3강 탐색적 자료분석(EDA)의 시각화 2

정보통계학과 이태림 교수

1. 탐색적 자료분석의 그래프에 의한 표현 이해
2. 히스토그램, 상자그림, 줄기그림 R에 의한 표현
3. 작성된 그래프에 의한 자료의 특징 파악

학습개요

히스토그램

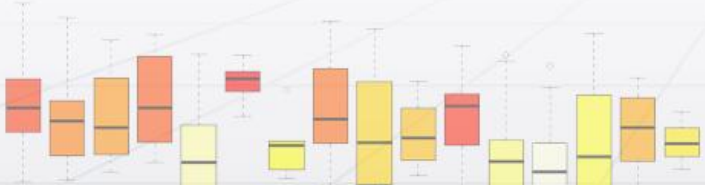
- 두 군의 분포 비교
- 임의수 생성에 따른 히스토그램

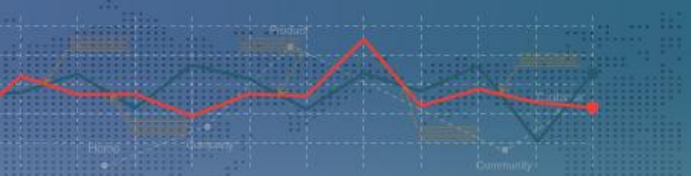
상자그림

- 상자그림 (Boxplot)
- 바이올린그림 (violin graph)

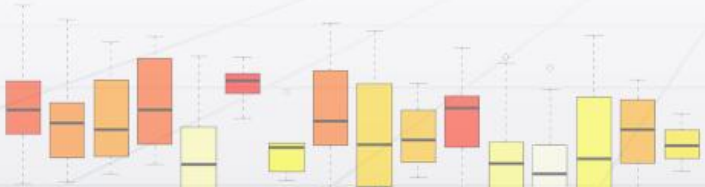
줄기 잎 그림

- 줄기그림 (stem plot)
- 줄기그림과 히스토그램 비교





1. 히스토그램



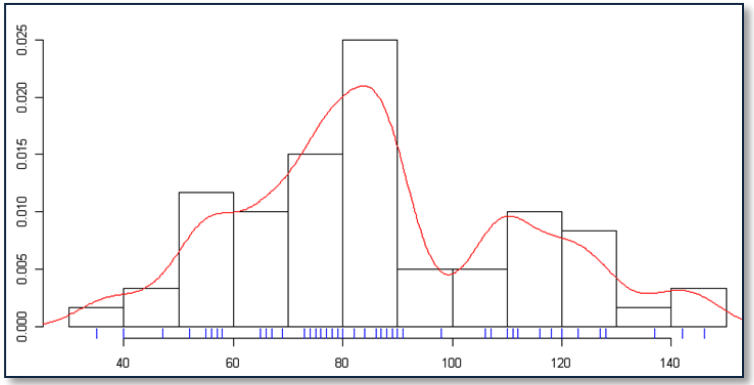
1 히스토그램의 정의

히스토그램 (Histogram) :

