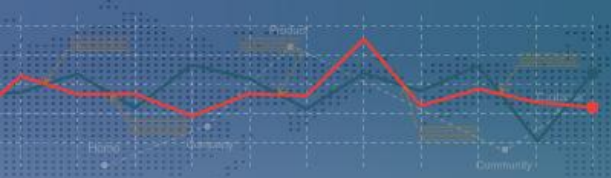


2강 탐색적 자료분석(EDA)의 시각화

정보통계학과 이태림 교수

1. 탐색적 자료분석의 개념과 배경을 이해
2. 데이터의 각종 그래프의 R에 의한 표현
3. 작성된 그래프에 의한 자료의 특징 파악

학습개요(EDA란 무엇인가?)

EDA의 정의

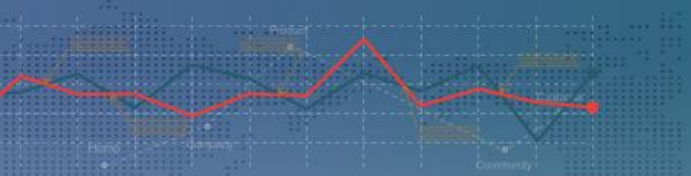
- 탐색적 자료분석이란?
- EDA의 역사

EDA의 네 가지주제

- 저항성(Resistance)
- 잔차의 해석(Residual)
- 자료의 재표현(Reexpression)
- 현시성(Graphic Representation)

작성된 그래프에 의한 자료의 특징 파악

- 원그래프
- 막대그래프



1. EDA의 정의

1 EDA의 정의

탐색적 자료분석(Exploratory Data Analysis) :

EDA

- 데이터의 특징과 내재하는 구조적인 관계를 알아내기 위한 분석기법

- 지금까지의 이론적 모형이나 틀에 바로 적용하기보다는 데이터를 있는 그대로 보려는데 중점을 두어 데이터 스스로 말하도록 유도하는 분석법

1 EDA의 정의

탐색적 자료분석(Exploratory Data Analysis) :

EDA

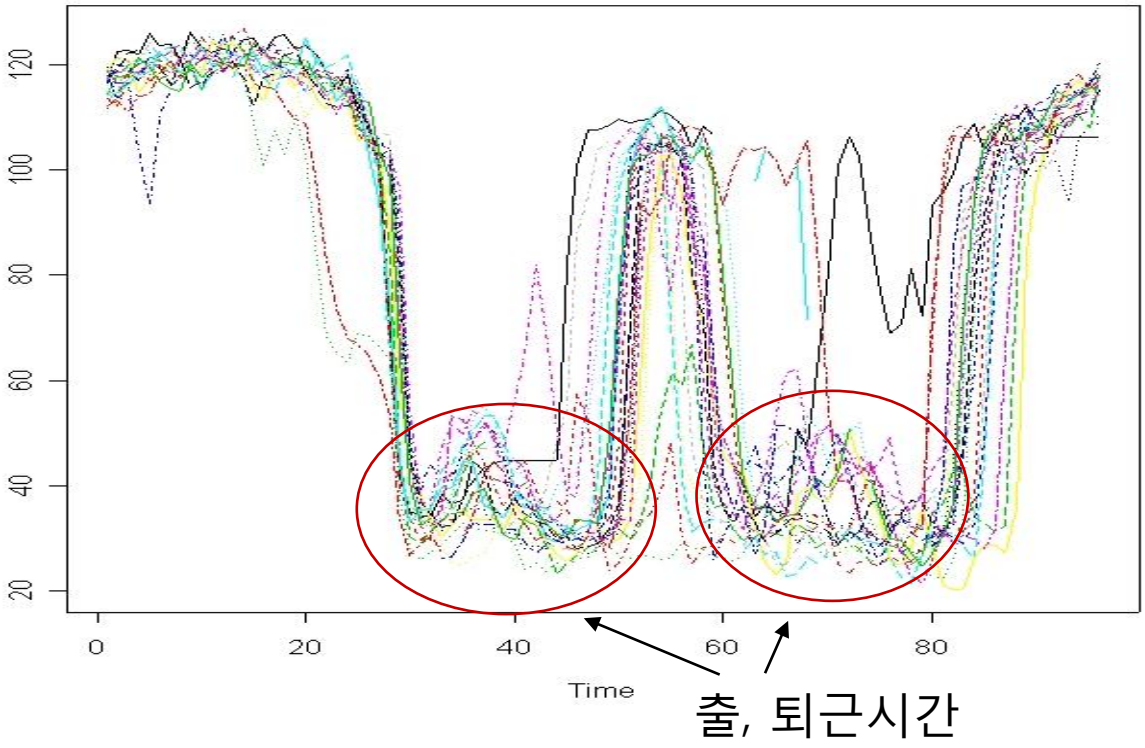
- 데이터의 특징과 내재하는 구조적인 관계를 알아내기 위한 분석기법

CDA

- 관측된 자료의 형태로 효과의 재현성을 평가하고 추정하는 전통적인 분석과정, 신뢰구간의 추정이나 유의성 검정에 의한 분석

한국도로공사 교통정보 제공 시스템 구축

서초–〉반포 일별 15분 단위 속도 (중앙값 사용)

