제5강
단일변수 자료의 탐색

Section 01

자료의 종류

1. 자료의 특성에 따른 분류

그림 5-1 자료의 특성에 따른 분류

1.1 범주형 자료

- 범주형 자료(categorical data)는 질적 자료(qualitative data)라고도 부르며, 성별과 같이 범주 또는 그룹으로 구분할 수 있는 값으로 구성된 자료

1. 자료의 종류

- 범주형 자료의 값들은 기본적으로 숫자로 표현할 수 없고, 대소(大小) 비교나 산술 연산이 적용되지 않음

범주형 자료	범주형 자료의 표현
성별	M, F, F, M, M, M, F
혈액형	A, B, O, AB, B, A, O
선호하는 색	빨강, 파랑, 노랑, 빨강, 초록, 검정
찬성 여부	YES, NO, NO, YES, NO

표 5-1 범주형 자료의 예

- 아래와 같이 범주형 자료를 숫자로 표기했다고 해서 계산 가능한 연속형 자료가 되는 것은 아님

- 성별 : 0, 1
- 혈액형 : 1, 2, 3, 4

1. 자료의 종류

1.2 연속형 자료

- 연속형 자료(**numerical data**)는 양적자료(**quantitative data**)라고도 부르며, 크기가 있는 숫자들로 구성된 자료
- 연속형 자료의 값들은 대소 비교가 가능하고, 평균, 최댓값, 최솟값과 같은 산술 연산이 가능

연속형 자료	연속형 자료의 표현
몸무게	57.4, 64.1, 71.0, 65.1, 90.1
키	162, 180, 174, 171, 181, 167
일평균 온도	19.1, 20.5, 20.5, 21.1, 22.0
자녀의 수	0, 2, 1, 3, 0, 1, 2

표 5-2 연속형 자료의 예

2. 변수의 개수에 따른 분류

- 통계학에서 말하는 변수는 우리가 **R**에서 배운 변수와는 의미상 다소 차이가 있음
- 통계학에서의 변수는 우리가 '연구, 조사, 관찰하고 싶은 대상의 특성'을 말하며, 키, 몸무게, 혈액형, 매출액, 습도, 미세먼지 농도 등

그림 5-2 변수의 개수에 따른 분류

- 단일변수 자료(**univariate data**): 하나의 변수로만 구성된 자료, '일변량 자료'라고도 부름
- 다중변수 자료(**multivariate data**): 두 개 이상의 변수로 구성된 자료, 다변량 자료라고 부름. 특별히 두 개의 변수로 구성된 자료를 이변량 자료(**bivariate data**)라고 함

1. 자료의 종류

몸무게
62.4
65.3
59.8
46.5
49.8
58.7

(a) 단일변수 자료

키	몸무게	성별
168.4	62.4	M
169.5	65.3	F
172.1	59.8	F
185.2	46.5	M
173.7	49.8	M
175.2	58.7	F

(b) 다중변수 자료

그림 5-3 단일변수 자료와 다중변수 자료

- **R에서는 단일변수 자료는 벡터에, 다중변수 자료는 매트릭스나 데이터 프레임에 저장하여 분석**
- 매트릭스 또는 데이터 프레임 형태의 자료에서 하나의 열(column)이 하나의 변수를 나타냄
- **열(column)의 개수 = 변수의 개수**

1. 자료의 종류

그림 5-4 변수의 개수와 자료의 특성에 따른 분류

- 변수의 개수와 자료의 특성에 따라 세분화된 분류가 가능
- 세분화된 분류에 따라 각각 서로 다른 분석 방법들이 존재

Section 02

단일변수 범주형 자료의 탐색

2. 단일변수 범주형 자료의 탐색

- 단일변수 범주형 자료(또는 일변량 질적 자료): 특성이 하나이면서 자료의 특성이 범주형인 자료
- 범주형 자료에 대해서 할 수 있는 기본적인 작업은 자료에 포함된 관측값들의 종류별로 개수를 세는 것
- 개수를 세면 종류별 비율을 알 수 있음
- 막대그래프나 원그래프의 작성이 가능
- 단일변수 범주형 자료의 예: 학생들이 선호하는 계절