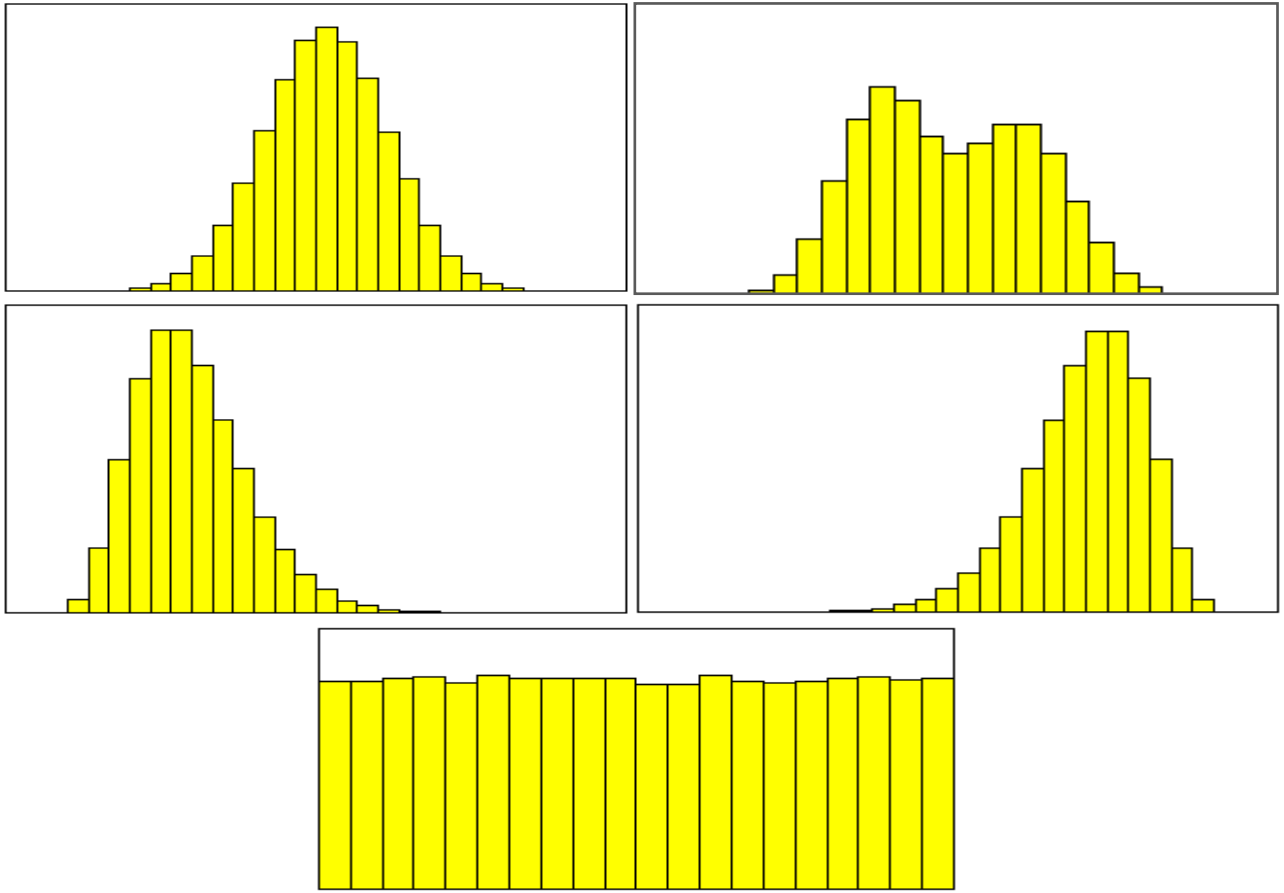
Histogram

- 연속형 관찰값의 구간별 도수를 상대적인 막대의 길이로 나타낸 그래프

- 분포의 개형을 파악하는데 도움

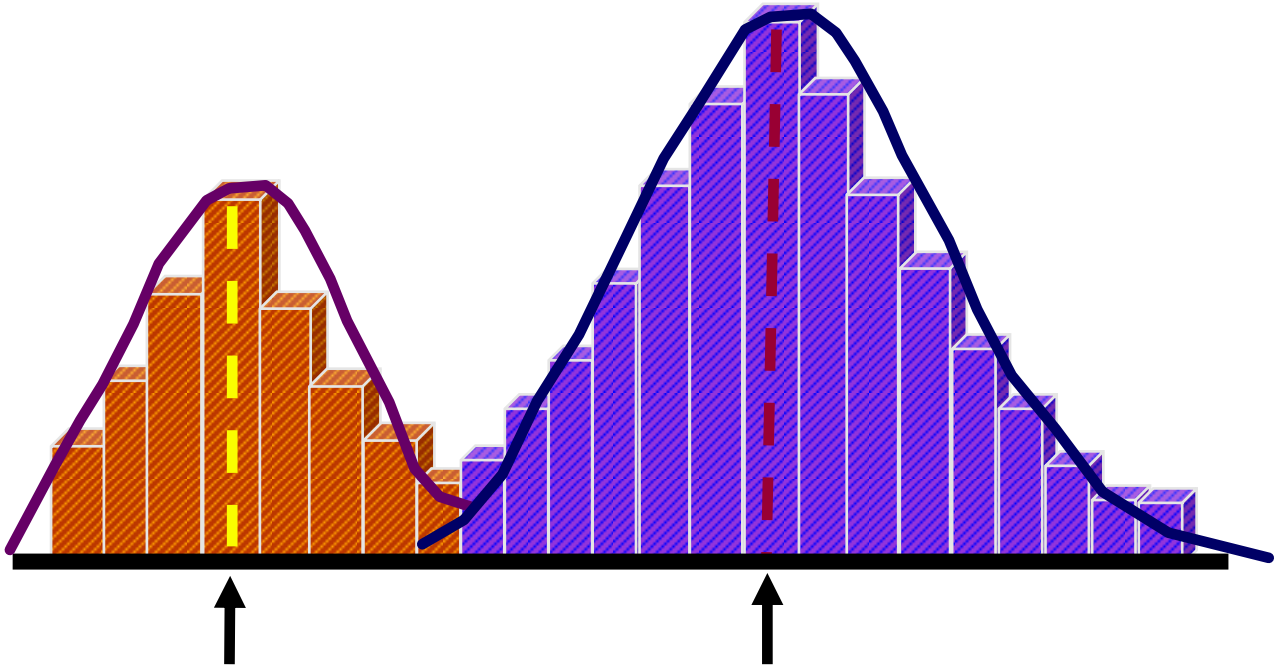


1 히스토그램의 유형



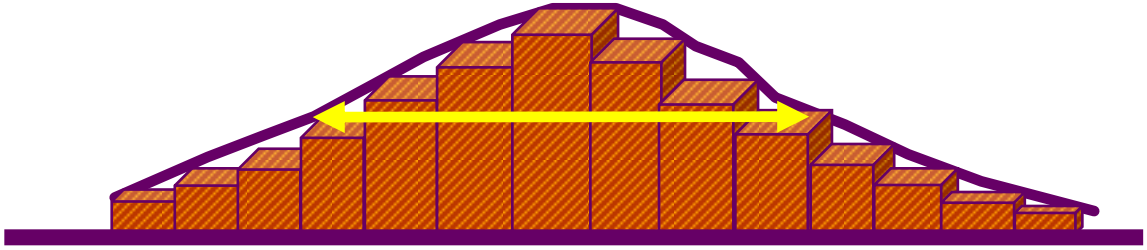
1 히스토그램의 검토요령

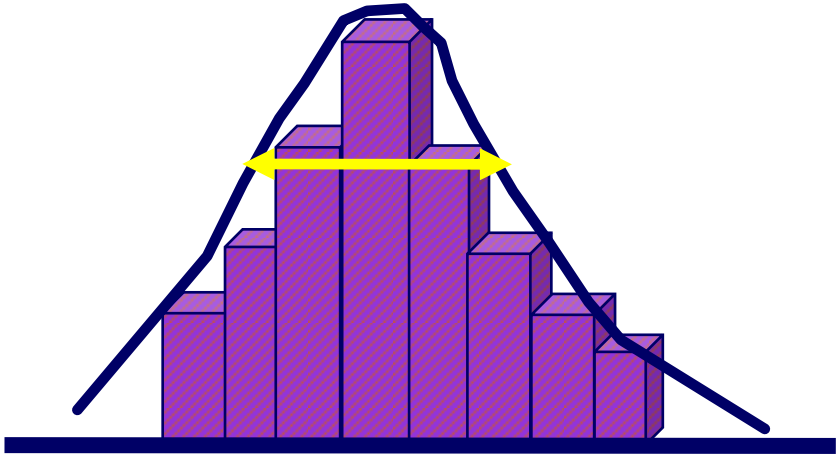
자료의 중심위치



1 히스토그램의 검토요령

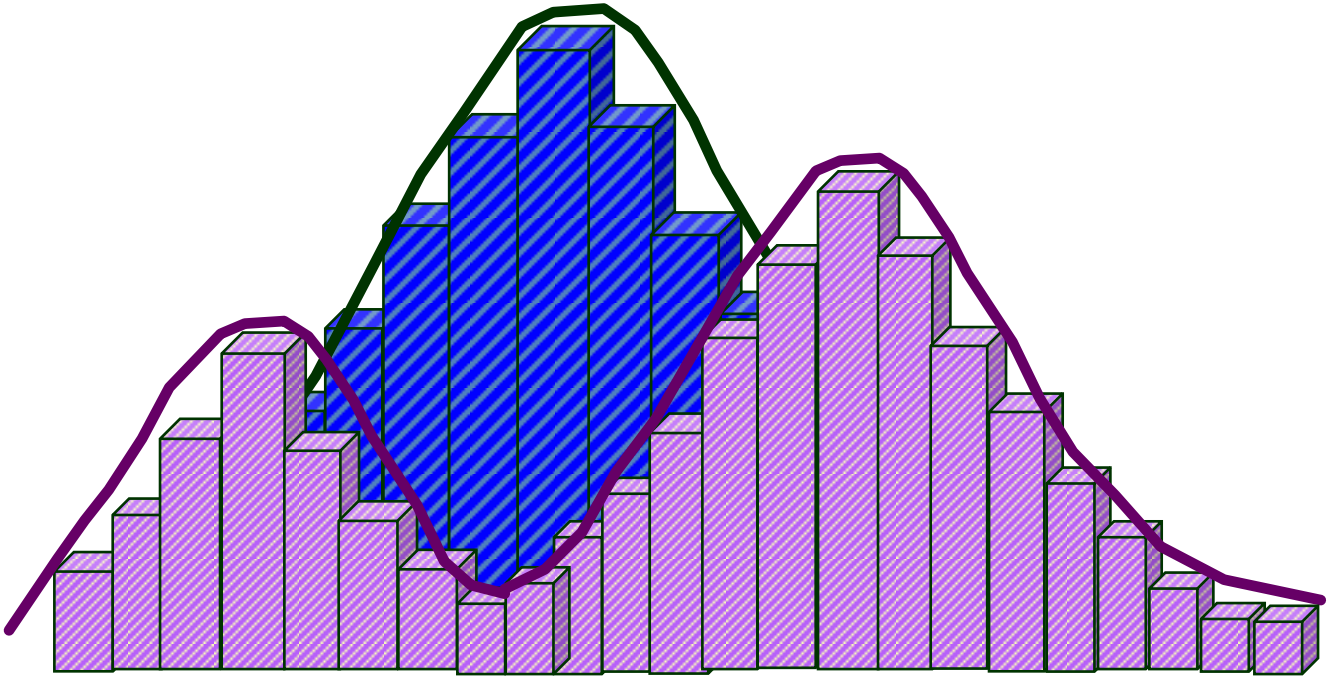
자료의 산포





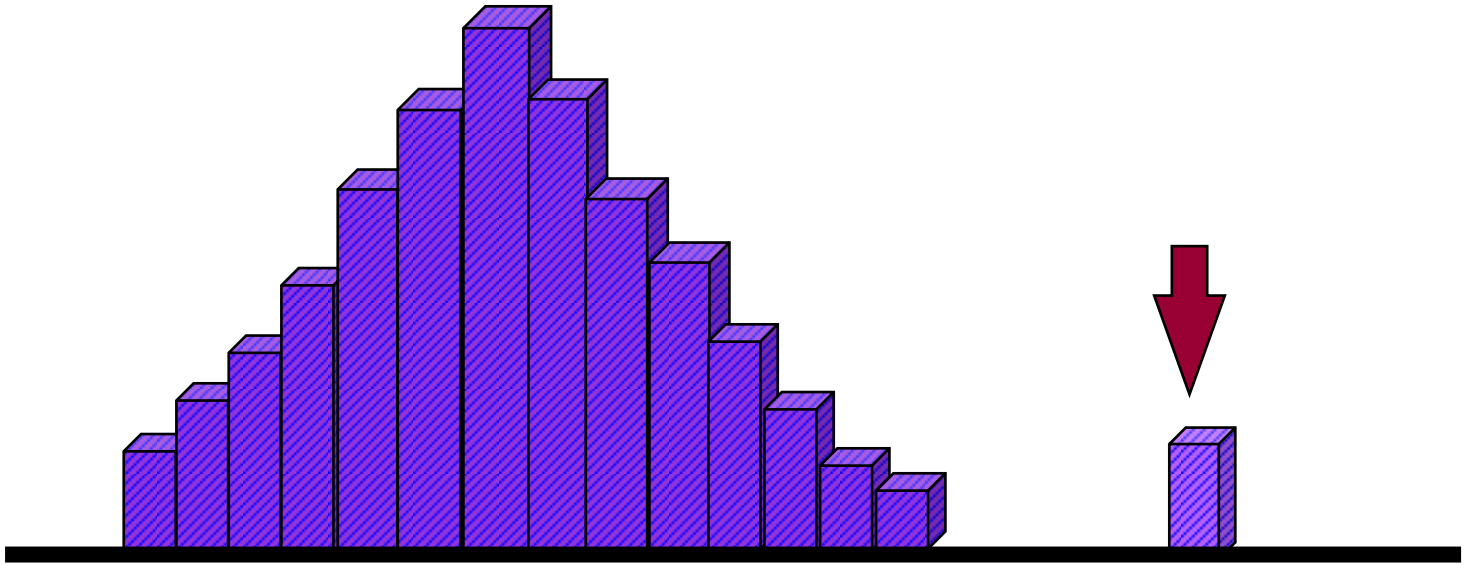
1 히스토그램의 검토요령

봉우리의 갯수



1 히스토그램의 검토요령

이상치의 점검



1 히스토그램 작성 예

담즙 과포화 비율 자료

자료

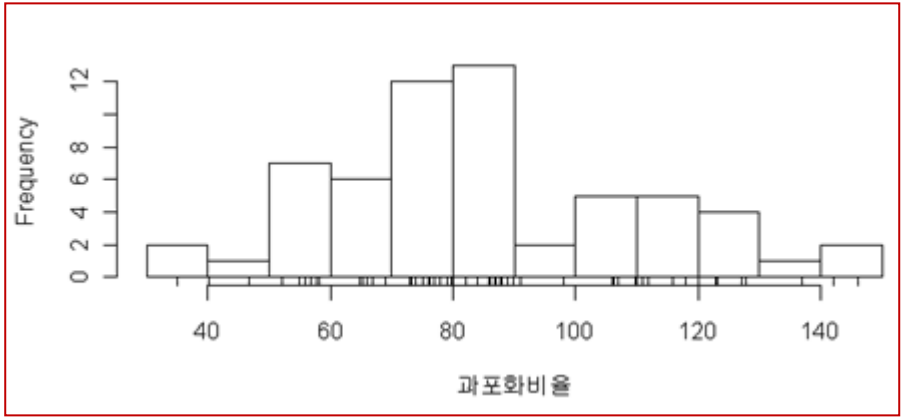
- 성인 남자 31명과 성인 여자 29명에 대한 담즙의 과포화 비율 자료

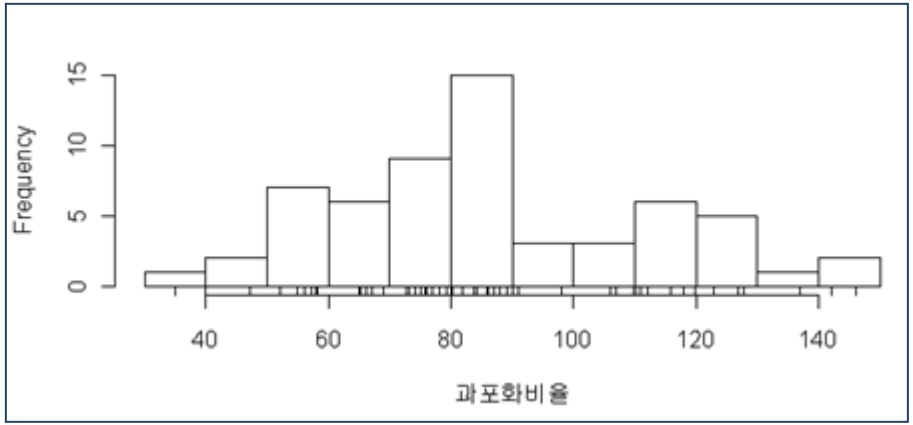
성별	담즙 과포화 비율
남자	40 86 111 86 106 66 123 90 112 52 88 137 88 80 65 79 87 56 110 106 110 78 80 47 74 58 88 73 118 67 57
여자	65 86 76 89 142 58 98 146 80 66 52 35 55 127 77 91 128 75 82 69 84 116 73 87 76 107 84 120 123

1 히스토그램 작성 R 프로그램

```
# 외부파일을 읽어 데이터프레임을 만들기
담즙과포화비율=read.table("k:₩₩WORK₩₩담즙과포화비율-자료.txt", header=T)
담즙과포화비율
attach(담즙과포화비율)
str(담즙과포화비율)
# 담즙과포화비율 자료의 크기
n=length(과포화비율)
n
# 담즙과포화비율 자료의 정렬
sort(과포화비율)
sort(과포화비율, decreasing=T)
# 담즙과포화비율 히스토그램과 개별자료
par(mfrow=c(2,1))
hist(과포화비율,breaks=class1,main=NULL)
rug(jitter(과포화비율))
hist(과포화비율,breaks=class1,right=F,main=NULL)
rug(jitter(과포화비율))
```

