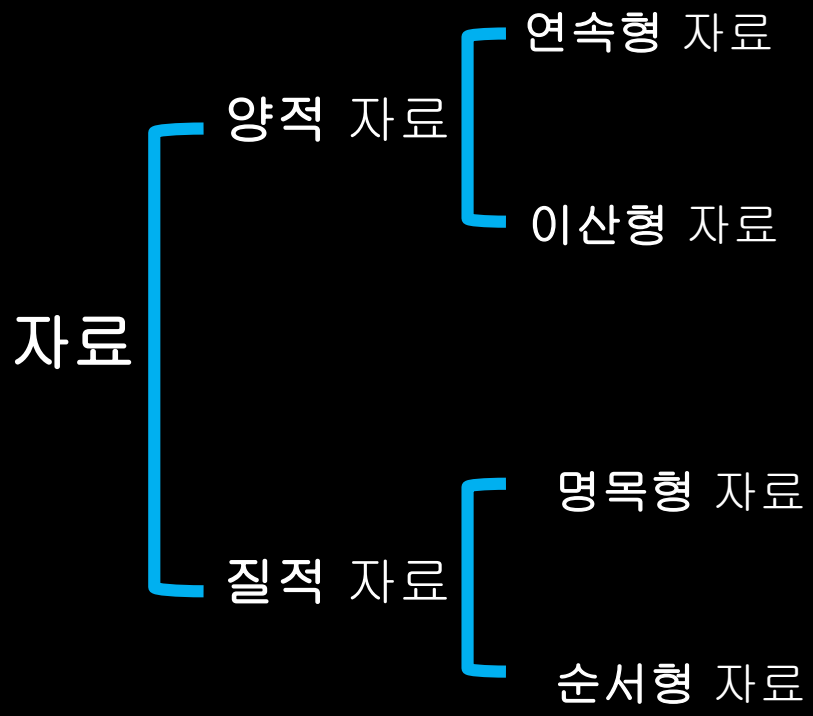
# 2. 자료의 정리 Description of Data

## 2.1 자료의 종류 Types of Data

“DATUM”
자료, 정보등의 의미

“DATA”
Datum의 복수형

**통계학**이란
**자료**라는 **재료** 를 이용하여
여러 가지 요리를 만들어내는 것과 같음


자료
양적 자료
연속형 자료
이산형 자료
질적 자료
명목형 자료
순서형 자료

## 양적 자료 (quantitative, numerical

자료 그 자체가 숫자와 일대일로 대응

**연속형 자료** (continuous data)

일정구간의 실수 값을 모두 취할 수 있는 자료

예) 혈압, 몸무게

**이산형 자료** (discrete data)

정수 값을 취하는 자료

예) 어느 학급의 여학생수, 연간 결혼 건수

## 질적 자료 (qualitative data)

자료 그 자체가 숫자의 개념을 가지는 것이 아닌 구분하는 개념을 가짐

**명목형 자료** (nominal data)
구분을 위해 숫자를 대응시킨 자료

예) 성별, 피부색
남→1, 여 →0
남 → −1, 여 → +1

**순서형 자료** (ordinal data)
범주들이 순서의 개념을 가지는 자료

예) 상>중>하: 1,2,3 또는 3,2,1 (O)
2,1,3 또는 3,1,2 (X)

# 2. 자료의 정리 Description of Data

## 2.2 표와 그래프 Table and Graphs

## 질적 자료 (qualitative data)

### 도수분포표 (frequency table)

예제 : 후보 A, B, C에 대해
2,800명이 투표한 결과

---

A : 1520표, B : 770표, C : 510표

| 후보자 | 도수 |
|---|---|
| A | 1,520 |
| B | 770 |
| C | 510 |
| 계 | 2,800 |

▶ R code

```
a <- rep("A", 1520)
b <- rep("B",770)
c <- rep("C",510)
x <- c(a,b,c)

table(x)

y <- as.matrix(table(x))
freq <- y[,1]

relative_freq <- freq/sum(y)
z <- cbind(freq, relative_freq)
z
```

## 질적 자료 (qualitative data)

### 파이차트 (pie chart)

▶ R code

```
x <- c(1520, 770, 510)
lab <- c("A", "B", "C")
y <- round(x/sum(x)*100,
digits=1)
w <- paste(lab, "(", y, "%", ")")

pie(x, labels=w, main="파이챠트")
```

