2. 데이터 분포경향 살펴보기

- 데이터 중심경향 알아보기
- 데이터 분포를 측정하는 통계치 알아보기
- 데이터 분포 시각화 histogram,boxplot 이해

I. 데이터 중심경향

➢ 수집된 데이터의 대푯값을 계산하기위한 통계치

- 평균 – 분석할 데이터의 중심위치를 측정할 때 사용(데이터 총합/데이터건수)
- 중앙값 – 평균의 단점을 보완하기 위한 측정방법으로 데이터를 크기순으로 나열하여 가장 가운데 오는 값
- 최빈수- 데이터 중 가장 많이 나타난 수(발생빈도가 높은 수)를 의미

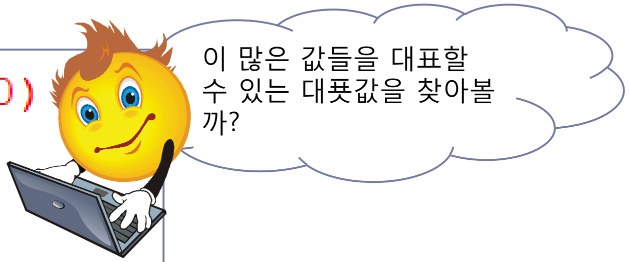
데이터가 65,75,80,80,80,85,95 가 존재할 때

평균=>(65+75+80+80+80+85+95)/7=>80

중앙값 = 정 가운뎃값 80

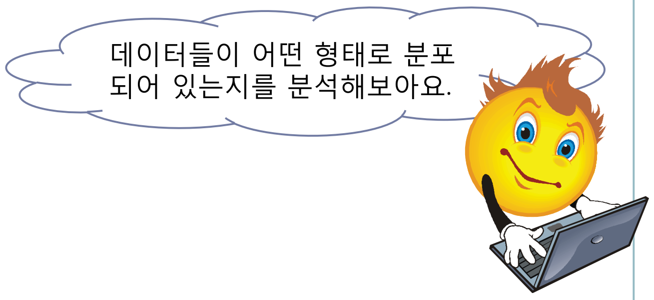
최빈수=80

```
> s_data<-c(23,32,5,17,100)
> mean(s_data)
[1] 35.4
> sort(s_data)
[1]   5  17  23  32 100
> sort_data<-sort(s_data)
> median(sort_data)
[1] 23
```

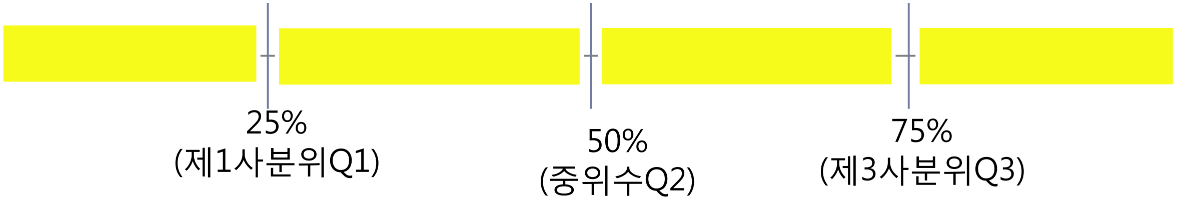


2. 데이터 분포형태 분석

➤ 전체 데이터의 분포양상을 살펴봄으로써 데이터 측정을 보다 명확하게 할 수 있음



- 산포도 – 데이터의 흩어짐(분포) 정도를 의미
- 범위 – 데이터의 최대값에서 최소값을 뺀 차이를 의미
- 사분위편차(quatile Deviation) –정렬된 자료 분포의 ¼에 해당하는 자료값과 ¾에 해당하는 자료값 차이를 반으로 나눠준 값을 의미



