통계처리입문

Week03

자료의 중심과 산포를 나타내는 수치적 측도

대푯값

- 대푯값
  - 자료들의 중심을 나타내는 값

- 평균
  - (산술)평균

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

  - 절사평균
    - 모든 관측값들을 크기 순으로 정렬한 후, 왼쪽 끝과 오른쪽 끝에 위치한 값들을 같은 개수로 제거한 후, 남은 관측값들을 이용하여 계산한 평균값
  - 가중산술평균

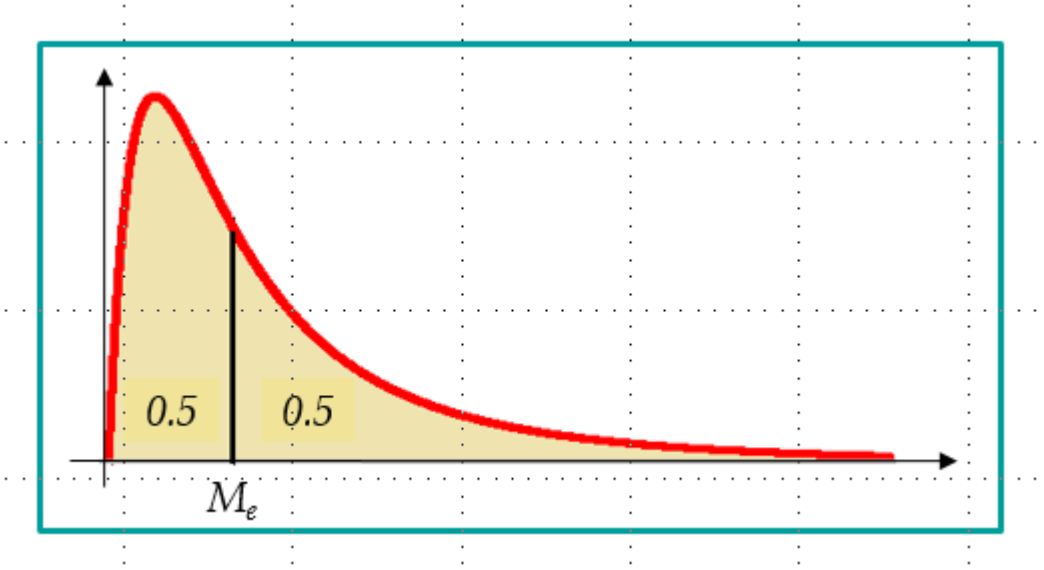
$$\bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

대푯값

- 중앙값(중위수)
  - 관측된 자료의 편중과는 상관없이 가장 작은 값에서 가장 큰 값까지 정렬했을 때 그 가운데 위치한 값

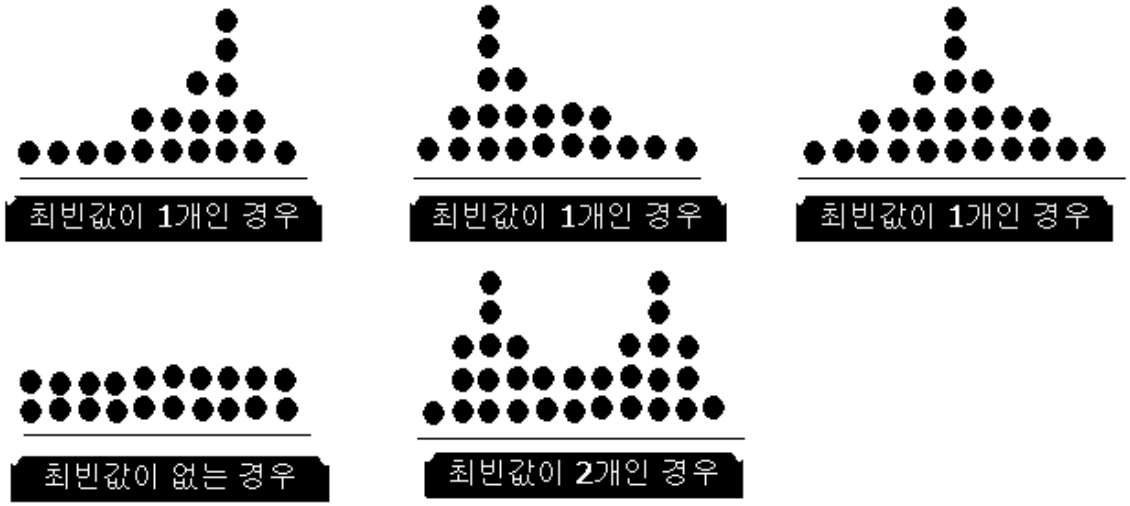
홀수일 때 : $x_{median} = x_{\left(\frac{n+1}{2}\right)}$