1 히스토그램 작성 결과


Frequency
0
2
4
6
8
12
40
60
80
100
120
140
과포화비율

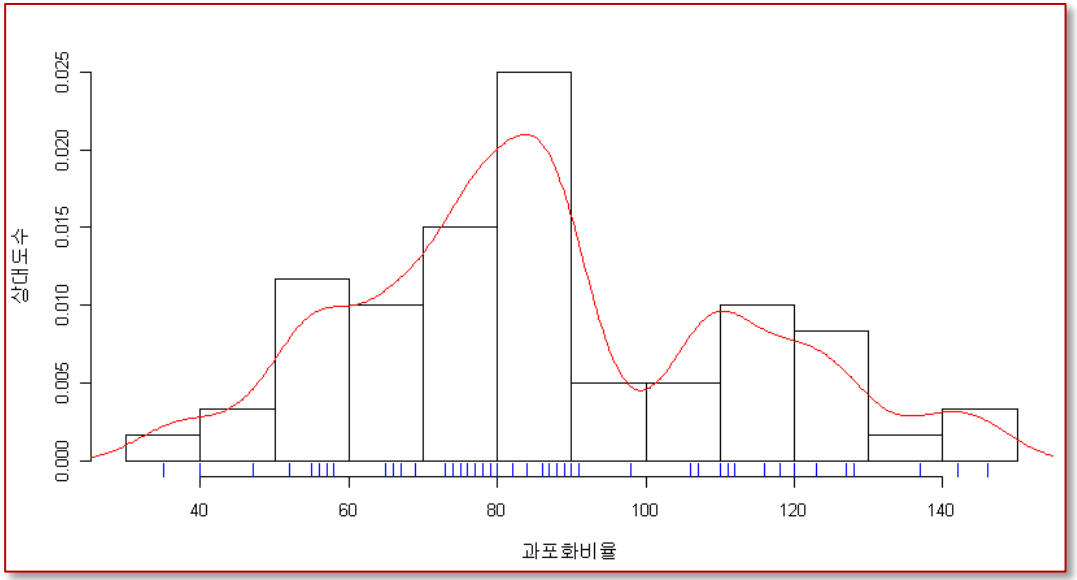

Frequency
0
5
10
15
40
60
80
100
120
140
과포화비율

1 히스토그램 작성 R 프로그램

```
# 담즙과포화비율 히스토그램과 개별 자료
par(mfrow=c(2,1))
hist(과포화비율, breaks=class1, main=NULL)
rug(jitter(과포화비율))
hist(과포화비율, breaks=class1, right=F, main=NULL)
rug(jitter(과포화비율))
# 상대도수밀도히스토그램, 커널밀도추정량과 개별 자료
m=matrix(c(1, 3, 2, 3), ncol=2, byrow=T)
layout(mat=m)
hist(과포화비율, prob=T,breaks=class1,right=F, ylab="상대도수",
     main=NULL)
lines(density(과포화비율,bw=5), col="red")
rug(과포화비율, col="blue")
```

1 히스토그램 작성 결과

담즙과포화 비율의 히스토그램



1 히스토그램 작성 예

간헐온천의 온천물 분출 지속시간 자료

https://www.youtube.com/watch?v=dAUlLQIj_wk

1 히스토그램 작성 예

간헐온천의 온천물 분출 지속시간 자료

자료

온천물 분출 지속시간을 분 단위로 잰 107개 자료

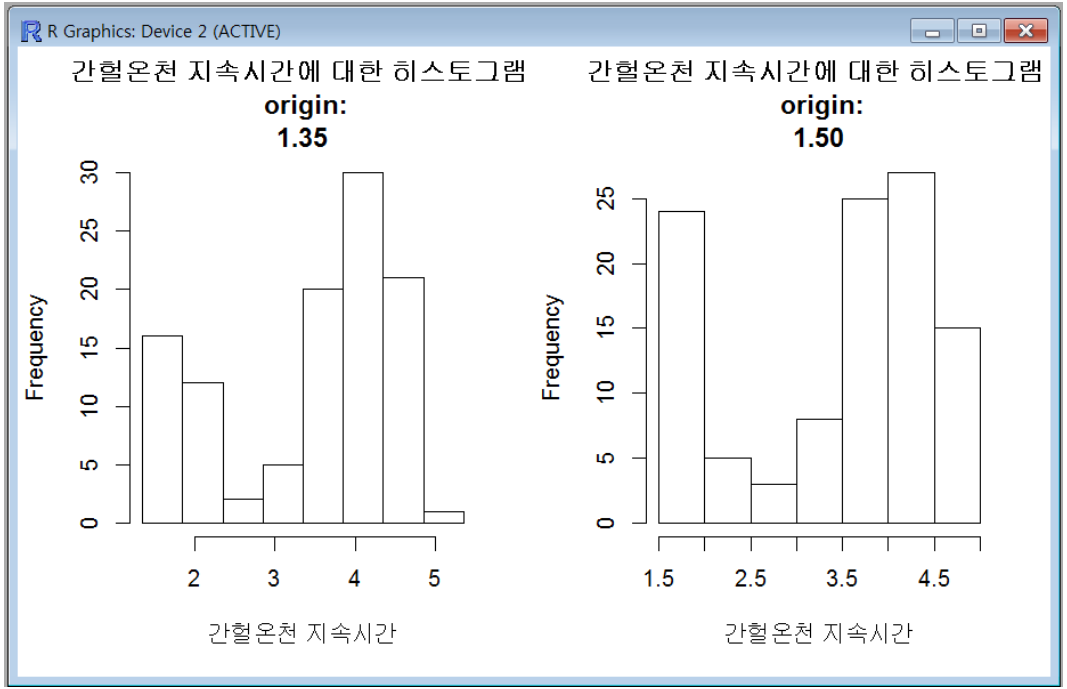
4.37	3.87	4.00	4.03	3.50	4.08	2.25	4.70	1.73	4.93
1.73	4.62	3.43	4.25	1.68	3.92	3.68	3.10	4.03	1.77
4.08	1.75	3.20	1.85	4.62	1.97	4.50	3.92	4.35	2.33
3.83	1.88	4.60	1.80	4.73	1.77	4.57	1.85	3.52	4.00
3.70	3.72	4.25	3.58	3.80	3.77	3.75	2.50	4.50	4.10
3.70	3.80	3.43	4.00	2.27	4.40	4.05	4.25	3.33	2.00
4.33	2.93	4.58	1.90	3.58	3.73	3.73	1.82	4.63	3.50
4.00	3.67	1.67	4.60	1.67	4.00	1.80	4.42	1.90	4.63
2.93	3.50	1.97	4.28	1.83	4.13	1.83	4.65	4.20	3.93
4.33	1.83	4.53	2.03	4.18	4.43	4.07	4.13	3.95	4.10
2.72	4.58	1.90	4.50	1.95	4.83	4.12			