예제 : 30페이지로 이루어진 보고서에서
각 페이지당 오자의 개수

▶ R code

```
x <-
c(1,1,1,3,0,0,1,1,1,0,2,2,0,0,0,1,2,1,2,0,0,1,6,4,3,3,1,2,4,0)
y <- as.matrix(table(x))

freq <- y[,1]
rel_freq <- freq/sum(freq)
csum <- cumsum(freq)
c_rel_freq <- csum/sum(freq)
z <- cbind(freq, rel_freq, csum, c_rel_freq)
```

## 양적 자료 (quantitative, numerical

히스토그램 (histogram)

줄기-잎 그림 (stem-and-leaf plot)

▶ R code

```
#히스토그램

data()
hist(faithful$waiting)


#줄기-잎 그림

stem(faithful$waiting)
```

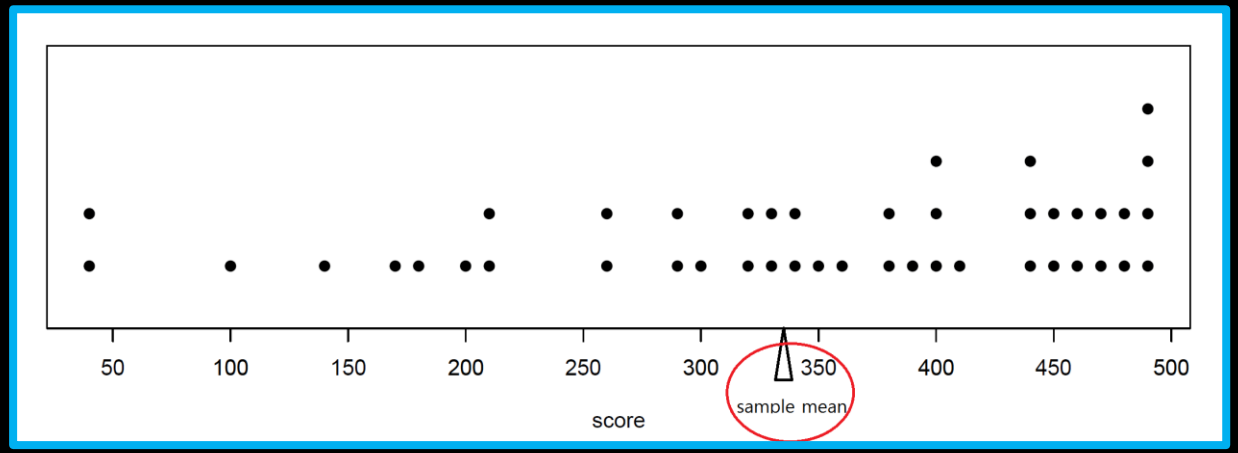
# 2. 자료의 정리 Description of Data

## 2.3 중심과 퍼짐 측도 Measures of Center and Dispersion

## (1) 표본 평균(sample mean)

$n$개의 자료 : $x_1, \cdots, x_n$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$ : n 개 값의 무게 중심


50
100
150
200
250
300
350
400
450
500
score
sample mean

## (2) 표본 중간값(sample median)

$n$개의 자료를 작은 것으로부터
크기 순으로 나열하였을 때
**가운데**에 있는 값

- n이 **홀수**일 때

⇒ 하나의 중간값이 유일하게 존재

- n이 **짝수**일 때

⇒ $\frac{n}{2}$번째와 $(\frac{n}{2}+1)$번째 값의

평균을 구하여 중간값으로 사용

## (2) 표본 중간값(sample median)

**자료** 1, 3, 4, 6, 6, 7, 8, 8, 9, 10, 15

**표본 평균** (1+3+4+6+6+7+8+8+9+10+15)/11 = 7

**표본 중간값** 11개 자료의 6번째 값인 7

만약 15가 150으로 바뀌게 되면
표본 중간값은 변하지 않지만
표본 평균은 19.27로 커지게 됨.

표본평균은 이상치(outlier)에 대해 민감(sensitive)하지만
표본 중간값은 이상치에 거의 영향을 받지 않음(robust).
자료에 이상치가 있을 경우 자료의 중심을 나타내는 값으로
표본평균보다 표본 중간 값이 더 좋은 측도

## (3) 표본 분위수(sample quantile)

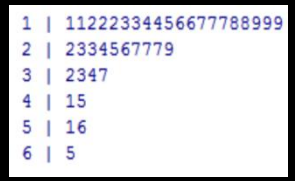
**표본 100p 백분율** (sample 100p-th percentile), $0 < p < 1$

: 100p% 개의 자료는 그 값보다 작거나 같고
100(1-p)% 개의 자료는 그 값보다 크거나 같음.

