EDA 2,3장 숙제

2016122043 김호성

2020 3 31

# 1. EDA의 네 가지 주제에 대하여 설명하라. 한 주제당 3줄 이하.

### 1) 저항성

EDA에서의 자료분석은 강건해야한다. 저항성을 가져야 한다는 의미이다. 이는 데이터의 파손, 결측, 표기 오류, 특이값 등의 영향을 적게 받도록 한다는 말이다. 예를 들어 대표값으로 평균보다 중위값을 사용하는 것이 자료의 저항성을 키울 수 있는 EDA의 방법일 것이다.

### 2) 잔차

잔차는 개별 자료의 관측값과 전체 데이터의 경향으로 예측한 값의 차이를 말한다. EDA에서는 각 관측갑의 잔차에 대한 의문을 풀도록 노력해야 한다.

### 3) 재표현

재표현은 변수를 변환하는 과정이다. 실제 있는 자료의 변수들의의 관계가 쉽게 눈에 보이지 않는 경우에도 변수들을 역수, 로그, 지수 등을 취하여 변환할 경우 변수들의 관계가 더 쉽게 보일 수도 있기 때문이다. 따라서 적절한 변수의 변환, 재표현을 통해 자료의 변수들간의 관계를 가장 잘 나타내는 표현방법을 찾아야 한다.

### 4) 그래픽 표현

그래픽 표현은 그래프를 그리는 것을 말한다. 수로 보아서는 안 보이던 정보들도 단순한 그림으로 그려봤을 때 숨겨진 정보를 찾아 낼 수도 있다. 이에 따라 줄기 그림등의 방법을 동원해 최적의 그래픽 표현을 찾아야 한다.

# 2. Ashwan에서 측정한 Nile 강의 유량 자료 (R datasets:: Nile) 줄기잎그림을 그리고 자료에 대하여 설명하여라.

우선 크게 두가지 방법이 있다. 5 num ssummary를 통해 수업시간에 배운 계산을 활요해 spread, skew, flatness 등을 구하는 방법이 있고 r에서 제공하는 함수를 통해 이를 구하는 방법이 있다. 두 가지를 모두 해보겠다.

Nile

## Time Series:  
## Start = 1871   
## End = 1970   
## Frequency = 1   
## [1] 1120 1160 963 1210 1160 1160 813 1230 1370 1140 995 935 1110 994 1020  
## [16] 960 1180 799 958 1140 1100 1210 1150 1250 1260 1220 1030 1100 774 840  
## [31] 874 694 940 833 701 916 692 1020 1050 969 831 726 456 824 702  
## [46] 1120 1100 832 764 821 768 845 864 862 698 845 744 796 1040 759  
## [61] 781 865 845 944 984 897 822 1010 771 676 649 846 812 742 801  
## [76] 1040 860 874 848 890 744 749 838 1050 918 986 797 923 975 815  
## [91] 1020 906 901 1170 912 746 919 718 714 740

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

lsum(Nile)

## letter depth lower mid upper spread  
## 1 M 50.5 893.5 893.5 893.5 0  
## 2 H 25.5 798.0 916.5 1035.0 237  
## 3 E 13.0 742.0 946.0 1150.0 408  
## 4 D 7.0 701.0 955.5 1210.0 509  
## 5 C 4.0 692.0 961.0 1230.0 538

skewness(Nile)

## [1] 0.3175462  
## attr(,"method")  
## [1] "moment"

kurtosis(Nile)

## [1] -0.3585392  
## attr(,"method")  
## [1] "excess"

먼저 Nile 자료를 가볍게 살펴보면 이것이 시계열 자료이며 1871년부터 1970년 까지 1년 간격으로 기록된 수치임을 알 수 있다.  
줄기잎그림을 봤을 때 백의 자리를 줄기로 선택하였고 10의 자리를 잎으로 그린 것을 알 수 있다.(1의자리에서 반올림) 총 자료 관측갑의 개수는 100개이며 그림에 따른 최소값은 460근처, 최대값은 1370근처이다. 기본 설정된 줄기개수의 경우, 줄기잎그림은 이봉분포의 모습을 볼 수 있다. 눈으로 보아서는 이봉분포를 따라 두 개 정도의 clusters 를 형성하는 듯 보인다. Center는 중위값인 893.5일 것이고 히스토그램 기준으로 살짝 우측, 줄기잎그림 기준으로 살짝 밑쪽으로 치우쳐 보인다. .이는 lsum함수의 mid값의 변화를 보면서 확실하게 확인할 수 있다. skewness함수로 이를 확인할 수도 있다.Kurtosis의 경우, 흔히 정규분포의 것과 빅교하는데 Nile자료의 첨도가 정규분포의 그것보다는 조금 더 크므로, 정규분포보다는 더 뾰족하고 특이값이 많다고 볼 수 있다. 수업시간에 배운대로 E-spred/H-spread (약 1.72)로 해봐도 정규분포의 그것(1.705)보다는 크다. 특징적으로 눈에 띄는 것은 최소값인 460이 다른 값들과는 상당히 동떨어진 거리에 있어 특이값일 확률이 높다는 점이다.