1 히스토그램 작성 R 프로그램

```
# eruption lengths(in minutes) of 107 eruptions of Old Faithful geyser
eruption.length=c(4.37,3.87,4.00,4.03,3.50,4.08,2.25,4.70,1.73,4.93,1.73,4.62,3.43,4.25,1.68,3.92,3.
68,3.10,4.03,1.77,4.08,1.75,3.20,1.85,4.62,1.97,4.50,3.92,4.35,2.33,3.83,1.88,4.60,1.80,4.73,1.77,4.5
7,1.85,3.52,4.00,3.70,3.72,4.25,3.58,3.80,3.77,3.75,2.50,4.50,4.10,3.70,3.80,3.43,4.00,2.27,4.40,4.05,
4.25,3.33,2.00,4.33,2.93,4.58,1.90,3.58,3.73,3.73,1.82,4.63,3.50,4.00,3.67,1.67,4.60,1.67,4.00,1.80,4.
42,1.90,4.63,2.93,3.50,1.97,4.28,1.83,4.13,1.83,4.65,4.20,3.93,4.33,1.83,4.53,2.03,4.18,4.43,4.07,4.1
3,3.95,4.10,2.72,4.58,1.90,4.50,1.95,4.83,4.12)
eruption.length
# 계급의 폭을 0.5로 하고 제 1계급의 하한값(원점이라고도 함)을 1.35로 하는 도수분포표
cat.class1=cut(eruption.length, breaks=class1)
t1=table(cat.class1)
T1
# 히스토그램
par(mfrow=c(1,2))
hist(eruption.length, breaks=class1, main="간헐온천 지속시간에 대한 히스토그램 ₩n origin:
1.35", xlab="간헐온천 지속시간")
hist(eruption.length, breaks=class2, main="간헐온천 지속시간에 대한 히스토그램 ₩n origin:
1.50", xlab="간헐온천 지속시간")
```

1 히스토그램 작성 결과

간헐온천 지속시간에 대한 히스토그램 결과

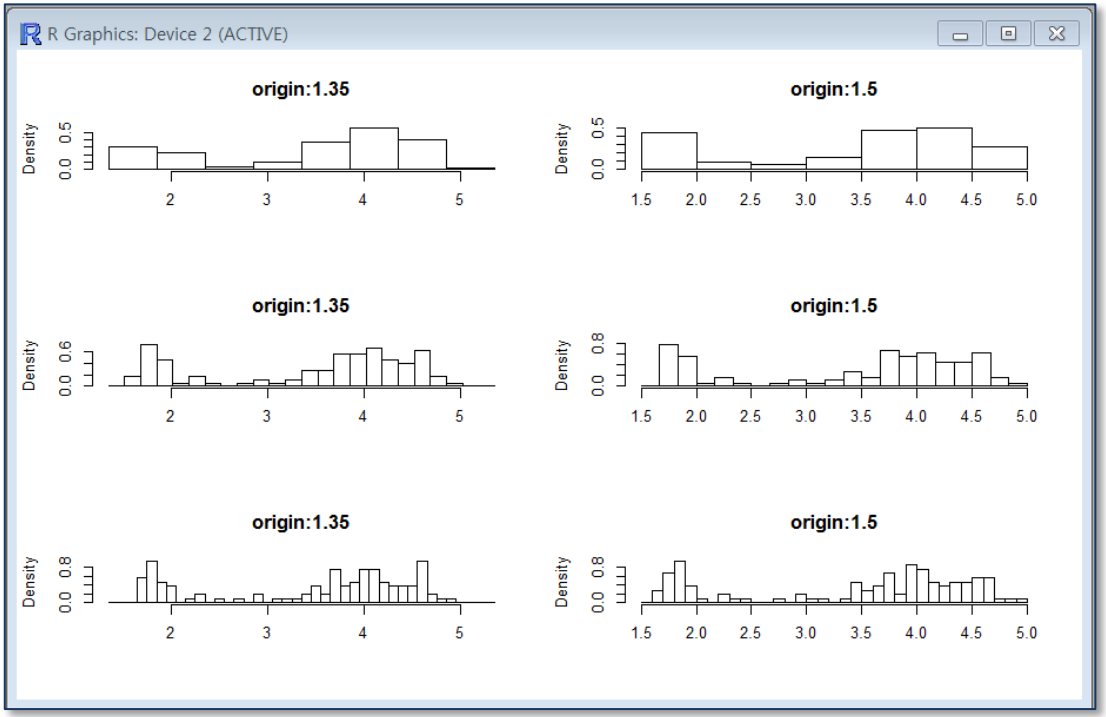


1 히스토그램 작성 R 프로그램

```
# 각각 제 1계급의 하한값을 1.35와 1.50으로 하는 히스토그램에서 계급의 폭 변동하기
hist.func2=function(n)
{par(mfrow=c(n,2))
for(i in 1:n)
{class1=seq(1.35,5.35,by=0.5/(2*i-1))
class2=seq(1.5,5.0,by=0.5/(2*i-1))
hist(eruption.length, breaks=class1, probability=T,main="origin:1.35",xlab=NULL)
hist(eruption.length, breaks=class2, probability=T,main="origin:1.5",xlab=NULL)}}
hist.func2(3)
# 도수분포표(계급의 폭: 0.1,0.2,0.3,0.4,0.5)
w=c(0.1,0.2,0.3,0.4,0.5)
for(i in 1:5)
{
class1=seq(1.5,5.1,by=w[i])
cat.class1=cut(eruption.length, breaks=class1)
table(cat.class1)
cat("계급의 폭=", w[i], "₩n")
print(table(cat.class1))
}
```