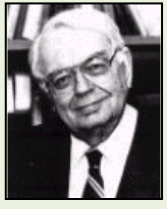


1 EDA의 정의

EDA의 역사

죤 튜키(J.W.Tukey) 1977

프린스턴 대학과 벨연구소에서 연구했으며 EDA의 첫저서인 Exploratory Data Analysis라는 책을 발간하면서 서문에 EDA의 중요성을 강조하여 통계학계의 충격과 함께 EDA가 학문적으로 출발



모스텔러(F.Mosteller) 1979

하바드의 모스텔러와 함께 자료분석과 회귀를 EDA의 입장에서 저술

모스텔러, 호글린 (D.C.Hoaglin) 1982

- 로버스트 통계와 탐색적 자료분석의 이해발간
- 자료 표 추세 및 분포의 탐색(Exploring Data Tables, Trends and Shapes) 발간

1 EDA의 정의

EDA의 목적

EDA의 목적은 자료의 구조 및 특징 파악을 위하여 효과적이고 신뢰성있는 자료의 요약과 그래프 기법의 활용

자료의 구조 및 특징 파악을 위하여 효과적이고 신뢰성 있는 자료의 요약과 그래프 기법의 활용

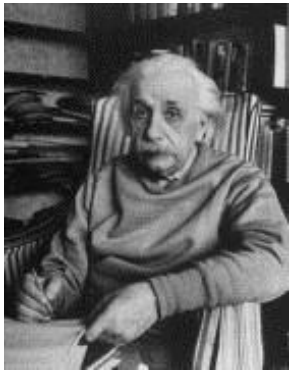
1 EDA의 정의

EDA : 자료분석의 기본철학

- 과학이라는 것은 아인슈타인과 같은 석학에 의해 이론적 통찰을 통해 상대성이론을 발표하거나 또는 퀴리부인의 발견과 같이 새로운 현상의 관찰로부터 혁신적으로 이루어지기도 하지만 대부분의 경우는 세심한 관찰, 실험과 분석의 반복을 통하여 가능
- 이러한 과학발전에 있어서 통계학 또는 통계분석가의 역할을 정리해보면 어떤 특정 모형에 대한 탐구와 함께 실험자료나 설문자료 관찰자료 등의 경험적 증거를 평가하는 역할을 담당

1 EDA의 정의

: 자료분석의 기본철학



- 최근 컴퓨터의 보급으로 데이터의 급속한 축적과 계산능력의 발전에 힘입어 통계인에게 보다 많은 과학적 기여 기회 부여
- 일종의 탐험가처럼 데이터를 샅샅이 뒤져서 가능한 모형을 제시하는 역할로 해당 분야에 대한 적절한 이해와 도전의식 해당분야 전문가와의 협업을 통해서 만이 좋은 결과 유도 가능

2. EDA의 네가지 주제

2 EDA의 네가지 주제

❖ EDA의 네 가지 주제

저항성
(Resistance)

EDA

2 EDA의 네가지 주제

❖ EDA의 네 가지 주제

저항성
(Resistance)

잔차의 해석
(Residual)

EDA

2 EDA의 네가지 주제

❖ EDA의 네 가지 주제

저항성
(Resistance)

자료의 재표현
(Re expression)

EDA

잔차의 해석
(Residual)

2 EDA의 네가지 주제

❖ EDA의 네 가지 주제

저항성
(Resistance)

잔차의 해석
(Residual)

