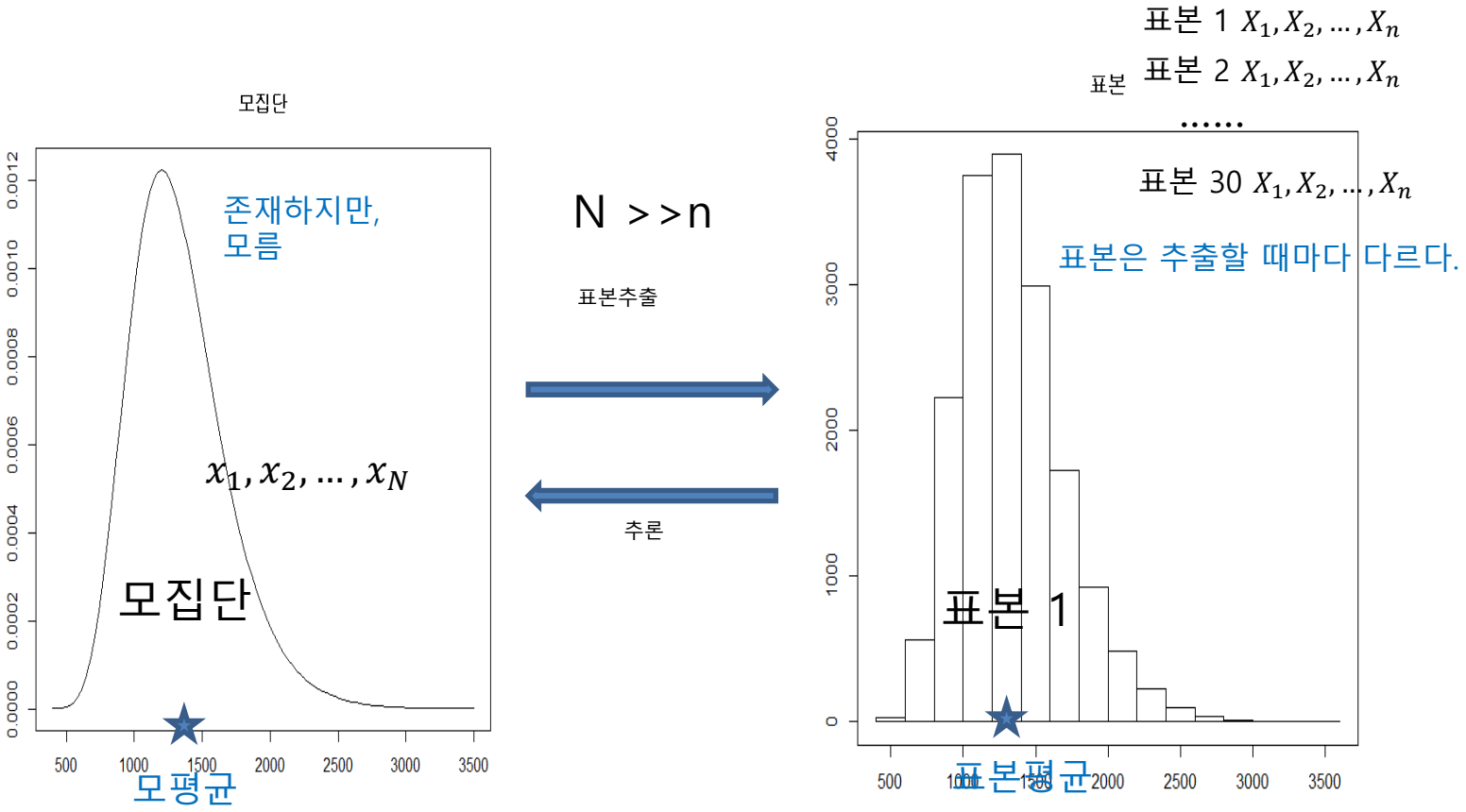
컴퓨터 응용통계

2장 일변량 기술통계

최경미

모집단과 표본
Population and sample



전국 20대 남성의 키

모집단(population)
모수(parameter)

- 모집단 전체 자료: $x_1, x_2, \cdots, x_{N-1}, x_N$, 모집단크기 N
- 모평균(Population mean):

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

- 모편차 (Population deviation):
  자료가 평균으로부터 떨어진 정도

$$x_i - \mu$$

  편차합 $= \sum_{i=1}^{n}(x_i - \mu) = \sum_{i=1}^{n} x_i - n\mu = n\mu - n\mu = 0$
- 모분산(Population variance):

