R 데이터 분석 입문

4주 차

일변량 자료 탐색

오 세 종

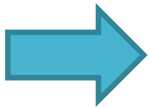
DANKOOK UNIVERSITY

Contents

1. 기초 통계 개념

- 통계 기법은 자료를 정리하고 분석할 수 있는 강력한 수단
- 데이터 분석 에서도 많은 부분에서 통계적 기법을 필요로 한다

  - 여론조사 결과 분석
  - 제조업 불량율 분석
  - 학습 효과 분석
  - …



데이터 분석가가 되기 위해서는 통계학을 알아야 한다.

1. 기초 통계 개념

- 질적 자료(qualitative data) 또는 범주형 자료(categorical data) : 원칙적으로 숫자로 표시될 수 없는 자료
  - 예) 교육수준 : 초졸, 중졸, 고졸, 대졸 / 성별 : M, F

```
> iris$Species
  [1] setosa     setosa     setosa     setosa     setosa     setosa     setosa     setosa     setosa
 [10] setosa     setosa     setosa     setosa     setosa     setosa     setosa     setosa     setosa
 [19] setosa     setosa     setosa     setosa     setosa     setosa     setosa     setosa     setosa
 [28] setosa     setosa     setosa     setosa     setosa     setosa     setosa     setosa     setosa
 [37] setosa     setosa     setosa     setosa     setosa     setosa     setosa     setosa     setosa
 [46] setosa     setosa     setosa     setosa     setosa     versicolor versicolor versicolor versicolor
 [55] versicolor versicolor versicolor versicolor versicolor versicolor versicolor versicolor versicolor
 [64] versicolor versicolor versicolor versicolor versicolor versicolor versicolor versicolor versicolor
 [73] versicolor versicolor versicolor versicolor versicolor versicolor versicolor versicolor versicolor
 [82] versicolor versicolor versicolor versicolor versicolor versicolor versicolor versicolor versicolor
 [91] versicolor versicolor versicolor versicolor versicolor versicolor versicolor versicolor versicolor
[100] versicolor virginica  virginica  virginica  virginica  virginica  virginica  virginica  virginica
[109] virginica  virginica  virginica  virginica  virginica  virginica  virginica  virginica  virginica
[118] virginica  virginica  virginica  virginica  virginica  virginica  virginica  virginica  virginica
[127] virginica  virginica  virginica  virginica  virginica  virginica  virginica  virginica  virginica
[136] virginica  virginica  virginica  virginica  virginica  virginica  virginica  virginica  virginica
[145] virginica  virginica  virginica  virginica  virginica  virginica
```

- 양적 자료(quantitative data) : 자료자체가 숫자로 표현됨.
  - 이산자료(discrete data) : 정수값을 취할 수 있는 자료(각 세대의 자녀수)
  - 연속자료(continuous data) : 실수 값을 취할 수 있는 자료(키, 몸무게, 온도)

```
> iris$Sepal.Length
  [1] 5.1 4.9 4.7 4.6 5.0 5.4 4.6 5.0 4.4 4.9 5.4 4.8 4.8 4.3 5.8 5.7 5.4 5.1 5.7 5.1 5.4 5.1 4.6 5.1 4.8 5.0
 [27] 5.0 5.2 5.2 4.7 4.8 5.4 5.2 5.5 4.9 5.0 5.5 4.9 4.4 5.1 5.0 4.5 4.4 5.0 5.1 4.8 5.1 4.6 5.3 5.0 7.0 6.4
 [53] 6.9 5.5 6.5 5.7 6.3 4.9 6.6 5.2 5.0 5.9 6.0 6.1 5.6 6.7 5.6 5.8 6.2 5.6 5.9 6.1 6.3 6.1 6.4 6.6 6.8 6.7
 [79] 6.0 5.7 5.5 5.5 5.8 6.0 5.4 6.0 6.7 6.3 5.6 5.5 5.5 6.1 5.8 5.0 5.6 5.7 5.7 6.2 5.1 5.7 6.3 5.8 7.1 6.3
[105] 6.5 7.6 4.9 7.3 6.7 7.2 6.5 6.4 6.8 5.7 5.8 6.4 6.5 7.7 7.7 6.0 6.9 5.6 7.7 6.3 6.7 7.2 6.2 6.1 6.4 7.2
[131] 7.4 7.9 6.4 6.3 6.1 7.7 6.3 6.4 6.0 6.9 6.7 6.9 5.8 6.8 6.7 6.7 6.3 6.5 6.2 5.9
```

1. 기초 통계 개념

- 일변량 자료(univariate data)
  - 분석대상이 되는 변수의 개수가 1개
  - ex) 단국대 학생들의 몸무게 분포를 분석해보자
  - vector 에 저장하여 분석

- 다변량 자료 (multivariate data)
  - 분석대상이 되는 변수의 개수가 2개 이상인 경우
  - 변수가 2개인 경우를 특별히 이변량 자료(bivariate data) 라고함
  - Ex) 출생 지역과 몸무게가 상관관계가 있는지 분석해보자
  - matrix 또는 data frame 에 저장하여 분석

이번 강의에서는 일변량 자료의 분석에 대해서 다룬다.
일변량 자료는 **벡터(vector)** 에 저장하여 분석 한다.

1. 기초 통계 개념

- 모집단 (population)
  - 관심을 가지는 조사대상 전체

- 표본(sample)
  - 모집단에서 실제 조사가 이루어지는 집단, 표본은 모집단의 부분집합

> 단국대 학생 중 100명을 선별하여 외국어 실력을 조사해보자

  - 모집단 : 단국대 학생 전체
  - 표본 : 선발된 100명

- 모수(parameter)
  - 모집단의 특성을 나타내는 척도로 보통 평균과 표준편차 등이 많이 사용됨

1. 기초 통계 개념

통계분석은 많은 경우 표본을 이용하여
모집단을 추정하는데 이용됩니다

1. 기초 통계 개념

- 변수의 개수와 형태에 따른 그래프의 종류

변수 개수	변수 형태	그래프
일변량 (변수 1개)	연속형 데이터	▪ 히스토그램 (Histogram) ▪ 커널 밀도 곡선 (Kernel Density Curve) ▪ 박스 그래프 (Box Plot) ▪ 바이올린 그래프 (Violin Plot)
	범주형 데이터 (명목형, 순서형)	▪ 막대 그림 (Bar Chart) ▪ 원 그림 (Pie Chart)
다변량 (변수 2개 이상)	**연속형 데이터**	▪ **산점도 (행렬) (Scatter Plot)** ▪ 선 그래프 (Line Plot) ▪ 시계열 그래프 (Time Series Plot)
	범주형 데이터	▪ 모자이크 그림 (Mosaic Chart)

참조: http://rfriend.tistory.com/72

2. 일변량 질적자료의 분석

- 도수분포표 작성

```
ans=c("Y","Y","N","Y","Y")
table(ans)                          # 도수분포표 출력
table(ans)/length(ans)              # 비율 출력
```

```
> ans=c("Y","Y","N","Y","Y")
> table(ans)                        # 도수분포표 출력
ans
N Y
1 4
> table(ans)/length(ans)            # 비율 출력
ans
  N   Y
0.2 0.8
```

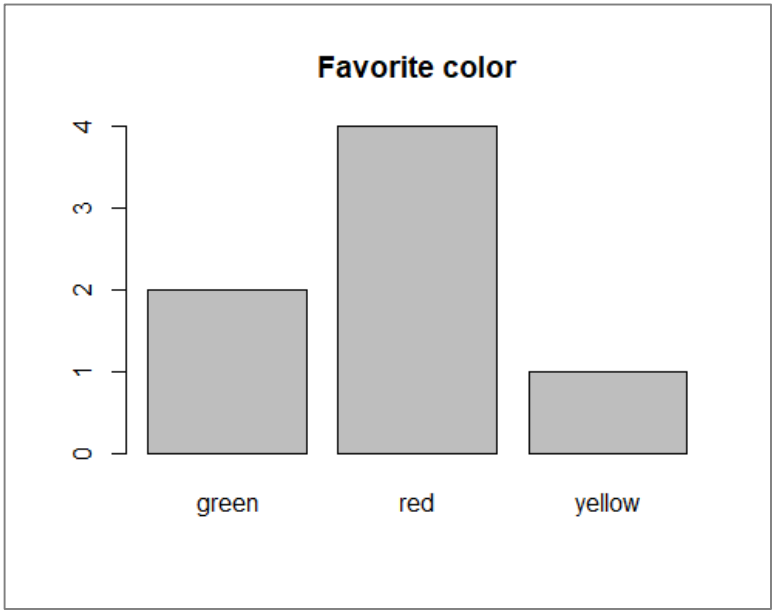
```
> table(iris$Species)

    setosa versicolor  virginica
        50         50         50
```