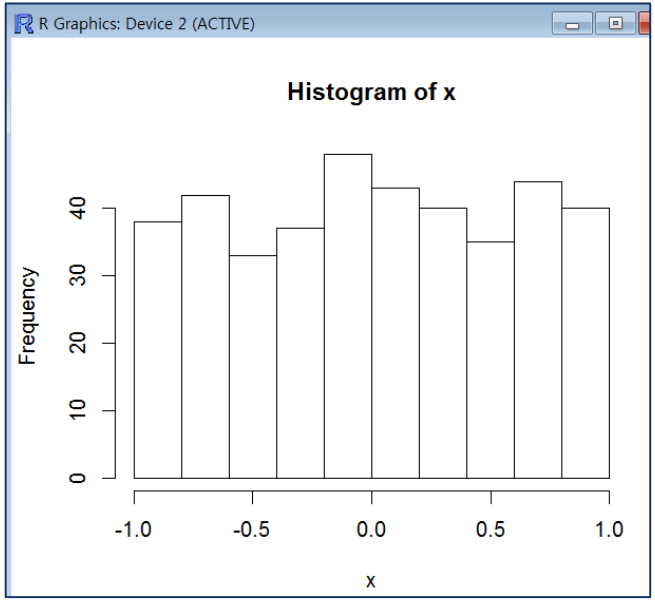
1 히스토그램 작성 결과

계급폭에 따른 간헐온천 지속시간의 히스토그램 유형



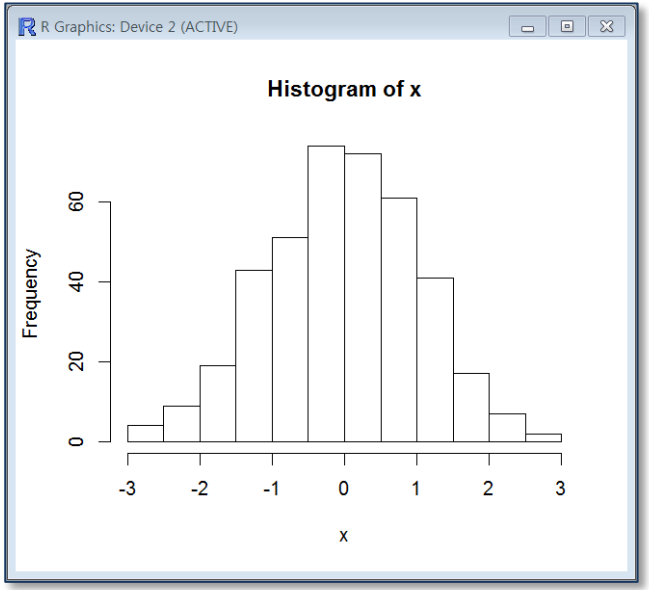
1 임의의 수 생성에 의한 히스토그램 작성

```
#임의수 생성
runif(n, min, max)
# 균일분포로부터 난수 생성 자료의 분포 히스토그램
>x<- runif(400, -1,1)
>hist(x)
```



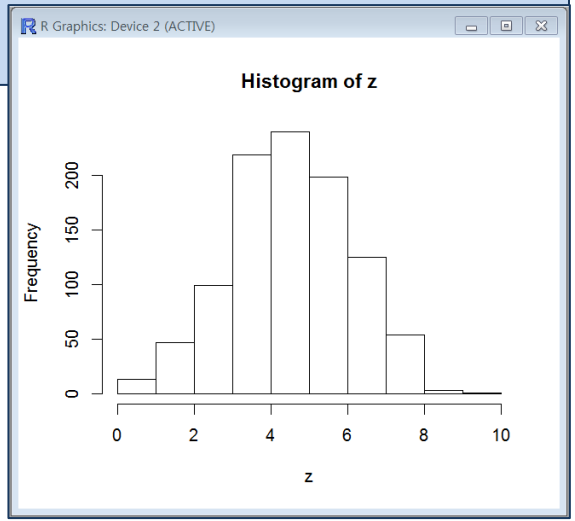
1 임의의 수 생성에 의한 히스토그램 작성

```
#정규분포로부터의 임의수 생성
rnorm(n, mean, sd)
# 정규분포로부터 난수 생성 자료의 분포 히스토그램
>x<- rnorm(400)
>hist(x)
```



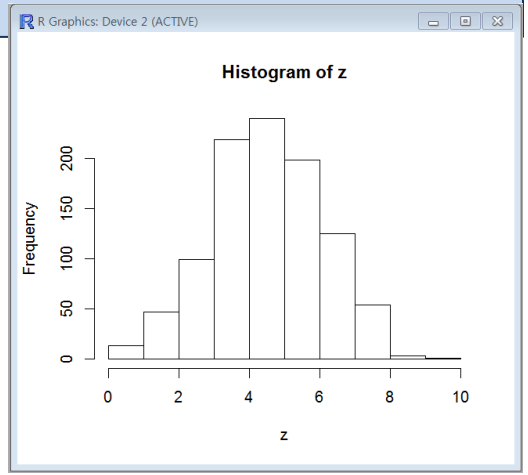
1 임의의 수 생성에 의한 히스토그램 작성

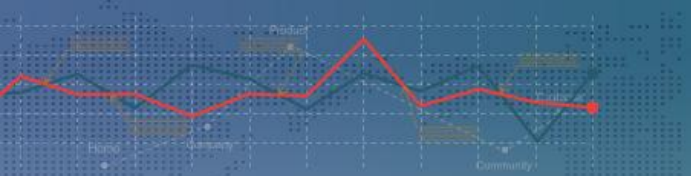
```
#이항분포로부터의 임의수 생성
rbinom(n, size,prob)
# 이항분포로부터 난수생성 자료의 분포 히스토그램
> z<- rbinom(1000, 10, 0.5)
> table(z)
z
0 1 2 3 4 5 6 7 8 9 10
2 11 47 99 219 240 199 125 54 3 1
> hist(z)
```



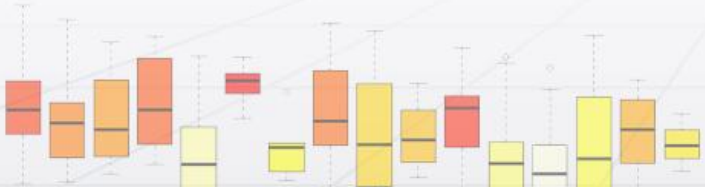
1 임의의 수 생성에 의한 히스토그램 작성

#포아송분포로부터의 임의수 생성
rpois(n, lamda)
 # 포아송분포로부터 난수생성 자료의 분포 히스토그램
〉 z〈− rpois(1000, 5)
〉 table(z)
z
0 1 2 3 4 5 6 7 8 9 10 11 12 14
5 39 85 165 164 165 151 101 61 32 21 5 5 1
〉 hist(z)





2. 상자그림 (Boxplot)

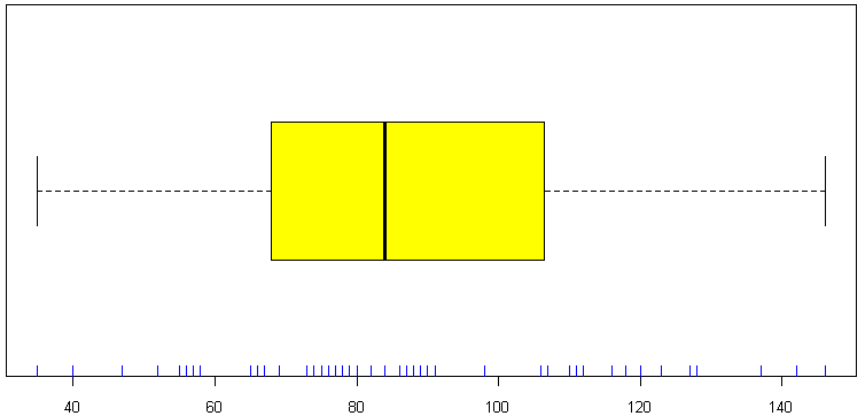


2 상자그림의 정의

상자그림(Boxplot) :

Boxplot

- 사분위수와 중앙값으로 상자를 만들고 최대 최소에 선을 연결한 자료의 퍼짐 정도를 파악하여 분포의 개형을 파악 하는데 도움을 주는 그래프

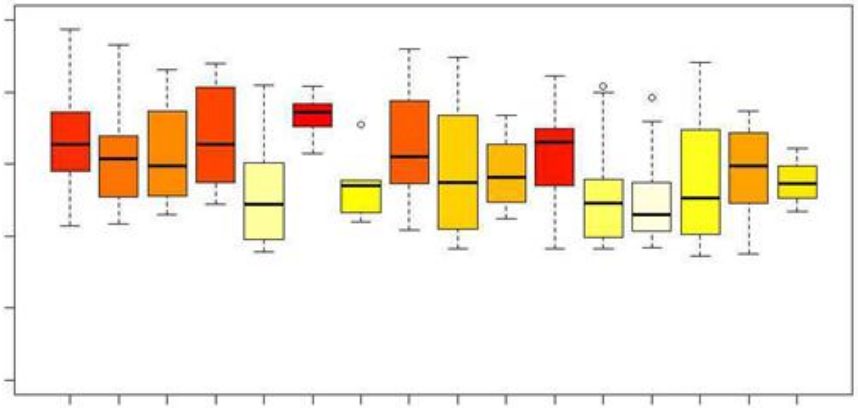


2 상자그림의 정의

상자그림(Boxplot) :