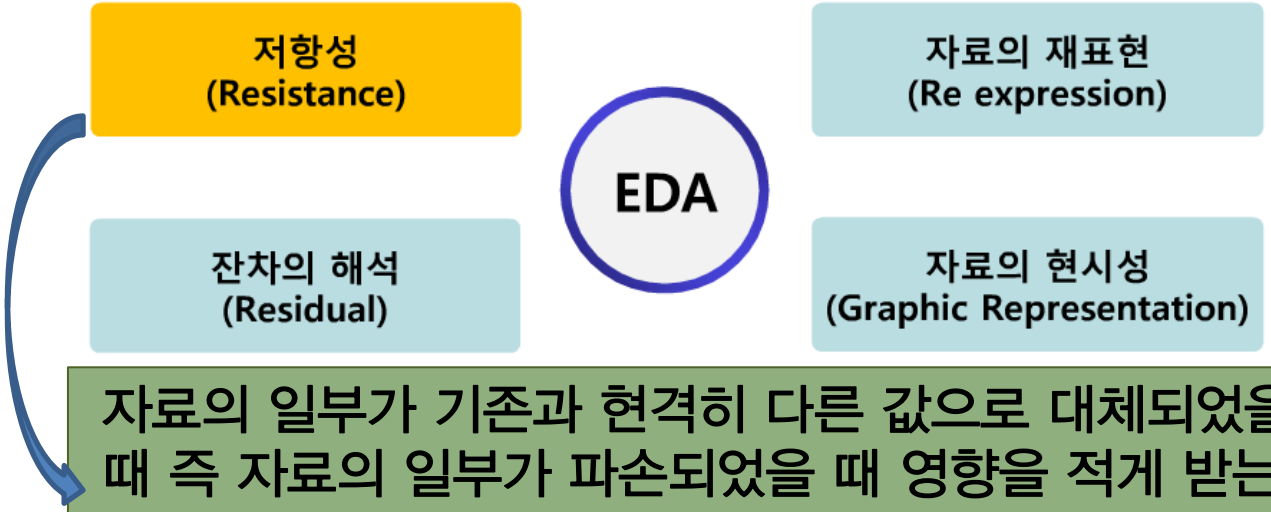
EDA

자료의 재표현
(Re expression)

자료의 현시성
(Graphic Representation)

2 EDA의 네가지 주제

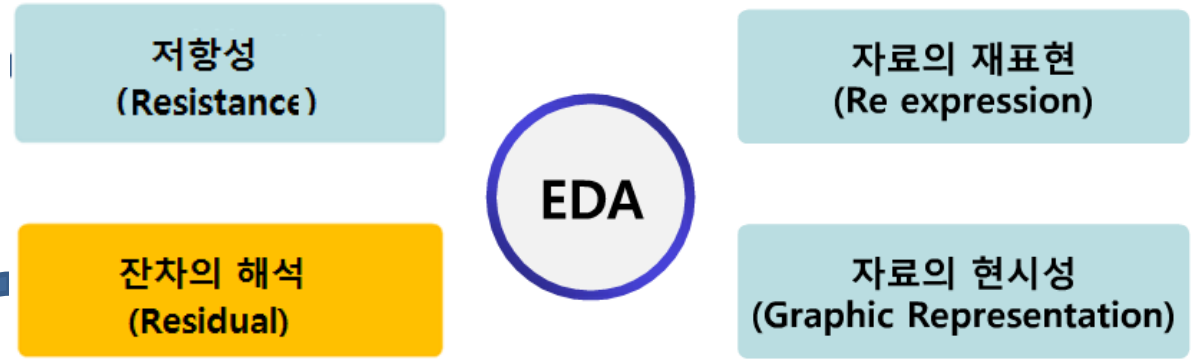
❖ EDA의 네 가지 주제



자료의 일부가 기존과 현격히 다른 값으로 대체되었을 때 즉 자료의 일부가 파손되었을 때 영향을 적게 받는 성질. 즉 저항성 있는 통계 또는 통계적 방법은 데이터의 부분적 변동에 민감하게 반응하지 않는다.