2. 일변량 질적자료의 분석

- 막대그래프 작성

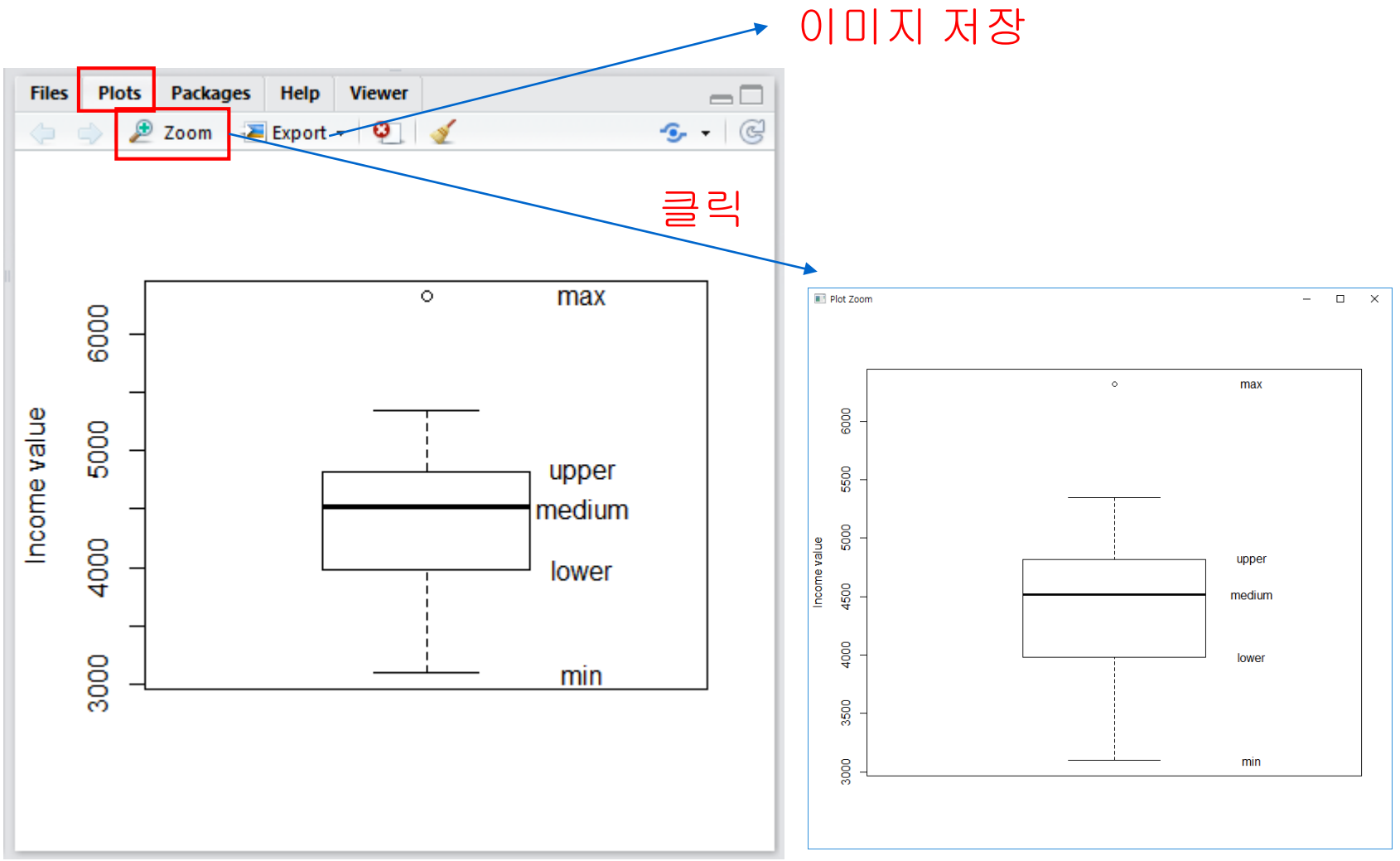
```
favorite.color <- c("red", "green", "yellow",
"red", "green", "red", "red")
sum <- table(favorite.color)      # 도수분포표
sum
barplot(sum, main="Favorite color")
```

```
> sum
favorite.color
 green    red yellow
     2      4      1
```



Note

- Rstudio 에서 그래프 보기



2. 일변량 질적자료의 분석

- 막대그래프 사례

```
head(mtcars)                          # 자동차 모델별 제원
carb <- mtcars[,"carb"]               # 기화기 수
table(carb)                           # 도수분포표
barplot(table(carb),
        main="Barplot of Carburetors",
        xlab="#of carburetors",
        ylab="frequency")
```

- table() 함수 : 주어진 자료로 부터 도수 분포표를 그려준다.

```
> table(mtcars$carb)

 1  2  3  4  6  8
 7 10  3 10  1  1
```

2. 일변량 질적자료의 분석

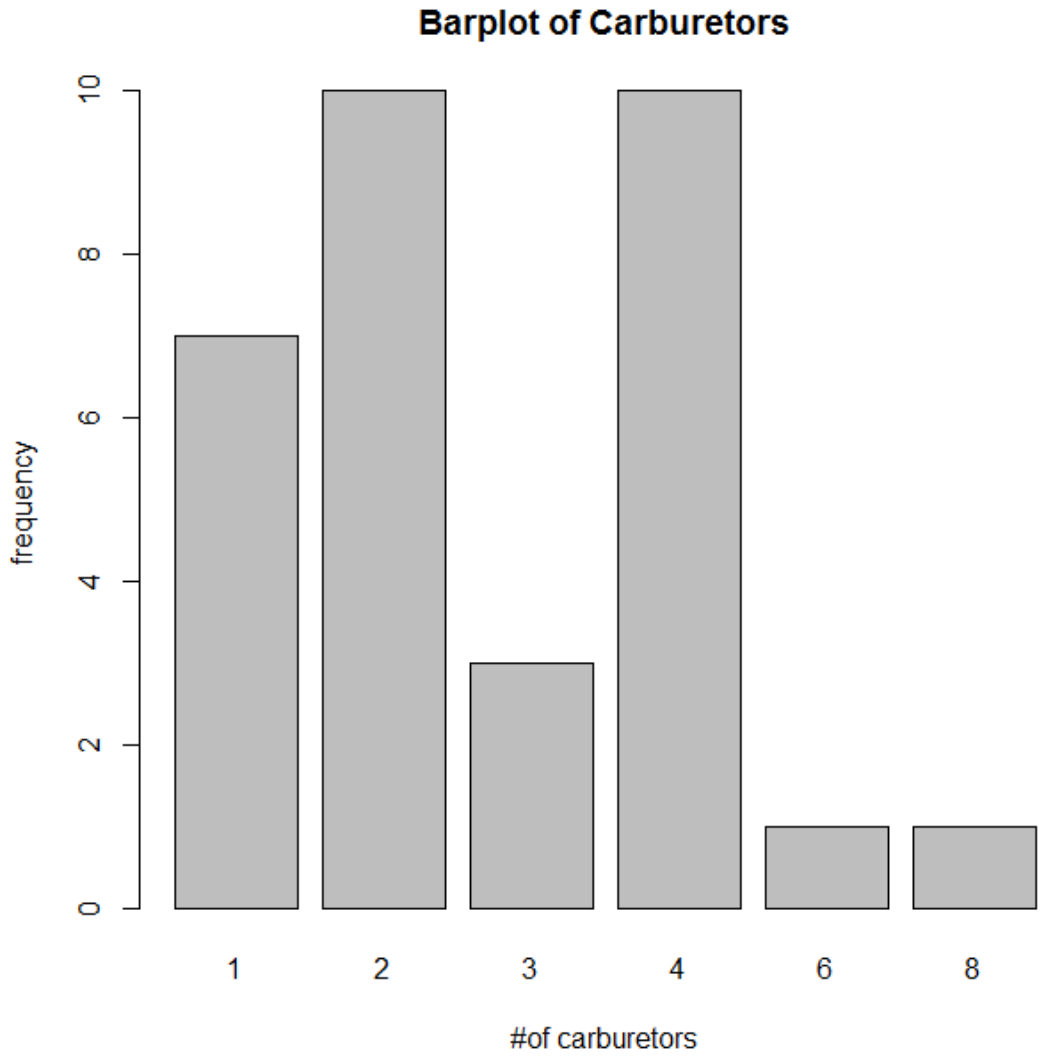
- barplot() 매개변수

- 옵션↓

인수	설명
angle, density, col	막대를 칠하는 선분의 각도, 선분의 수, 선분의 색 지정
legend	오른쪽 상단에 범례추가
names	각 막대의 라벨을 정하는 문자열 벡터를 지정
width	각 막대의 상대적인 폭을 벡터로 지정
space	각 막대 사이의 간격을 지정
beside	TRUE를 지정하면 각각의 값마다 막대를 그림
horiz	TRUE를 지정하면 막대를 옆으로 눕혀서 그림