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

  분산이 크면, 자료가 평균으로부터 멀리 흩어짐.
  분산이 작으면, 자료가 평균 주변에 모여 있음
- 모표준편차(Population standard deviation):

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

- 돈과 시간이 부족하여 모집단 전체 측정 불가능.
- 모수 μ, σ^2, σ 는 상수값으로 존재하지만 알 수 없음.
- 관측값에 대해서는 소문자 $\bar{x}$과 s^2 사용

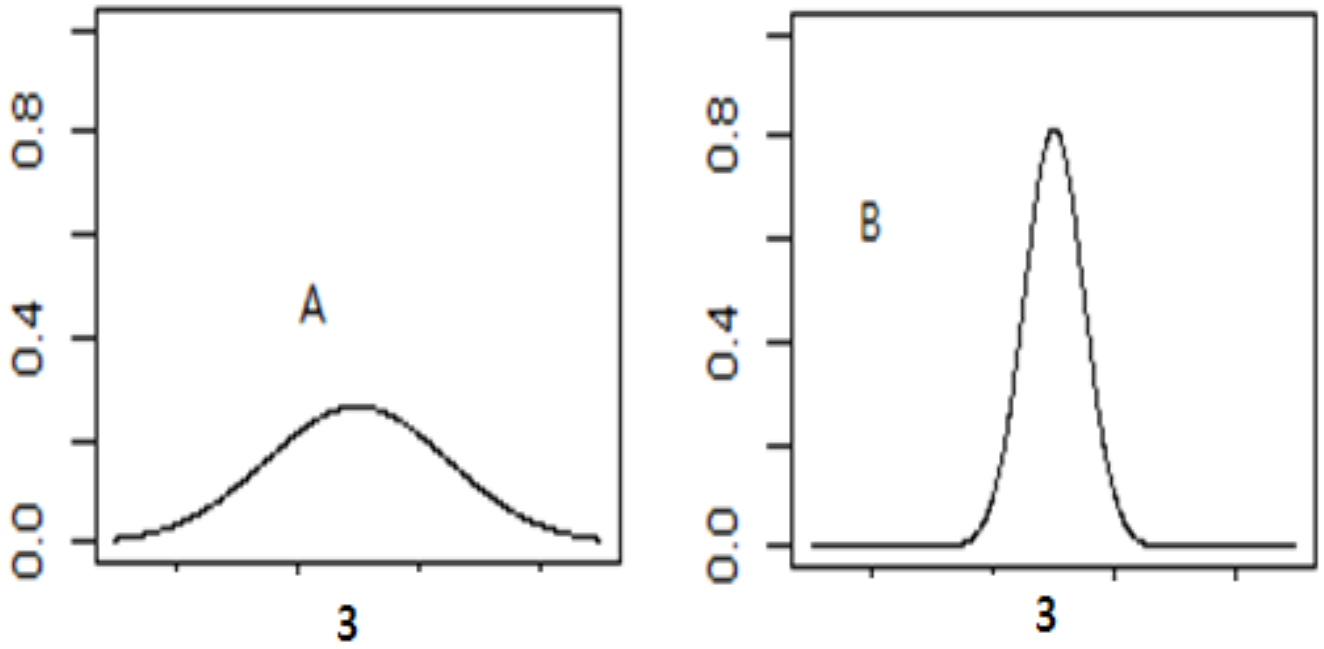
표본(sample)
추정치(estimator)

- 표본: $X_1, X_2, \cdots, X_{n-1}, X_n$
  표본크기 n ≪ N.
- 표본을 자료(data)라고 부름
- 표본은 추출할 때마다 달라짐.
  예제: 전국 20대 남성 키를 추정하기 위해서
  n=30명 표본1 추출.
  다시 n=30명 표본2 추출.
  표본1과 표본2는 다른 사람들
- 표본평균 $\hat{\mu} = \bar{X}$
- 표본분산 $\hat{\sigma}^2 = S^2$

$n \to \infty$일 때, $\bar{X} \to \mu$, $S^2 \to \sigma^2$

분산의 중요성

병 뚜껑 지름의 분포라면, 어느 것이 더 좋은가?



자료의 3가지 측정 척도

명목척도(nominal scale),
순서척도(ordinal scale),
비척도(ratio scale)

(1) 명목척도: 순서가 없는 변수
범주형 자료(categorical data)
집단(group)을 표현

예제 성별, 질병여부, 흡연여부, 음주여부, 불량여부, 직업, 브랜드, 처리종류 등

(2) 순서척도 (ordinal scale)
리커트 척도 (Likert scale)
의미척도 (semantic difference scale)

선호도 조사, 순서는 의미가 있지만, 순서의 차이는 측정 불가능함
부정확, 경향만 살핌

예제
매우 나쁨, 나쁨, 보통, 좋음, 매우 좋음
나쁨 – 매우 나쁨 ≠ 매우 좋음- 좋음

(3) 비척도(ratio scale):
0이 물리적인 의미를 가짐.
차이와 비율이 모두 측정 가능함.