2 EDA의 네가지 주제

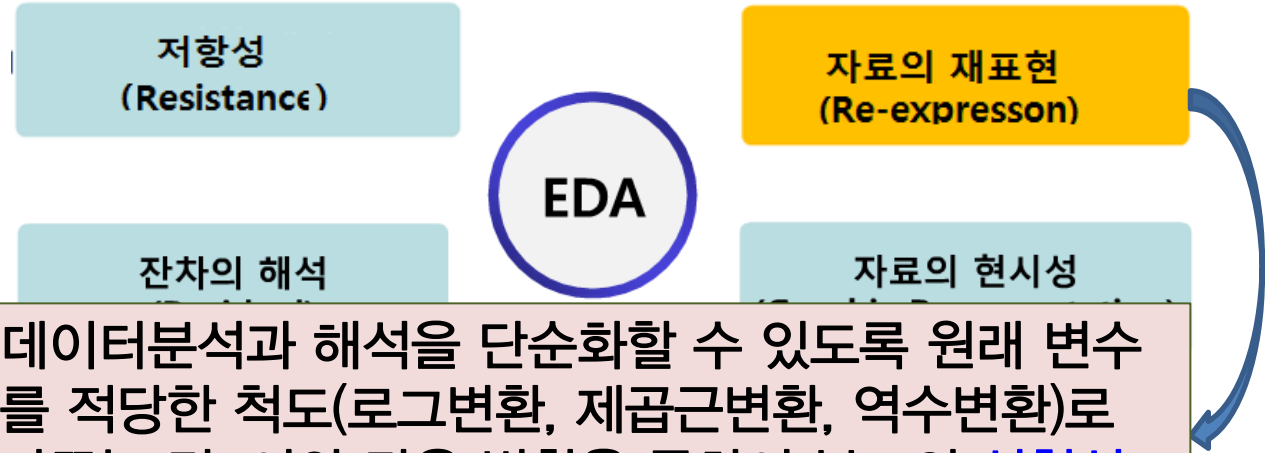
❖ EDA의 네 가지 주제



잔차란 관찰값들이 주경향으로부터 얼마나 벗어났는지를 말해준다. 즉 잔차를 구해봄으로써 데이터의 보통과 다른 특징을 찾아내야 한다.

2 EDA의 네가지 주제

❖ EDA의 네 가지 주제

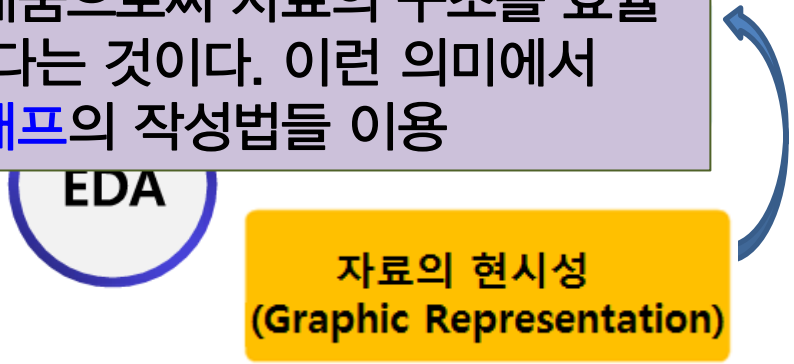


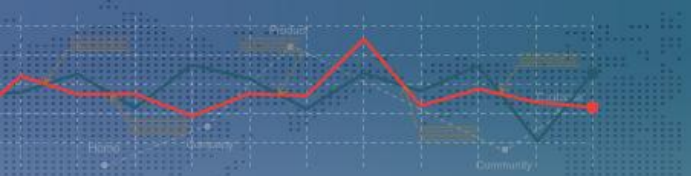
데이터분석과 해석을 단순화할 수 있도록 원래 변수를 적당한 척도(로그변환, 제곱근변환, 역수변환)로 바꾸는 것. 이와 같은 변환을 통하여 분포의 선형성, 분산의 안정성, 관련변수의 가법성, 분포의 대칭성 등 데이터 구조파악과 해석에 도움

2 EDA의 네가지 주제

❖ EDA의 네 가지 주제

자료의 그래프에 의한 표현. 즉 자료 안에 숨어있는 정보를 시각적으로 나타내줌으로써 자료의 구조를 효율적으로 잘 파악하게 된다는 것이다. 이런 의미에서 EDA에서는 다양한 그래프의 작성법들 이용



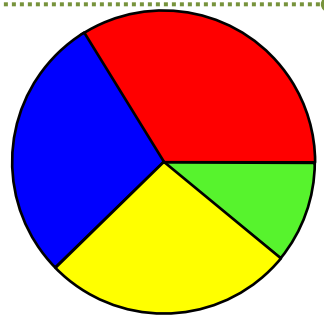


3. 자료의 그래프에 의한 표현

3 자료의 그래프에 의한 표현

R에 의한 원그래프 작성:

- 원그래프 : 전체에서 각 항목이 차지하는 비율을 파악 하기 위한 그래프



- 시각화의 목표 : 데이터의 통계적 정보를 그림의 형태로 나타내어 분포의 구성을 상대적으로 비교하는 데 유용

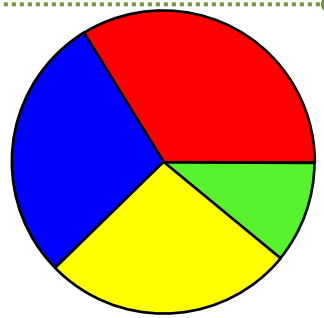
- 예 : 선거에서 후보별 득표수, 정당별 선호도, 방송대 재학생의 연령별 구성

3 자료의 그래프에 의한 표현(원그래프 pie chart)

자료: 대학생 100명의 혈액형 자료

Table 1. 대학생 100명에 대한 혈액형

A	B	B	A	A	O	A	AB	O	O
O	A	A	B	AB	A	O	B	A	B
B	A	B	A	B	AB	B	A	O	AB
O	B	A	B	A	O	B	A	A	A
A	O	A	O	O	B	B	O	AB	A
B	AB	B	O	O	O	AB	O	O	B
A	A	O	A	B	O	A	O	B	O
A	B	O	AB	B	B	A	O	B	A
B	B	O	AB	B	A	AB	A	B	A
A	O	O	A	A	O	AB	A	A	O