짝수일 때 : $x_{median} = \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}$



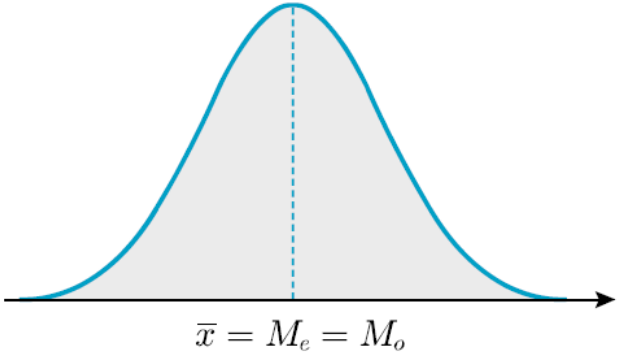
대표값

- 최빈값
  - 표본에서 가장 많이 나타나는 관측치

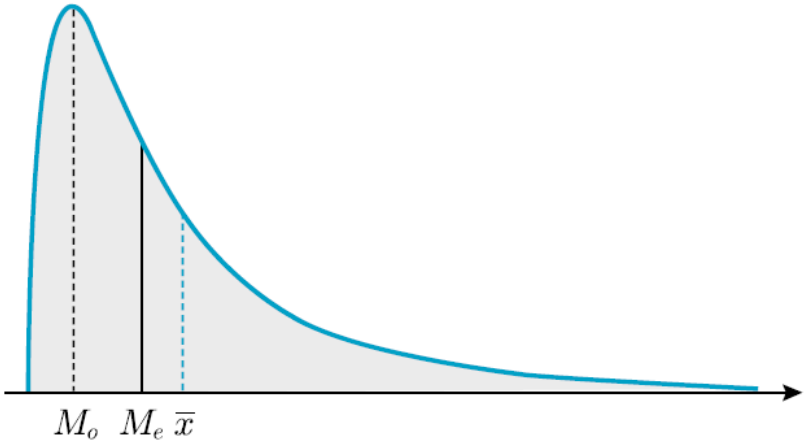


산술평균, 중앙값, 최빈값의 관계

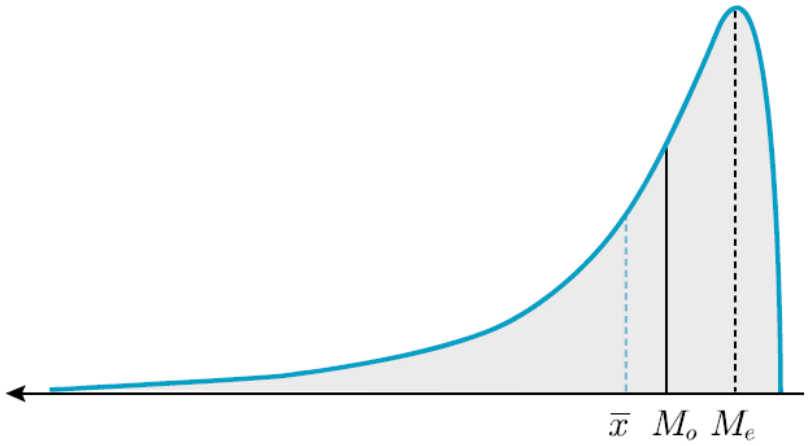
대칭인 분포모양인 경우



양의 비대칭인 분포모양인 경우



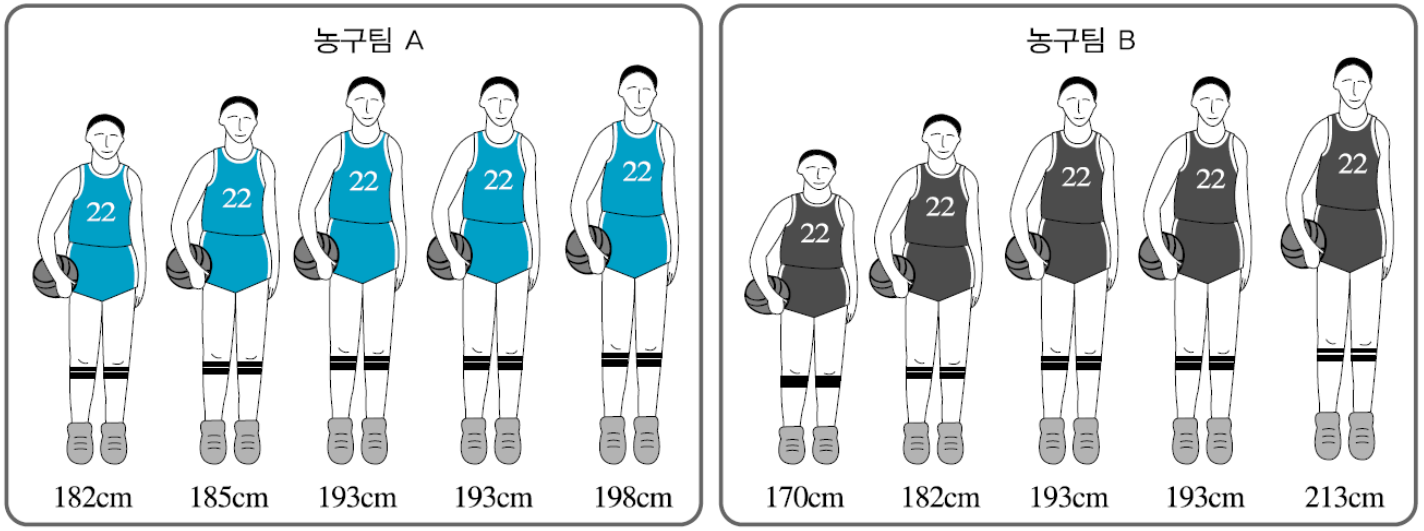
음의 비대칭인 분포모양인 경우



산포의 척도

-예

- 농구팀 A [182, 185, 193, 193, 198]과
  농구팀 B [170, 182, 193, 193, 213]
- 평균 키는 190.2, 중위수는 193, 최빈값이 193으로 동일하다.
- 점도표를 그리면, 농구팀 A의 분포는 집중되지만 농구팀 B는 넓게 퍼짐



편차

- 편차
  - 자료와 자료 무게중심 사이의 거리
    - 자료의 무게중심으로는 주로 산술평균 이용

$$d_i = x_i - \bar{x}$$

분산

- **모분산**population variance : 각 자료값과 모평균과의 편차제곱의 평균

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

- **표본분산**sample variance : 크기 n인 표본의 각 자료값과 표본 평균과의 편차 제곱의 합을 $n - 1$로 나눈 값

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

표준편차

- **모표준편차**population standard deviation : 모분산의 양의 제곱근

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

- **표본표준편차**sample standard deviation : 표본분산의 양의 제곱근

$$s\sqrt{} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

평균편차

- **평균편차**mean deviation : 각 자료값과 표본평균과의 편차의 절댓값에 대한 산술평균

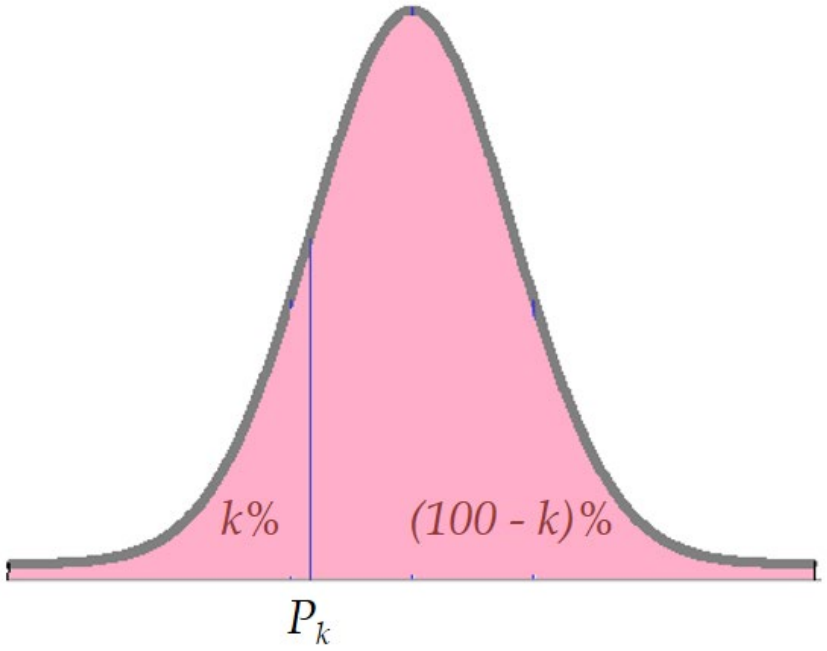
$$MD = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \overline{x}\right|$$

범위

- **범위**range : 최댓값과 최솟값의 차이
  1. 계산하기 쉽다.
  2. 특이점에 매우 큰 영향을 받는다.
  3. 개개의 자료값에 대한 정보를 반영하지 못한다.
  4. 자료의 개수가 많은 경우에 부적절하다.