# 3. R의 warpbreaks 자료를 이용하여 wool A, B 각기 따로 breaks의 stem and leaf display 그리고 분포의 특징에 대하여 비교 설명하여라.

subA=warpbreaks %>% filter(wool=='A')%>% select(breaks)# dplyr 활용해서 warpbreaks에서 wool A의 breaks자료 분리 한것  
stem(subA$breaks)

##   
## The decimal point is 1 digit(s) to the right of the |  
##   
## 1 | 0257888  
## 2 | 114566689  
## 3 | 00566  
## 4 | 3  
## 5 | 124  
## 6 | 7  
## 7 | 0

subB=warpbreaks %>% filter(wool=='B')%>% select(breaks)# 마찬가지로 wool B자료 분리 한 것  
stem(subB$breaks)

##   
## The decimal point is 1 digit(s) to the right of the |  
##   
## 1 | 34  
## 1 | 5566799  
## 2 | 00114  
## 2 | 6788999  
## 3 | 1  
## 3 | 99  
## 4 | 124

자동으로 그리면 두 그림의 줄기 갯수가 다르다. A의 경우 10단위로 줄기 하나씩 그려진 반면 B는 10단위에 줄기가 두 개씩, 즉 5단위로 하나씩 그려진 것으로 볼 수 있다. 그리고 그냥 봤을 때는 B는 잘 모르겠지만 A는 오른쪽으로 치우친 듯 보인다. . 좀 더 보기 편하게 하기 위해 A자료를 줄기를 두배로 늘려서 보겠다

stem(subA$breaks, 2)

##   
## The decimal point is 1 digit(s) to the right of the |  
##   
## 1 | 02  
## 1 | 57888  
## 2 | 114  
## 2 | 566689  
## 3 | 00  
## 3 | 566  
## 4 | 3  
## 4 |   
## 5 | 124  
## 5 |   
## 6 |   
## 6 | 7  
## 7 | 0

stem(subB$breaks)

##   
## The decimal point is 1 digit(s) to the right of the |  
##   
## 1 | 34  
## 1 | 5566799  
## 2 | 00114  
## 2 | 6788999  
## 3 | 1  
## 3 | 99  
## 4 | 124

이렇게 보니 두 자료의 중심은 20 중반 근처로 비슷해 보이지만, A자료가 수치가 큰 값들이 많아 더 극단적으로 오른쪽으로 치우친 것을 알 수 있다. 실제 여러 수치를 확인해보자

median(subA$breaks)

## [1] 26

median(subB$breaks)

## [1] 24

skewness(subA$breaks)

## [1] 0.9689172  
## attr(,"method")  
## [1] "moment"

skewness(subB$breaks)

## [1] 0.5632993  
## attr(,"method")  
## [1] "moment"

kurtosis(subA$breaks)

## [1] 0.03535152  
## attr(,"method")  
## [1] "excess"

kurtosis(subB$breaks)

## [1] -0.9124127  
## attr(,"method")  
## [1] "excess"

중위값은 A,B 각각 24, 26으로 별 차이 없다. B는 살짝 치우친 반면 A는 상당히 많이 오른쪽으로 치우쳐있다. 첨도에서는 A가 B보다 훨씬 뾰족한 것으로 나온다. 특이한 점은 r에서 제공하는 skewness와 수업에서 배운 fivenum을 이용해서 구한 skewness가 받대방향인 경우가 있다는 것이다. 위의 r skewness 함수에서는 B도 약간이나마 오른쪽으로 치우친 것으로 나오지만 fivenum을 이용해서 배운대로 EDA를 진행하여 skewness를 구하면 {(Upper hinge- med)-(med-lower hinge)}/ {(Upper hinge-med)=(med-lower hinge)}의 경우 약 -0.1가량으로 미묘하게 약간 왼쪽으로 치우친 것으로 나온다. 이는 사실상 왜도가 0에 가까운 wool B 자료 특성상 약간의 오류가 난 것으로 보인다. 결론적으로 B는 거의 대칭인 듯 하다.

fivenum(subB$breaks)

## [1] 13 18 24 29 44

# R에서 Seatbelt 파일에 대한 설명은 ‘?Seatbelts’ 하면 알 수 있다.

### (1) 운전자 1000명당, 운행거리 10000km 당 사망운전사수(killed)를 계산하여라.

Seatbelt는 69년 부터 84년까지 매달 영국 운전자, 벨트 착용여부, 사망자 수 dnsgodrjfl등을 기록해놓은 시계열 자료이다. 원하는 자료를 구하기 위해 행과 열을 선택하고 계산해 보겠다.