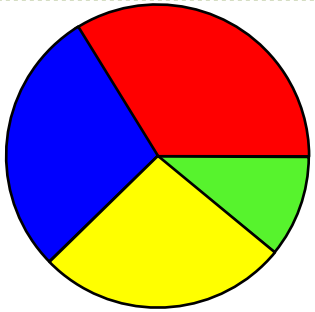


교재 페이지 33 예제 2-1

3 자료의 그래프에 의한 표현(원그래프 pie chart)

원그래프 작성 R프로그램 :

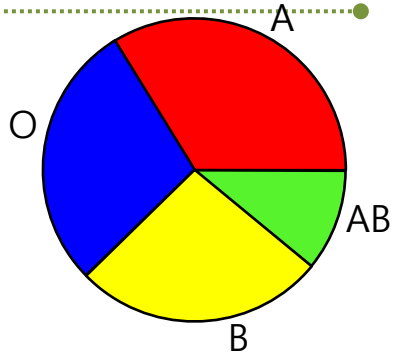
```
# 원그래프 그리기
혈액형=c( "A" , "B" , "B" , "A" , "A" , "O" , "A" , "AB" , "O" , "O" ,
"O" , "A" , "A" , "B" , "AB" , "A" , "O" , "B" , "A" , "B" ,
"B" , "A" , "B" , "A" , "B" , "AB" , "B" , "A" , "O" , "AB",
"O" , "B" , "A" , "B" , "A" , "O" , "B" , "A" , "A" , "A" ,
"A" , "O" , "A" , "O" , "O" , "B" , "B" , "O" , "AB" , "A",
"B" , "AB", "B" , "O" , "O" , "O" , "AB" , "O" , "O" , "B" ,
"A" , "A" , "O" , "A" , "B" , "O" , "A" , "O" , "B" , "O" ,
"A" , "B" , "O" , "AB" , "B" , "B" , "A" , "O" , "B" , "A" ,
"B" , "B" , "O" , "AB" , "B" , "A" , "AB" , "A" , "B" , "A" ,
"A" , "O" , "O" , "A" , "A" , "O" , "AB" , "A" , "A" , "O" )
```



3 자료의 그래프에 의한 표현(원그래프 pie chart)

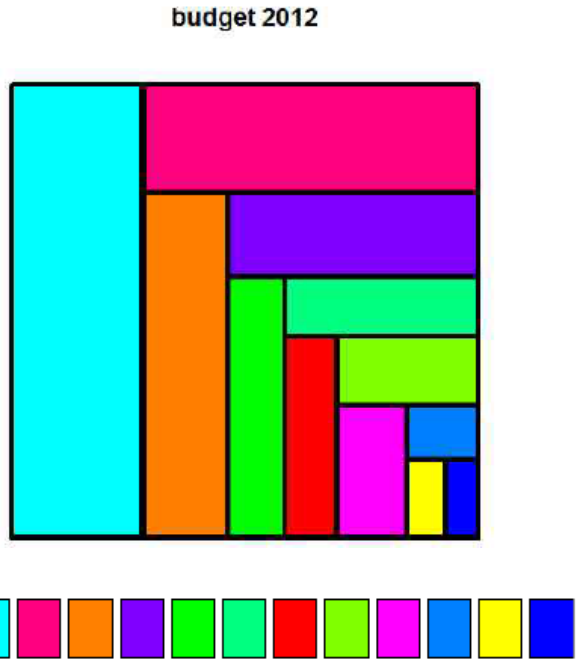
원그래프 작성 R프로그램 :

```
혈액형
정렬.혈액형=sort(table(혈액형),decreasing=T)
정렬.혈액형
par(mfrow=c(1,2))
slices=c("red","blue","yellow","green")
pie(정렬.혈액형,col=slices,radius=1,main="원그래프")
```

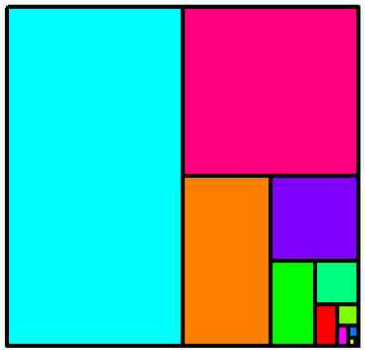


3 자료의 그래프에 의한 표현

R에 의한 사각파이 그래프 작성:




```
square.pie( )
    p = 구성 비율)
    color = 컬러 배당
    title = 그림 제목
```