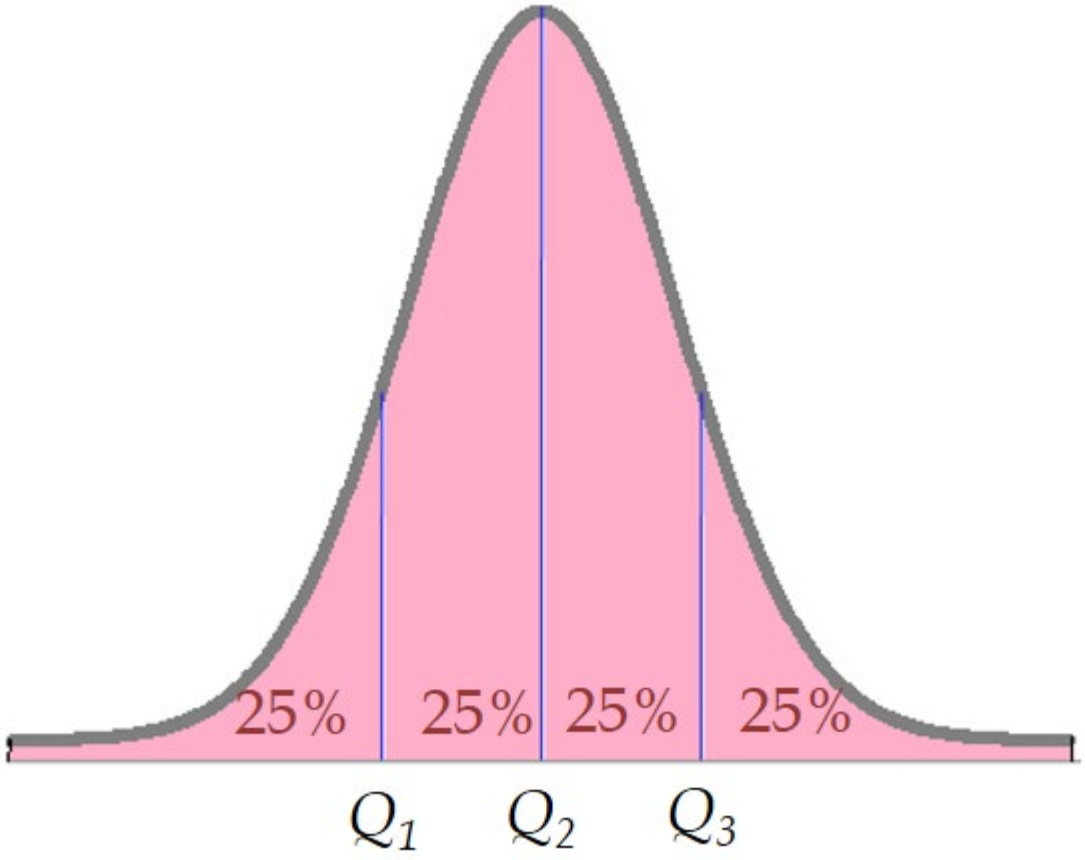
백분위수

- **백분위수**percentiles : 크기 순서로 나열된 자료집단을 100등 분하는 척도
  $P_1, P_2, \cdots, Q_{99}$



사분위수

- **사분위수**quartiles : 크기 순서로 나열된 자료집단을 4등분하 척도 Q_1, Q_2, Q_3이다.



변동계수

- **변동계수**coefficient of variation : 평균을 중심으로 한 상대적인 산포의 척도
  - 모집단의 변동계수

$$CV_P = \frac{\sigma}{\mu} \times 100(\%)$$

  - 표본의 변동계수

$$CV_S = \frac{s}{\bar{x}} \times 100(\%)$$

  - 평균이 큰 차이를 보이는 두 자료 집단 또는 측정 단위가 서로 다른 두 자료 집단에 대한 산포의 척도를 비교할 때 많이 사용

상자그림

- **상자그림**box plot : 사분위수를 이용하여 수집한 자료에 포함된 극단값을 알려주는 그림
  - **안울타리**inner fence : 사분위수 Q_1과 Q_3에서 각각 $1.5IQR$ 만큼 떨어져 있는 값

$$f_l = Q_1 - 1.5IQR, \qquad f_u = Q_3 + 1.5IQR$$

  - **바깥울타리**outer fence : 사분위수 Q_1과 Q_3에서 각각 $3IQR$ 만큼 떨어져 있는 값

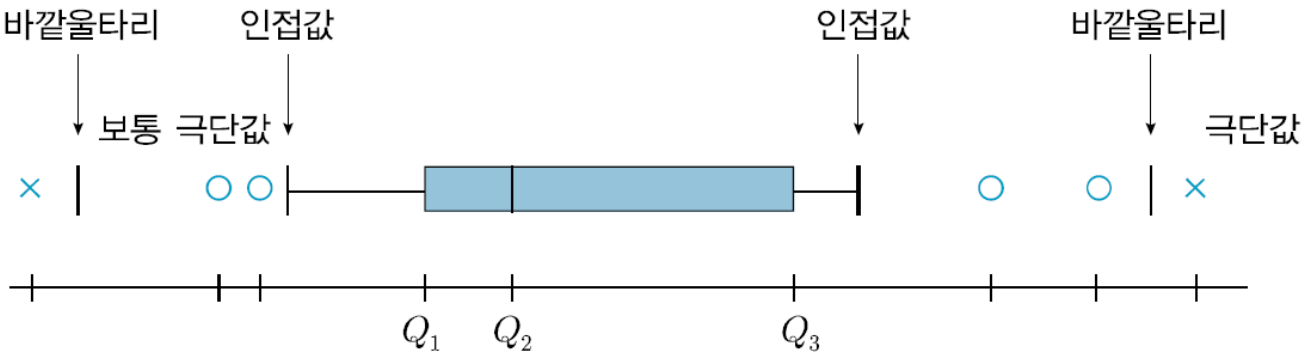
$$F_l = Q_1 - 3IQR, \qquad Fu = Q_3 + 3IQR$$

  - **인접값**adjacent value : 안울타리 안에 놓이는 가장 극단적인 자료값, 즉 아래쪽 안울타리보다 큰 가장 작은 자료값과 위쪽 안울타리보다 작은 가장 큰 자료값
  - **보통 극단**mild outlier : 안울타리와 바깥울타리 사이에 놓이는 자료값
  - **극단값**extreme outlier : 바깥울타리 외부에 놓이는 자료값