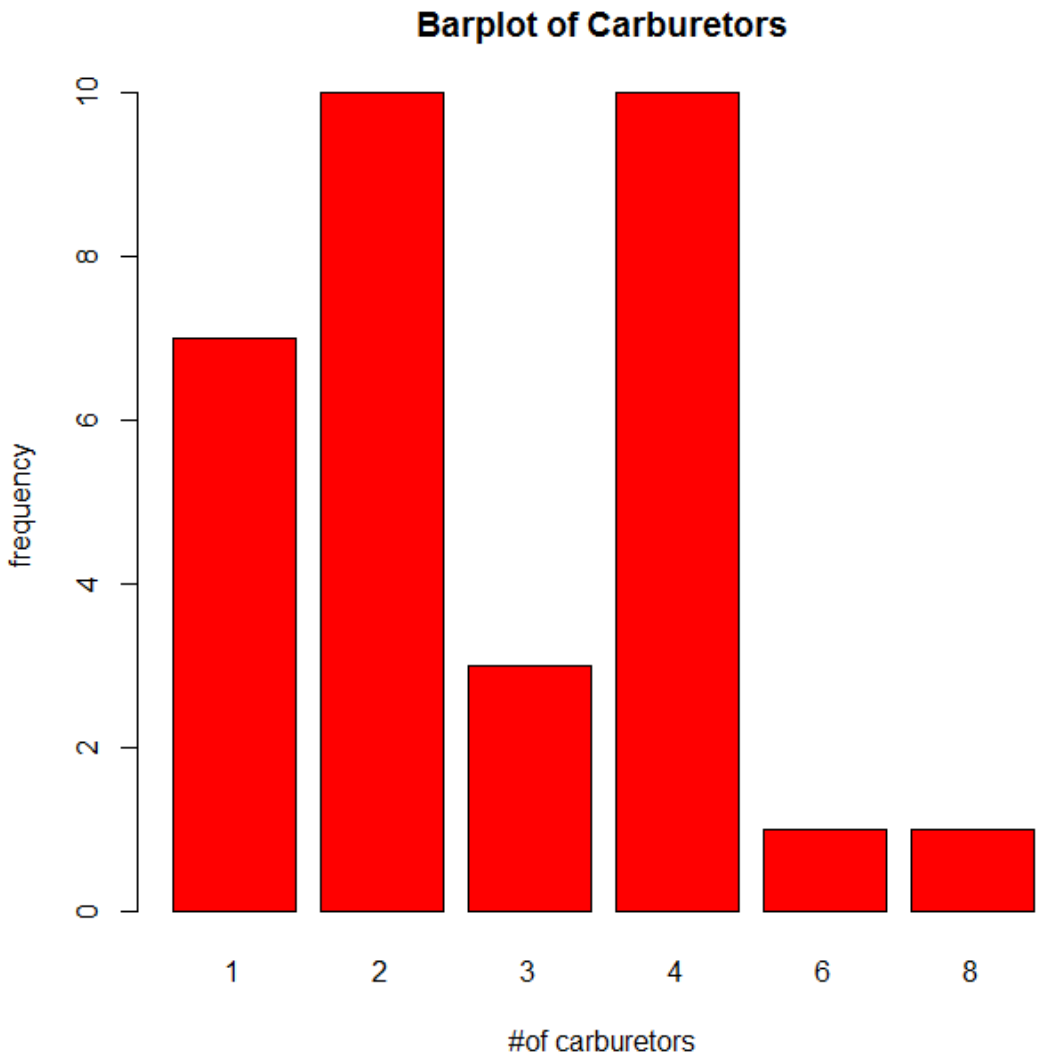
2. 일변량 질적자료의 분석

```
> barplot(table(carb),
+         main="Barplot of Carburetors",
+         xlab="#of carburetors",
+         ylab="frequency")
```



2. 일변량 질적자료의 분석

```
> barplot(table(carb),
+         main="Barplot of Carburetors",
+         xlab="#of carburetors",
+         ylab="frequency",
+         col="red")
```

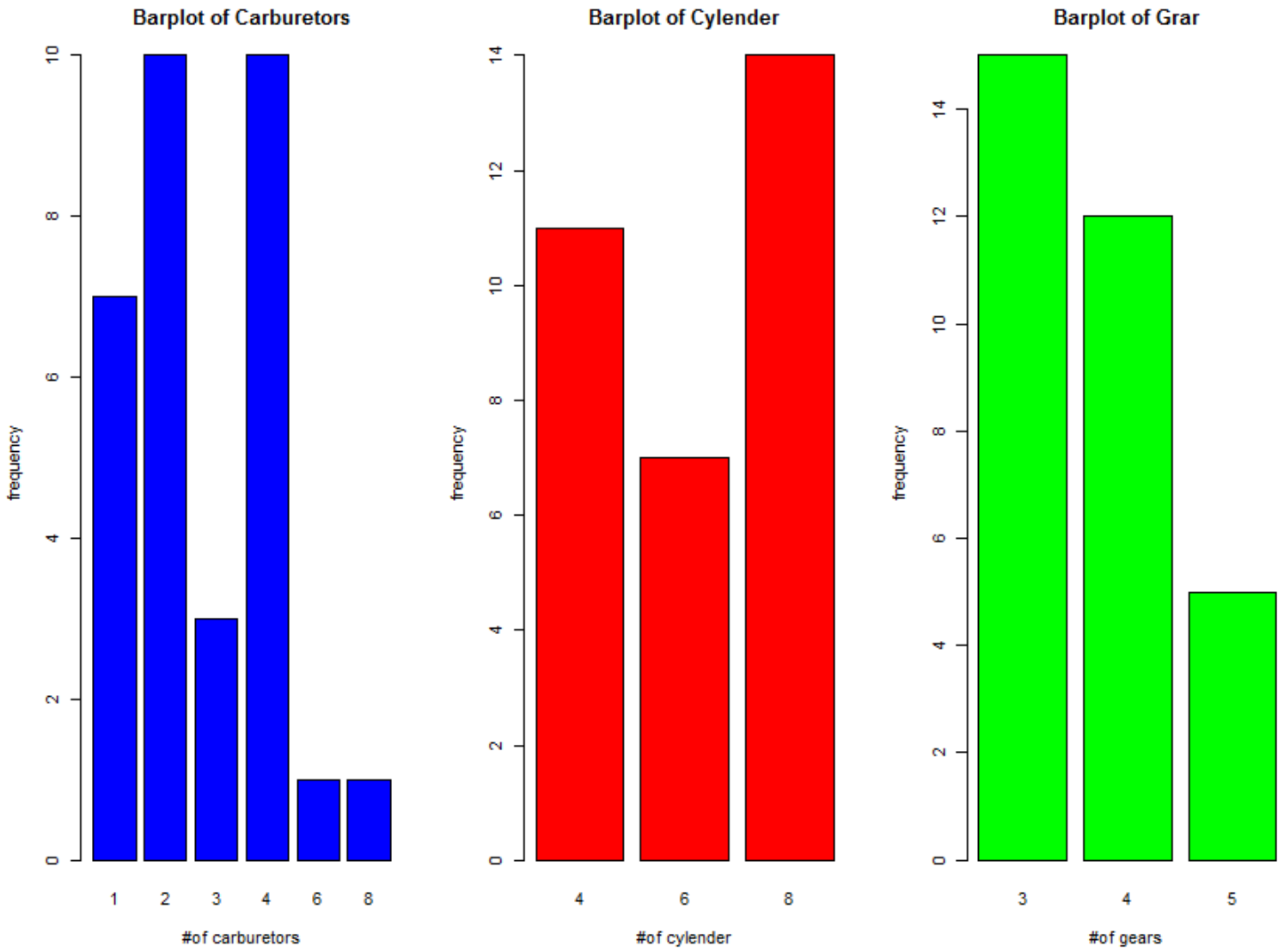


2. 일변량 질적자료의 분석

- 한 화면에 그래프 여러 개 그리기

```
par(mfrow=c(1,3))              # 1x3 윈도우 생성
barplot(table(mtcars$carb),
        main="Barplot of Carburetors",
        xlab="#of carburetors",
        ylab="frequency",
        col="blue")
barplot(table(mtcars$cyl),
        main="Barplot of Cylender",
        xlab="#of cylender",
        ylab="frequency",
        col="red")
barplot(table(mtcars$gear),
        main="Barplot of Grar",
        xlab="#of gears",
        ylab="frequency",
        col="green")
```

