준편차는 43, 13이다. 따라서, metres의 분산은 7^2=49, feet의 분산은 13^2=169임을 알 수 있다.

2. In 1960s in the United States of America 16 states owned the retail liquor stores while in 26 the stores were privately owned. (Some are omitted for technical reasons.) The table shows the price in dollars of a fifth of Seagram 7 Crown Whisky in two sets of states in 1961. Do the distributions of prices of two group follow the same distribution?

16 monopoly states:

4.65, 4.55, 4.11, 4.15, 4.20, 4.55, 3.80, 4.00, 4.19, 4.75,
4.74, 4.50, 4.10, 4.00, 5.05, 4.20

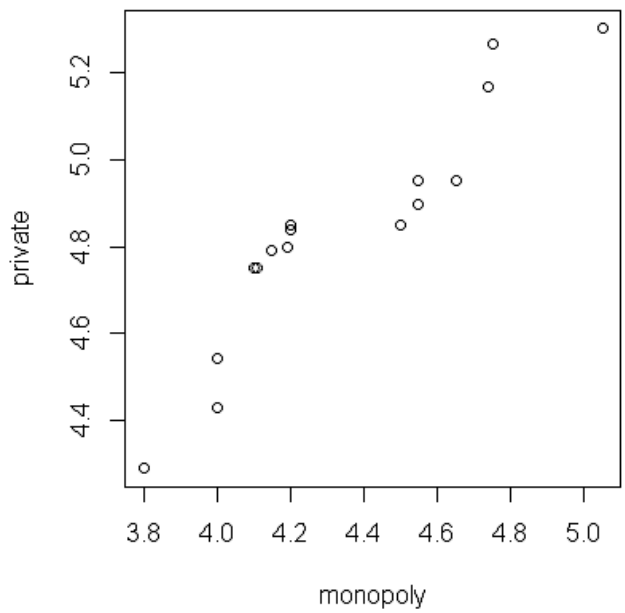
26 private-ownership states:

4.82, 5.29, 4.89, 4.95, 4.55, 4.90, 5.25, 5.30, 4.29, 4.85,
4.54, 4.75, 4.85, 4.85, 4.50, 4.75, 4.79, 4.85, 4.79, 4.95,
4.95, 4.75, 5.20, 5.10, 4.80, 4.29

```
> monopoly = c(4.65,  4.55,  4.11,  4.15,  4.20,
 4.55,  3.80,  4.00, 4.19,  4.75,
+              4.74,  4.50,  4.10,  4.00,  5.05,
 4.20)
> private = c(4.82,  5.29,  4.89,  4.95,  4.55,
 4.90,  5.25,  5.30,  4.29,  4.85,
+             4.54,  4.75,  4.85,  4.85,  4.50,
 4.75,  4.79,  4.85,  4.79,  4.95,
+             4.95,  4.75,  5.20,  5.10,  4.80,
 4.29)
> length(monopoly)
[1] 16
>
> length(private)
[1] 26
```

상단의 코드를 실행하여 monopoly와 private에 할당해주었고 데이터의 길이는 16과 26으로 같지 않다는 사실을 알 수 있었다. 두 데이터의 크기가 다른 경우라도 크기가 큰 자료의 값들이 작은 자료의 크기에 맞추어 interpolated 되기 때문에 qqplot을 그려서 두 데이터를 비교해보았다.

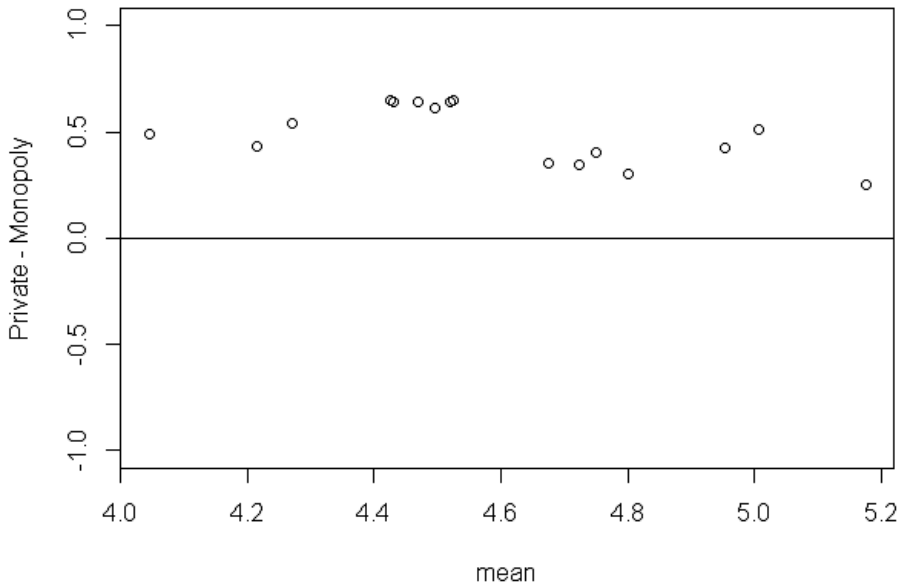
```
> line(qqplot(monopoly, private))

Call:
line(qqplot(monopoly, private))

Coefficients:
[1]  2.1013  0.6382
```