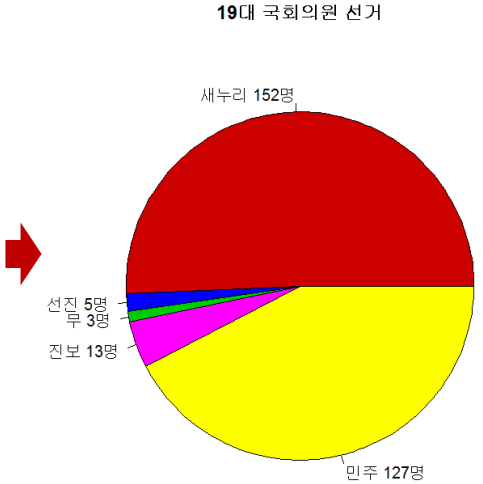
보건복지–행정–교육–국방–사회자본–농림–연구개발–산업–안전–환경–문화–외교

3 자료의 그래프에 의한 표현(원그래프 pie chart)

자료: 국회의원 정당별 분포 자료

2012 19대 국회의원(총의석 300석)에서 새누리당 152석, 민주통합당 127석, 통합진보당 13석, 선진당 5석, 무소속 3석의 자료로 전체 의석수에서 각 정당이 차지하는 비율을 보여주는 원그래프를 작성해보자.

```
require(grDevices)
pie.vote <- c(0.5067,0.0167,0.0100,0.0433,0.4233)
names(pie.vote) <- c("새누리 152명", "선진 5명", "무 3명", "진보 13명", "민주 127명")
pie(pie.vote)
pie(pie.vote,
col = c("red3", "blue", "green3", "magenta", "yellow"),
main = "19대 국회의원 선거")
```

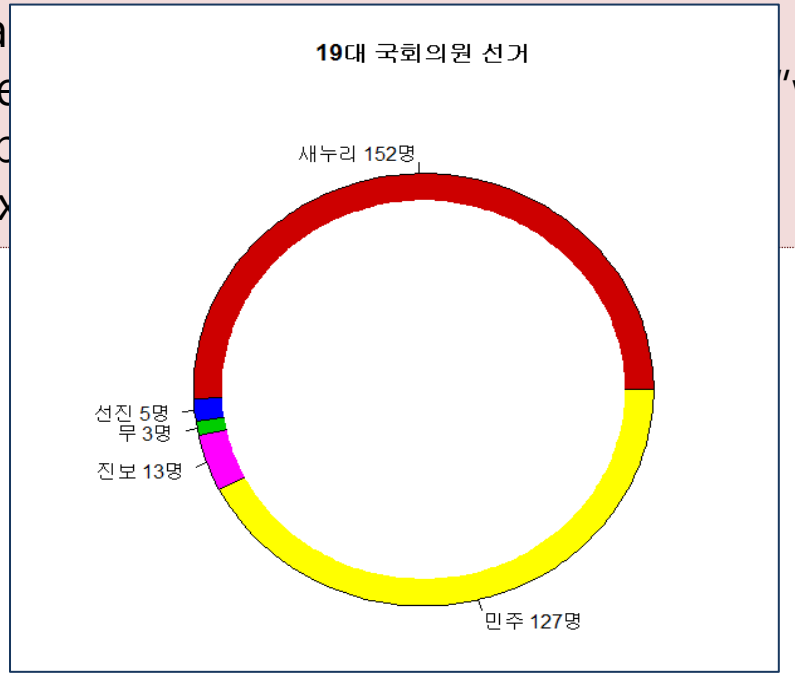


3 자료의 그래프에 의한 표현(원그래프 pie chart)

원띠그래프 작성 R프로그램 :

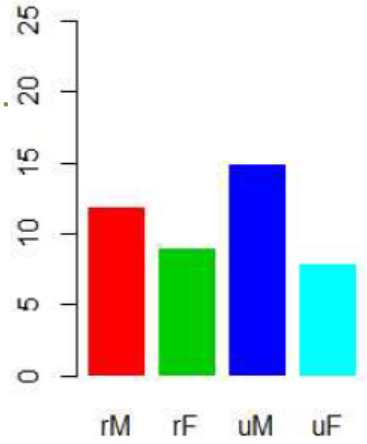
pa
pie
lab
tex

'white",



3 자료의 그래프에 의한 표현

R에 의한 막대그래프 작성:



- 막대그래프 : 항목별 도수를 막대의 상대적인 길이로 나타낸 그래프
- 시각화의 목표 : 전체의 구성비 보다는 어느 항목의 막대가 제일 긴지 보여준다.
- 예 : 선거에서 정당별 의원수, 매체별 선호도, 방송대 재학생의 연령별 분포

3 자료의 그래프에 의한 표현(막대그래프 bar chart)