2. 일변량 질적자료의 분석



2. 일변량 질적자료의 분석

- Barplot에 대한 보다 상세한 옵션을 보려면

```
? barplot
```

또는 Rstudio 의 help 탭에서 barplot 검색

- 다양한 막대 그래프 예제
  http://www.theanalysisfactor.com/r-11-bar-charts/

- R 에서 지원하는 color 이름
  http://www.stat.columbia.edu/~tzheng/files/Rcolor.pdf

[연습 1]

1. R에서 제공하는 infert 데이터셋의 education 컬럼에는 각 사람이 교육 받은 기간이 범주형 자료 형태로 저장되어 있다. infert 데이터셋의 education 컬럼 값을 잘라내어 edu 에 저장한뒤 edu 의 값을 보이시오

2. edu 에 있는 값들을 중복을 제거하고 보이시오

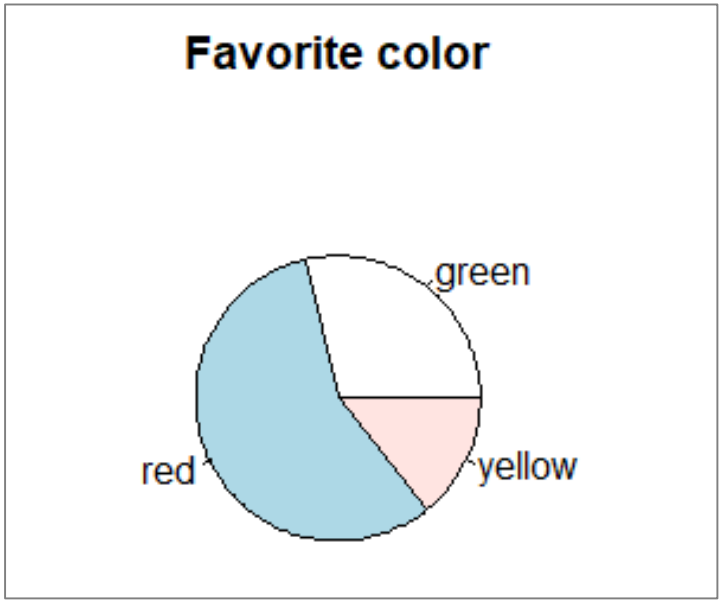
3. edu 에 있는 값들에 대해 도수 분포표를 작성하여 보이시오

4. edu 에 있는 값들에 대해 막대 그래프를 작성하여 보이시오

2. 일변량 질적자료의 분석

- 원 그래프 작성

```
favorite.color <- c("red", "green", "yellow",
"red", "green", "red", "red")
sum <- table(favorite.color )        # 도수분포표
pie(sum, main="Favorite color")
```



원그래프 참고사이트
http://www.statmethods.net/graphs/pie.html

3. 일변량 양적 자료의 분석

- 양적자료는 질적 자료에 비해 분석 방법이 많다
  - 평균/중앙값
  - 4분위수
  - 분산, 표준편차
  - Boxplot
  - Histogram
  - 나무-잎 그림

3. 일변량 양적 자료의 분석 : 평균

모든 국민들의 소득자료를 가지고 있다. 이 자료를 요약해서 설명할수 있는 값은 무엇이 있을까?

- 평균(mean)
  - 균형점, 무게중심

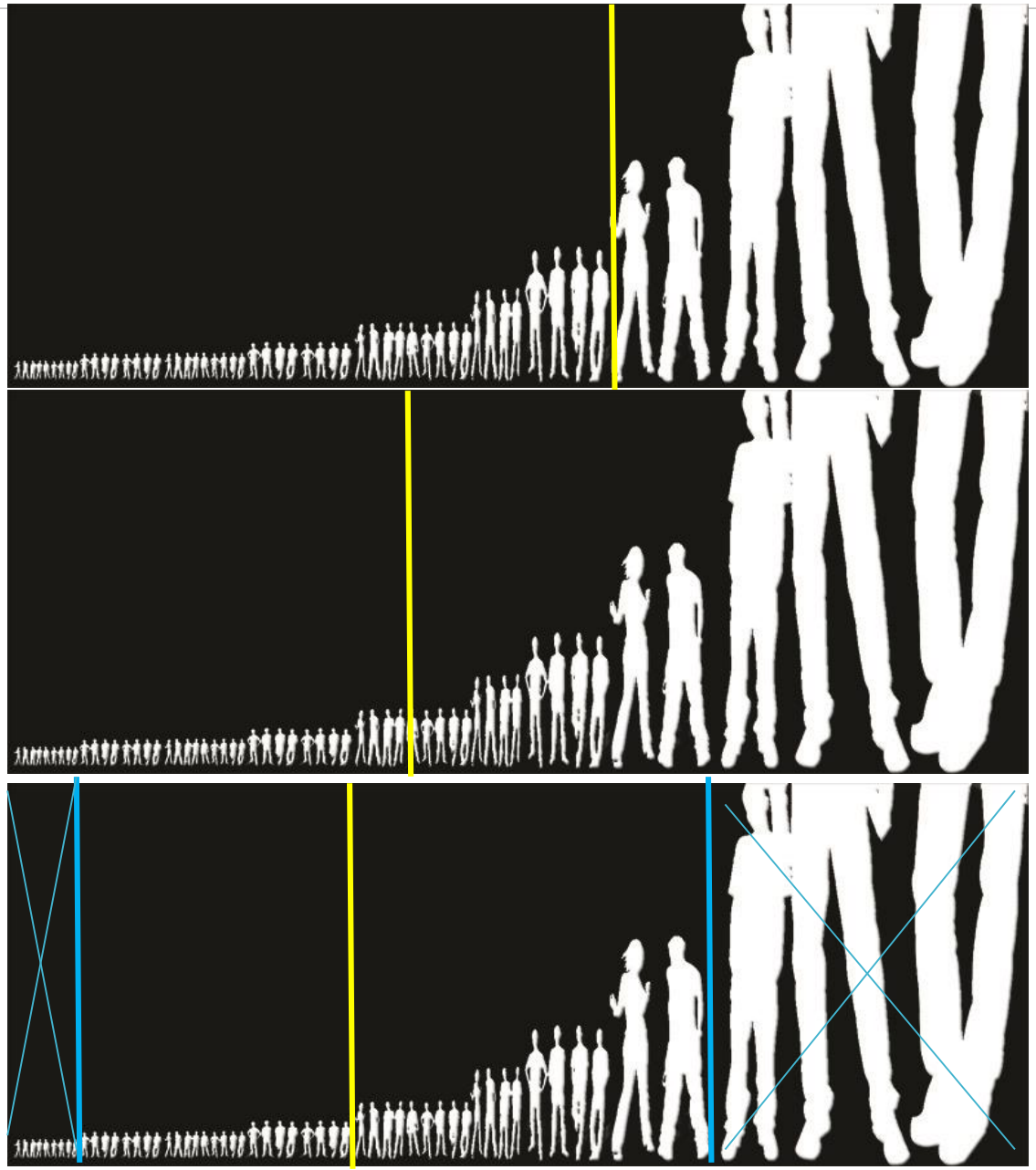
$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

- 중앙값(median)
  - 어떤 주어진 값들을 정렬했을 때 가장 중앙에 위치하는 값을 의미

- 절사평균(trimmed mean)
  - 표본중에서 작은값 n% 와 큰값 n%를 제외하고 나머지 (100-2n)% 의 자료만 사용하여 구한 평균