Boxplot

- 사분위수와 중앙값으로 상자를 만들고 최대 최소에 선을 연결한 자료의 퍼짐 정도를 파악하여 분포의 개형을 파악 하는데 도움을 주는 그래프



2 상자그림

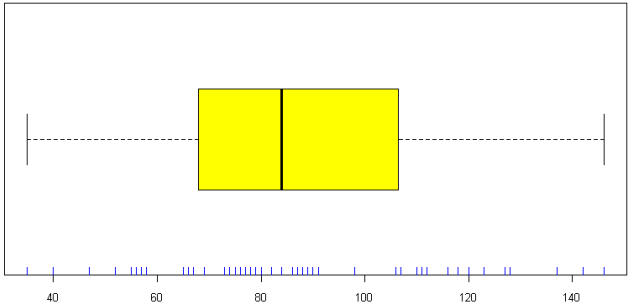
상자그림의 작성

상자그림 작성은 다음과 같은 절차로 작성된다.

① 다섯수치요약(min, Q_1, M, Q_3, max)을 구한다.

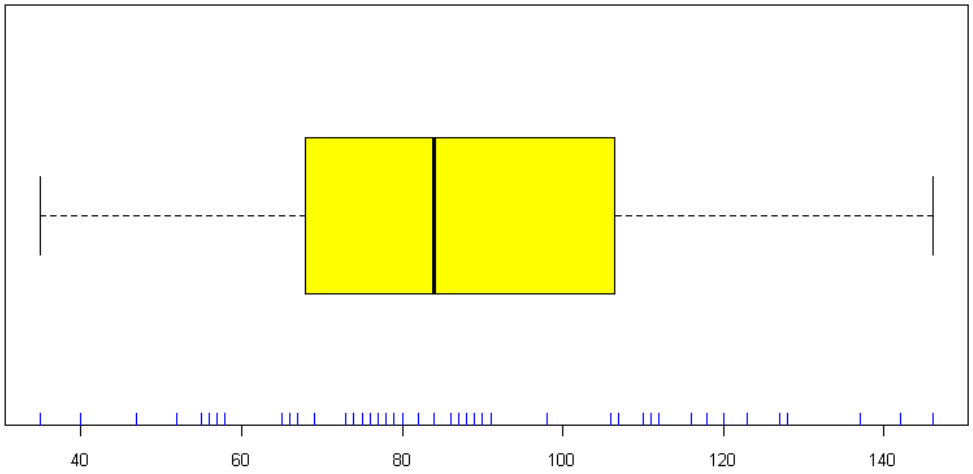
② 위.아래 4 분위수(Q)에 해당하는 수직선상의 위치에 네모형 상자의 양끝이 오도록 하고, 중위수(M)에 해당하는 위치에 +표시를 한다.

③ 최소값과 최대값의 위치에 점을 찍고, 이 점을 상자의 양끝과 연결하는 선분을 그려 넣는다.



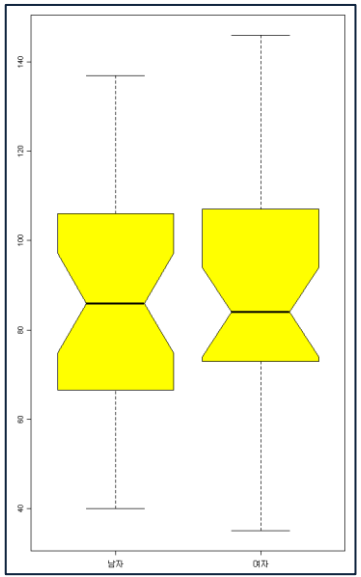
2 상자그림 작성

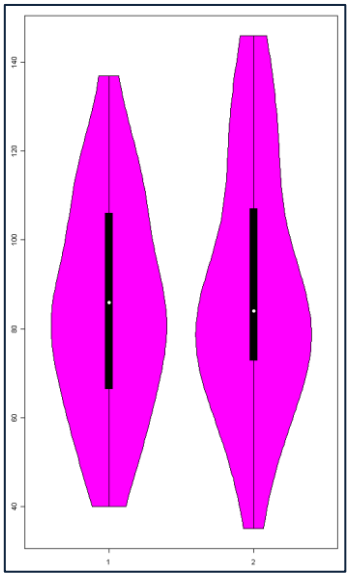
```
# 상자그림 작성
m=matrix(c(1,3,2,3),ncol=2,byrow=T)
layout(mat=m)
boxplot(과포화비율, col="yellow", horizontal=T, main=NULL)
rug(과포화비율,col="blue")
```



2 성별에 따른 상자그림 바이올린그림 작성

```
# 상자그림과 바이올린그림
par(mfrow=c(1,3))
require(vioplot)
boxplot(과포화비율~성별, notch=T, col="yellow", main=NULL)
vioplot(male$과포화비율, female$과포화비율)
```



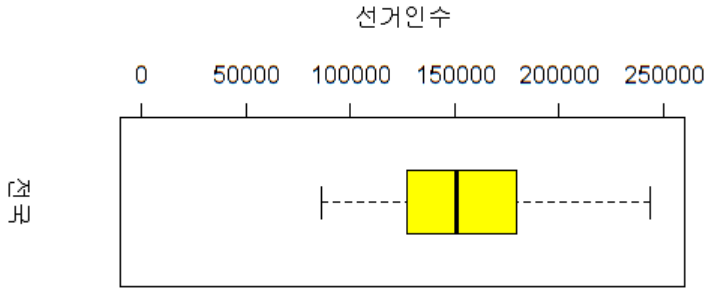


2 상자그림 작성 예

자료

우리나라 18대 국회의원 245개 선거구 선거인수 자료

```
electorate <- read.csv("국회의원 선거구 유권자 수.csv",
header=T)
str(electorate); attach(electorate)
summary(선거인.수)
windows(height=6, width=2.5)
boxplot(선거인.수, col="yellow", ylim=c(0,250000),
xlab="전국", ylab="선거인.수")
```