예제 키, 몸무게, 길이, 무게, 면적, 체적, 점수, 밀도 등 연속값을 갖는 변수

(187-177) cm= (193-183) cm
80cm/100cm =0.8

크게 둘로 나누면...

- 범주형 자료 (집단 표현 O)
- 연속형 자료 (집단 표현 X)
- 자료의 측정 척도에 따라서, 사용할 수 있는 통계분석방법이 다르다.

예제 mtcars in R

Names	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21	6	160	110	3.9	2.62	16.46	0	1	4	4
Mazda RX4 Wag	21	6	160	110	3.9	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108	93	3.85	2.32	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
…	…	…	…	…	…	…	…	…	…	…	…

32개 관측치 11개 변수
명목척도: vs, am
순서척도: cyl, gear, carb
비척도: mpg, cyl, disp, hp, drat, wt, qsec, gear, carb

예제 mtcars in R

- BUT....
- 순서척도, 비척도에서 경우의 수가 적으면,
  명목척도로 취급되어, 범주형자료가 될 수 있음.

  cyl=4 소형차, cyl=6 중형차, cyl=8 대형차

범주형 자료와 연속형 자료
범주형 자료의 분포 (Distribution)

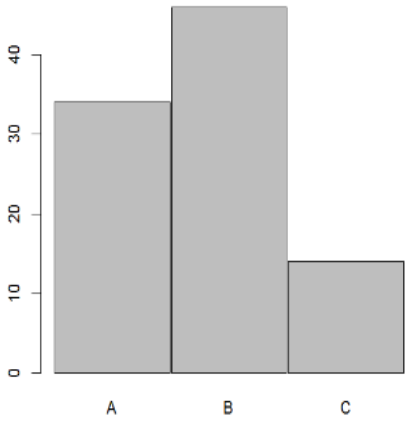
표와 그래프

막대그래프(Bar chart)
명목척도 또는 순서척도 자료에 적용
막대의 높이가 빈도 또는 상대빈도에 해당함

예제 빈도표(Frequency table)
상대 빈도표(Relative Frequency table)

A	B	C
34 (0.36)	46 (0.49)	14 (0.15)

막대 높이
=빈도 또는 상대빈도



R

```
# 범주형 자료 분석

am <- mtcars$am
table(am)
barplot(table(am), space=0, xlab=c("자동", "수동"))
pie(table(am),c("a","m"),col = c("orange", "blue"))

data <- subset(mtcars, cyl==4, "am")
am <- data$am
table(am)
barplot(table(am), space=0)
pie(table(am),c("a","m"),col = c("orange", "blue"))
```

각자 실습
6기통 차들의 엔진타입에 대한 표, 막대그래프, 파이차트 그리기

범주형 자료와 연속형 자료
연속형 자료의 분포 (Distribution)

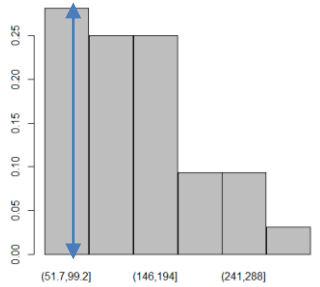
- **히스토그램(Histogram)**
  연속형 자료에 적용
  막대의 면적이 상대빈도에 해당함
  총면적이 1 (100%).

- **예제** hp of the mtcars
  (52 62 65 66 66 91 93 95 97] [105 109 110 110 110 113 123 123]
  (150 150 175 175 175 180 180 180] (205 215 230] (245 245 264] (335]

- 6개 구간에 대한 상대빈도(Relative frequency)

(51,99]	(99,146]	(146,194]	(194,241]	(241,288]	(288,335]
0.28125	0.25000	0.25000	0.09375	0.09375	0.03125