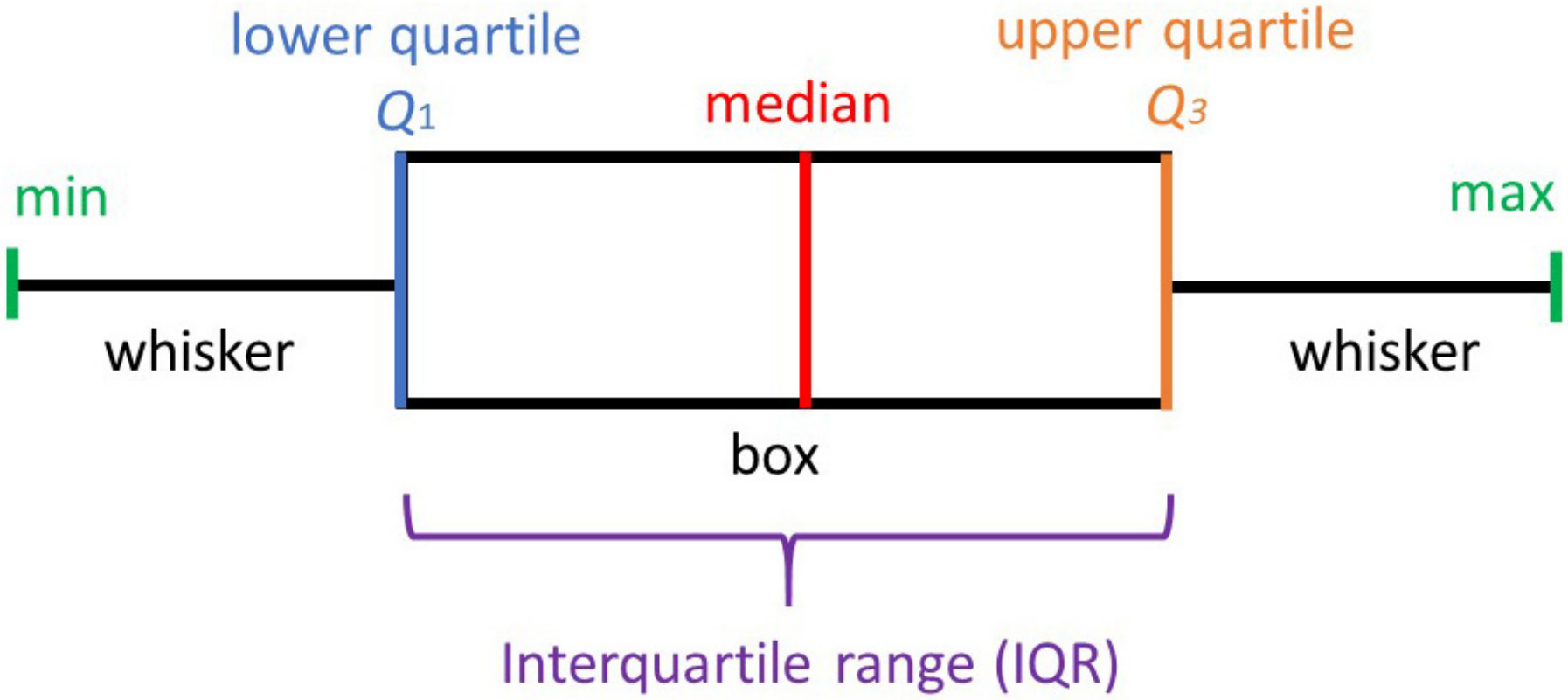
상자그림

- **사분위수범위**interquartile range : 제1사분위수에서 제3사분위수까지의 범위

$$IQR = Q_3 - Q_1$$



상자그림


lower quartile
Q_1
median
upper quartile
Q_3
min
max
whisker
whisker
box
Interquartile range (IQR)

그룹화 자료에서의 평균과 분산

- 그룹화 자료의 평균

$$\bar{x} = \frac{\sum_{i=1}^{k} m_i f_i}{n}$$

- 그룹화 자료의 분산

$$s^2 = \frac{\sum_{i=1}^{k} (m_i - \bar{x})^2 f_i}{n-1}$$

R Programming

도수분포표

- 도수분포표

```
> 지문=c(27,480,439)
> names(지문)=c("궁상문", "제상문", "와상문")
> 지문
궁상문 제상문 와상문
    27    480    439
> prop.table(지문)
     궁상문      제상문      와상문
0.02854123 0.50739958 0.46405920
```

도수분포표

```
> data=read.csv('./2015_인구주택총조사.csv')
> colnames(data)<-c('성별', '만나이', '가구주관계', '교육정도', '출생지', '경제활동상태',
 '혼인상태')
> data$성별<- factor(data$성별, levels=c(1, 2), labels=c("남자", "여자"))
> data$가구주관계 <- factor(
+   data$가구주관계,
+   levels=1:14,
+   labels=c("가구주", "가구주의 배우자", "자녀", "자녀의 배우자", "가구주의 부모",
 "배우자의 부모", "손자녀, 그 배우자", "증손자녀, 그 배우자", "조부모", "형제자매, 그
 배우자", "형제자매의 자녀, 그 배우자", "부모의 형제자매, 그 배우자", "기타 친인척",
 "그외같이사는사람")
+ )
> data$교육정도 <- factor(
+   data$교육정도,
+   levels=1:8,
+   labels=c("안 받았음", "초등학교", "중학교", "고등학교", "대학-4년제 미만", "대학-4년제
 이상", "석사과정", "박사과정")
+ )
> data$경제활동상태 <- factor(data$경제활동상태, levels=1:4, labels=c("주로 일","다른 활동
을 하면서 일","휴직","무직"))
> data$혼인상태 <- factor(data$혼인상태, levels=1:4, labels=c("미혼","기혼","사별","이혼"))
> data$출생지 <- factor(data$출생지, levels=1:4, labels=c("거주지","같은 시군구 내 다른
 집","다른 시군구","해외"))
```

도수분포표

```
> table1=table(data$만나이)
> table1

   0    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16
3758 3714 3906 4255 4122 4074 3949 4231 4275 3985 3898 4243 4313 4397 5059 5522 5207
  17   18   19   20   21   22   23   24   25   26   27   28   29   30   31   32   33
5492 5611 5048 4540 4592 5442 5704 5386 5273 5139 5105 5156 5198 5411 5737 6251 6982
  34   35   36   37   38   39   40   41   42   43   44   45   46   47   48   49   50
6973 7242 6821 6230 6557 6174 6802 7365 7463 7685 7822 8056 8056 7658 7306 7098 7802
  51   52   53   54   55   56   57   58   59   60   61   62   63   64   65   66   67
7319 7665 8018 8411 8162 8062 7681 7009 6970 6962 6010 5434 5743 4665 4968 4827 4960
  68   69   70   71   72   73   74   75   76   77   78   79   80   81   82   83   84
4857 4180 3853 3852 4097 4714 3898 3810 3670 3368 3202 2958 2800 2463 1987 1911 1463
  85
5881
```

도수분포표

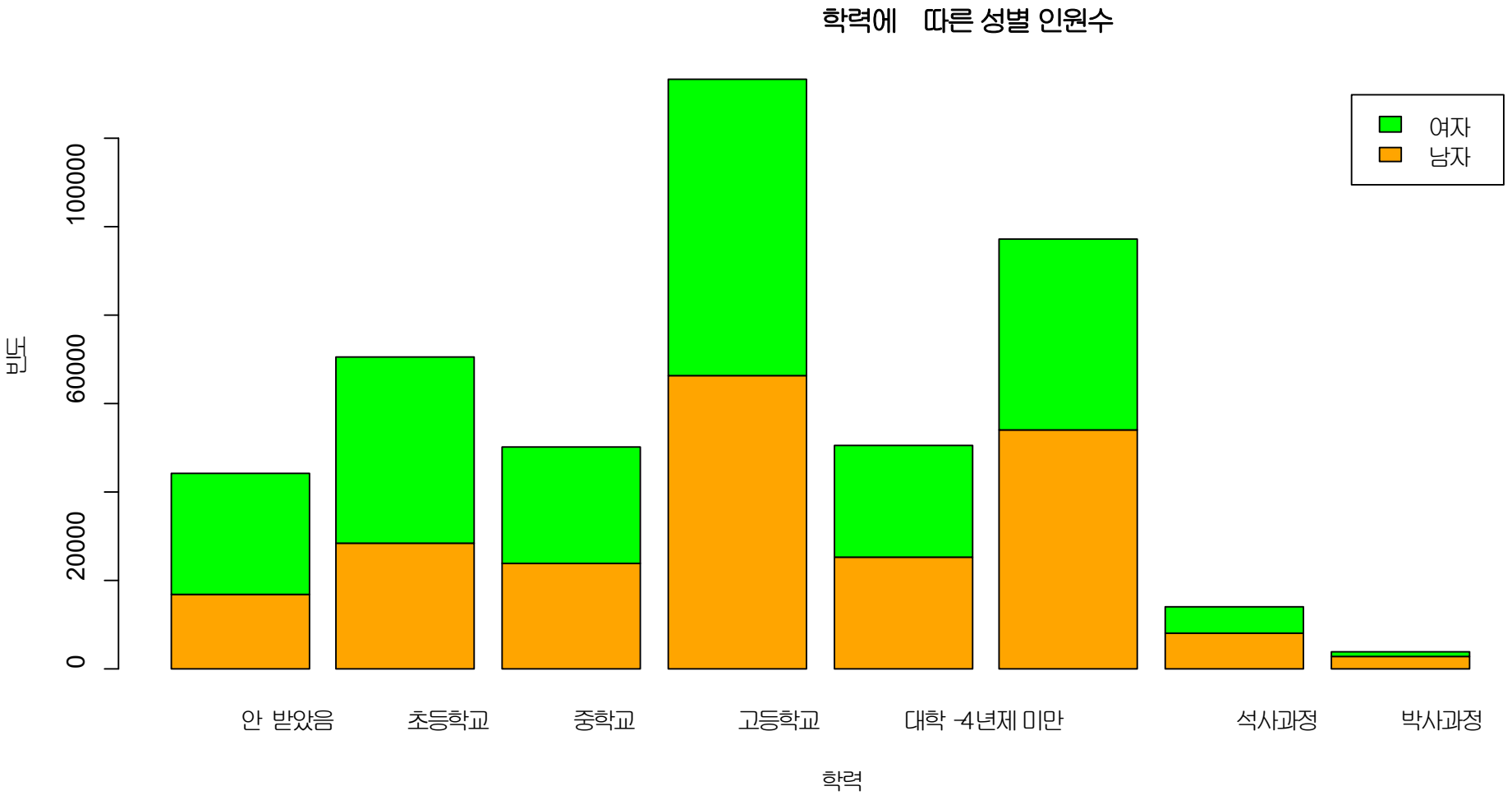
```
> table2=table(data$성별, data$교육정도)
> table2

     안 받았음 초등학교 중학교 고등학교 대학-4년제 미만 대학-4년제 이상 석사과정
남자     16667    28562  23710    66543           25263          54080     7972
여자     27562    42004  26455    66903           25318          43343     5861

     박사과정
남자     2597
여자     1075
```

