

빅데이터 분석

[R 데이터 탐색 - 단일변수 자료]

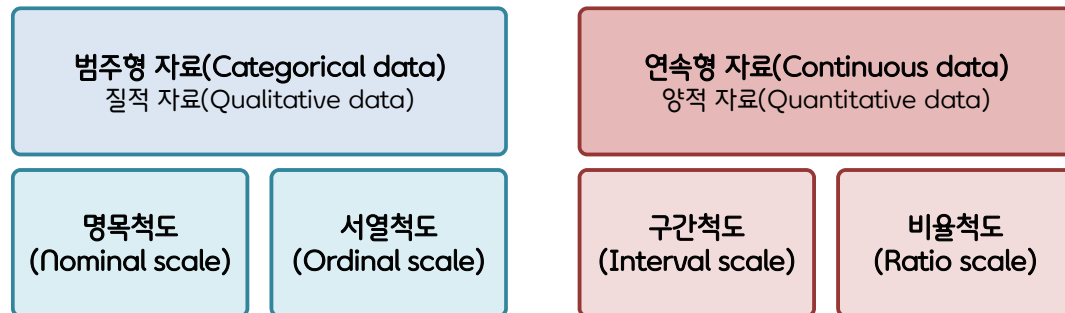
데이터 구조 파악



데이터 구조 파악

데이터 구조 파악 - 자료 특성 기반 분류

- 데이터 분석 대상의 자료 특성에 따라 범주형 자료와 연속형 자료로 구분



데이터 구조 파악

범주형 데이터

명목척도

- 자료를 이름이나 명칭으로 구분하고서 숫자를 부여한 척도
ex) 혈액형의 A = 1, B = 2, AB = 3, O = 4는 구분을 위해 부여된 숫자
- 순서를 매길 수 없고, 연산할 수 없으며, **단순히 식별을 위해서 숫자를 부여**
ex) O형이 4이고 B형이 2이기 때문에 O형이 더 “크다”고 표현할 수 없음



단순히 분류 & 식별을 위해 부여한 숫자이기 때문에,

본 학습자료는 저작권자의 동의없이 무단 복제 및 배포할 수 없습니다.

데이터 구조 파악

범주형 데이터

서열척도

- 개체 간의 특정 속성을 바탕으로 서열 관계를 가지는 척도
ex) 국가 GDP 순위의 미국 = 1, 중국 = 2, 일본 = 3은 “GDP 규모”라는 속성을 통해 서열을 가짐
- 상대적인 정보를 지니고 있기 때문에, **순서만 중요할 뿐 연산을 할 수 없음**
ex) 미국이 중국보다 앞선다고 할 수 있지만, “미국 + 중국 = 일본”처럼 연산을 할 수는 없음



미국(1)



중국(2)



일본(3)

데이터 구조 파악

연속형 데이터

구간척도

- 연속적인 숫자로 수량화할 수 있으며, 그 숫자들 간의 간격이 동일한 척도
ex) 서울 기온 23도, 도쿄 기온 28도, 하와이 기온 30도
- 수량화할 수 있기 때문에 00만큼 크고 작다를 표현할 수 있지만,
절대적인 ‘0’ 값을 가지고 ‘없다’라고 표현할 수 없음
ex) 만일 시베리아 기온이 0도라고 할 때, ‘기온이 없다’라고 할 수 없음(‘0도’도 엄연히 기온!)

데이터 구조 파악

연속형 데이터

비율척도

- 연속적인 숫자로 수량화할 수 있으며, 그 숫자들 간의 비율이 동일한 척도
ex) A회사 인원 45명, B회사 인원 60명, C회사 인원 90명
- 특정 숫자 간의 비율이 동일하기 때문에, 대소 비교 뿐만 아니라 사칙연산도 가능하며
절대적인 '0' 값을 가지고 '없다'라고 표현할 수 있음
ex #1) 45명 \times 2 = 90명이기 때문에, C회사는 A회사보다 2배 인원이 있다고 할 수 있음
ex #2) 만일 어떤 회사 인원이 0명이라면, 그 회사에는 인원이 '없다'라고 할 수 있음