WINTER	SUMMER	SPRING	SUMMER	SUMMER
FALL	FALL	SUMMER	SPRING	SPRING

2. 단일변수 범주형 자료의 탐색

1. 도수분포표의 작성

코드 5-1

```
favorite <-c('WINTER', 'SUMMER', 'SPRING', 'SUMMER', 'SUMMER',
             'FALL', 'FALL', 'SUMMER', 'SPRING', 'SPRING')
favorite                                  #favorite의 내용 출력
table(favorite)                           #도수분포표 계산
table(favorite)/length(favorite)          #비율 출력
```

```
> favorite <- c('WINTER', 'SUMMER', 'SPRING', 'SUMMER', 'SUMMER',
+              'FALL', 'FALL', 'SUMMER', 'SPRING', 'SPRING')
> favorite
 [1] "WINTER" "SUMMER" "SPRING" "SUMMER" "SUMMER" "FALL"   "FALL"
 [8] "SUMMER" "SPRING" "SPRING"
> table(favorite)                    # 도수분포표 계산
favorite
  FALL SPRING SUMMER WINTER
     2      3      4      1
> table(favorite)/length(favorite)   # 비율 출력
favorite
  FALL SPRING SUMMER WINTER
   0.2    0.3    0.4    0.1
```

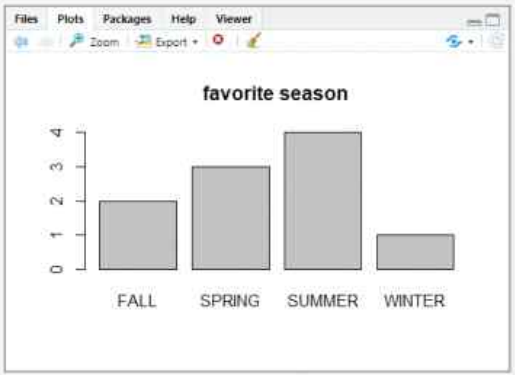
2. 단일변수 범주형 자료의 탐색

2. 막대그래프의 작성

코드 5-2

```
ds <-table(favorite)
ds
barplot(ds, main='favorite season')
```

```
> ds <- table(favorite)
> ds
favorite
  FALL SPRING SUMMER WINTER
     2      3      4      1
> barplot(ds, main='favorite season')
```



2. 단일변수 범주형 자료의 탐색

3. 원그래프의 작성

코드 5-3

```
ds <-table(favorite
ds
pie(ds, main='favorite season')
```

```
> ds <- table(favorite)
> ds
favorite
  FALL SPRING SUMMER WINTER
     2      3      4      1
> pie(ds, main='favorite season')
```

2. 단일변수 범주형 자료의 탐색

4. 숫자로 표현된 범주형 자료

- 숫자 형태의 범주형 자료도 문자 형태의 범주형 자료와 마찬가지로 도수분포를 계산한 후 막대그래프와 원그래프를 그려서 자료의 내용을 확인
- 학생 **15**명이 선호하는 색깔을 조사한 자료

2, 3, 2, 1, 1, 2, 2, 1, 3, 2, 1, 3, 2, 1, 2

(1=초록, 2=빨강, 3=파랑)

코드 5-4

```
favorite.color <-c(2, 3, 2, 1, 1, 2, 2, 1, 3, 2, 1, 3, 2, 1, 2)
ds <-table(favorite.color)
ds
barplot(ds, main='favorite color')
colors <-c('green', 'red', 'blue')
names(ds) <-colors          #자료값 1,2,3을 green, red, blue로 변경
ds
barplot(ds, main='favorite color', col=colors)   #색 지정 막대그래프
pie(ds, main='favorite color', col=colors)       #색 지정 원그래프
```

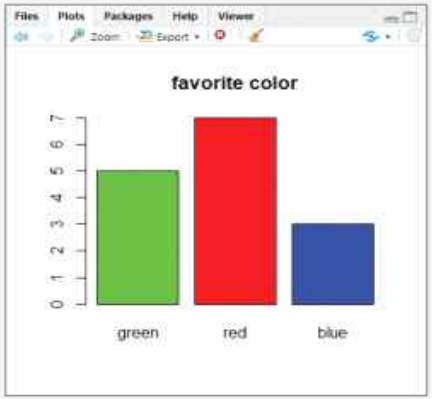
2. 단일변수 범주형 자료의 탐색

```
> favorite.color <- c(2, 3, 2, 1, 1, 2, 2, 1, 3, 2, 1, 3, 2, 1, 2)
> ds <- table(favorite.color)
> ds
favorite.color
1 2 3
5 7 3
> barplot(ds, main='favorite color')
```

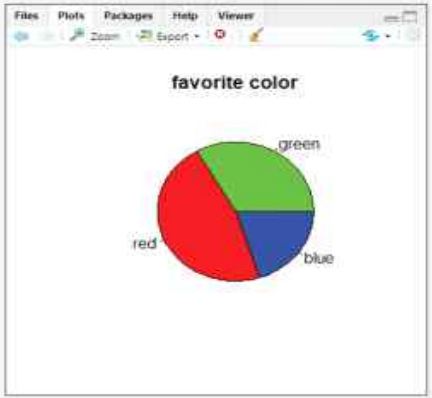

```
> colors <- c('green', 'red', 'blue')
> names(ds) <- colors        # 자료값 1,2,3을 green, red, blue로 변경
> ds
```

