EDA1LDS

DSL

March 21, 2019

# 탐색적자료분석 1차과제

## 1번 Ashwan에서 측정한 Nile 강의 유량 자료 (R datasets:: Nile) 줄기그림을 그리고 자료에 대하여 설명하여라. depth 등이 들어간 ‘완벽한’ 줄기그림을 손으로 작성하여라. R이 만들어 준 부분을 오려 붙이고 나머지 연필로 완성하여도 된다.

summary(Nile)

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

stem(Nile)

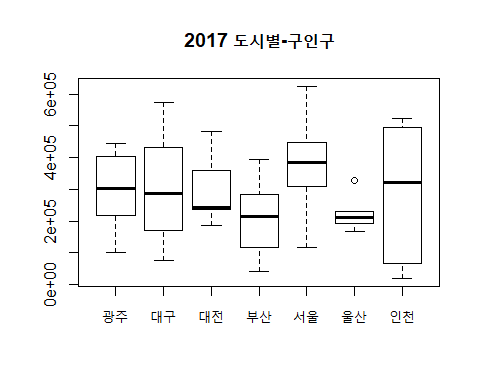
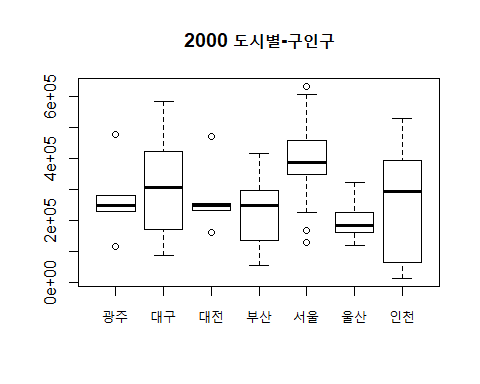
##   
## The decimal point is 2 digit(s) to the right of the |  
##   
## 4 | 6  
## 5 |   
## 6 | 5899  
## 7 | 000123444455667778  
## 8 | 000011222233344555556667779  
## 9 | 0011222244466678899  
## 10 | 0122234455  
## 11 | 00012244566678  
## 12 | 112356  
## 13 | 7

강 유량의 범위는 400~1300으로 퍼져 있으며, 유량 800대로 센터링 되어있으며, 중앙값은 893.5로 positive skewness를 보이고 있다. 줄기 개수를 수정한다면 클러스터가 두개 있다고 볼 수 있다. 이때 나타나는 800대가 아닌 다른 유량 집중 구간은 1100대이다. 유량이 600대가 드물고, 500대는 완전히 없지만 400대에 한 번 관측된 것으로 보아 이를 irregularity로 볼 수 있다,

### ２번 줄기그림과 히스토그램의 차이, 장단점 등을 간단히 표로 작성하여라. 교과서 등 참조

|  |  |  |
| --- | --- | --- |
|  | 줄기그림 | 히스토그램 |
| 공통점1 | 관측 빈도에 비례하는 길이의 막대 기둥이 그려진다 |  |
| 공통점2 | 따라서 자료의 분포를 눈으로 확인할 수 있다 |  |
| 차이점1 | 자료들이 숫자로 나타나진다 | 상자가 쌓인 형태로 표현된다 |
| 차이점2 | 중간값 계산이 가능하다 | 중간값 계산이 어렵다 |
| 차이점3 | 원자료와 무관하게 기존의 줄기그림을 활용할 수 있다 | 변형을 위해서는 원자료를 사용해야한다. |
| 차이점4 | 구간 폭이 1/2/5 \* 10^정수로 고정적으로 사용하여야 한다 | 구간 폭을 임의로 지정하여 사용할 수 있다. |

### ３ 인터넷 검색 또는 출판 자료에서 가장 최근의 인구주택총조사 자료를 구하여 81쪽의 상자그림을 그리고, 81쪽의 상자 그림과 비교하여 구별 인구의 변동 추이를 설명하여라.



구별 인구의 변동 추이를 살펴보면 다음과 같다. 이때, 중앙값의 위치 변화로 positive skewness가 negative skewness로 변하였다면 이는 인구가 특정 구로 몰렸음을 뜻하고, spread와 whisker의 증가 또한 인구 몰림을 설명한다.