EBS – 평균의 함정(https://youtu.be/Pp_Pd6GZLOE)

3. 일변량 양적 자료의 분석 : 평균



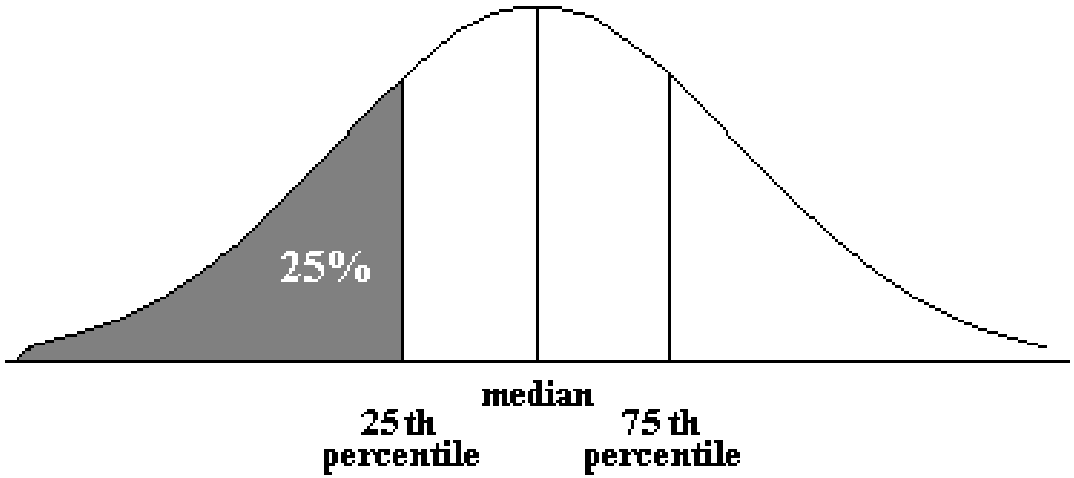
평균

중앙값

절사평균

3. 일변량 양적 자료의 분석 : 4 분위 수

- 4분위수 (quartile)
  - **측정값을 4등분하는 백분위수**
  - 제 1 사분위수 (Q1) : 제 25 백분위수
  - 제 2 사분위수 (Q2) : 제 50 백분위수, 중앙값
  - 제 3 사분위수 (Q3) : 제 75 백분위수



3. 일변량 양적 자료의 분석

- mean(), median(), quantile(), summary()

```
mydata = c(50,60,100,75,200)
mydata.big = c(mydata, 50000)
mean(mydata)                    # 평균
mean(mydata.big)
median(mydata)                  # 중앙값
median(mydata.big)
mean(mydata, trim=0.2)          # 절사평균
mean(mydata.big, trim=0.2)
quantile(mydata)                # 사분위수
quantile(mydata, (0:10)/10)
summary(mydata)
fivenum(mydata)                 # quantile()과 비슷
```

3. 일변량 양적 자료의 분석

- quantile()

```
> quantile(mydata)
  0%  25%  50%  75% 100%
  50   60   75  100  200
```

최소 중앙 최대

```
> quantile(mydata, (0:10)/10)
  0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
  50   54   58   63   69   75   85   95  120  160  200
```

- summary()

```
> summary(mydata)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     50      60      75      97     100     200
```

3. 일변량 양적 자료의 분석 : 산포(distribution)

- 산포
  - 데이터가 퍼져 있는 정도, 흩어져 있는 정도
  - 분산과 표준편차를 가지고 표현

- 분산 (variance)

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

- 표준편차(standard deviation)

$$S = \sqrt{(\text{분산})}$$

3. 일변량 양적 자료의 분석

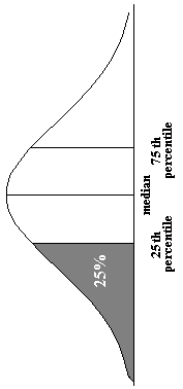
- diff(), var(), sd()

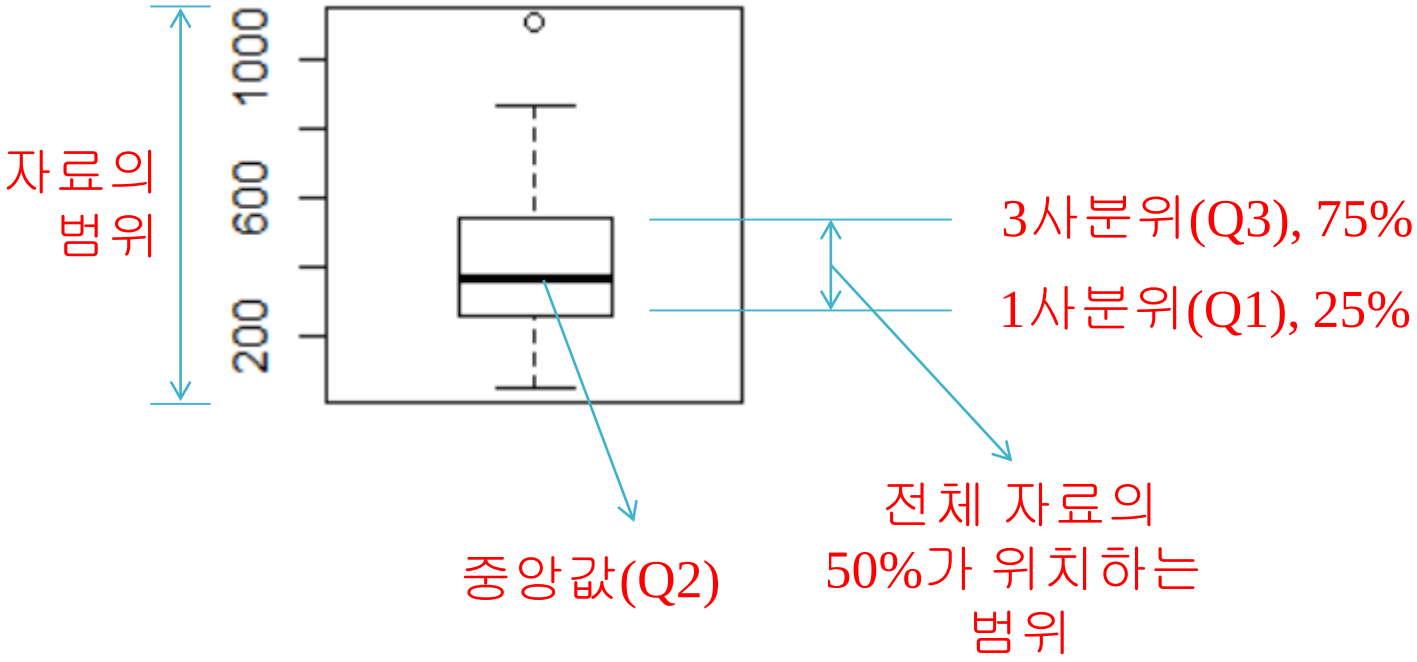
```
diff(range(mydata))        # 최대값-최소값
var(mydata)                # 분산
sd(mydata)                 # 표준편차
```

```
> diff(range(mydata))
[1] 150
> var(mydata)
[1] 3670
> sd(mydata)
[1] 60.58052
```

3. 일변량 양적 자료의 분석 : 상자 그림(box plot)

- Box plot 또는 Box whisker plot 이라고 불리움



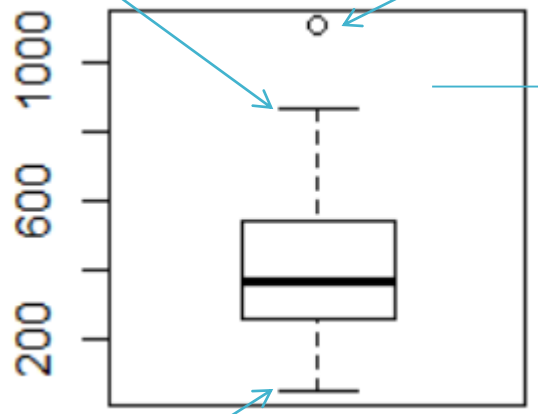


☞ box 의 넓이는 아무 의미가 없음

3. 일변량 양적 자료의 분석 : 상자 그림(box plot)

이상치를 제외한 값중
최대값

정상범위 밖에 존재하는
데이터 표시. 이상치.특이값(outlier)