데이터 구조 파악

데이터 구조 파악 - 변수 개수 기반 분류

- 데이터 분석에서 '변수'는 연구, 조사, 관찰하고 싶은 대상의 특성을 의미
- 키, 몸무게, 심박수, 수면시간 등을 의미

단일변수 자료(Univariate data)

다중변수 자료(Multivariate data)

데이터 구조 파악

단일변수 & 다중변수 자료

- 단일변수 자료는 하나의 변수로만 구성된 자료를 의미(벡터가 한 개)
- 다중변수 자료는 두 개 이상의 변수로 구성된 자료를 의미(벡터가 여러 개)

단일변수 자료

성별
F
M
M
F

다중변수 자료

키	몸무게	심박수
170	55	90
187	89	120
172	72	111
161	48	72

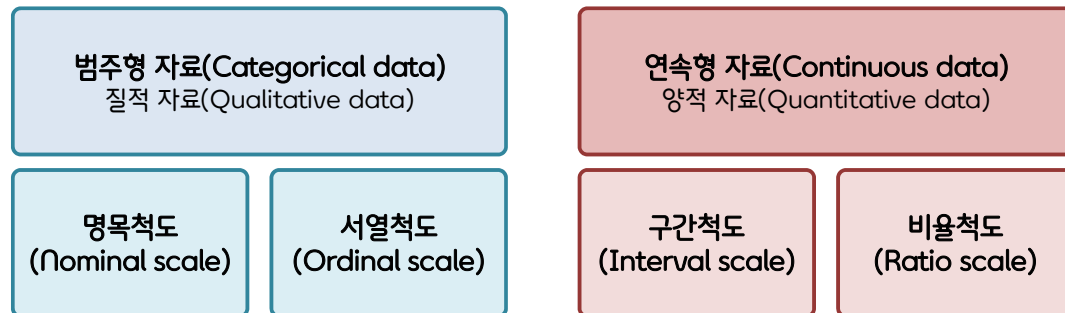
단일변수 범주형 자료 탐색



데이터 구조 파악

데이터 구조 파악 - 자료 특성 기반 분류

- 데이터 분석 대상의 자료 특성에 따라 범주형 자료와 연속형 자료로 구분



단일변수 자료탐색 - 범주형 자료

데이터 탐색

- 범주형 자료는 크기를 갖지 않기 때문에 연산 불가능
- 범주형 자료에서 가능한 기본 작업이란, 자료에 포함된 관측값들을 종류별로 세는 것
- 이를 통해 종류별 비율을 알 수 있고, 이 결과를 바탕으로 막대 그래프 또는 원 그래프 등을 그릴 수 있음

상명대	대전대	동의대	단국대	단국대
동의대	우송대	홍익대	상명대	원광보건대

단일변수 자료탐색 - 범주형 자료

데이터 탐색

상명대	대전대	동의대	단국대	단국대
동의대	우송대	홍익대	상명대	원광보건대

- 본 수업을 듣는 학생들이 재학 중인 학교의 표본을 추출한 결과값
- '재학 중인 학교'라는 단일 특성에 대해 수집한 자료 -> 단일변수 자료
- 상명대, 동의대, 원광보건대, 단국대는 크기를 측정할 수 없음
-> 범주형 자료

단일변수 자료탐색 - 범주형 자료

도수분포표

상명대	대전대	동의대	단국대	단국대
동의대	우송대	홍익대	상명대	원광보건대

- 다음과 같은 형태의 자료를 분석하기 위해, 종류별 개수를 세고 비율 계산
- table() 함수로 벡터 내 범주형 자료의 종류별 도수분포표 계산 가능

단일변수 자료탐색 - 범주형 자료

table() 함수

- table() 함수는 데이터 빈도 분할표를 자동으로 만든다
- table(데이터프레임1, 데이터프레임2)

```
> gender <- c("M", "F", "M", "M", "F", "M", "F", "M", "F")
> table(gender)
gender
F M
4 5
> religion <- c("Ch", "I", "N", "B", "N", "Ch", "Ca", "B", "N")
> table(religion)
religion
B Ca Ch I N
2 1 2 1 3
> table(gender, religion)
      religion
gender B Ca Ch I N
F      0  1  0  1  2
M      2  0  2  0  1
```

gender 데이터 빈도를 보여준다.

religion 데이터 빈도를 보여준다.