그 결과, 16개의 점들이 찍히는 것을 확인할 수 있었고 전체적으로 값들이 직선을 이루고 있음을 알 숭 있었다. 결론적으로 두 데이터의 분포는 같은 모양임을 확인할 수 있었다. 즉, monopoly와 private한 상점에서의 위스키 가격의 분포는 같다는 결론을 내릴 수 있다.

Tukey Mean-Difference Plot
두 분포가 동일한 모양을 가지더라도 위치에 차이가 있을 수 있다. 따라서, 이를 확인하기 위해 Tukey의 Mean-Difference Plot을 그려서 확인해보았다.

```
#tukey mean-diff
qq.x = qqplot(monopoly,private)$x
qq.y = qqplot(monopoly,private)$y
plot((qq.x+qq.y)/2, qq.y-qq.x, ylim=c(-1, 1),
     ylab="Private - Monopoly", xlab="mean")
abline(0,0)
```



코드를 실행한 결과 상단의 plot을 확인할 수 있었다. 점들이 수평으로 직선을 이룰 때가 가장 이상적이다. 그러나, 상단의 결과에서 점들은 수평으로 직선을 이루지만 0이 아닌 0.5에서 직선을 이룬다.
즉 Monopoly와 Private에서 평균적으로 0.5의 차이가 발생한다는 사실을 알 수 있다. 평균적으로 Private 상점의 위스키 가격이 약 0.5달러 정도 더 비싸다.

다음으로, Kolmogorov-Smirnov test을 수행해보았다.
귀무가설은 "monopoly와 private의 위스키 가격의 CDF는 같다."이며 양측검정(two.sided)이다. 결과적으로 p값이 0.0001956이 나와 유의수준 0.05를 기준으로 귀무가설을 기각했다. 즉 두 분포의 CDF는 다르다.

