EDA 4장 숙제

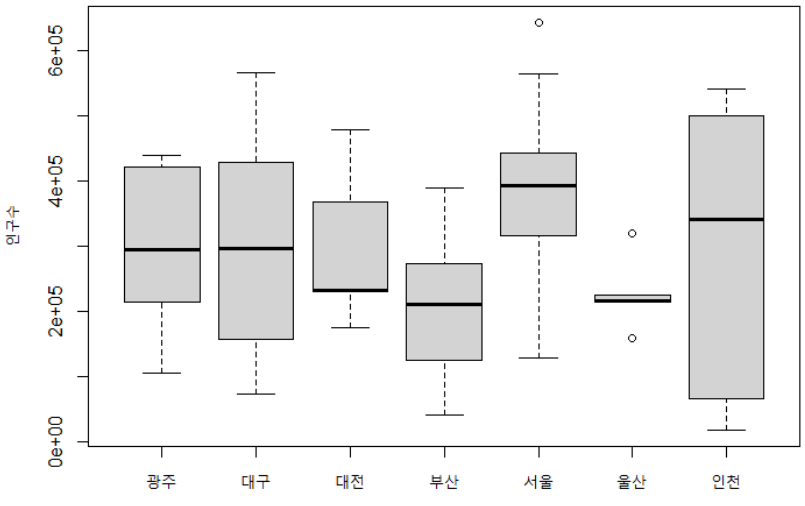
2017122007 정헌영

(1)

C: 2015년 주택인구총조사 자료

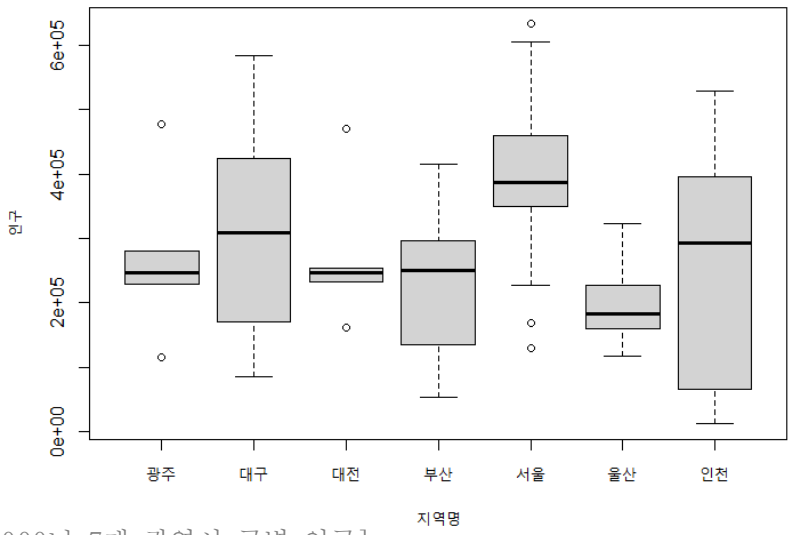
```
> C
      시 인구수
1   서울 151291        20 서울 392772        39 부산 171461
2   서울 128744        21 서울 502641        40 부산 216350
3   서울 225882        22 서울 401749        41 부산 169465        58 인천  66020
4   서울 291918        23 서울 508135        42 대구  74278        59 인천  19292
5   서울 353967        24 서울 643288        43 대구 333294        60 광주 106005
6   서울 351057        25 서울 444434        44 대구 170437        61 광주 294854
7   서울 385663        26 부산  41439        45 대구 145588        62 광주 214860
8   서울 438833        27 부산 105303        46 대구 446466        63 광주 439352
9   서울 299535        28 부산  87246        47 대구 412912        64 광주 422502
10  서울 315979        29 부산 113224        48 대구 566712        65 대전 232559
11  서울 511982        30 부산 351403        49 대구 261013        66 대전 231959
12  서울 463102        31 부산 263345        50 인천 138586        67 대전 478629
13  서울 317209        32 부산 269111        51 인천  61285        68 대전 368895
14  서울 365612        33 부산 280177        52 인천 390260        69 대전 176393
15  서울 439068        34 부산 389535        53 인천 528927        70 울산 214147
16  서울 564854        35 부산 312057        54 인천 500812        71 울산 320175
17  서울 435442        36 부산 237219        55 인천 292852        72 울산 159000
18  서울 249419        37 부산 136734        56 인천 541534        73 울산 217051
19  서울 403619        38 부산 204947        57 인천 405886        74 울산 225050
```

```
boxplot(인구수 ~ 시, data = C)
```



[2015년 7개 광역시 구별 인구]