data=Seatbelts[,c(1,2,5)]   
sum(data[,1])/{sum(data[,2])/1000} # 1000명당 사망자수

## [1] 73.52065

sum(data[,1])/{sum(data[,3])/10000} # 10000km 당 사망자 수

## [1] 81.90298

### (2) 사망운전사수의 줄기 그림을 그리고 간단히 서술하여라.

stem(data[,1])

##   
## The decimal point is 1 digit(s) to the right of the |  
##   
## 6 | 0  
## 7 | 99  
## 8 | 12244467799  
## 9 | 02224455556677789  
## 10 | 000012222333334445666667788888899  
## 11 | 0000111122233344444555556777778899  
## 12 | 00001222222233555567778899  
## 13 | 001122344444466677788  
## 14 | 0000003446788  
## 15 | 01223333444667899  
## 16 | 11235889  
## 17 | 018  
## 18 | 0337  
## 19 | 08

최소값은 60 최대값은 198이며 110대 사이에서 centering을 이루고 있다. 전체적으로 크게 치우치자 않았으며 크게 보면 단봉분포이지만 150대에 한번의 peak가 더 있다고 볼 수도 있다. 만약 줄기를 조금 더 많게 한다면,살짝 오른쪽으로 치우친듯 보일 것 같고 60이 특잇값으로 보일 수도 있다.

### (3) 안전띠 법이 시행되기 전과 후의 사망운전사수의 줄기 그림을 각각 그리고 비교하여라.

안전띠 법은 169행부터 시행된 것으로 보인다. 따라서 그 행을 기준으로 위 아래를 나누어서 비교하겠다.

law0=data[c(1:169),1] # 법 시행전 사망자 수   
law1=data[c(170:192),1] # 법 시행후 사망자 수  
stem(law0)

##   
## The decimal point is 1 digit(s) to the right of the |  
##   
## 7 | 9  
## 8 | 247  
## 9 | 2244555677789  
## 10 | 00012222333334445666667788888899  
## 11 | 00001111222334444455555677777899  
## 12 | 0001222223355557778899  
## 13 | 00112234444446667788  
## 14 | 0000003446788  
## 15 | 0122333344667899  
## 16 | 11235889  
## 17 | 018  
## 18 | 0337  
## 19 | 08

stem(law1)

##   
## The decimal point is 1 digit(s) to the right of the |  
##   
## 6 | 09  
## 8 | 124467990256  
## 10 | 038  
## 12 | 02267  
## 14 | 4

fivenum(law0)

## [1] 79 108 121 140 198

fivenum(law1)

## [1] 60 85 92 119 154

kurtosis(law0)

## [1] -0.1336242  
## attr(,"method")  
## [1] "excess"

kurtosis(law1)

## [1] -0.3881812  
## attr(,"method")  
## [1] "excess"

두 그림을 보면 안전띠 법이 시행되고 사망운전자수가 줄어드는 경향을 보인다고 해석할 수 있다. 우선 두 자료의 크기가 많이 다르긴하다. 시행전은 자료가 169개인 반면 시행 후는 23개 밖에 되지 않는다. 그러나 최소값은 시행후가 더 낮고 전체적인 자료의 중심도 약 20가량 더 낮은 쪽에 있다. 두 그림 모두 오른쪽으로 살짝 치우쳐 보이지만 시행전의 그림은 그 정도가 매우 크다. 실제 최고값은 거의 50씩 차이가 난다. 5 number summary로 본 spread나 skew는 눈으로 본 것과 큰 차이가 나지 않았다. spread로 시행 후의 skew를 구하면 0.6가량으로 꽤 크게 나오기는 하지만 적은 표본수를 고려하면 그럴 수 도 있다. 함수로 구해보나 kurtosis는 A가 조금 더 뾰족하게 나왔다.

# 5. 줄기그림과 히스토그램의 차이, 장단점 등을 간단히 표로 작성하여라. 강의노트, 책, 허명회 교수의 R 프로젝트 파일 참조.

|  |  |  |
| --- | --- | --- |
| 구분 | 줄기그림 | 히스토그램 |
| 장점 | 그림만으로도 비교적 정확하게 중간값을 알 수 있다. 정보의 손실이 적다,원자료의 의존성이 적다. | 크기가 큰(개수가 많은) 자료를 그리기 쉽다, 구간의 폭을 임의로 정할 수 있다. |
| 단점 | 구간 폭이 1,2,5,등으로 제한된다, 많은 수의 자료 그리기 힘들다 | 원자료의 의존성이 높다. 졍확한 중간을 알기 어렵다. 한 상자 내부의 정보 손실이 일어난다. |