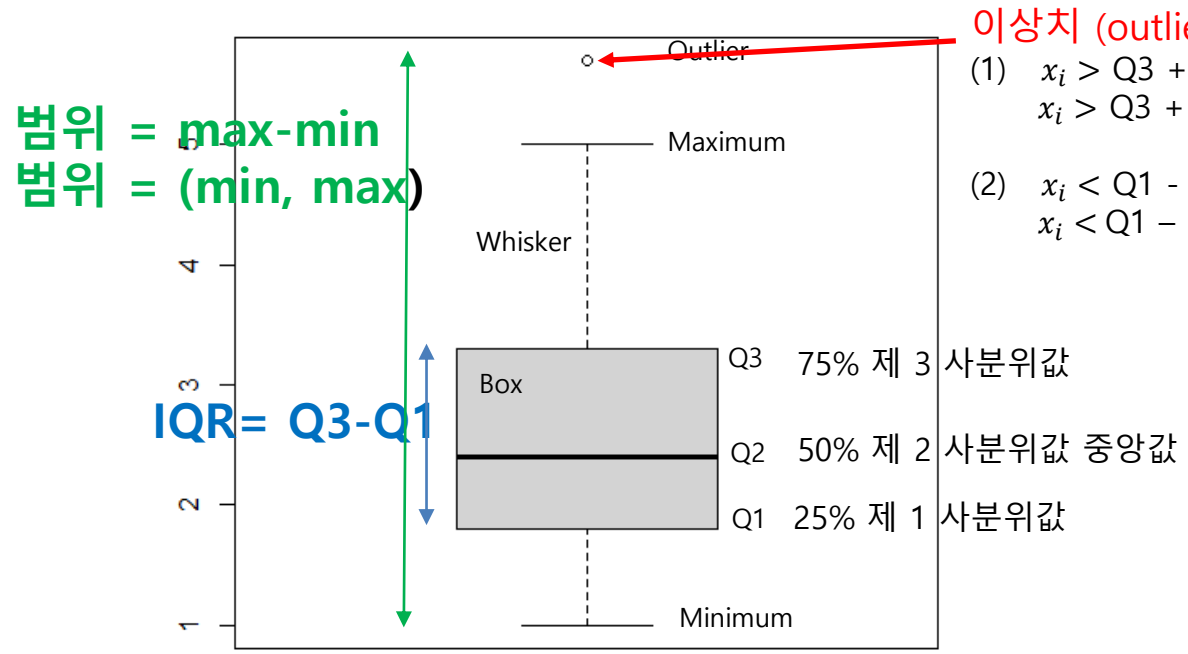
h=0.28125/(99.2-51.7)



연속형 자료의 분포

상자도표에서만 이상치와 최대값를 자동으로 구분함.
실제로는 이상치가 최대값임.

- **상자도표(Boxplot)** (Box and whisker plot):
- **예제** Data: 110 110 93 110 175 105 245 62 95 123 123 180 180 180 205 215 230
  66 52 65 97 150 150 245 175 66 91 113 264 175 335 109



이상치 (outlier)

(1) $x_i > Q3 + 1.5\ IQR$
$x_i > Q3 + 3.0\ IQR$

(2) $x_i < Q1 - 1.5\ IQR$
$x_i < Q1 - 3.0\ IQR$

Q3 75% 제 3 사분위값

Q2 50% 제 2 사분위값 중앙값

Q1 25% 제 1 사분위값

연속형 자료
분포의 중심 (Central tendency)

- 평균(Mean) $\bar{x}$: 모든 값을 더해서 표본크기로 나눔

  $\bar{x} = \frac{1}{n}\,(\,x_1 + x_2 + \,\ldots + x_n)$

- 중앙값(Median) m : 표본의 가운데 값.

  $x_{(1)} \leq \; x_{(2)} \leq \; \ldots \leq m \leq \cdots \leq \; x_{(n)}$

  $x_{(1)} \leq \; x_{(2)} \leq \; \ldots \leq m \leq \cdots \leq \; x_{(7)} \qquad m = x_{(4)}$

  $x_{(1)} \leq \; x_{(2)} \leq \; \ldots \leq m \leq \cdots \leq \; x_{(8)} \qquad m = (x_{(3)} + x_{(4)})/2$

- pth 백분위수(percentile): 표본 중 하위 100p%와 상위100(1-p)%

  Q1=25th 백분위수, 1st quartile

  Q2=50th 백분위수, the 2nd quartile, 중앙값, 중위수 (median)

  Q3=75th 백분위수, the 3rd quartile

연속형 자료
분포의 흩어짐(Dispersion)

- 모평균 $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$
- 모편차 $x_i - \mu$
- 편차합 =0
- 모분산(Population variance):

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2$$

- 모표준편차(Population standard deviation):

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

- 표본편차 $(x_i-\bar{x})$

$$\sum_{i=1}^{n}(x_i-\bar{x}) = \sum_{i=1}^{n} x_i - n\bar{x} = n\bar{x} - n\bar{x} = 0$$

- 표본분산(Sample variance):