Q3 + 1.5 * IQR

정상적인 데이터가
분포할 것으로 기대되는 범위
(표시되지 않음)

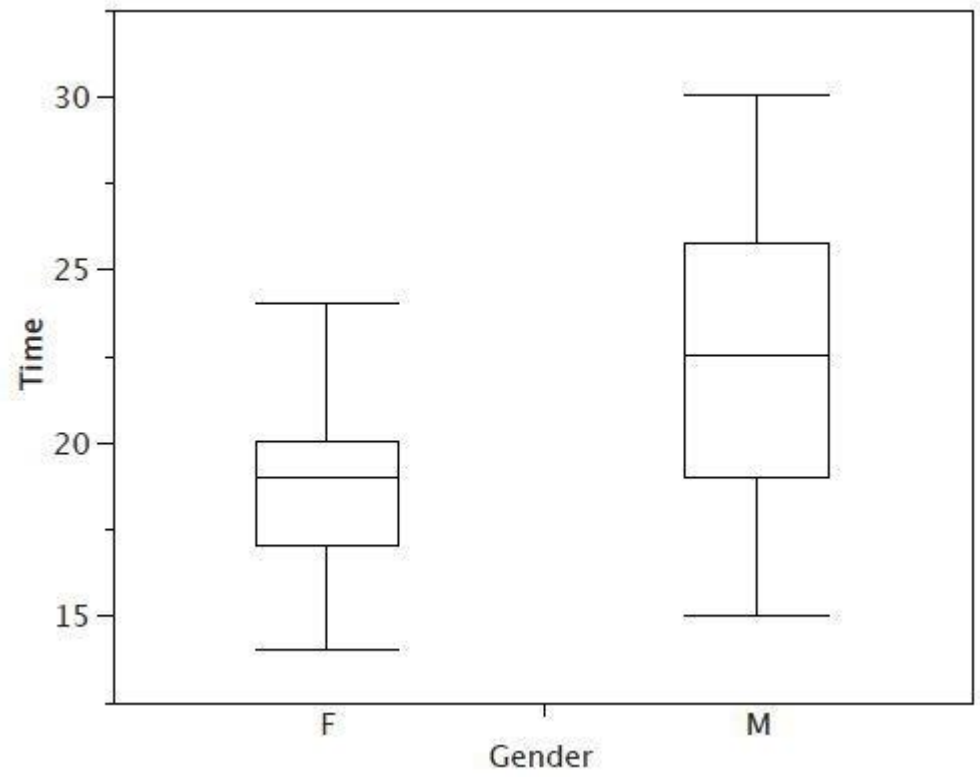
Q1 - 1.5 * IQR

이상치를 제외한 값중
최소값

(IQR = Q3-Q1)

3. 일변량 양적 자료의 분석 : 상자 그림(box plot)

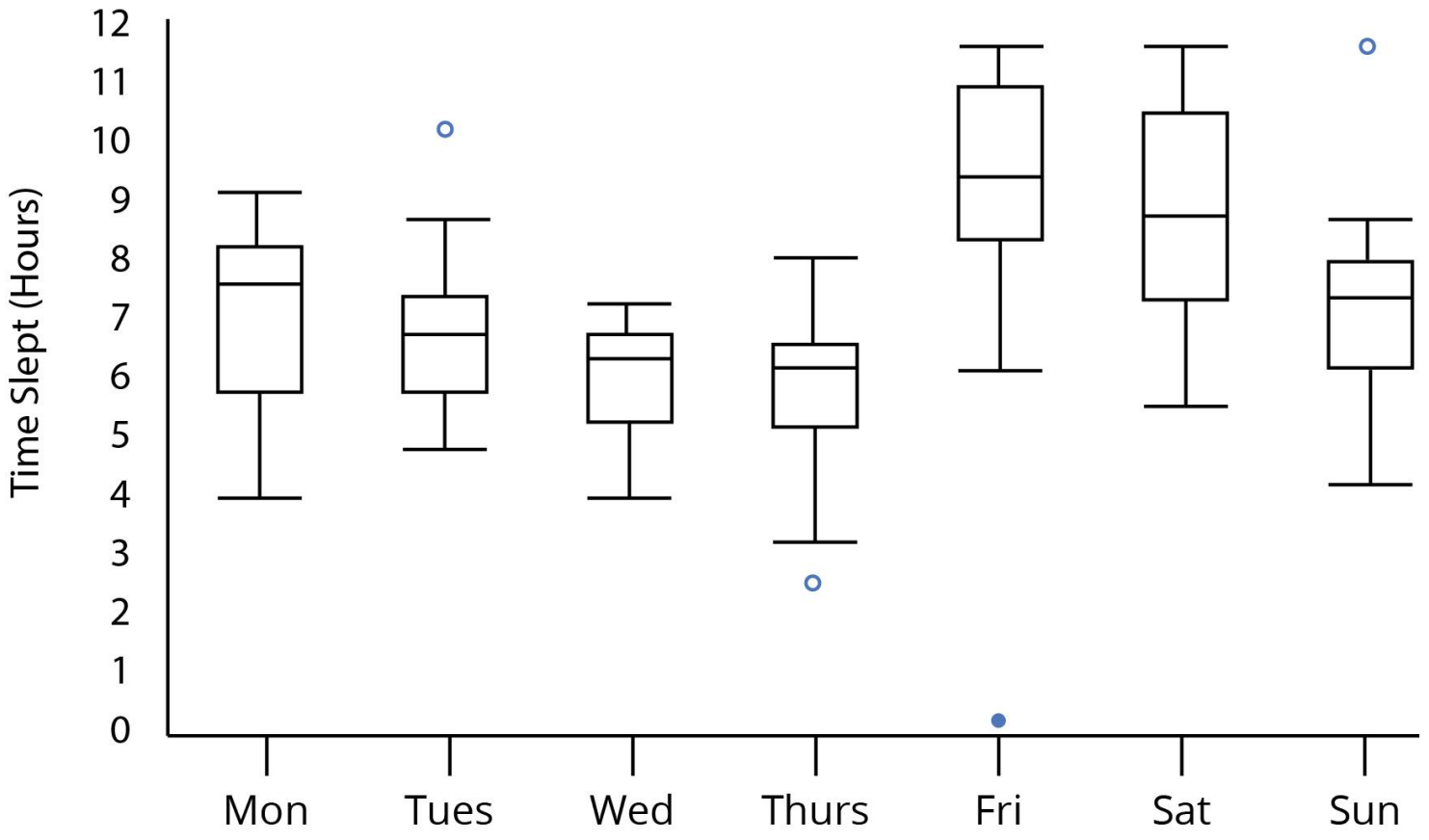
- 사례1



시간제 일자리 근무시간(남녀)

3. 일변량 양적 자료의 분석 : 상자 그림(box plot)