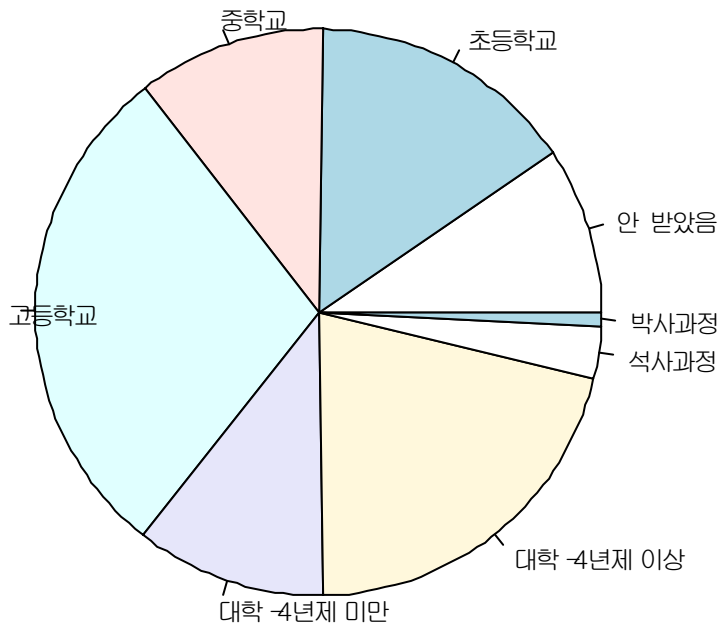
막대그래프

```
> barplot(table2, col=c("orange", "green"), main="학력에 따른 성별 인원수", xlab="학력", yl
ab="빈도", legend.text=T)
```



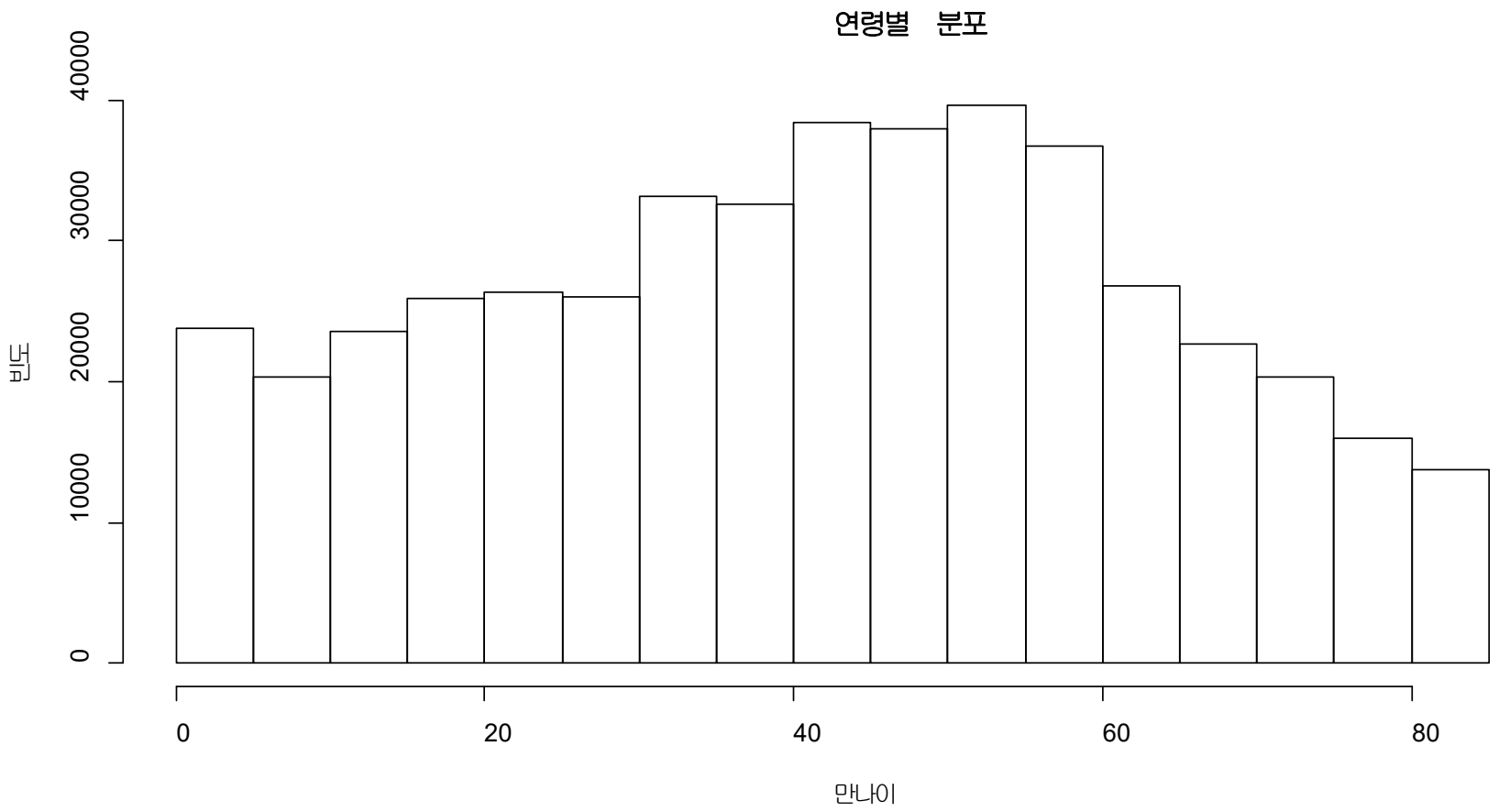
원그래프

```
> table3=table(data$교육정도)
> pie(table3, main="학력수준별 비중", cex=0.8)
```