데이터가 3,3,6,7,7,10,10,10,11,15,20 이 존재할 때

범위=20-3 즉, 17

하한사분위수=데이터 개수(N)/4 총 11개 데이터/4➔반올림하여 3 즉, 3번째 위치한 6

상한 사분위수=3*데이터갯수(N)/4 3*11/4→8.25 →반올림하여 9 즉, 9번째 위치한 11

사분위범위=상한 사분위수-하한사분위수

11-6=5

```
> n_data<-c(27,28,29,29,29,30,31,31,32,32,32,34,34,34,35,35,36)
> quantile(n_data)
  0%  25%  50%  75% 100%
  27   29   32   34   36
> IQR(n_data)
[1] 5
> range(n_data)
[1] 27 36
```

사분위 범위➔ IQR()함수를 이용하여 계산

3. 데이터 시각화 –히스토그램과 상자도표

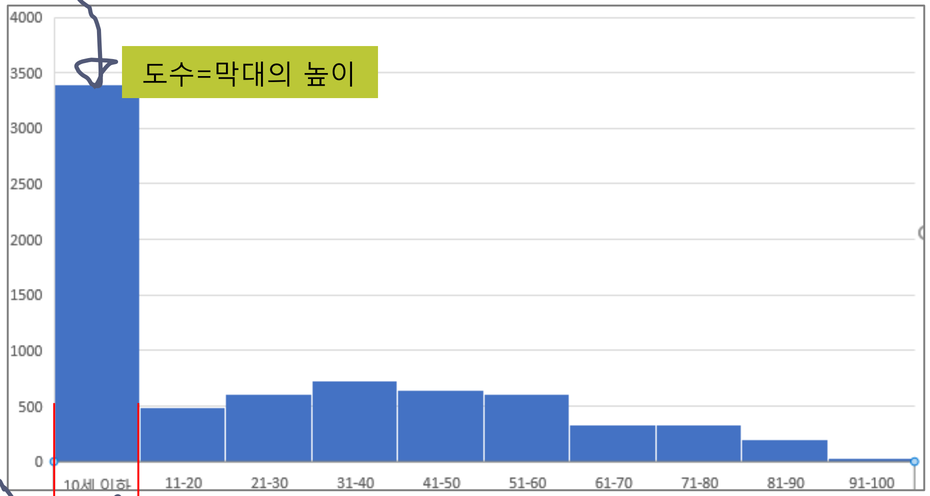
➢ 수치 데이터에 대한 분석결과를 시각화하면 데이터의 분포형태를 신속하게 이해하는데 용이

히스토그램

연령대가 일정 범위에 따라 그룹으로 나뉜 데이터셋이예요. 연령대 값을 연속적 수치눈금으로 표현하여 데이터를 시각화 할 수 있어요 그게 바로 '히스토그램'이예요. R에서는 hist()함수를 이용하여 히스토그램을 구현해요.

연령대	피해자수
0-9세	3338
10-19세	510
20-29세	607
30-39세	738
40-49세	626
50-59세	608
60-69세	352
70-79세	325
80-89세	216
90-99세	41
100-109세	2