(81쪽의 상자그림)



[2000년 7개 광역시 구별 인구]

변동 추이 분석:

중앙값을 보면 광주,서울,인천,울산은 증가했고 대구,대전은 감소했으며 부산은 거의 변화가 없다.
막대의 길이를 보면 광주,대전이 산포도가 크게 증가했고, 부산, 울산은 산포도가 크게 줄었다.
또한 대전과 울산은 Median이 상자 밑부분으로 크게 내려갔는데 이는 평균보다 적은 인구수의 시가 많아졌음을 의미한다.

skewness 계산:

```
(*C2: 2000년 주택인구총조사 자료)
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)))

}

s15 <- skew(C[(1:25),2])
s00 <- skew(C2[(1:25),2])
b15 <- skew(C[(26:41),2])
b00 <- skew(C2[(26:41),2])

matrix(nrow=2, c(s15, s00, b15, b00),
       dimnames = list(c("2015", "2000"), c("서울", "부산")))
```

```
#           서울       부산
#2015 -0.1956405 -0.1316329
#2000  0.2990875 -0.4183731
```

[서울과 부산의 연도별 skewness]

skewness 비교:

서울은 skewness가 2000년 양수에서 2015년 음수로 전환했다. 즉 2000년 왼쪽에 시가 많이 분포했으나 2015년은 인구수가 많은 오른쪽에 시가 많이 분포한다.
부산은 skewness가 둘 다 음수였으나 절댓값은 2015년 더 낮았다. 즉 2000년과 2015년 모두 오른쪽에 더 시가 많이 분포하지만 2015년엔 그 정도가 줄어들었다.
2015년 서울과 부산은 모두 skewness가 음수이나 서울의 절댓값이 더 크다. 즉 양쪽 모두 인구수가 많은 오른쪽에 시가 많이 분포했으나 정도는 서울이 더 높았다.
2000년 서울은 skewness가 양수이나 부산은 음수이다. 즉 서울은 인구수가 많은 오른쪽에 시가 더 많이 분포하고 부산은 왼쪽에 더 많이 분포한다.

kurtosis 계산:

```
kurto <- function(x) {
  E_L = quantile(x, 0.125)
  E_U = quantile(x, 0.875)
  H_L = quantile(x, 0.25)
  H_U = quantile(x, 0.75)
  print((E_U-E_L)/(H_U-H_L)-1.705)
}
s15k <- kurto(C[(1:25),2])
s00k <- kurto(C2[(1:25),2])
b15k <- kurto(C[(26:41),2])
b00k <- kurto(C2[(26:41),2])
matrix(nrow=2, c(s15k, s00k, b15k, b00k),
       dimnames = list(c("2015", "2000"), c("서울", "부산")))
```

```
#           서울       부산
#2015 0.3090594 -0.1879963
#2000 0.4592568  0.3235200
```

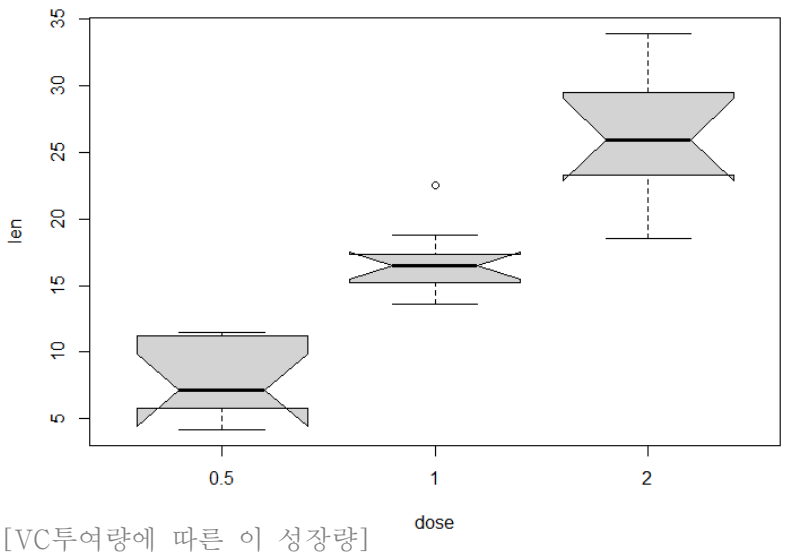
[서울과 부산의 연도별 kurtosis]

kurtosis 비교:

서울과 부산 모두 2015년이 2000년에 비해 첨도가 줄었다. 즉 2015년이 중앙과 꼬리 부분에 도시가 적게 분포한다.
부산의 2015년 첨도가 음수인것은 정규분포보다 중앙과 꼬리부분에 도시가 적게 분포함을 의미한다.
2015년과 2000년 모두 서울이 부산보다 첨도가 높다. 즉 서울이 부산보다 중앙과 꼬리 부분에 자료가 많이 분포한다.