※ 기독교 Ch, 이슬람 I, 카톨릭 Ca, 불교 B, 무교 N

단일변수 자료탐색 - 범주형 자료

도수분포표

```
univ <- c("상명대", "대전대", "동의대", "단국대", "단국대", "동의대",
"우송대", "홍익대", "상명대", "원광보건대")
table(univ)
table(univ) / length(univ)
```


단일변수 자료탐색 - 범주형 자료

도수분포표

```
univ <- c ("상명대", " 대전대 ", " 동의대", "단국대",
"단국대", "동의대", " 우송대", " 홍익대", "상명대",
"원광보건대")
```

```
print(univ)    #univ 벡터 출력
```

```
table(univ)    #도수분포표 계산
```

```
table(univ) / length(univ) #비율 출력
```

단일변수 자료탐색 - 범주형 자료

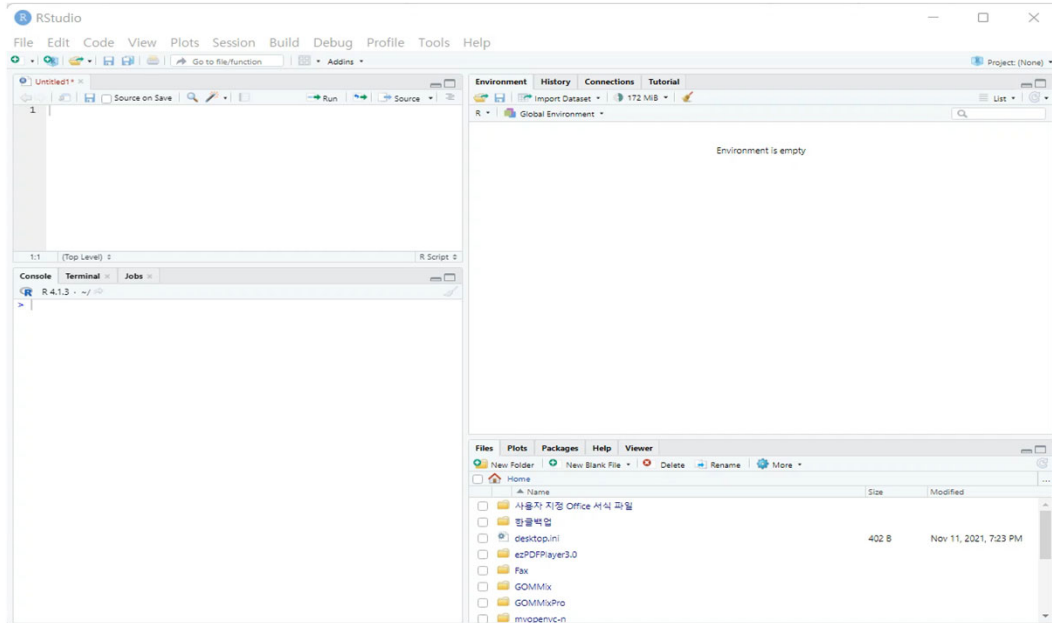
도수분포표

```
> univ <- c("상명대", "대전대", "동의대", "단국대", "단국대", "동의대", "우송대", "홍익대", "상명대", "원광보건대")
> table(univ)
univ
단국대    대전대    동의대    상명대    우송대    원광보건대    홍익대
      2         1         2         2         1         1         1
> table(univ) / length(univ)
univ
단국대    대전대    동의대    상명대    우송대    원광보건대    홍익대
  0.2      0.1      0.2      0.2      0.1      0.1      0.1
> |
```

univ라는 벡터 안에 “동의대”가 2개 있다는 것을 확인할 수 있습니다.

univ라는 벡터 속 전체 요소들 중에 “단국대”가 0.2(20%) 차지한다는 것을 알 수 있습니다.