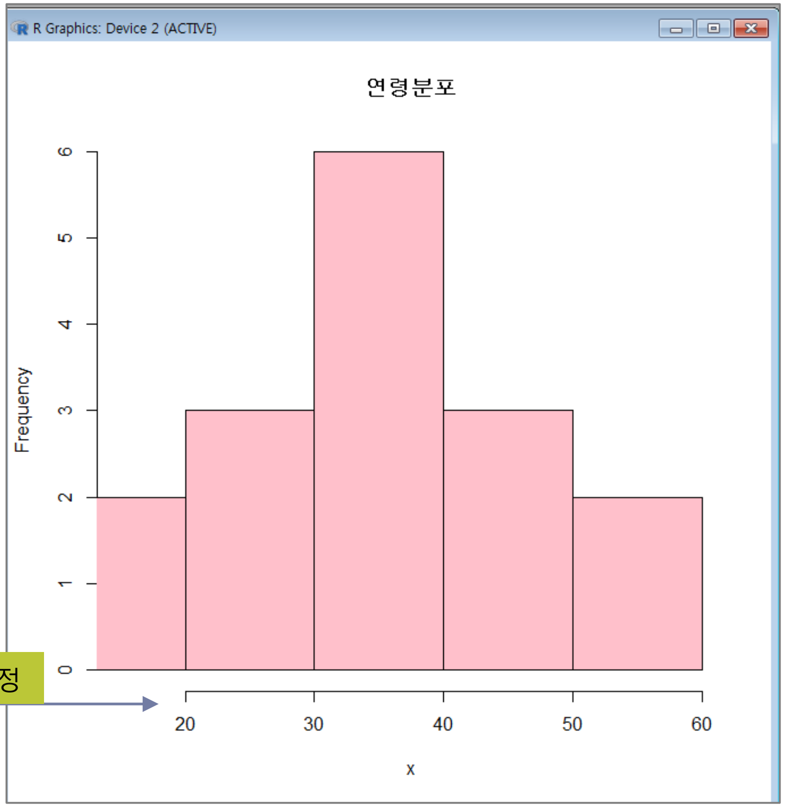
도수



막대 폭=연령대

```
> plot.new()
> x<-c(27,37,45,37,18,38,46,50,35,15,38,52,53,37,26,28)
> hist(x,main="연령분포",xlim=c(15,60),col="pink")
```