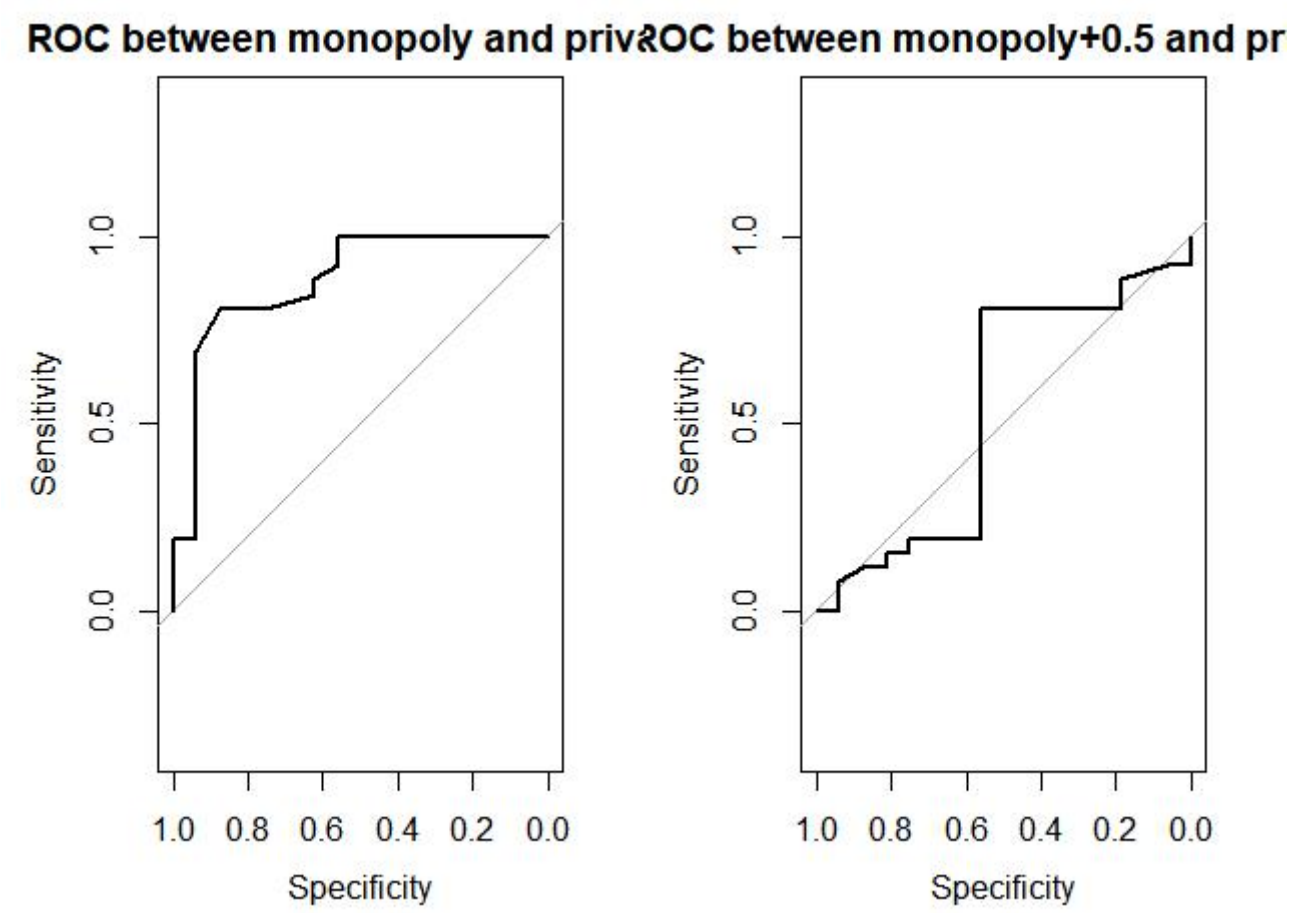
```
> ks.test(monopoly, private, alternative="two.sided")

        Two-sample Kolmogorov-Smirnov test

data:  monopoly and private
D = 0.68269, p-value = 0.0001956
alternative hypothesis: two-sided
```

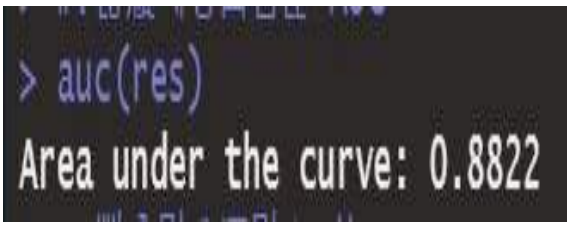
이번에는 monoploy에 0.5를 더하여 KS검정을 수행했다. p값은 0.1324로 귀무가설을 기각할 수 없다. 즉 두 분포의 CDF는 같다고 말할 수 있다.

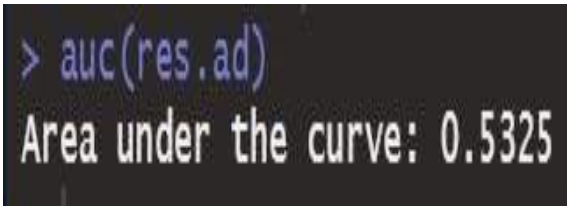
```
> ks.test(monopoly+0.5, private, alternative="two.sided")

        Two-sample Kolmogorov-Smirnov test

data:  monopoly + 0.5 and private
D = 0.37019, p-value = 0.1324
alternative hypothesis: two-sided
```

결론적으로, monopoly와 private 상점에서의 위스키 가격의 분포 모양은 동일하지만 private의 자료가 monopoly보다 0.5정도 오른쪽으로 이동해있다고 해석할 수 있다.

```
#AUC
library(pROC)
ylab=factor(c(rep(0,length(monopoly)), rep(1,length(private))))
res = roc(ylab, c(monopoly, private))
res.ad = roc(ylab, c(monopoly+0.5, private))

par(mfrow=c(1,2))
plot(res, xlim=c(1,0), main="ROC between monopoly and private")
plot(res.ad, xlim=c(1,0), main="ROC between monopoly+0.5 and private")
```