상명대	대전대	동의대	단국대	단국대
동의대	우송대	홍익대	상명대	원광보건대



단일변수 자료탐색 - 범주형 자료

도수분포표로 막대 그래프 그리기

```
univ <- c("상명대", "대전대", "동의대", "단국대", "단국대",  
"동의대", "우송대", "홍익대", "상명대", "원광보건대")
```

```
table <- table(univ)
```

```
barplot(table, main = "재학 대학 분포")
```

단일변수 자료탐색 - 범주형 자료

도수분포표로 막대 그래프 그리기

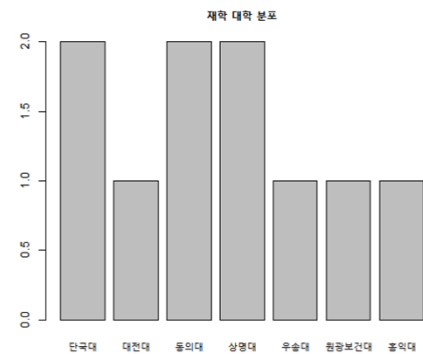
- 도수분포표를 작성하는 것만으로도 단일변수 범주형 자료가 포함되는 정보 파악 가능
- 더욱 쉽게 정보를 파악할 수 있도록 막대 그래프로 시각화

```
> table(univ)
```

```
univ
단국대    대전대    동의대    상명대    우송대    원광보건대    홍익대
      2         1         2         2         1         1         1
```

```
> table(univ) / length(univ)
```

```
univ
단국대    대전대    동의대    상명대    우송대    원광보건대    홍익대
  0.2     0.1     0.2     0.2     0.1     0.1     0.1
```



단일변수 자료탐색 - 범주형 자료

도수분포표로 막대 그래프 그리기

```
univ <- c("상명대", "대전대", "동의대", "단국대",
"단국대", "동의대", "우송대", "홍익대", "상명대",
"원광보건대")
```

```
table <- table(univ)
```

```
pie(table, main = "재학 대학 분포")
```

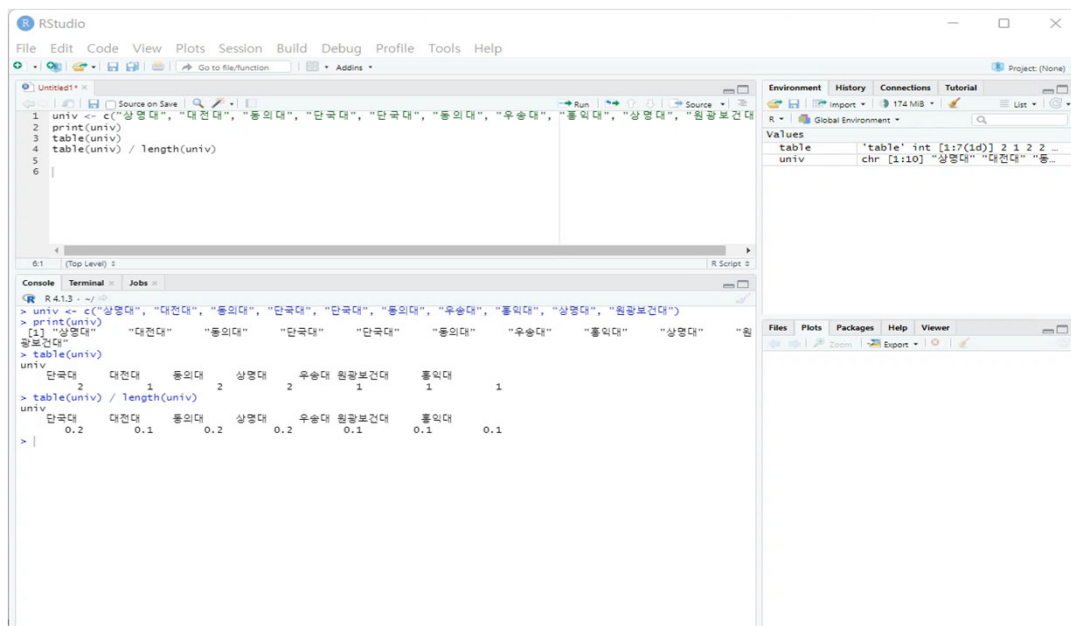
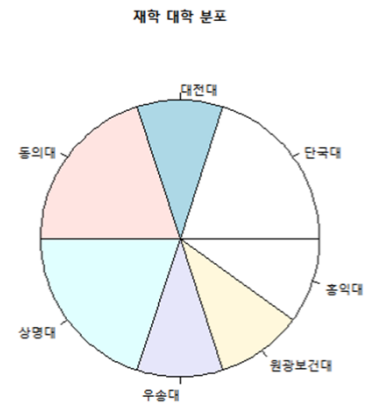
단일변수 자료탐색 - 범주형 자료