- $p = 0.25$ ⇒ 25% 백분율 = 제 **1** 분위수 (1st quartile) = $Q_1$
- $p = 0.50$ ⇒ 50% 백분율 = 제 **2** 분위수 (2nd quartile) = $Q_2$ *표본 중간값*
- $p = 0.75$ ⇒ 75% 백분율 = 제 **3** 분위수 (3rd quartile) = $Q_3$

# (4) 분포의 형태

### 오른편으로 긴 꼬리 형태

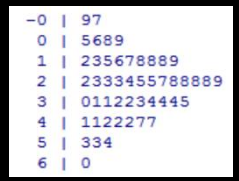



표본 평균 : 24.35897

표본 중간값 : 19

표본 평균 > 표본 중간값

### 대칭에 가까운 형태

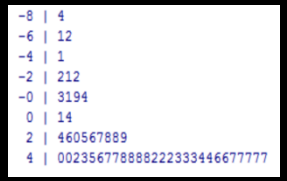



표본 평균 : 27.32653

표본 중간값 : 28

표본 평균 ≒ 표본 중간값

### 왼편으로 긴 꼬리 형태




표본 평균 : 26.32653

표본 중간값 : 42

표본 평균 < 표본 중간값

## (1) 표본 분산(sample variance)

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

: 표본 분산 (sample variance)

## (2) 표본 범위(sample range)

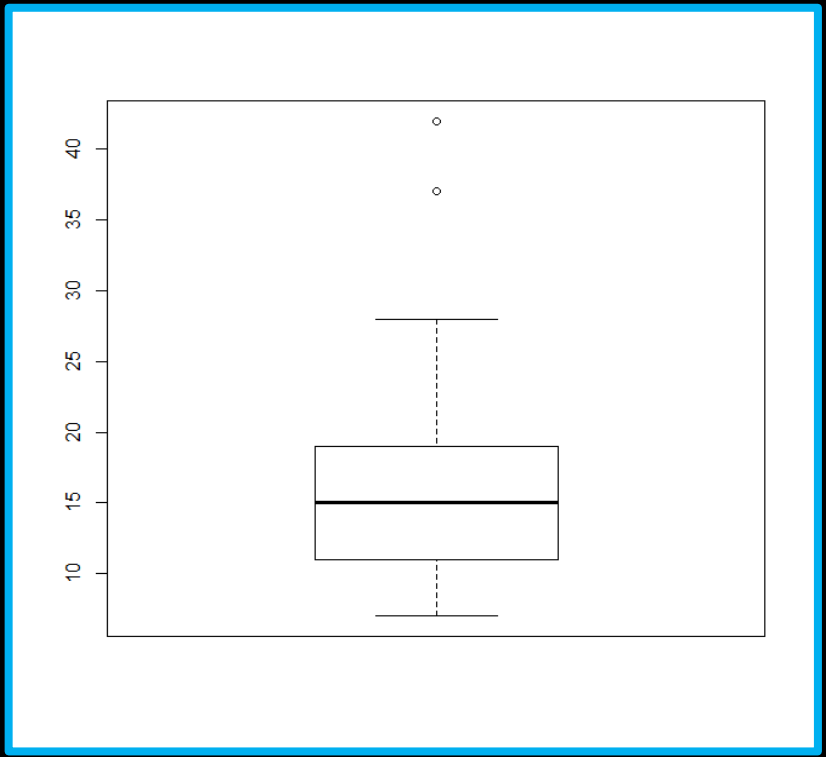
R = 최댓값 − 최솟값

: 표본 범위 (sample range)