* 광주의 경우 중앙값은 늘었고, spread가 커졌으며, whisker가 길어졌다. 아웃라이어는 사라졌으며, positive skew에서 negative skew로 변하였다. 이는 전체적으로 광주 인구의 증가를 뜻한다. 또, 인구가 과거에는 구별로 고루 퍼져있던 것에 비해 특정 지역으로 인구가 몰리는 현상임을 뜻한다. 다만, 2000년에는 한 구에 극단적으로 인구가 집중되어 있던 구가 있었던 것에 비해 지금은 상대적으로 여러 구에 인구가 집중되고 있음을 알 수 있다.
* 대구, 부산의 경우 2000년과 2017년 구별 인구수의 분포에 크게 변화가 없다. 다만 중앙값이 조금 내려간 것과 spread가 커진 것으로 보아 경미한 수준의 인구 몰림 현상이 나타났다.
* 대전의 경우 중앙값이 내려갔고, negative skew에서 positive skew로 변하였으며, spread가 커졌다. 또, whisker길이 또한 길어졌다. 이런 점들은 대전의 경우 특정 구의 인구 몰림이 심해졌음을 보여준다.
* 서울의 경우 중앙값은 경미하게 내려갔으며 negative skewness로 변화했다.또, spread와 whisker가 증가했다.
* 울산의 경우 타 구역과 다르게 spread가 줄었으며 whisker가 짧아졌고, outlier 하나를 제외하면 상대적으로 symmetric해졌다. 이는 구별 인구수가 2000년에 비해 고루 분산됐음을 알 수 있다. 다만 outlier값 한 구의 인구는 여전히 높이 유지된다.
* 인천의 경우 상당히 유사하지만 중앙값이 조금 올랐고, spread가 증가했다. 이는 큰 변화는 없었으나 인구 몰림 현상이 나타났다고 이해할 수 있다.

종합해보면, 울산을 제외한 6개 시에서 인구 몰림 현상이 나타났음을 상자그림 분석을 통해 알 수 있다.

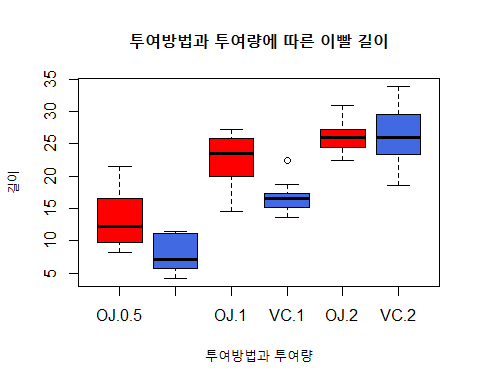
### ４ (R datasets:: ToothGrowth) 이의 성장에 비타민 C의 영향이 있는지 상자그림으로 분석하여라.

summary(ToothGrowth)

## len supp dose   
## Min. : 4.20 OJ:30 Min. :0.500   
## 1st Qu.:13.07 VC:30 1st Qu.:0.500   
## Median :19.25 Median :1.000   
## Mean :18.81 Mean :1.167   
## 3rd Qu.:25.27 3rd Qu.:2.000   
## Max. :33.90 Max. :2.000

우선, 다루게 될 ToothGrowth 데이터의 개수는 총 60개로, len은 기니피그의 이빨 길이, supp은 비타민C 투여방법으로 OJ는 orange juice를 통한 비타민c 투여방법, VC는 약물을 통한 비타민c 투여방법이다. Dose는 비타민 c 투여량으로 총 0.5, 1, 2 씩 투여하는 실험군이 있다.

boxplot(len ~ supp\*dose, data=ToothGrowth, col=(c("red","royalblue")),ylab="길이", xlab="투여방법과 투여량", main="투여방법과 투여량에 따른 이빨 길이")

 투여방법과 투여량 둘 모두를 고려하여 상자그림을 그려 분석하는 것이 가능하다. 이 그림에서 보면 비타민 투여양이 0.5와 1일때, 약물을 통해 비타민 C를 섭취한 기니피그(VC)의 이빨길이 분포가 좀 더 precise 하지만, 오렌지 주스를 통해 비타민 C를 섭취한 기니피그(OJ)의 이빨길이보다 짧게 나타난다. 그러나, 비타민 투여양이 2일때는 OJ와 VC 중앙값이 거의 일치하게 나타나는 반면, OJ가 더 precise하게 나타난다. 이는 비타민C 투여량이 2가 되기 전까지는 오렌지주스에 있는 다른 요인이 이빨길이 성장에 영향을 주지만, 투여량이 2가 되면 이빨길이에 영향을 주는 타 요인이 있다 하더라도 큰 영향을 못 끼친다고 해석할 수 있다.

### ５ R에서 summary()와 fivenum()의 차이를 알아 보아라.

fivenum(Nile)

## [1] 456.0 798.0 893.5 1035.0 1370.0

summary(Nile)

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

R에서 fivenum(x, na.rm=T) 명령어는 결측치를 포함하는 숫자형 자료에 대해 투키의 다섯수치요약인 최소값, lower Hinge, 중앙값, upper Hinge, 최대값을 구해준다. 이때 결측치는 default 조건인 na.rm=T로 인해 제거된다. 결과값은 특별한 설명 없이 숫자 벡터로 반환된다. R에서 summary() 명령어는 다섯수치요약에 있는 최소, 최대, 중앙값을 반환해주는 동시에 다섯수치요약과는 달리 평균과 1사분위값, 3사주분위값을 반환해준다. 추가적으로 summary 함수는 숫자형이 아닌 문자열의 경우 개수를 세어 명시해준다. summary() 함수는 fivenum()과 같이 결측치를 제하고 결과값을 알려주며, 결측치의 개수를 따로 명시해주고 있다.