ROC curve와 AUC를 통해서도 확인해볼 수 있다. ROC 곡선이 가운데 대각선으로부터 멀리 떨어져 있을

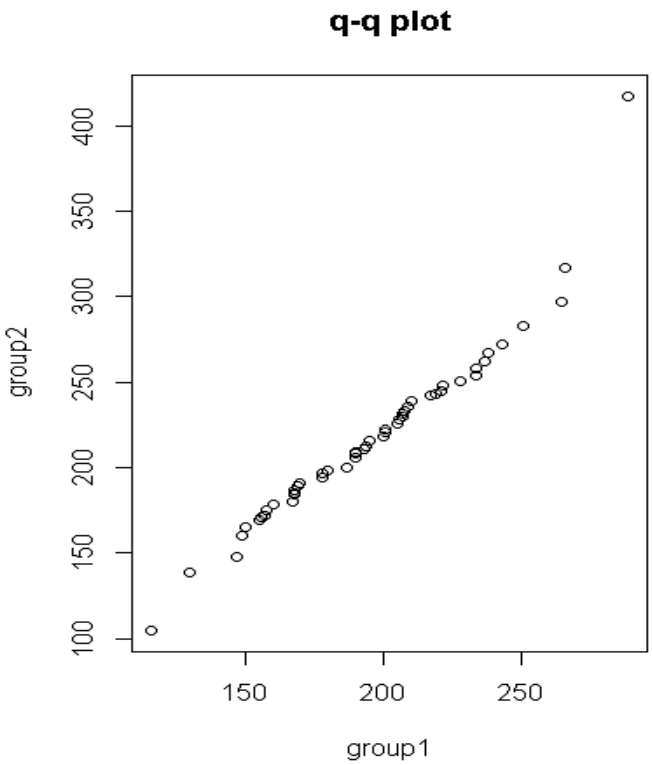
수록 두 데이터의 분포는 다르다는 결론을 내릴 수 있다. 왼쪽의 그림은 스케일링을 하지 않은 그대로의 monopoly와 private의 ROC 곡선이다. 오른쪽은 monopoly에 0.5를 더해준 후에 ROC 곡선을 그린 것이다. 이를 auc() 함수를 실행하여 수치적으로도 확인해보았다. AUC(Area Under Curve)라고 불리는 통계량은 ROC 곡선의 아래의 면적을 나타내는 수치이다. 이 값이 1에 가까울수록 두 데이터의 분포는 다르다고 말할 수 있으며 0.5에 가까울수록 두 데이터의 분포는 같다고 말할 수 있다. 왼쪽은 0.8822, 오른쪽은 0.5325로서 0.5를 더해준 이후의 두 데이터의 분포는 거의 같음을 알 수 있다. 결론적으로, monopoly와 private 상점에서의 위스키 가격의 분포는 유사하지만 private의 자료가 monopoly보다 0.5정도 크다는 사실을 알 수 있다.

3. [BLOODFAT.DAT] 혈액 속의 지방과 심장병의 관련성을 연구하기 위한 자료이다. 혈관이 좁아진 사람들(2그룹; 320명)과 그렇지 않은 사람들(1그룹; 51명)의 혈중 지방 분포가 같은지 분석하여라. 그룹의 구분은 빈칸으로 되어있다. 콜레스테롤은 홀수 컬럼, 트라이글리세라이드(triglyceride)(지방의 일종)는 짝수 컬럼이다. 메모장, 아래한글, 엑셀 등으로 그룹과 지방의 종류가 구분될 수 있도록 편집한 후 R로 읽던지, 처음부터 R로 읽은 후 조합하여라. 1그룹과 2그룹의 콜레스테롤의 분포가 같은지 분석하여라.

```
D:/2022-1(3-2)/2022-01_탐자분/Data/
> setwd("D:\\2022-1(3-2)\\2022-01_탐자분\\Data")
> data <- read.table("BLOODFAT.DAT", header=FALSE, fill = TRUE)
>
> data = data[,c("V1","V3","V5","V7","V9","V11","V13")]
>
> group1 <- data[0:8,c("V1","V3","V5","V7","V9","V11","V13")]
> group1 <- group1[!is.na(group1)]
> length(group1)
[1] 51
>
> group2 <- data[9:54,c("V1","V3","V5","V7","V9","V11","V13")]
> group2 <- group2[!is.na(group2)]
> length(group2)
[1] 320
>
```