- 사례2



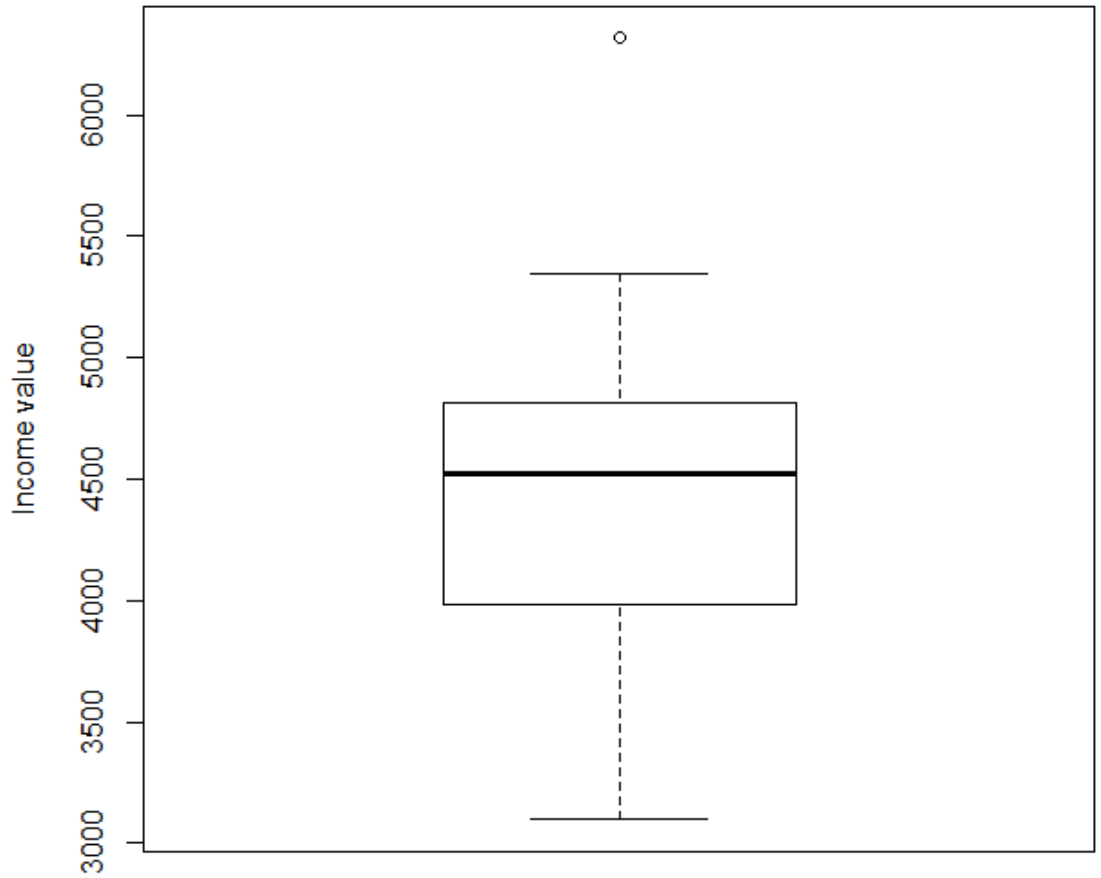
20명의 학생에 대한 수면시간

3. 일변량 양적 자료의 분석 : 상자 그림(box plot)

- state.x77 데이터셋에서 주별 소득에 대해 boxplot 을 그려보자

```
head(state.x77)
st.income <- state.x77[,"Income"]
boxplot(st.income, ylab="Income value")
```

3. 일변량 양적 자료의 분석 : 상자 그림(box plot)



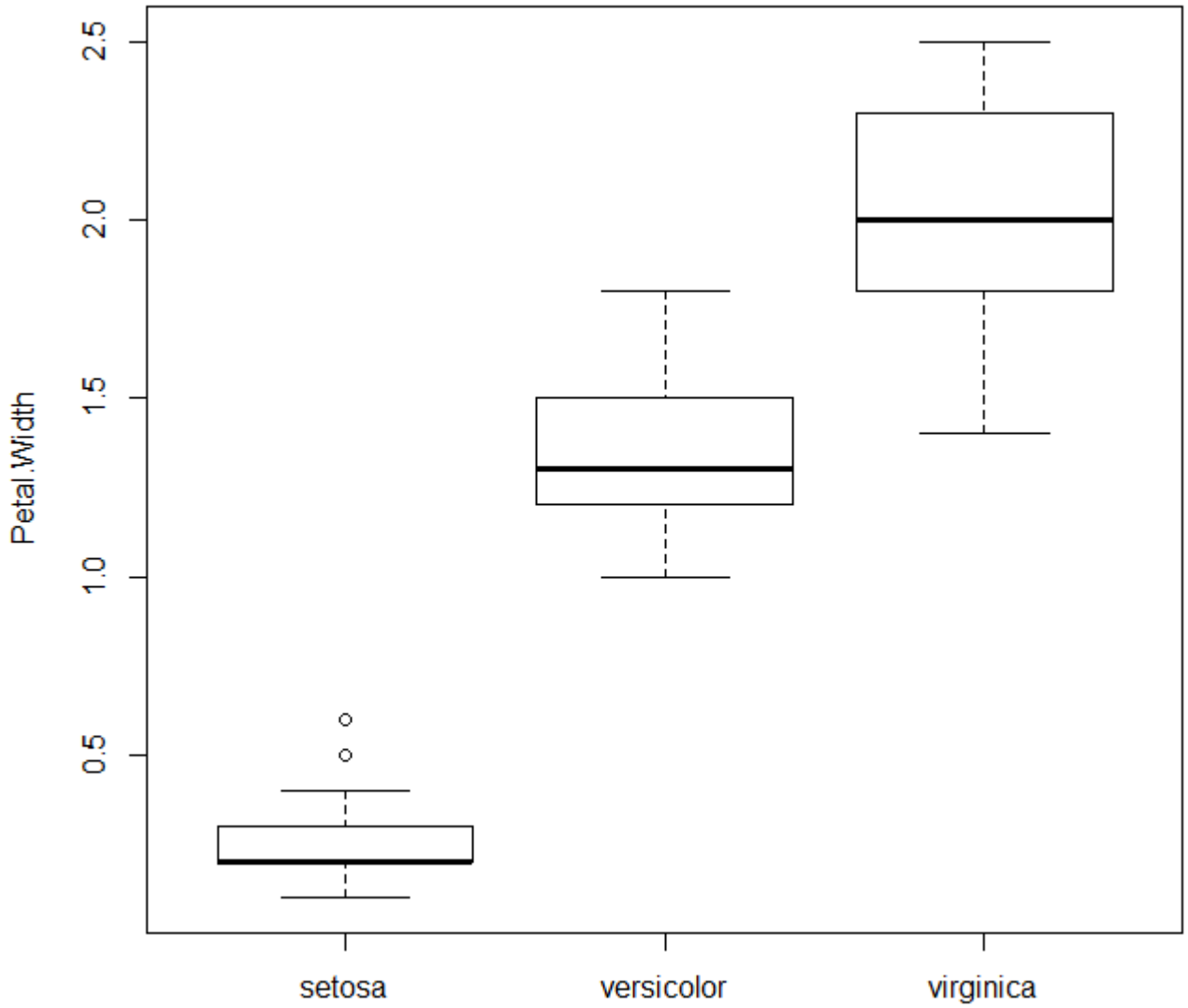
3. 일변량 양적 자료의 분석 : 상자 그림(box plot)

- iris dataset에서 품종(Species)에 따른 Petal.Width 자료에 대한 boxplot 을 그려 보시오 (데이터에 그룹이 있는 경우)

```
boxplot(Petal.Width~Species,data=iris,
          ylab="Petal.Width")
```

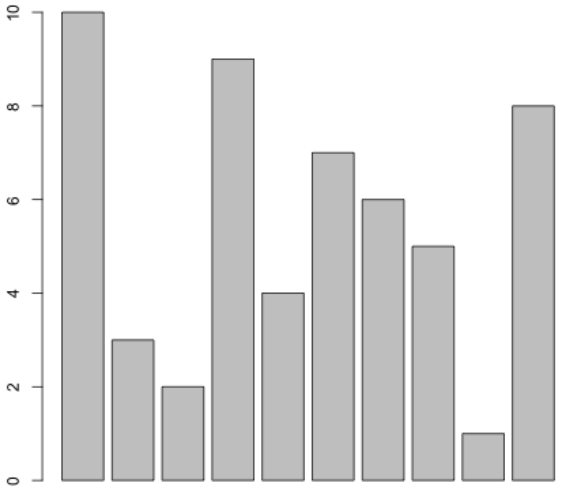
```
> head(iris)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
```

3. 일변량 양적 자료의 분석 : 상자 그림(box plot)