2. 단일변수 범주형 자료의 탐색

```
green   red  blue
    5     7     3
> barplot(ds, main='favorite color', col=colors)      # 색 지정 막대그래프
```




```
> pie(ds, main='favorite season', col=colors)         # 색 지정 원그래프
```



여기서 잠깐! 플롯 창의 Zoom 아이콘과 Export 아이콘

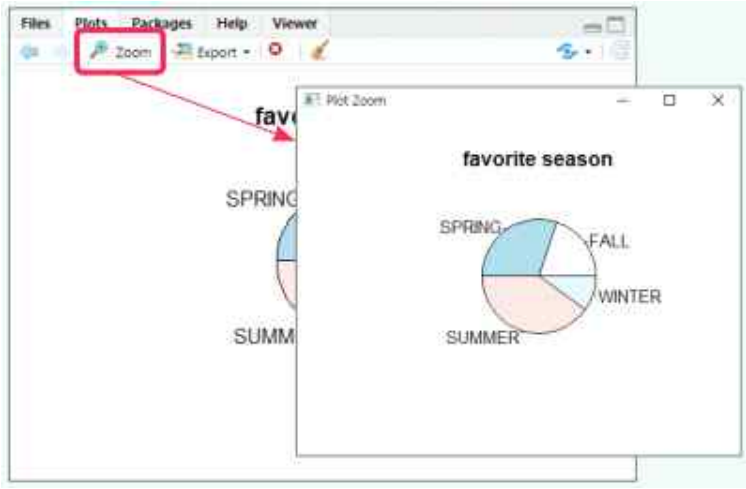


그림 5-5 [Zoom] 아이콘의 활용

Section 03

단일변수 연속형 자료의 탐색

3. 단일변수 연속형 자료의 탐색

1. 평균과 중앙값

- 연속형 자료는 관측값들이 크기를 가지기 때문에 범주형 자료에 비해 다양한 분석 방법이 존재
- 평균, 중앙값 : 전체 데이터를 대표할 수 있는 값
- 평균

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

중앙값과 평균은 일치할 수 도 있지만 대부분 틀리다.

- 중앙값(median) : 자료의 값들을 크기순으로 일렬로 줄 세웠을 때, 가장 중앙에 위치하는 값

좌측 그림은 120값 때문에 평균이 우측으로 치우쳐 있다.이런 경우는 절사 평균으로 계산하는 것이 올바른 방법이다.

그림 5-6 평균과 중앙값

- 절사평균(trimmed mean)은 자료의 관측값들 중에서 작은 값들의 하위 n%와 큰 값들의 상위 n%를 제외하고 중간에 있는 나머지 값들만 가지고 평균을 계산

3. 단일변수 연속형 자료의 탐색

코드 5-5

```
weight <-c(60, 62, 64, 65, 68, 69)
weight.heavy <-c(weight, 120)
weight
weight.heavy

mean(weight)                          #평균
mean(weight.heavy)                    #평균

median(weight)                        #중앙값
median(weight.heavy)                  #중앙값

mean(weight, trim=0.2)                #절사평균(상하위 20% 제외)
mean(weight.heavy,trim=0.2)           #절사평균(상하위 20% 제외)
```

평균, 절사평균, 중앙값은 각각의 특징이 존재하므로 본인이 분석하고자 하는 자료에 어떤 방법을 적용하는 것이 좋을지는 스스로가 판단해서 분석해야 할 것이다.

3. 단일변수 연속형 자료의 탐색

```
> weight <- c(60, 62, 64, 65, 68, 69)
> weight.heavy <- c(weight, 120)
> weight
[1] 60 62 64 65 68 69
> weight.heavy
[1]  60  62  64  65  68  69 120
> mean(weight)                    # 평균
[1] 64.66667
> mean(weight.heavy)              # 평균
[1] 72.57143
> median(weight)                  # 중앙값
[1] 64.5
> median(weight.heavy)            # 중앙값
[1] 65
> mean(weight, trim=0.2)          # 절사평균(상하위 20% 제외)
[1] 64.75
> mean(weight.heavy,trim=0.2)     # 절사평균(상하위 20% 제외)
[1] 65.6
```

매개변수 trim은 상하위 몇 %정도 제외 후 평균을 구할 것인지를 지정한다.