자료: 대학생 100명의 혈액형 자료

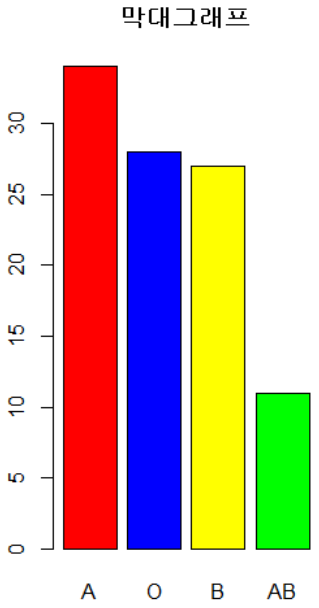
Table 1. 대학생 100명에 대한 혈액형

A	B	B	A	A	O	A	AB	O	O
O	A	A	B	AB	A	O	B	A	B
B	A	B	A	B	AB	B	A	O	AB
O	B	A	B	A	O	B	A	A	A
A	O	A	O	O	B	B	O	AB	A
B	AB	B	O	O	O	AB	O	O	B
A	A	O	A	B	O	A	O	B	O
A	B	O	AB	B	B	A	O	B	A
B	B	O	AB	B	A	AB	A	B	A
A	O	O	A	A	O	AB	A	A	O

3 자료의 그래프에 의한 표현(막대그래프 bar chart)

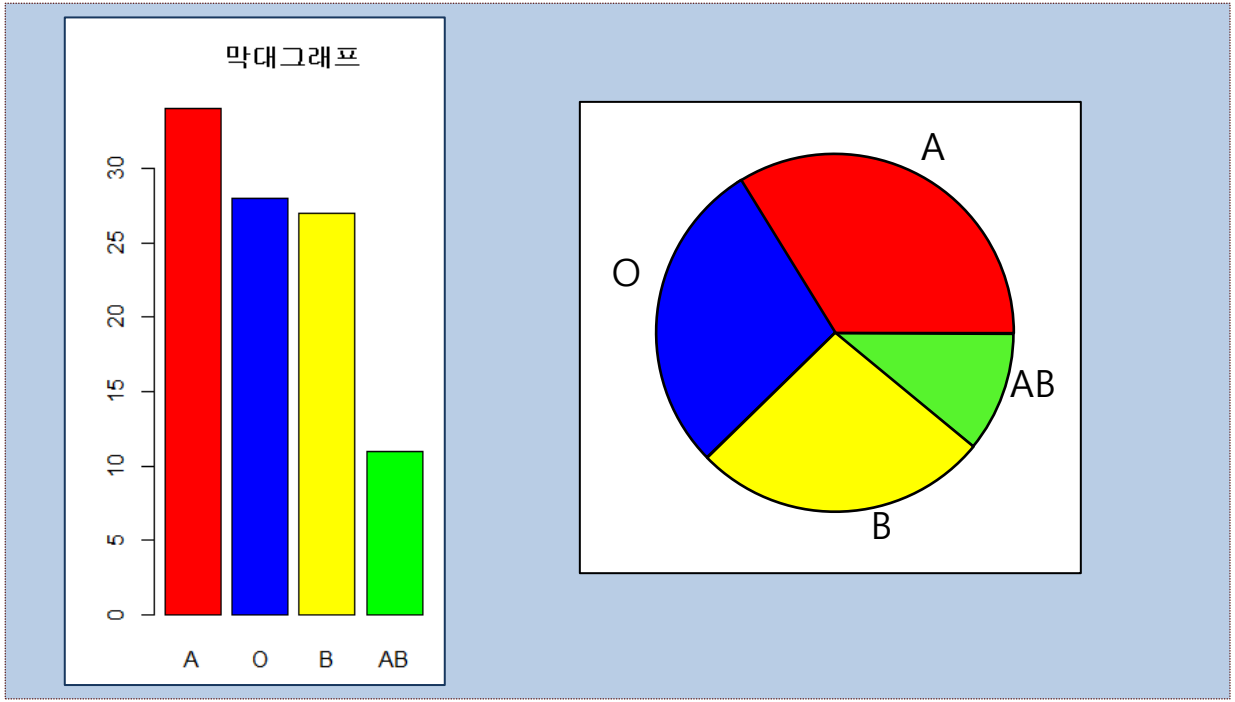
막대그래프 작성 R프로그램 :

```
# 자료입력
혈액형=c( "A" , "B" , "B" , "A" , "A" , "O" , "A" , "AB" , "O" , "O" ,
……..
"A" , "O" , "O" , "A" , "A" , "O" , "AB" , "A" , "A" , "O" )
혈액형
정렬.혈액형=sort(table(혈액형),decreasing=T)
정렬.혈액형
par(mfrow=c(1,2))
slices=c("red","blue","yellow","green")
# 막대그래프 그리기
barplot(정렬.혈액형,col=slices,main="막대그래프")
```



3 자료의 그래프에 의한 표현(막대그래프 bar chart)

막대그래프 작성 R프로그램 :



3 자료의 그래프에 의한 표현(막대그래프 bar chart)

자료: 국회의원 정당별 분포 자료

2012 19대 국회의원(총의석 300석)에서 새누리당 152석, 민주통합당 127석, 통합진보당 13석, 선진당 5석, 무소속 3석의 자료로 전체 의석수에서 각 정당이 차지하는 비율을 보여주는 막대그래프를 작성해보자.