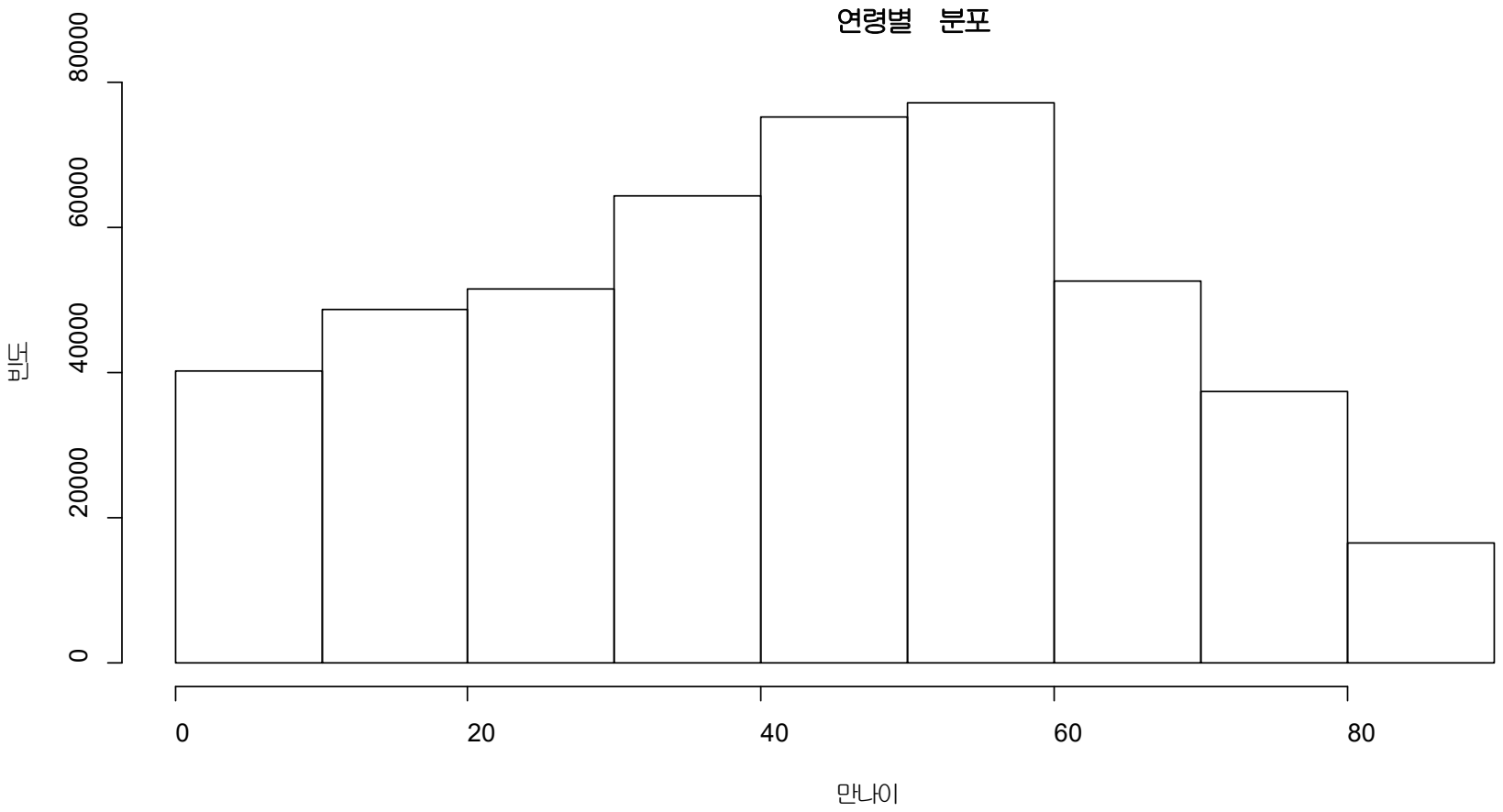
히스토그램

```
> hist(data$만나이,  main="연령별 분포", xlab="만나이", ylab="빈도" )
```



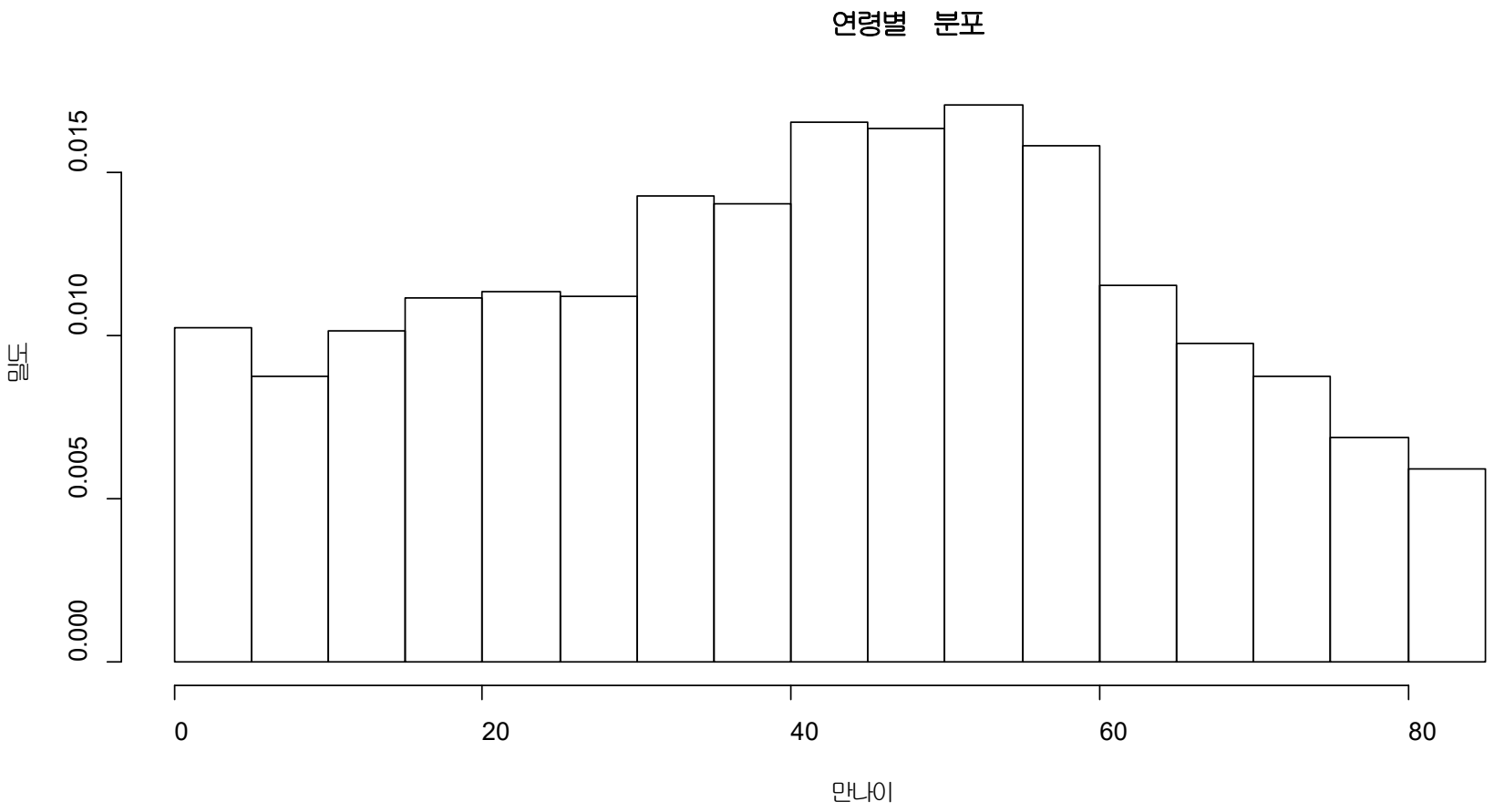
히스토그램

```
> hist(data$만나이, breaks=c(seq(0, 90, 10)), right=F, main="연령별 분포", xlab="만나이", ylab
="빈도")
```