$$IQR = Q_3 - Q_1$$

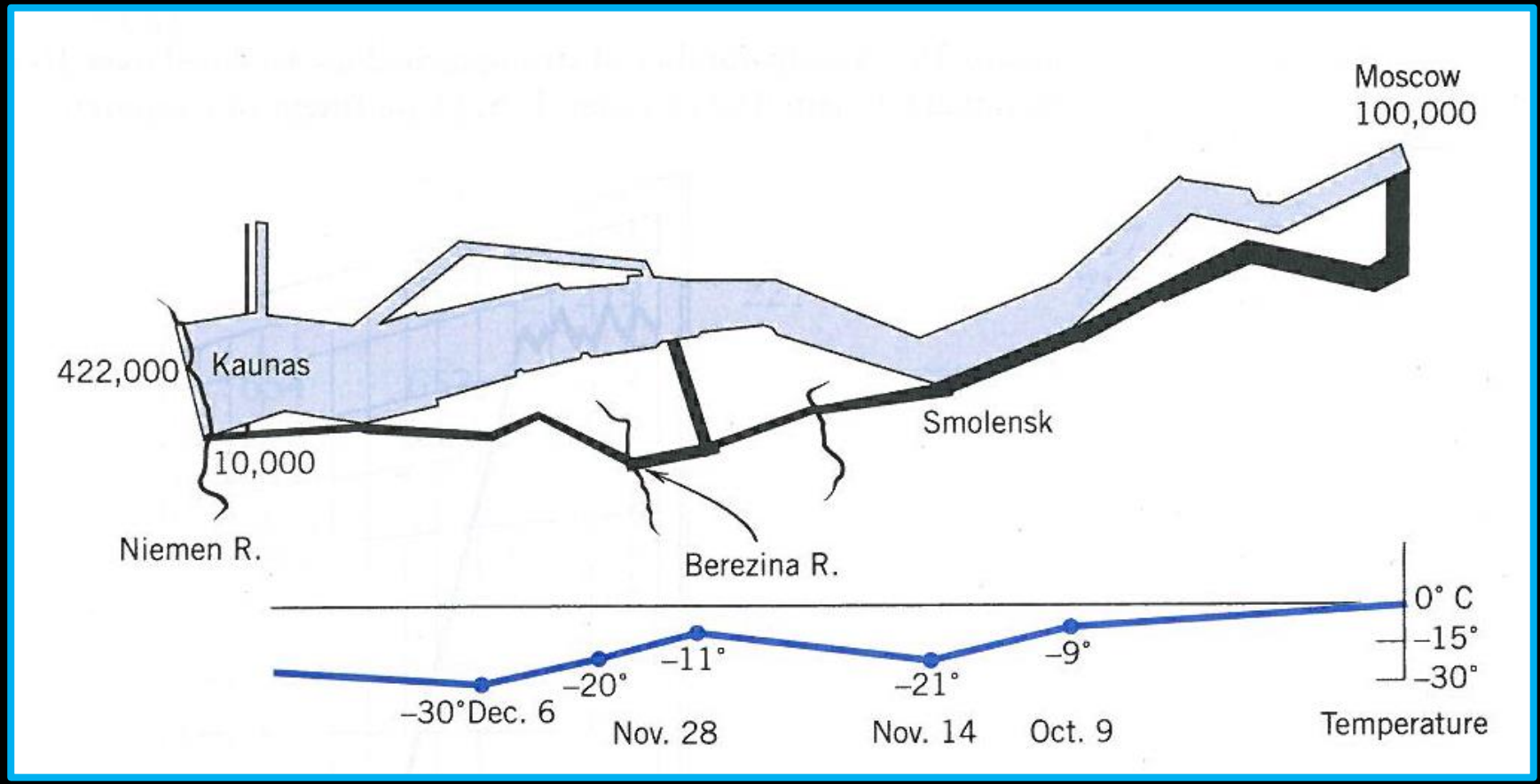
: 표본 사분위수 범위 (sample interquartile range)

# (3) 상자 그림(box plot)



▶ R code

```
data()
x <- stackloss$stack.loss
x

mean(x)

var(x)

sd(x)

quantile(x, c(0.1,0.25,0.5,0.95))

fivenum(x)

summary(x)

boxplot(x)
```


Moscow
100,000
422,000
Kaunas
10,000
Niemen R.
Berezina R.
Smolensk
0° C
-15°
-30°
-30°
Dec. 6
-20°
Nov. 28
-11°
-21°
Nov. 14
-9°
Oct. 9
Temperature

# 2. 자료의 정리 Description of Data

## 2.4 이변량 자료와 상관계수

Bivariate Data and Correlation Coefficient

“일변량 자료”
(univariate data)

하나의 변수에
대한 자료

“이변량 자료”
(bivariate data)

두개의 변수에
대한 자료

확장

“다변량 자료”
(multivariate data)

여러 개의 변수에
대한 자료

**두 변수가 모두 질적 자료인 경우**

첫 번째 자료는 $r$개의 범주, 두 번째 자료는 $c$개의 범주.