3. 일변량 양적 자료의 분석 : 히스토그램

- 막대 그래프는 도수 분포표를 만들 수 있는 정수형, 문자형 자료의 경우에 사용하고, 실수형 자료에 대해서는 히스토그램을 사용한다



barplot

histogram

막대그래프는 막대간 간격이 있고 막대의 면적은 의미가 없다
히스토그램은 막대가 붙어 있고 막대의 면적이 의미가 있다

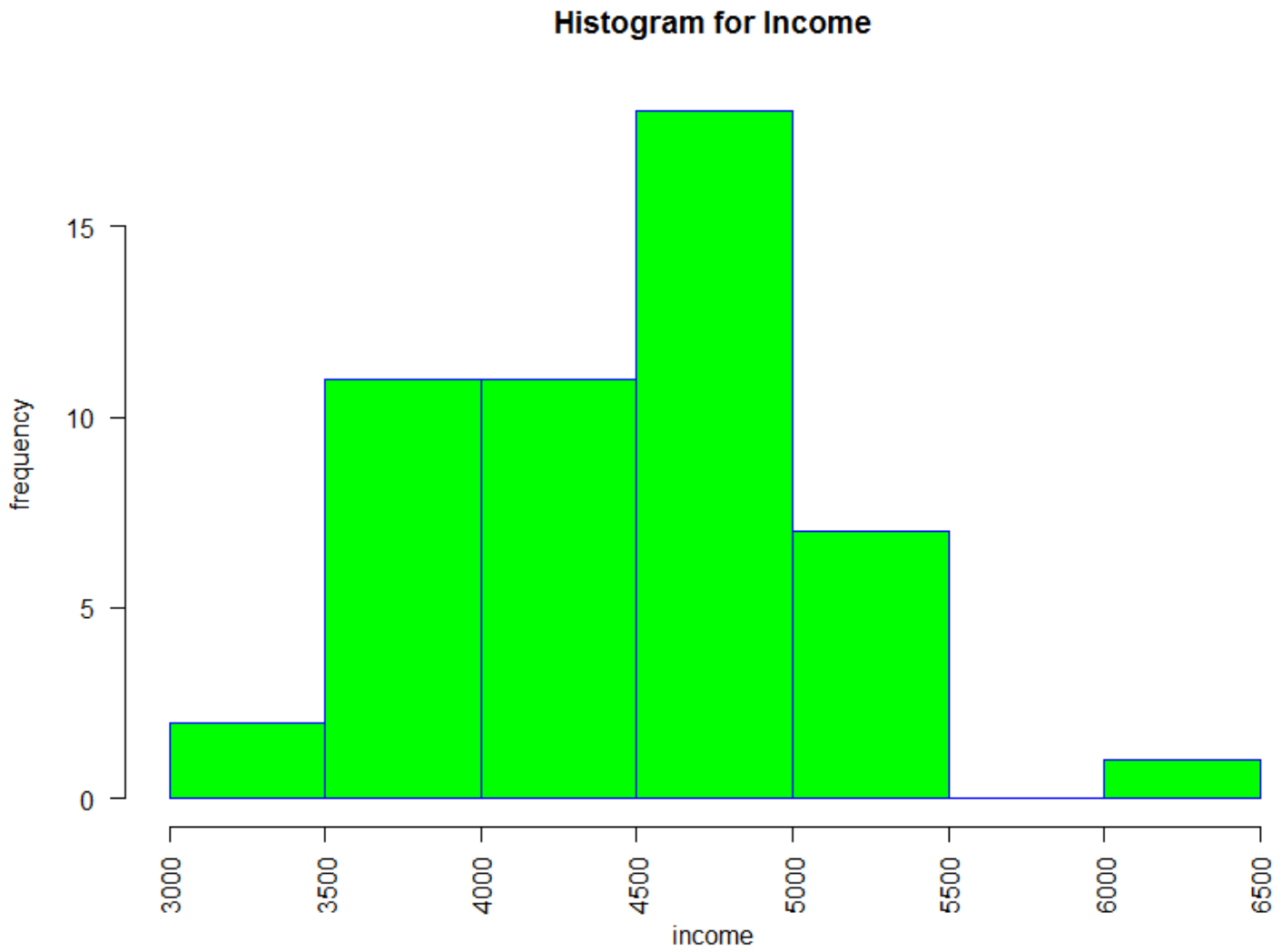
3. 일변량 양적 자료의 분석 : 히스토그램

- 히스토그램 그리기

```
st.income <- state.x77[,"Income"]
hist(st.income,                                  # data
     main="Histogram for Income",                # 제목
     xlab ="income",          # x축 레이블
     ylab="frequency",        # y축 레이블
     border="blue",           # 막대 테두리색
     col="green",             # 막대 색
     las=2,                   # x축 글씨방향(0~3)
     breaks=5)                # x축 막대 개수 조절
```

breaks=n 일때 막대 개수는 $\log_2(n)+1$ 로 계산
N이 커질수록 막대의 개수가 늘어난다

3. 일변량 양적 자료의 분석 : 히스토그램



3. 일변량 양적 자료의 분석 : 줄기-잎 그림

- 줄기-잎 그림

```
score <- c(40,55,90,75,59,60,63,65,69,71)
stem(score, scale=2)
```

```
> score <- c(40,55,90,75,59,60,63,65,69,71)
> stem(score, scale=2)

  The decimal point is 1 digit(s) to the right of the |

  4 | 0
  5 | 59
  6 | 0359
  7 | 15
  8 |
  9 | 0

>
```

Scale 값이 커지면 줄기의 수 증가 (줄기당 잎 수 감소)

3. 일변량 양적 자료의 분석 : 줄기-잎 그림

```
> stem(score,scale=0.5)

The decimal point is 1 digit(s) to the right of the |

0 | 00058
2 | 13334588890355789
4 | 11133456678111222333444456688
6 | 14777933478
8 | 2909
```

```
> stem(score,scale=2)

The decimal point is 1 digit(s) to the right of the |

0 | 00
0 |
1 | 0
1 | 58
2 | 13334
2 | 58889
3 | 03
3 | 55789
4 | 111334
4 | 56678
5 | 111222333444
5 | 56688
6 | 14
```

[연습2]

- 홍길동군의 과목별 성적은 다음과 같다.

KOR	ENG	MATH	HIST	SOC	MUSIC	BIO	EARTH	PHY	ART
90	85	73	80	85	65	78	50	68	96

1. 이 데이터를 score 벡터에 저장하시오. (과목명은 데이터 이름으로 저장하시오)
2. score 벡터의 내용을 보이시오
3. 전체 성적의 평균은 얼마인가
4. 전체 성적의 중앙값은 얼마인가
5. 전체 성적의 표준편차를 보이시오
6. 가장 성적이 높은 과목의 이름을 보이시오
7. 성적에 대한 boxplot 을 그리시오. 이상치에 해당하는 과목이 있으면 제시하시오
8. 성적에 대한 histogram 을 그리되 다음조건을 만족하도록 하시오
   (그래프 title : Hong's score, 막대색: 보라색)

[연습3]

- mtcars 데이터셋을 이용하여 다음 문제를 해결하시오

1. 중량(wt) 의 평균값, 중앙값, 절사평균값(절사범위:15%), 표준편차 를 구하시오
2. 중량(wt)에 대해 summary() 함수의 적용 결과를 보이시오
3. 실린더수(cyl) 에 대해 도수분포표를 구하시오
4. 앞에서 구한 도수분포표를 막대그래프로 그려 보시오
5. 중량(wt)의 히스토그램, 실린더(cyl), 기어(gear) 에 대한 막대 그래프를 한 화면에 보이게 작성하시오
6. 중량(wt)에 대해 boxplot 을 그려 보시오. Boxplot 으로 부터 관찰할 수 있는 정보를 적으시오
7. 배기량(disp)에 대해 boxplot 을 그려 보시오. Boxplot 으로 부터 관찰할 수 있는 정보를 적으시오

[tip] 문자열 함수

- paste() 함수 : 여러 문자열을 연결하여 하나로 만들 때 사용

```
paste("Good", "Morning", "Tom", sep=" ")
paste("Good", "Morning", "Tom", sep="/")
paste(1:10, "is good", sep=" ")
```

sep : 연결하는 단어 사이사이에 넣을 값을 지정

```
> paste("Good", "Morning", "Tom", sep=" ")
[1] "Good Morning Tom"
> paste("Good", "Morning", "Tom", sep="/")
[1] "Good/Morning/Tom"
> paste(1:10, "is good", sep=" ")
 [1] "1 is good"  "2 is good"  "3 is good"  "4 is good"  "5 is good"
 [6] "6 is good"  "7 is good"  "8 is good"  "9 is good"  "10 is good"
```

[tip] 문자열 함수

- substr() : 문자열 자르기
- nchar() : 문자열 길이

```
str <- "Good Morning"
substr(str, 1,4)
substr(str, 6,nchar(str))
```

```
> str <- "Good Morning"
> substr(str, 1,4)
[1] "Good"
> substr(str, 6,nchar(str))
[1] "Morning"
>
```

[tip] 문자열 함수

- gsub() : 문자열 바꾸기(replace)

```
str <- "Good Morning"
gsub ("Good", "nice", str)
str <- gsub(" ", "/", str)
str
```

```
> str <- "Good Morning"
> gsub ("Good", "nice", str)
[1] "nice Morning"
> str <- gsub (" ", "/", str)
> str
[1] "Good/Morning"
>
```