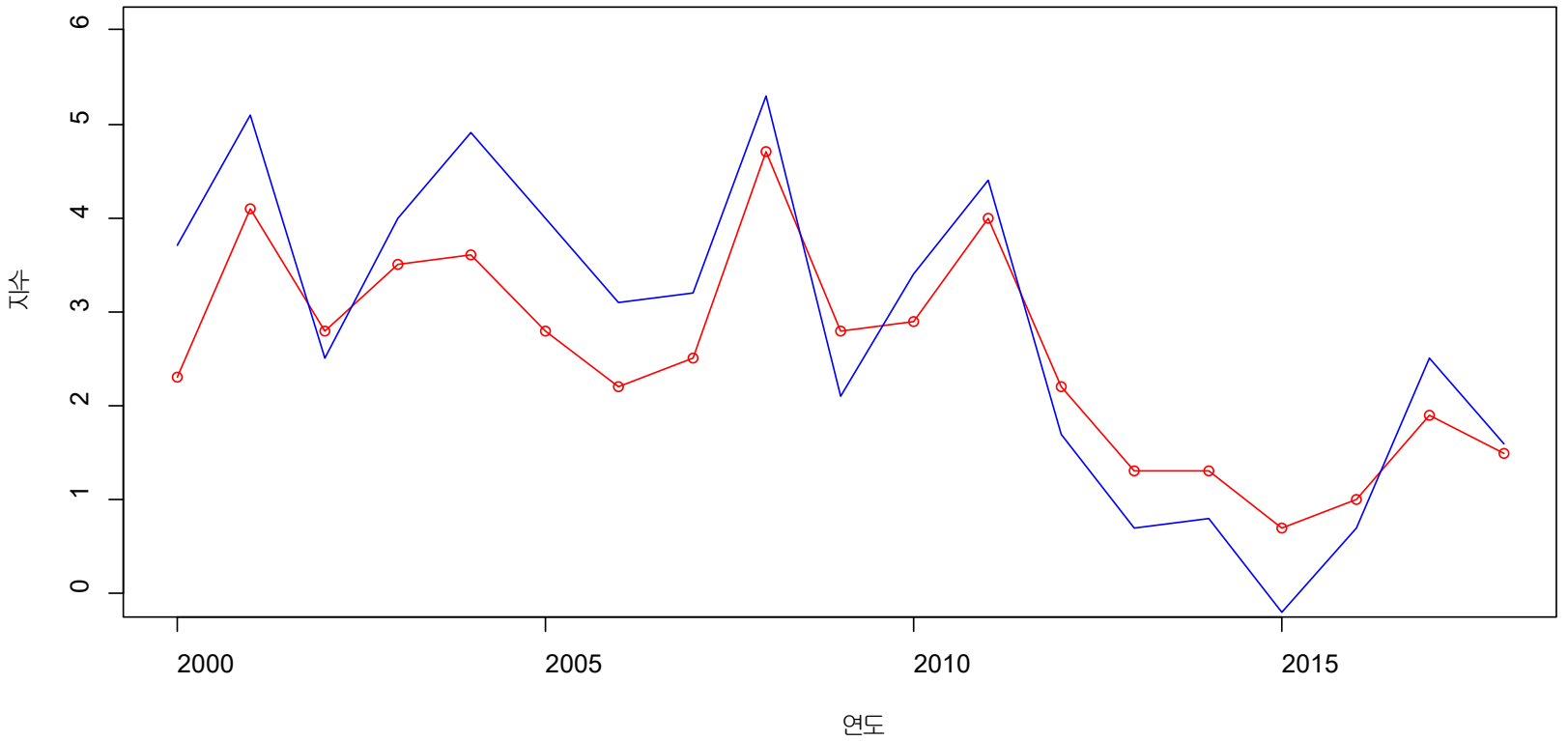
히스토그램

```
> hist(data$만나이, probability=T,  main="연령별 분포", xlab="만나이", ylab="밀도" )
```



막대그래프

```
> data=read.csv('./연도별_소비자물가_등락률.csv')
> plot(data$연도, data$총지수, col = "red", type="o", ylim=c(0,6), xlab="연도", ylab="지수")
> lines(data$연도, data$생활물가지수, col = "blue")
```