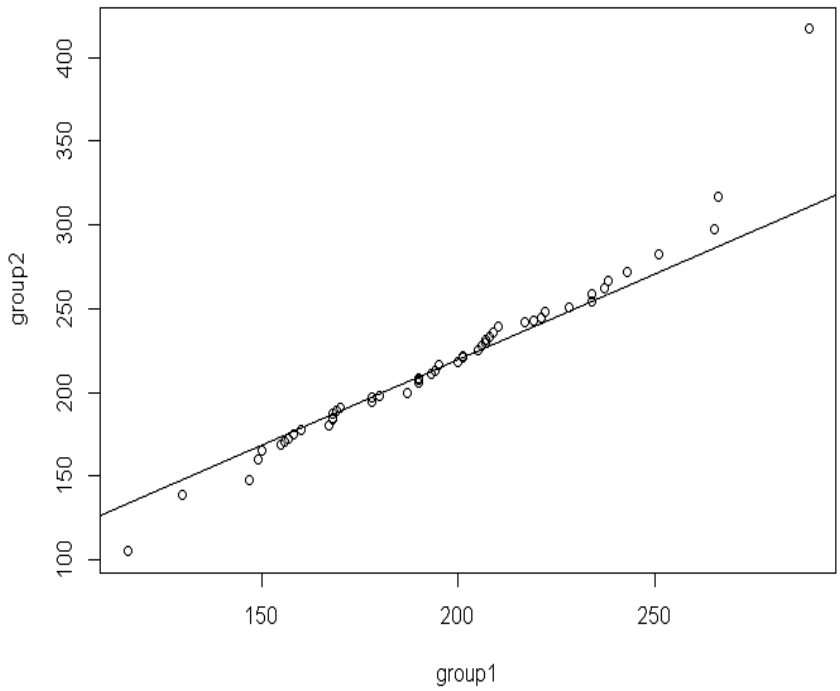
상단의 코드를 실행하여 데이터 전처리를 완료해주었다.

qqplot(group1, group2, main="q-q plot")



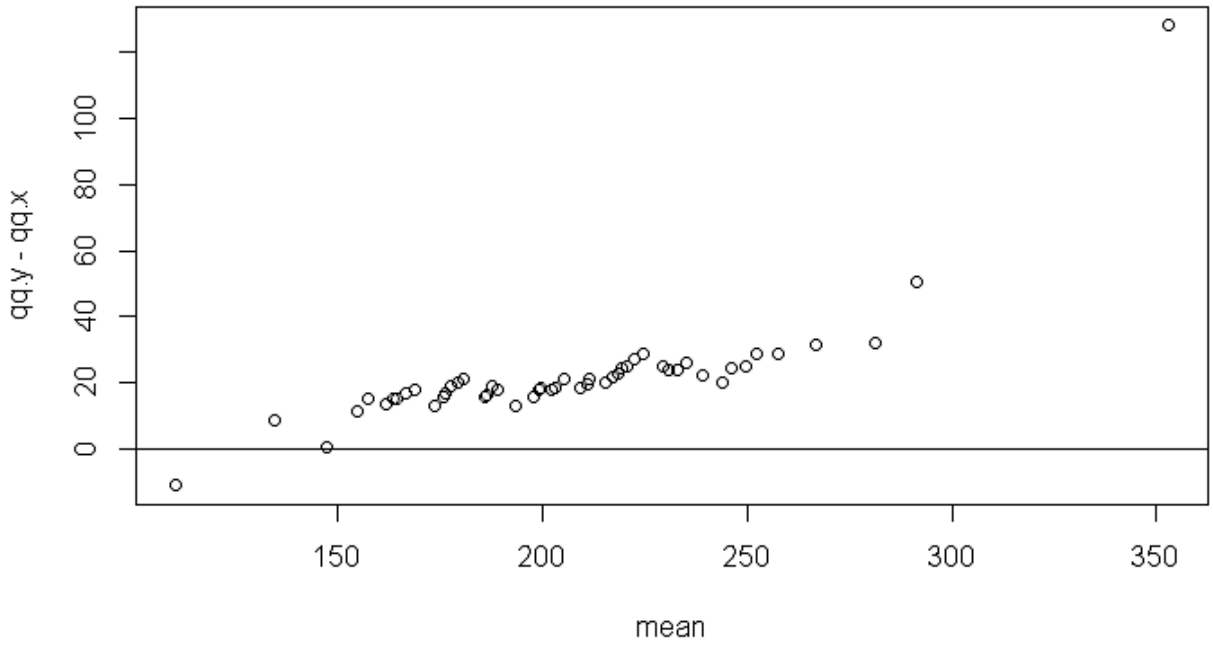
상단의 코드를 실행한 뒤, 아래의 코드를 추가로 실행하여 line을 그려보았다.

```
> qqplot(group1, group2, main="q-q plot")
> line(qqplot(group1,group2))

Call:
line(qqplot(group1, group2))

Coefficients:
[1]  14.208   1.027

> abline(line(qqplot(group1,group2)))
```