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

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

- n이 아니라, $n-1$로 나눌 때, s^2이 σ^2에 더 잘 수렴한다.
- 표본표준편차(Sample standard deviation):

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

- 범위(Range) = max-min
- 사분위수범위 IQR = Q3 –Q1

예제
hp of the data mtcars

52 62 65 66 66 91 93 95 97 105 109 110 110 110
113 123 123 150 150 175 175 175 180 180 180 205 215
230 245 245 264 335

Mean $\bar{x}$ =146.7

Median m=123

Variance s^2= 4700.867

SD s = 68.56287

Range= (52,335) or 335-52=283

Q1=96.5, Q2=123, Q3=180

IQR= 180-96.5=83.5

Outlier 355 > Q3 + 3.0*IQR = 180 + 1.5*83.5= 305.25

Example

```
# barplot.R

x<-mtcars$hp
hist(x)
boxplot(x)

mean(x)
median(x)
quantile(x)

range(x)
var(x)
sd(x)

# 335
max(hp)
boxplot(x)$out
which(hp>180+ 1.5*83.5)
# 31st observation
```

예제

Data: 1,2,2,3,3,3,4,4,5
Mean $\bar{x}$=(1+2+2+3+3+3+4+4+5)/9=3
Median m=3
Deviation -2,-1,-1,0,0,0,1,1,2
Variance

$s^2 = \frac{1}{9-1}\{(-2)^2 + (-1)^2 + \ldots + 1^2 + 2^2\}$= 1.5

SD s = 1.224745
Range (1,5) or 5-1=4
Q1=2, Q2=3, Q3=4
IQR=Q3-Q1=2
Outlier 없음
Q3 + 3.0*IQR = 4 + 3.0*2=10

Example

```
# barplot.R

x<-c(1,2,2,3,3,3,4,4,5)
table(x)
hist(x)

boxplot(x)

mean(x)
median(x)
quantile(x)

range(x)
var(x)
sd(x)
```

예제

Data: 1,2,2,3,3,3,4,4,555
Mean $\bar{x}$=(1+2+2+3+3+3+4+4+555)/9=64.11111
Median m=3
Variance s^2= 33887.61
SD s = 184.0859
Range= (1,555) or 555-1=554
Q1=2, Q2=3, Q3=4
IQR=Q3-Q1=2
Outlier 555 > Q3 + 3.0*IQR = 4 + 3.0*2=10

Example

```
# barplot.R

x<-c(1,2,2,3,3,3,4,4,555)
table(x)
hist(x)

boxplot(x)

mean(x)
median(x)
quantile(x)

range(x)
var(x)
sd(x)
```

실습+ 과제2장 2, 3 번

2. R mtcars 기통수 (cyl)에 대하여 다음 물음에 답하라.

```
6  6  4  6  8  6  8  4  4  6  6  8  8  8  8  8  8  4  4  4  4  8  8  8  8  4  4  4  8
6  8  4
```

```
# 자료 가져오기
x<- mtcars$cyl      # 자료 가져오기
x                   # 자료 확인하기
```

(1) 빈도표와 상대빈도표를 작성하라.

(2) 막대그래프를 그려라.

(3) 원그래프를 그려라.

(4) 소형차(4기통), 중형차(6기통), 대형차(8기통)의 비율(%)은 얼마인가?

3. 붓꽃의 한 종류에 대하여 꽃받침 길이를 측정한 후, 다음과 같은 자료 x를 얻었다 (Fisher의 iris 데이터 중 setosa 일부).

```
5.1  4.9  4.7  4.6  5.0  5.4  4.6  5.0  4.4  4.9  5.4  4.8  4.8  4.3  5.8  5.7  5.4  5.1  5.7
5.1  5.4  5.1  4.6  5.1  4.8  5.0  5.0  5.2  5.2  4.7  4.8  5.4  5.2  5.5  4.9  5.0  5.5  4.9
4.4  5.1  5.0  4.5  4.4  5.0  5.1  4.8  5.1  4.6  5.3  5.0
```

```
# 자료 가져오기.

iris                  # 행렬(matrix)로 저장된 자료 보기

x<- iris[1:50, 1]     # 1행부터 50행까지 선택, 1열 선택
```

(1) 평균, 분산, 표준편차를 구하라.

(2) 중앙값, 최대값, 최소값, 범위, $Q1, Q2, Q3, IQR$을 구하라.

(3) 상자도표와 히스토그램을 그려라.

(4) $Q3 + 1.5 * IQR$을 벗어나는 이상치가 있는가? $Q1 - 1.5 * IQR$을 벗어나는 이상치가 있는가?