2 상자그림 작성 예

자료

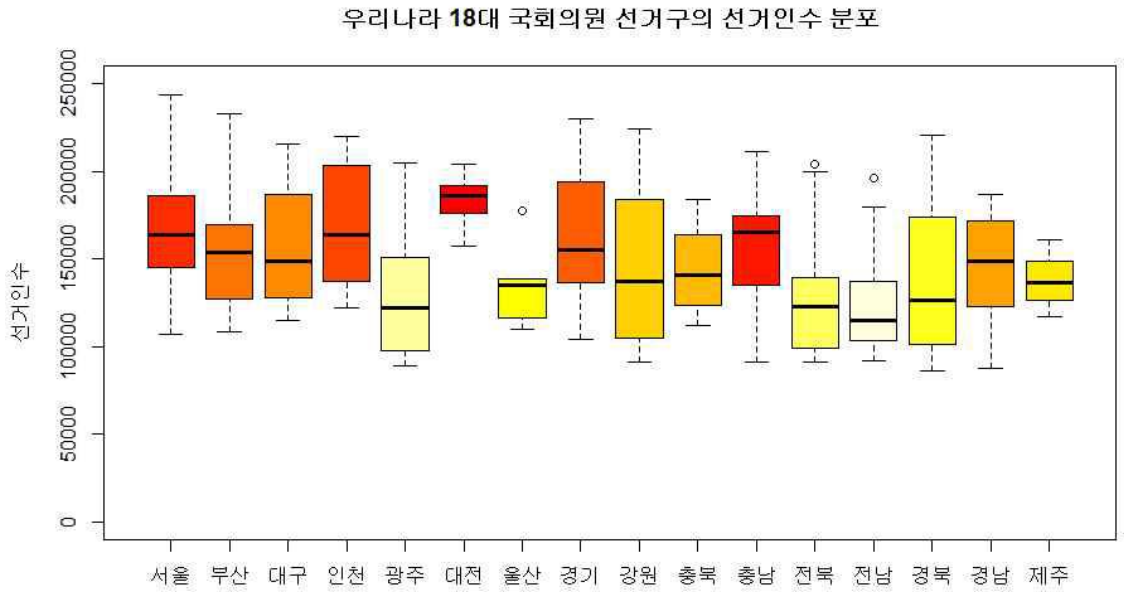
우리나라 18대 국회의원 245개 선거구 선거인수 자료

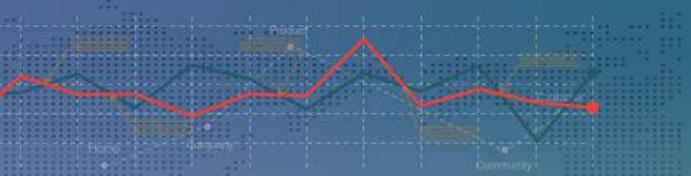
```
시도.순서
<- reorder(시도,시도번호)
bp <- boxplot(선거인.수~시도,순서)
order <- rank(-bp$stats[3,])
windows(height=6, width=10)
boxplot(선거인.수~시도,순서, col=heat.colors(16)[order],
ylim=c(0,250000), ylab="선거인.수",
main="우리나라 18대 국회의원 선거구의 선거인수 분포")
```

2 상자그림 작성 예

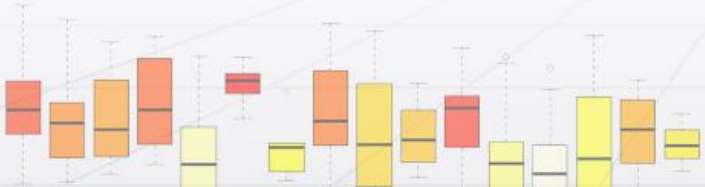
결과

우리나라 18대 국회의원 245개 지역별 선거구 선거인수 자료분포





3. 줄기 잎 그림 (Stem-Leaf Plot)



3 줄기 잎 그림

줄기 잎 그림의 정의

수치로 된 자료를 줄기와 잎으로 분류하여 자료의 분포개형을 파악하는 그림

예 $54 = 5 \times 10 + 4$ ➡ 5④, 5⑤, 5⑦

↑ ↑

줄기 잎 50 →줄기 ④,⑤,⑦ → 잎

3 줄기 잎 그림

줄기 잎 그림의 정의

수치로 된 자료를 줄기와 잎으로 분류하여 자료의 분포개형을 파악하는 그림

① 자료의 줄기부분을 선택하고 나머지 부분을 잎으로 정한다.

② 줄기 값을 크기 순으로 세로로 나열하고 그 옆에 수직선을 긋는다.

③ 각 줄기에 해당되는 자료의 잎 부분을 줄기의 오른쪽에 가로로 나열한다.

④ 각 줄기에서 잎의 값을 크기 순으로 재배열한다.

⑤ 줄기 열 왼쪽에 각 줄기의 도수를 세어 기록한다.

⑥ 누적도수를 그림에 써넣는다.

3 줄기 잎 그림

줄기 잎 그림의 작성

수치로 된 자료를 줄기와 잎으로 분류하여 자료의 분포개형을 파악하는 그림

깊이	도수	줄기	잎			
3	3	0	0	5	7	8
(4)	4	1	0	4	8	
4	2	2	5	9		
2	2	3	1	3		
	11					

3 줄기 잎 그림 작성 예

수리통계학 점수 자료

자료

교재 P 58 수리통계학 점수 자료

54	67	55	23	29	58	67	35
51	64	90	51	25	38	61	53
52	43	15	10	23	73	69	47
82	74	54	78	41	45	77	56
37	73	52	48	89	28	54	99
41	33	52	30	00	43	35	24
41	51	18	39	21	23	67	00
46	28	53	44	53	46	56	28
58							

3 줄기 잎 그림 작성 예

```
#나무 줄기그림 작성
> exam1 <- read.table("exam1.txt", header=T)
> str(exam1)
> attach(exam1)
> stem(score)
> stem(scor, scale=2)
```

0	00
1	058
2	1333458889
3	0355789
4	11133456678
5	11122233344456688
6	147779
7	33478
8	29
9	09

3 줄기 잎 그림 작성 예

```
#나무 줄기그림 작성
> exam1 <- read.table("exam1.txt", header=T)
> str(exam1)
> attach(exam1)
> stem(score)
> stem(scor, scale=2)
```

0	00
1	058
2	1333458889
3	0355789
4	11133456678
5	11122233344456688
6	147779
7	33478
8	29
9	09

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
6	7779
7	334
7	78
8	2
8	9
9	0
9	9

3 줄기 잎 그림과 히스토그램의 비교

공통점

- 외양적인 테두리가 동일하다.
- 자료의 분포 개형을 파악하는데 중요한 정보를 제공한다.
- 각 구간에 속하는 자료 점의 도수에 비례하는 막대기둥을 갖는다.

차이점

- 줄기그림은 구간내의 자료들의 도수가 아닌 숫자로 구별되어 있으므로 도수로 나타낸 히스토그램이 정보의 손실을 가져온다.
- 줄기그림에서는 작성된 원 줄기그림을 이용하여 좀 더 효율적으로 쉽게 줄기 수를 조절할 수 있다.
- 줄기그림은 구간폭이 정수이어야 하기 때문에 히스토그램처럼 구간폭을 임의로 정할 수 없다.

3 줄기 잎 그림과 히스토그램의 비교

줄기그림에서 관찰할 수 있는 자료의 특징

- 자료가 몇 개의 그룹으로 나누어지는가?
- 분포의 중앙이 어디인가?
  즉, 자료가 어느 줄기에 집중적으로 몰려있는가에 따라
  집중도가 높은 구간을 알 수 있다.
- 좌우대칭을 평가하여 자료가 어느 한쪽으로 몰려 있는가를
  파악한다.
- 자료의 퍼짐 정도 등을 비교한다.
- 특이점의 존재여부를 파악한다.

다음시간안내

이변량 자료의 시각화 1

산점도와 회귀모형