도수분포표로 원 그래프 그리기

- 도수분포표를 작성하는 것만으로도 단일변수 범주형 자료가 포함되는 정보 파악 가능
- 더욱 쉽게 정보를 파악할 수 있도록 원 그래프로 시각화

```
> table(univ)
univ
단국대  대전대  동의대  상명대  우송대  원광보건대  홍익대
      2       1       2       2       1       1       1

> table(univ) / length(univ)
univ
단국대  대전대  동의대  상명대  우송대  원광보건대  홍익대
 0.2    0.1    0.2    0.2    0.1    0.1    0.1
```



단일변수 연속형 자료 탐색



단일변수 자료탐색 - 연속형 자료

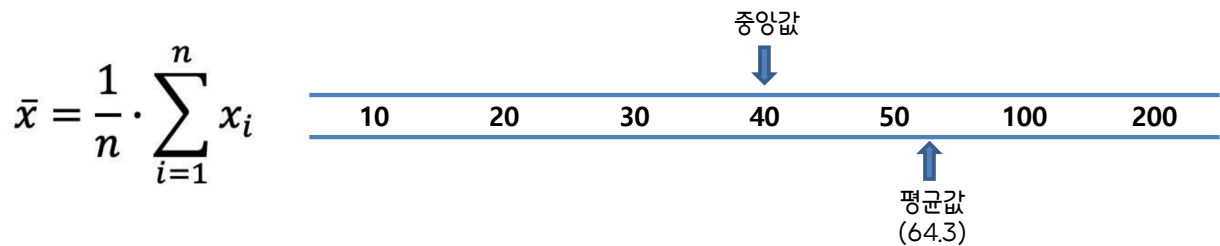
데이터 탐색

- 관측값들이 크기를 가지기 때문에, 범주형 자료에 비해서 다양한 분석이 방법 존재
 - 평균
 - 중앙값
 - 사분위수
 - 산포(분산, 표준편차, 값의 범위, 최대값, 최소값) 등 계산

단일변수 자료탐색 - 연속형 자료

평균과 중앙값

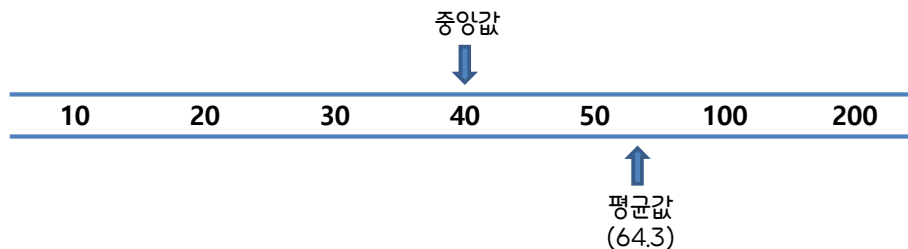
- 평균은 하나의 값으로 전체를 대표할 수 있는 값의 의미를 포함
- 평균(mean)을 계산하는 방법 -> 자료의 값을 모두 합산하고 값들의 개수로 나누기