기울기는 1.027정도 그리고 절편은 14.208 정도임을 알 수 있었다. 전체적으로 봤을 때 값들이 직선을 이루고 있기 때문에 두 그룹의 분포는 같은 모양이라고 말할 수 있다. 즉 혈관이 좁아진 사람들(2그룹; 320명)과 그렇지 않은 사람들(1그룹; 51명)의 분포는 같은 모양이라고 말할 수 있다.

```
#tukey mean-diff
qq.x = qqplot(group1,group2)$x
qq.y = qqplot(group1,group2)$y
plot((qq.x+qq.y)/2, qq.y-qq.x, xlab="mean")
abline(0,0)
```



코드를 실행한 결과 상단의 plot을 확인할 수 있었다. 점들이 수평으로 직선을 이룰 때가 가장 이상적이다. 그러나, 상단의 결과에서 점들은 수평으로 직선을 이루지만 0이 아닌 20에서 직선을 이룬다.
즉 두 그룹에서 평균적으로 20의 차이가 발생한다는 사실을 알 수 있다. 혈관이 좁아진 사람들(2그룹; 320명)이 그렇지 않은 사람들(1그룹; 51명)에 비해 콜레스테롤이 20 정도 높다는 사실을 알 수 있다.

다음으로, Kolmogorov-Smirnov test을 수행해보았다.

귀무가설은 "두 그룹의 CDF는 같다."이며 양측검정(two.sided)이다. 결과적으로 p값이 0.009185이 나와 유의수준 0.05를 기준으로 귀무가설을 기각했다. 즉 두 분포의 CDF는 다르다.

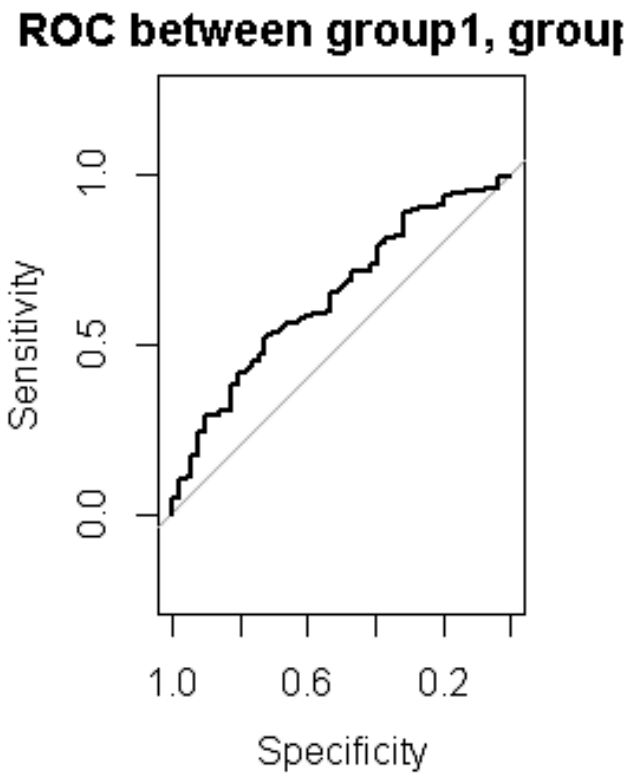
```
> ks.test(group1, group2, alternative="two.sided")

        Two-sample Kolmogorov-Smirnov test

data:  group1 and group2
D = 0.24737, p-value = 0.009185
alternative hypothesis: two-sided
```