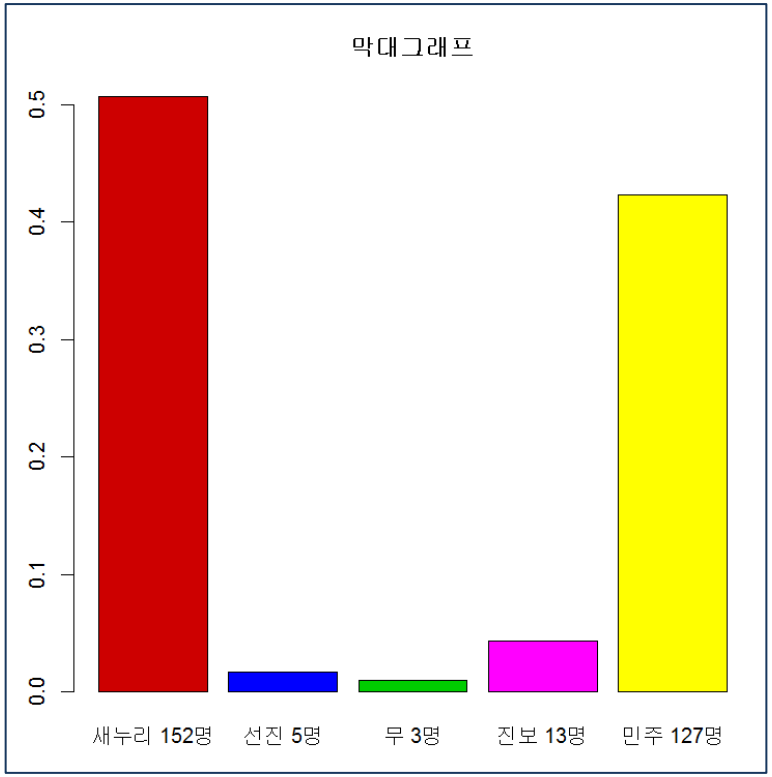
3 자료의 그래프에 의한 표현(막대그래프 bar chart)

자료: 국회의원 정당별 분포 자료

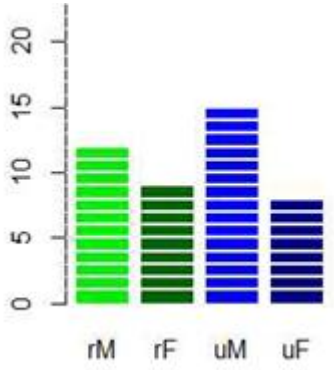
```
# 막대그래프 그리기
require(grDevices)
pie.vote <- c(0.5067,0.0167,0.0100,0.0433,0.4233)
names(pie.vote) <- c("새누리 152명", "선진 5명", "무 3명", "
진보 13명", "민주 127명")
pie(pie.vote)
pie(pie.vote,
col = c("red3", "blue", "green3", "magenta", "yellow"),
main = "19대 국회의원 선거")
barplot(pie.vote,col = c("red3", "blue", "green3",
"magenta", "yellow"),main="막대그래프")
```

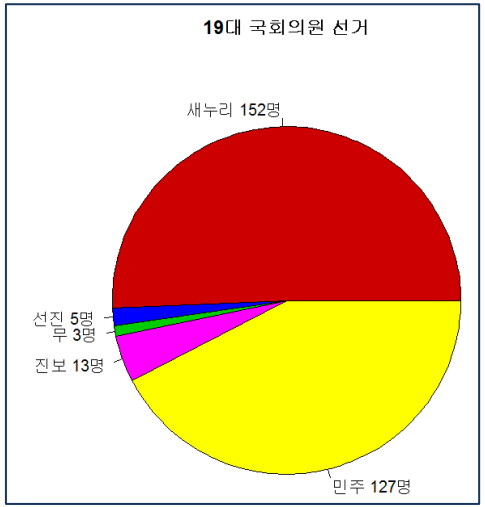
3 자료의 그래프에 의한 표현(막대그래프 bar chart)

R에 의한 막대그래프 작성:



abline( ) 함수를 써서
평행선을 일정 간격으로 그은 것





다음시간안내

탐색적 자료분석의 시각화2

히스토그램 상자그림 줄기 잎 그림