단일변수 자료탐색 - 연속형 자료

평균과 중앙값

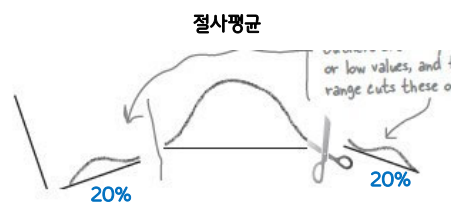
- 중앙값(median)은 자료의 값들을 크기순으로 일렬로 나열했을 때, 중앙에 위치하는 값을 의미



단일변수 자료탐색 - 연속형 자료

평균과 중앙값

- 절사평균 : 평균이 자료 내에 있는 너무 크거나 작은 관측값에 영향을 받는 것을 완화하기 위해서 제안(상하위 20% 값 제거 후 평균 계산)
- 평균값, 중앙값, 절사평균은 각각 특징이 다르기 때문에, 분석하고자 하는 자료에 어떤 방법을 적용하는 게 효과적일지 스스로 판단하는 역량 필요



단일변수 자료탐색 - 연속형 자료

평균과 중앙값 계산

12.1	12.8	15	19	22	28	31.1
------	------	----	----	----	----	------

다음과 같은 체질량지수(BMI, Body Mass Index) 를 담은 벡터가 있다고 할 때, 간단한 함수로 아래 연산을 수행

- 1) 평균
- 2) 중앙값
- 3) 절사평균

단일변수 자료탐색 - 연속형 자료

평균과 중앙값 계산

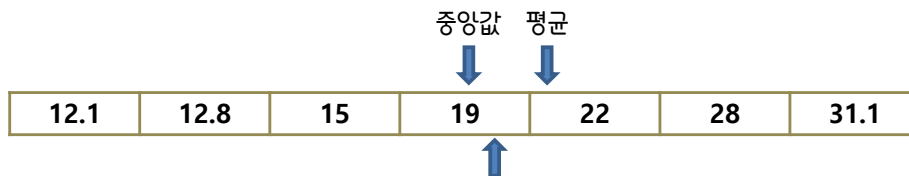
```
bmi <- c(12.1, 12.8, 15, 19, 22, 28, 31.1)
```

```
mean(bmi) #평균
```

```
median(bmi) #중앙값
```

```
mean(bmi, trim = 0.2) #절사평균(상하위 20% 값 제외)
```

```
> bmi <- c(12.1, 12.8, 15, 19, 22, 28, 31.1)
> mean(bmi)
[1] 20
> median(bmi)
[1] 19
> mean(bmi, trim = 0.2)
[1] 19.36
> |
```



단일변수 자료탐색 - 연속형 자료

사분위수

- 사분위수(quartile)은 주어진 자료에 값들을 크기순으로 나열했을 때, 4등분하는 지점에 있는 값들을 의미
- 자료의 값을 4등분을 하면 등분점(값이 나뉘는 곳)이 3개가 생기며, 각각 1사분위수(Q1), 2사분위수(Q2), 3사분위수(Q3)라고 부름
- 2사분위수는 중앙값과 동일 -> 주어진 자료의 값들을 절반으로 나눈 구간이기 때문
- 전체 데이터를 4개로 나눴기 때문에 4개 구간에는 25%의 자료가 존재



단일변수 자료탐색 - 연속형 자료

사분위수

- 예시. 헬스케어 빅데이터구조 수업을 듣는 100명의 학생들 몸무게를 사분위수로 계산
- $Q1 = 60\text{kg}$, $Q2 = 72\text{kg}$, $Q3 = 82\text{kg}$ 이라고 가정
- 해당 사분위수 결과로 유추할 수 있는 사실은 다음과 같음
 - 25명의 학생들의 몸무게는 60kg 미만이다
 - 25명의 학생들의 몸무게는 60~72kg 사이이다
 - 25명의 학생들의 몸무게는 72~82kg 사이이다
 - 25명의 학생들의 몸무게는 82kg 이상이다
- 이와 같이 평균이나 중앙값에 비해 사분위수는 보다 많은 결과 해석이 가능