이러한 자료를 행렬의 형태로 요약한 표를

$$r \times c\text{분할표 } (r \times c \text{ contingency table})$$

라 부른다.

통계학 개론을 수강하는 400명의 학생들에게 시험을 본 후 문제수준에 대하여 조사

| | 어렵다 | 보통이다 | 쉽다 | 합 |
|---|---|---|---|---|
| 남 | 112 | 36 | 28 | 176 |
| 여 | 84 | 68 | 72 | 224 |
| 합 | 196 | 104 | 100 | 200 |

$2 \times 3$ 분할표

**두 변수 모두 양적 자료인 경우**

통계학과 1학년 남학생 15명의 IQ와 통계학 개론 중간고사 성적

$$(x_1, \quad y_1), \cdots, (x_{15}, \quad y_{15})$$

$x_i$는 $i$번째 학생의 IQ, $y_i$ 는 $i$번째 학생의 통계학 개론 중간고사 성적

$\Longrightarrow$ $n$개의 이변량 자료는

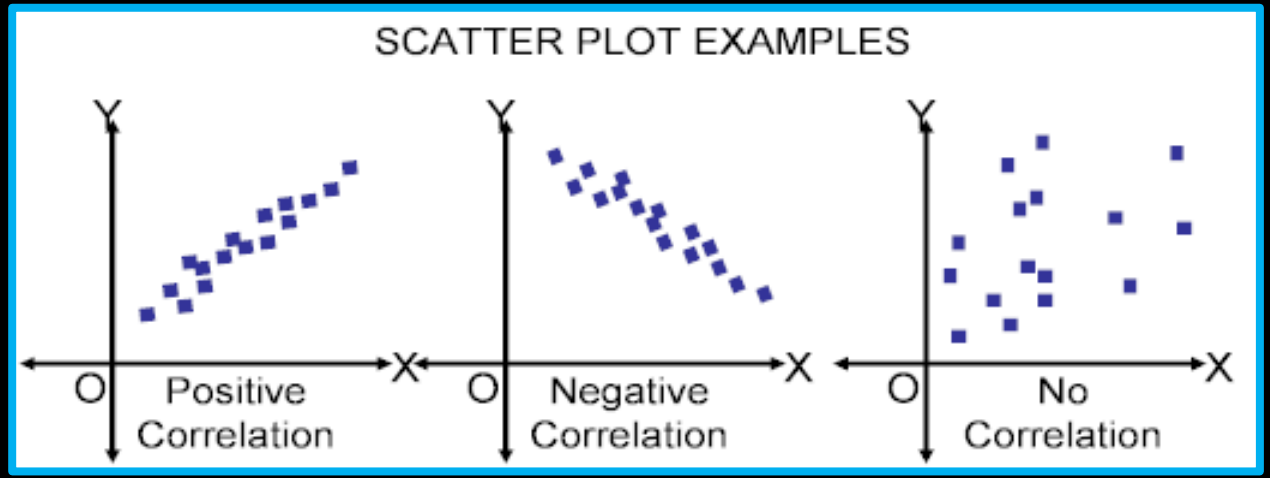
$$(x_1, \ y_1), \cdots, (x_n, \ y_n)$$

으로 표현

## (1) 산점도(scatter plot)

산점도 : 이차원 평면에 각 변수의 값에 해당되는 점을 찍은 그림

흔히, 산점도는 하나의 변수 값이 증가할 때,
다른 변수의 값이 증가 (또는 감소)하는 추세에 있거나
별 다른 함수관계를 보이지 않는 경우도 있음.



## (3) 허위상관과 잠복변수(spurious correlation and lurking

자료 : $(x_1, y_1), \cdots, (x_n, y_n)$

$x_i$ : $i$번째 도시의 연간 강력범죄 발생 건수

$y_i$ : $i$번째 도시의 교회 수

⇒ 매우 높은 양의 상관관계

⇒ “인구” 라는 잠복변수 인하여 범죄건수와 교회 수가 양의 상관관계를 가지는 것처럼 잘 못 판단 할 수 있음

⇒ 허위상관

## (2) 표본 상관계수(sample correlation

:두 변수의 선형적 함수 관계를 나타내는 측도

$$r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$$

$$= \frac{X, Y\text{의 표본 공분산}}{\sqrt{X\text{의 분산}}\sqrt{Y\text{의 분산}}}$$

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

▶ R code

```
x <- faithful$eruptions
y <- faithful$waiting

plot(x,y)

cor(x,y)
```