평균은 120라는 데이터로 차이가 많이 나지만, 중앙값은 특이값 즉 120에 영향을 상대적으로 덜 받는 것을 알 수가 있다.

2. 사분위수

- **사분위수(quatile)란 주어진 자료에 있는 값들을 크기순으로 나열했을 때 이것을 4등분하는 지점에 있는 값들을 의미**
- 자료에 있는 값들을 4등분하면 등분점이 3개 생기는데, 앞에서부터 '제1사분위수(Q1)', '제2사분위수(Q2)', '제3사분위수(Q3)'라고 부르며, 제2사분위수(Q2)는 중앙값과 동일
- 전체 자료를 4개로 나누었기 때문에 4개의 구간에는 각각 25%의 자료가 존재

그림 5-7 사분위수의 예

평균이나 중앙값이 하나의 값으로 전체의 특성을 추정해 볼 수 있는 도구인 것처럼 사분위수는 세 개의 값으로 전체의 특성을 추정하는데 사용되며 하나의 값보다는 세 개의 값으로 전체의 특성을 추정하므로 보다 많은 정보를 줄 수 있다.

3. 단일변수 연속형 자료의 탐색

- **100명의 학생을 대상으로 영어시험을 본 결과에 대해 사분위수를 구하였더니 Q1=60, Q2=80, Q3=90이라고 가정하면 →**

  25명의 학생은 성적이 60점 미만이다.

  25명의 학생은 성적이 60점~80점 사이이다.

  25명의 학생은 성적이 80점~90점 사이이다.

  25명의 학생은 성적이 90점 이상이다.

  90점 이상인 학생이 25명이나 되기 때문에 이번 영어시험은 매우 쉬웠다.

  전체 50%의 학생이 80점 이상의 성적을 받았다.

3. 단일변수 연속형 자료의 탐색

코드 5-6

```
mydata <-c(60, 62, 64, 65, 68, 69, 120)
quantile(mydata)                    #25% 단위로 구간을 나누어 계산
quantile(mydata, (0:10)/10)         #10% 단위로 구간을 나누어 계산
summary(mydata)
```

몇 개의 구간으로 나눌지 결정하는 인자값

```
> mydata <- c(60, 62, 64, 65, 68, 69, 120)
> quantile(mydata)
  0%  25%  50%  75% 100%
60.0 63.0 65.0 68.5 120.0
> quantile(mydata, (0:10)/10)        # 10% 단위로 구간을 나누어 계산
  0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
60.0 61.2 62.4 63.6 64.4 65.0 66.8 68.2 68.8 89.4 120.0
> summary(mydata)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  60.00   63.00   65.00   72.57   68.50  120.00
```

0%는 최소값, 100%는 최대값을 나타냄.

통상, 사분위수를 구할 때, 가장 일반적으로 많이 사용하는 함수가 summary()이다. 최소값, 최대값, 중앙값, 평균이 함께 출력되어 편리하다.

3. 산포

- **산포(distribution)란 주어진 자료에 있는 값들이 퍼져 있는 정도(흩어져 있는 정도)**
- **산포는 수학시간에 배운 분산(variance)과 표준편차(standard deviation)를 가지고 파악**
- **분산 : 주어진 자료의 각각의 값들이 평균으로부터 떨어져 있는 정도를 계산 후 합산한 후, 값들의 개수로 나누어 계산함.**

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

- **표준편차 : 분산의 제곱근으로 계산함.**

$$s = \sqrt{s^2}$$

- **자료의 분산과 표준편차가 작다는 의미는 자료의 관측값들이 평균값 부근에 모여 있다는 뜻이며, 반대라면 관측값들이 평균값에서 멀리 흩어져서 분포함을 의미한다.**

3. 단일변수 연속형 자료의 탐색

코드 5-7

```
mydata <-c(60, 62, 64, 65, 68, 69, 120)
var(mydata)                    #분산
sd(mydata)                     #표준편차
range(mydata)                  #값의 범위
diff(range(mydata))            #최댓값, 최솟값의 차이
```

```
> var(mydata)                  # 분산
[1] 447.2857
> sd(mydata)                   # 표준편차
[1] 21.14913
> range(mydata)                # 값의 범위
[1] 60 120
> diff(range(mydata))          # 최댓값, 최솟값의 차이
[1] 60
```

diff()함수는 두 값의 차이를 알려주는 용도이다. 여기선 매개변수값이 최소값과 최대값이므로 그 차이가 60인 것이다. 이 값이 의미가 있는 이유는 최대값과 최소값의 차이가 크면 관측값들이 넓게 퍼져있다는 의미이고 반대라면 관측값이 좁게 모여있다는 뜻인 것이다.

감사합니다.