단일변수 자료탐색 - 연속형 자료

사분위수 계산

```
bmi <- c(12.1, 12.8, 15, 19, 22, 28, 31.1, 34)
```

```
quantile(bmi) #사분위수 계산
```

```
summary(bmi) #요약통계량
```

```
> bmi <- c(12.1, 12.8, 15, 19, 22, 28, 31.1, 34)
> quantile(bmi)
 0%    25%    50%    75%   100%
12.100 14.450 20.500 28.775 34.000
> summary(bmi)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 12.100  14.450  20.500  21.750  28.775  34.000
> |
```



단일변수 자료탐색 - 연속형 자료

산포

- 산포(distribution)는 주어진 자료의 값들이 흩어져 있는 정보를 의미
- 분산(variance)과 표준편차(standard deviation)로 파악

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

s = 표준편차
 s^2 = 분산
 n = 값들의 개수
 x_i = 개별 값
 \bar{x} = 평균

- 분산과 표준편차가 작다 -> 자료들의 관측값들이 평균값 주변에 모여있다
- 분산과 표준편차가 크다 -> 자료들의 관측값들이 평균값에서 멀리 떨어져있다

단일변수 자료탐색 - 연속형 자료

산포

```
bmi <- c(12.1, 12.8, 15, 19, 22, 28, 31.1, 34)
```

```
var(bmi) #분산
```

```
sd(bmi) #표준편차
```

```
range(bmi) #값의 범위
```

```
diff(bmi) #최대값과 최소값 간의 차이
```

```
> bmi <- c(12.1, 12.8, 15, 19, 22, 28, 31.1, 34)
> var(bmi)
[1] 71.85143
> sd(bmi)
[1] 8.476522
> range(bmi)
[1] 12.1 34.0
> diff(bmi)
[1] 0.7 2.2 4.0 3.0 6.0 3.1 2.9
> |
```