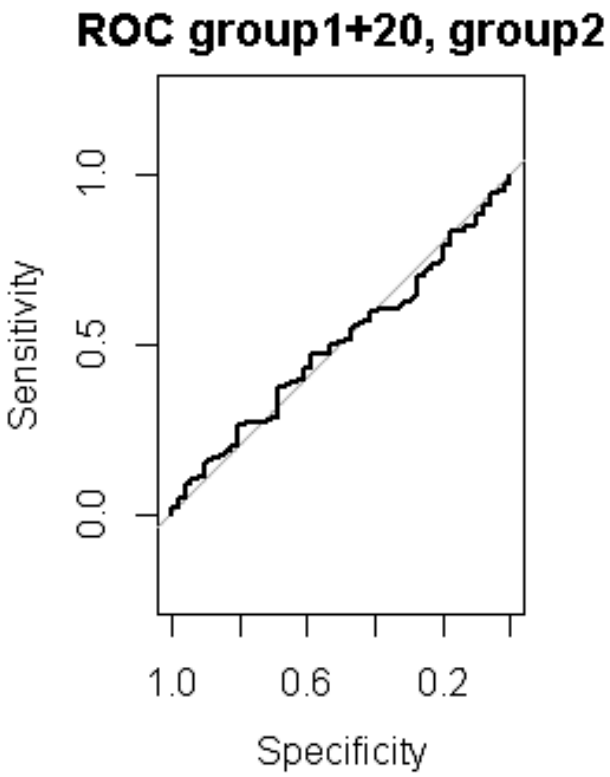
다음으로에는 1그룹에 20를 더하여 KS검정을 수행했다. p값은 0.953으로 귀무가설을 기각할 수 없다. 즉 두 분포의 CDF는 같다고 말할 수 있다.

```
> ks.test(group1+20, group2, alternative="two.sided")

        Two-sample Kolmogorov-Smirnov test

data:  group1 + 20 and group2
D = 0.077757, p-value = 0.953
alternative hypothesis: two-sided
```

결론적으로, 두 그룹의 콜레스테롤 분포 모양은 동일하지만 혈관이 좁아진 사람들(2그룹)의 데이터가 1그룹에 비해 20 정도 오른쪽으로 이동해있다고 해석할 수 있다.

```
library(pROC)
ylab=factor(c(rep(0,length(group1)), rep(1,length(group2))))
res = roc(ylab, c(group1, group2))
res.ad = roc(ylab, c(group1+20, group2))

par(mfrow=c(1,2))
plot(res, xlim=c(1,0), main="ROC between group1, group2")
plot(res.ad, xlim=c(1,0), main="ROC group1+20, group2")

auc(res)
auc(res.ad)
```





```
> auc(res)
Area under the curve: 0.6452
> auc(res.ad)
Area under the curve: 0.5007
>
```

ROC curve와 AUC를 통해서도 확인해보았다. ROC 곡선이 가운데 대각선으로부터 멀리 떨어져 있을수록 두 데이터의 분포는 다르다는 결론을 내릴 수 있다. 왼쪽의 그림은 스케일링을 하지 않은 그대로의 두 그룹의 ROC 곡선이다. 오른쪽은 1그룹에 20을 더해준 후에 ROC 곡선을 그린 것이다. 이를 auc() 함수를 실행하여 수치적으로도 확인해보았다. AUC(Area Under Curve)라고 불리는 통계량은 ROC 곡선의 아래의 면적을 나타내는 수치이다. 이 값이 1에 가까울수록 두 데이터의 분포는 다르다고 말할 수 있으며 0.5에 가까울수록 두 데이터의 분포는 같다고 말할 수 있다. 왼쪽은 0.6452, 오른쪽은 0.5007로서 20을 더해준 이후의 두 데이터의 분포는 거의 같음을 알 수 있다. 결론적으로, 두 그룹의 콜레스테롤 분포는 유사하지만 혈관이 좁아진 사람들(2그룹)의 콜레스테롤 데이터가 1그룹에 비해 20 정도 크다는 사실을 알 수 있다.