대푯값, 산포의 척도

- 평균

```
> mean(data$총지수)
[1] 2.531579
```

- 중앙값

```
> median(data$총지수)
[1] 2.5
```

- 분산 및 표준편차

```
> var(data$총지수)
[1] 1.234503
> sd(data$총지수)
[1] 1.111082
```

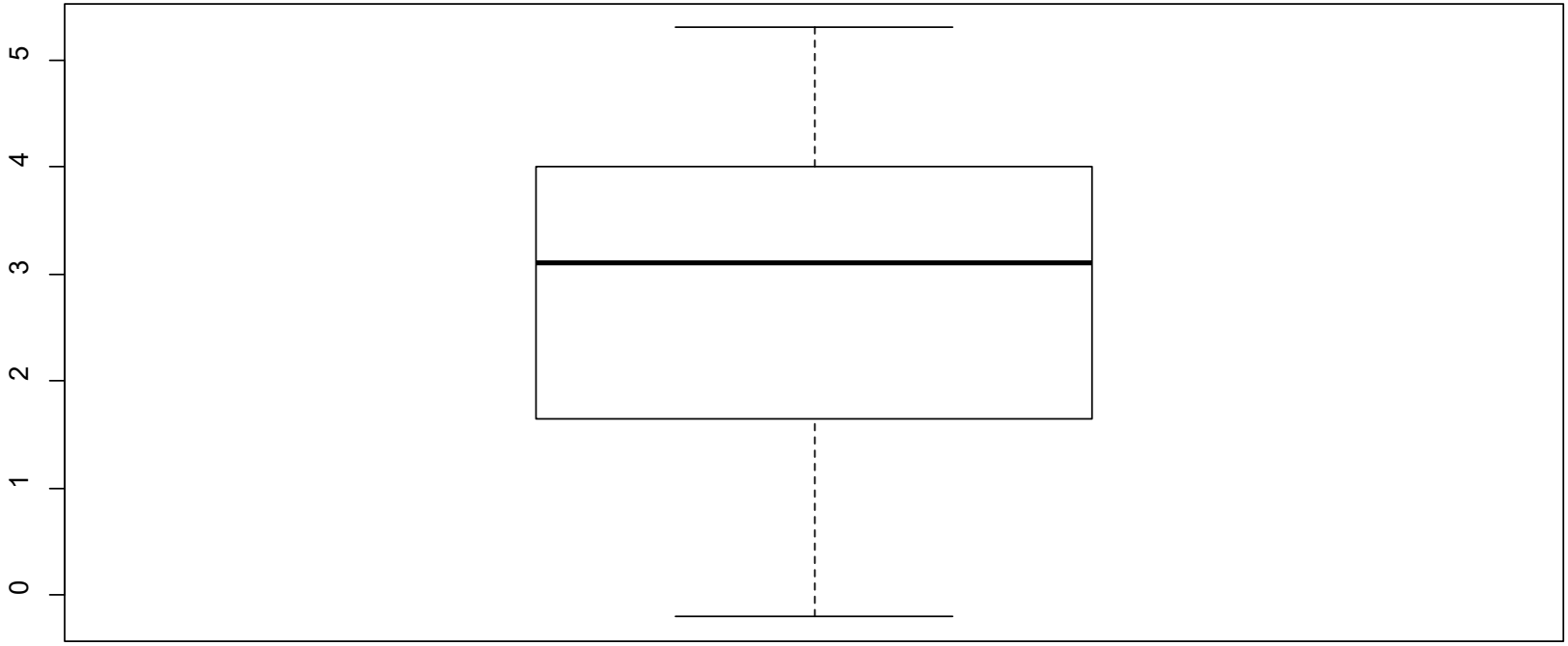
대푯값, 산포의 척도

- 최소, 최대, 사분위수, 백분위수

```
> min(data$총지수)
[1] 0.7
> max(data$총지수)
[1] 4.7
> quantile(data$총지수)
  0%  25%  50%  75% 100%
 0.7  1.7  2.5  3.2  4.7
> quantile(data$총지수, probs = seq(0,1,0.1))
  0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
0.70 1.24 1.42 2.02 2.22 2.50 2.80 2.86 3.54 4.02 4.70
```

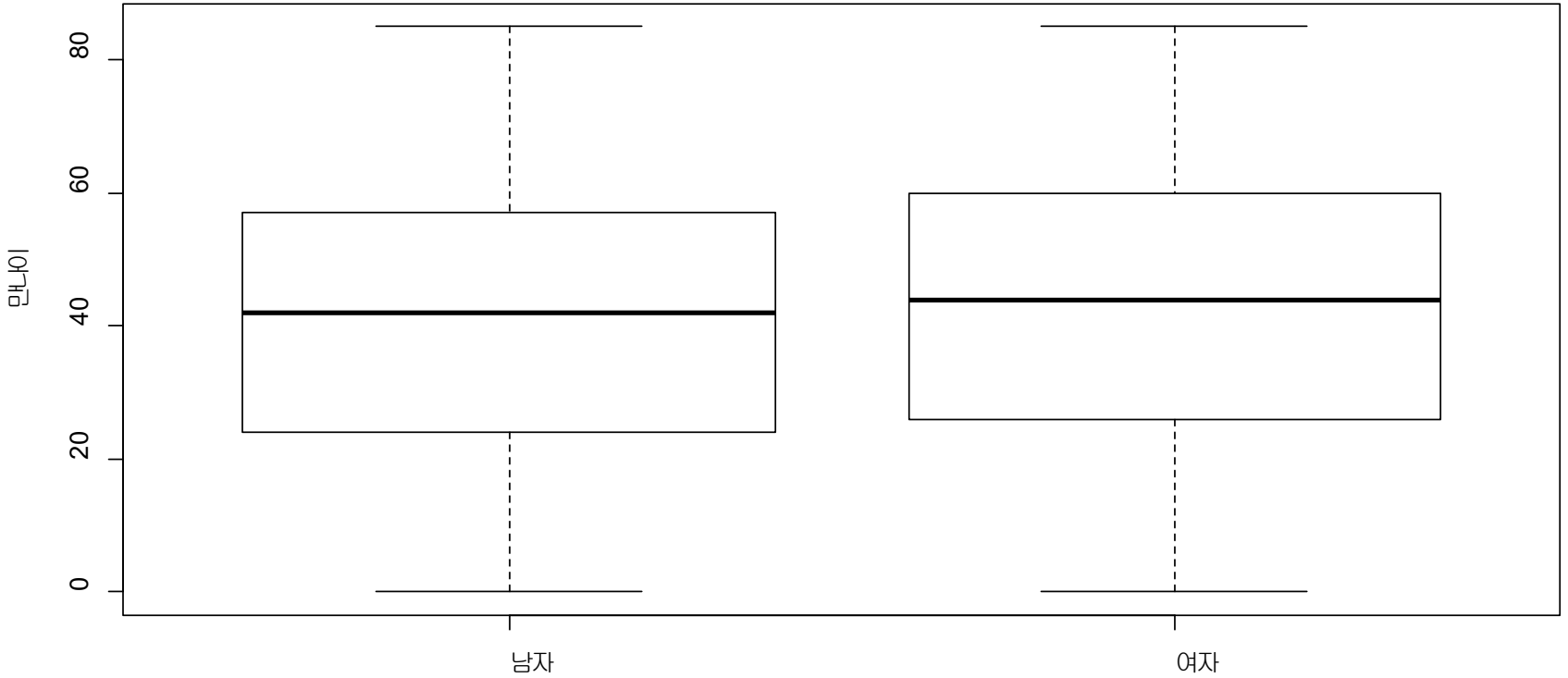
상자그림

```
> boxplot(data$생활물가지수)
```



상자그림

```
> boxplot(만나이~성별,data=data, ylab="만나이")
```



함수 정리

	R 함수	설명
자료의 개수(n)	length()	자료의 개수
최솟값($\min(x)$, $x_{(1)}$)	min()	자료 중 가장 작은 값
최댓값($\max(x)$, $x_{(n)}$)	max()	자료 중 가장 큰 값
범위	range	최댓값-최솟값으로 전체 자료가 분포하는 범위를 나타냄
최빈값	table()로 확인	자료 중 빈도수가 가장 많은 값 혹은 구간
평균	mean()	전체 자료의 무게중심이 되는 값으로 양 끝 값의 변화에 민감한 단점을 갖고 있음
중앙값	median()	자료를 순서대로 나열했을 경우 중앙이 되는 값으로 순위가 정해진 후에는 각 값이 갖고 있는 값은 무시됨
표준편차	sd()	평균을 중심으로 자료가 퍼진 정도를 나타내는 값
제1사분위수	quantile()	자료를 순서대로 나열했을 경우 25% 위치의 값
제3사분위수		자료를 순서대로 나열했을 경우 75% 위치의 값
사분위수 범위	IQR()	(제3사분위수－제1사분위수)

과제

과제 3

- 통계청에서 제공하는 데이터를 이용해 그래프 그리기
  - 직접 선택한 데이터를 바탕으로 의미 있는 정보를 이해하기 쉽게 볼 수 있는 서로 다른 종류의 그래프를 3개 그리기
    - 보여주려는 정보에 따라서 서로 다른 형태의 그래프를 그리고 그러한 그래프를 선택한 이유를 작성
  - 통계청
    - http://kostat.go.kr/portal/korea/index.action