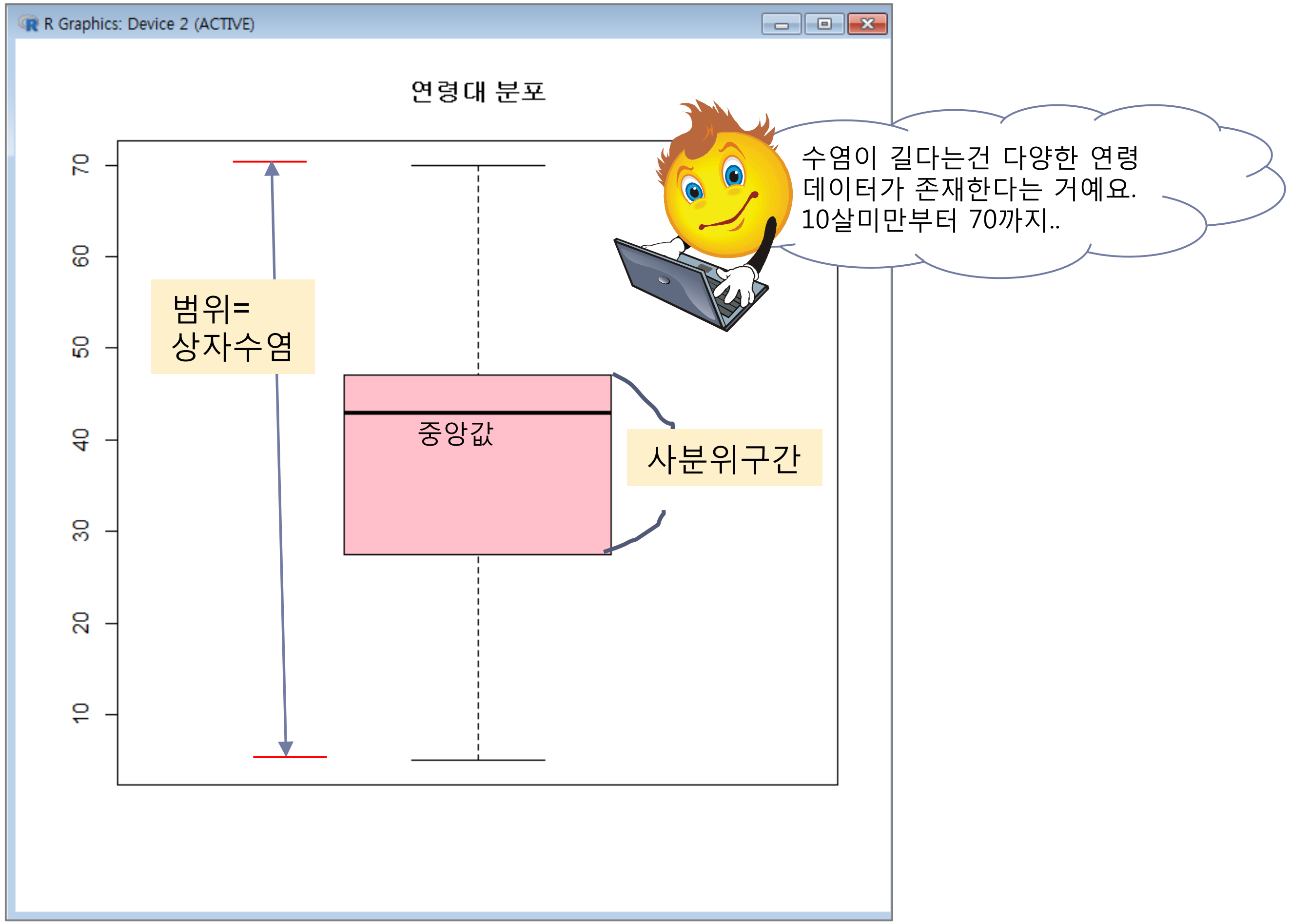
x축의 값 범위를 지정



상자도표(상자수염그림)

앞서 공부한 사분위수를 시각적으로 가장 쉽게 이해할 수 있는 그래프가 바로 상자도표예요.
상자도표에는 데이터셋의 범위, 사분범위, 중앙값등을 알 수 있어요. R에서는 boxplot()함수를 이용하여 상자도표를 구현할 수 있어요.

```
>
> plot.new()
> x<-c(63,43,45,37,28,28,46,70,25,5,48,62,43,27,16)
> boxplot(x,names="age",col='pink',main="연령대 분포")
>
```



4. 데이터 분포 변이 측정하기

➤ 데이터 분포에서 한단계 더 나아가 해당 데이터가 평균값으로부터 얼마나 멀리 떨어져있는가를 측정

- 분산 – 평균으로부터 떨어져있는 각 데이터 사이의 거리를 제곱한 평균
- 표준편차 – 데이터가 평균값으로부터 얼마나 떨어져있는가를 측정

데이터가 65,75,80,80,80,85,95 가 존재할 때

평균=(65+75+80+80+80+80+85+95)/7 →80

분산=$(65-80)^2+(75-80)^2+(85-80)^2+(95-80)^2/7$ ➔71.4

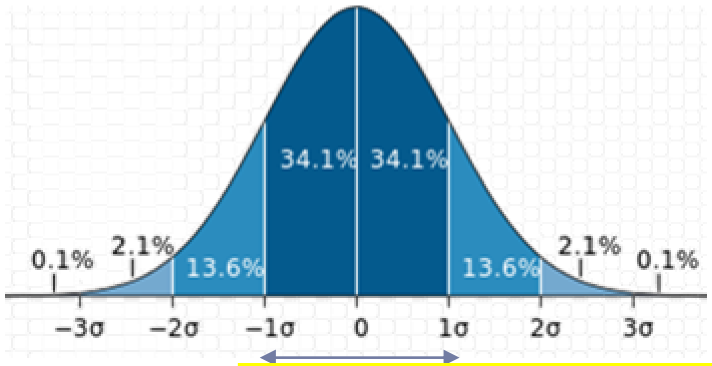
표준편차=$\sqrt{71.4}$ →8.4

```
> n_data<-c(27,28,29,29,29,30,31,31,32,32,32,34,34,34,35,35,36)
> var(n_data)
[1] 7.367647
> sd(n_data)
[1] 2.714341
>
```

분산값 계산

표준편차 계산

참고) 평균값으로부터 표준편차 의미



평균이 0이고 표준편차가 1이내에 존재할 경우 데이터의 약68.2%가 이 영역에 분포