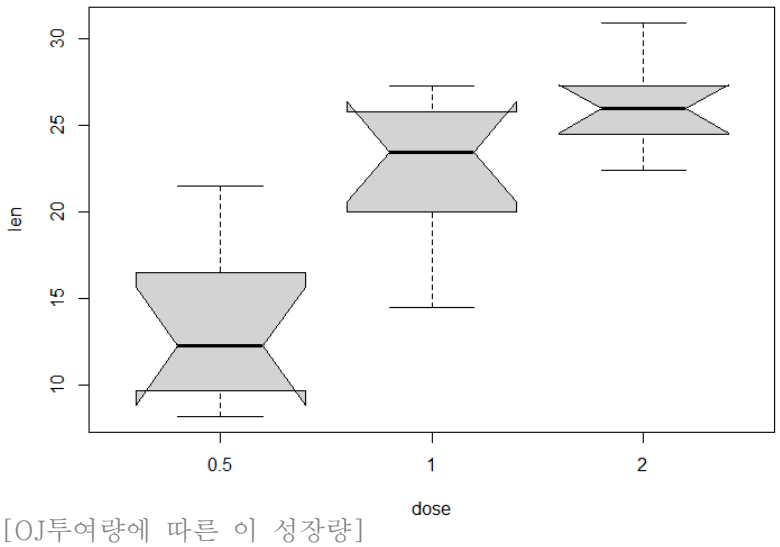
(2)

boxplot(len~dose, data=ToothGrowth[(1:30),], notch=TRUE)



[VC투여량에 따른 이 성장량]

boxplot(len~dose, data=ToothGrowth[(31:60),], notch=TRUE)



[OJ투여량에 따른 이 성장량]

분석:

각각의 box들은 dose가 0.5, 1, 2로 증가할 때마다 median이 증가한다. 이때 VC투여의 경우 각각은 notch의 범위가 겹치지 않는다는 점을 확인할 수 있다. 이는 각 median이 5% 수준에서 significantly different하다는 것을 의미한다. 즉, VC의 경우 이의 성장을 촉진시키는 비타민C의 영향이 있는 것으로 보인다.

OJ투여의 경우 dose가 0.5에서 1로 증가할 때는 notch의 범위가 겹치지 않으나, 1에서 2로 증가할 때에는 notch의 범위가 상당 부분 겹친다. 즉 dose가 0.5에서 1일때까지만 5% 수준에서 significantly different 하고 이는 OJ의 경우 dose를 1로 늘릴 때 까지는 이의 성장을 촉진시키는 비타민C의 영향이 있는 것으로 볼 수 있으나 그 이상은 영향이 있다고 볼 수 없음을 의미한다.

(3)

summary(Nile)

```
#    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#   456.0   798.5   893.5   919.4  1032.5  1370.0
```

fivenum(Nile)

```
# [1]  456.0  798.0  893.5 1035.0 1370.0
```

summary()와 fivenum()의 차이를 보면
우선 summary()는 Mean값을 출력하지만 fivenum()은 출력되지 않으며, summary()는 각 데이터의 이름이 Min. 1st Qu. Median Mean 3rd Qu. Max. 로 윗줄에 표시된다.
또한 summary()는 fivenum()과 1사분위수, 3사분위수의 값이 0.5씩의 차이를 갖는다.
이는 분위수를 계산할 때 소수점 번째의 자료를 구하는 방법에 차이가 있기 때문일 것으로 생각된다.

(4)

# 100			Mid	Spread
M 50h	893.5		893.5	
H 25h	797.83	1035.83	916.83	238
E 13	741.75	1151.25	946.5	409.5
D 7	699.81	1210	954.91	510.19
C 4	683.5	1240.63	962.07	557.13
B 2h	629.9	1270.89	950.40	640.99
A 1h	478.62	1357.11	917.87	878.49
Z 1	456	1370	913	914

C까지 Mid값이 커지므로 왼쪽에 데이터가 많이 분포하고 오른쪽으로 꼬리가 긴 분포이다. (skew to the right)

또한 첨도는 $\frac{1151.25-741.75}{1035.83-797.83} - 1.704 = \frac{238}{409.5} - 1.704 = -1.12$ 이다.

이는 정규분포보다 중앙과 꼬리부분에 데이터가 많이 존재하는 분포임을 의미한다.