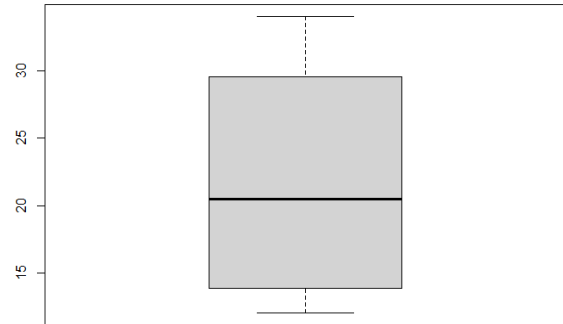
단일변수 자료탐색 - 연속형 자료

산포를 상자그래프로 나타내기

```
bmi <- c(12.1, 12.8, 15, 19, 22, 28, 31.1, 34)
```

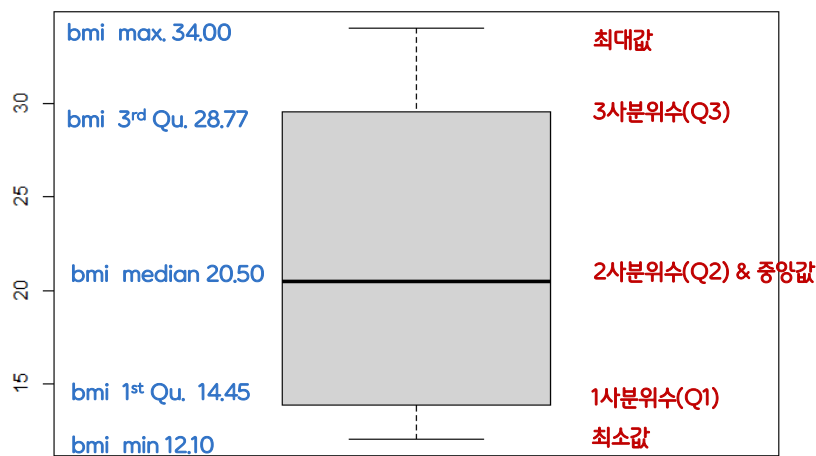
```
boxplot(bmi)
```

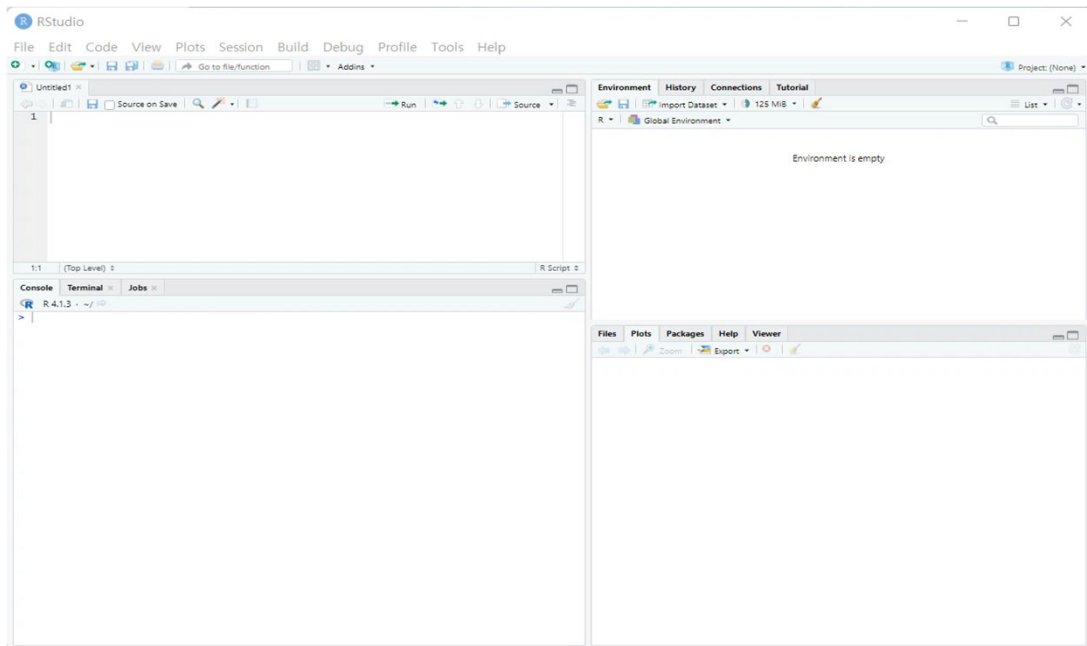
```
> bmi <- c(12.1, 12.8, 15, 19, 22, 28, 31.1, 34)
> boxplot(bmi)
> summary(bmi)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
12.10  14.45   20.50   21.75  28.77   34.00
```



단일변수 자료탐색 - 연속형 자료

TIP : 상자그래프 보는 법





SUMMARY

- 데이터 구조 파악 - 자료 특성 기반 분류
 - 범주형 자료 : 명목척도, 서열척도
 - 연속형 자료 : 구간척도, 비율척도
- 데이터 구조 파악 - 변수 개수 기반 분류
 - 단일변수 자료
 - 다중변수 자료

SUMMARY

▪ 단일변수 범주형 자료 탐색

- 범주형 자료는 크기를 갖지 않기 때문에 연산 불가능
- 자료에 포함된 관측값들을 종류별로 세는 것
- table() 함수는 데이터 빈도 분할표를 자동으로 만든다
 - ✓ table(데이터프레임1, 데이터프레임2)
- 그래프
 - ✓ barplot(table)
 - ✓ pie(table)

SUMMARY

▪ 단일변수 연속형 자료 탐색

- 관측값들이 크기를 가지기 때문에 다양한 분석이 방법 존재
- 자료 분석

<ul style="list-style-type: none"> ✓ 평균 mean() ✓ 사분위수 계산 quantile() ✓ 분산 var() ✓ 값의 범위 range() 	<ul style="list-style-type: none"> ✓ 중앙값 median() ✓ 요약통계량 summary() ✓ 표준편차 sd() ✓ 최대값과 최소값의 차이 diff()
--	---
- 그래프
 - ✓ boxplot()

