3강

데이터 요약 II

통계·데이터과학과 박서영 교수

한국방송통신대학교

데이터 요약 방법

- 질적 데이터의 요약
  - 막대그래프
- 양적 데이터의 요약
  - 히스토그램
  - 점도표
  - 평균, 분산, 표준편차
  - 상자그림
  - 중앙값, 사분위수 범위

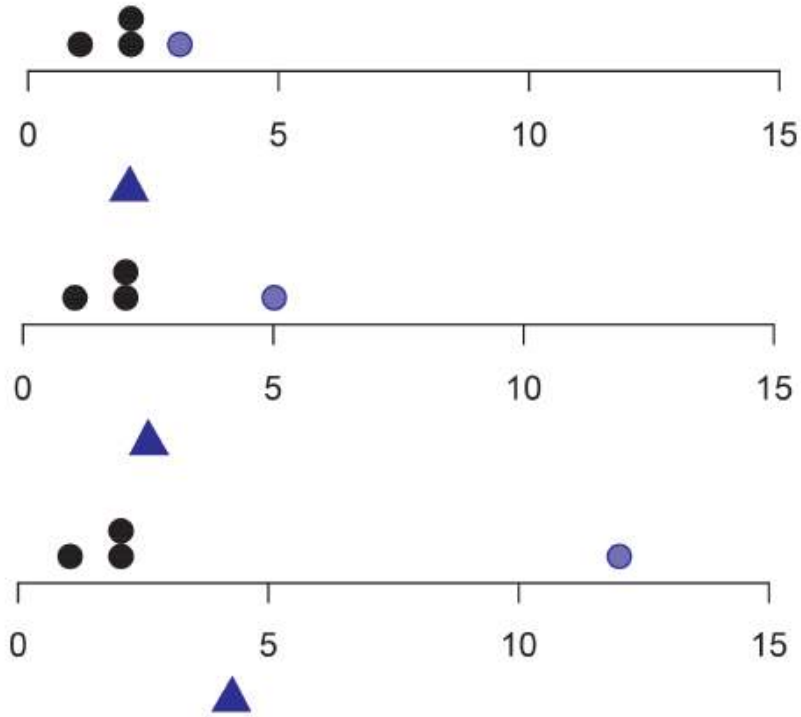
목차

01

상자그림, 중앙값, 사분위수 범위

평균의 특징

- 데이터의 분포가 좌우 대칭인 경우 평균은 분포의 가운데에 위치한다
- 데이터 중 하나라도 한쪽으로 치우치면 평균은 특이점 쪽으로 움직이게 된다



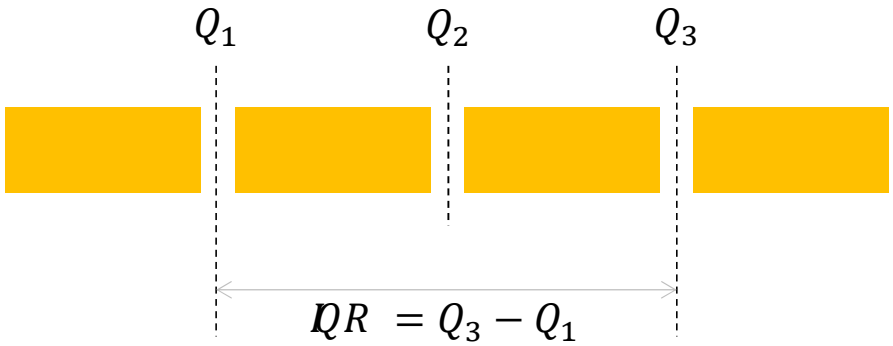
중앙값(median)

- 데이터를 크기 순서대로 늘어놓았을 때 정확히 중앙에 위치하는 값
  - 관찰값의 개수가 홀수일 때: 중앙에 위치하는 관찰값
    - 예: 1, 2, 3, 4, 5 중앙값=3
  - 관찰값의 개수가 짝수일 때: 중앙에 위치하는 2개 관찰값의 평균
    - 예: 1, 2, 3, 4, 5, 6 중앙값 = (3+4)/2=3.5
- 특이점의 영향을 거의 받지 않는다
- 분포가 한쪽으로 쏠려 있거나, 특이점이 존재하는 데이터를 요약할 때 주로 사용된다

사분위수(quartiles)

- 크기 순서대로 늘어놓은 데이터를 4등분하는 값
  - 1사분위수(Q_1): 전체 데이터 중 값이 낮은 1/4과 나머지를 가르는 값
  - 2사분위수(Q_2): 전체 데이터 중 값이 낮은 2/4와 나머지를 가르는 값=중앙값
  - 3사분위수(Q_3): 전체 데이터 중 값이 낮은 ¾과 나머지를 가르는 값
- 사분위수 범위(InterQuartile Range: IQR):

  3사분위수(Q_3)-1사분위수(Q_1)



사분위수 예제

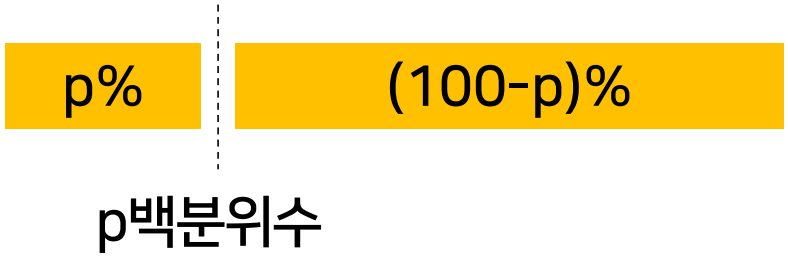
- 예제 2-13: 학생 10명이 1분당 할 수 있는 윗몸일으키기 개수

  8, 23, 25, 28, 32, 35, 37, 41, 42, 52

  - 1사분위수 = 25
  - 2사분위수(중앙값) = 33.5
  - 3사분위수 = 41

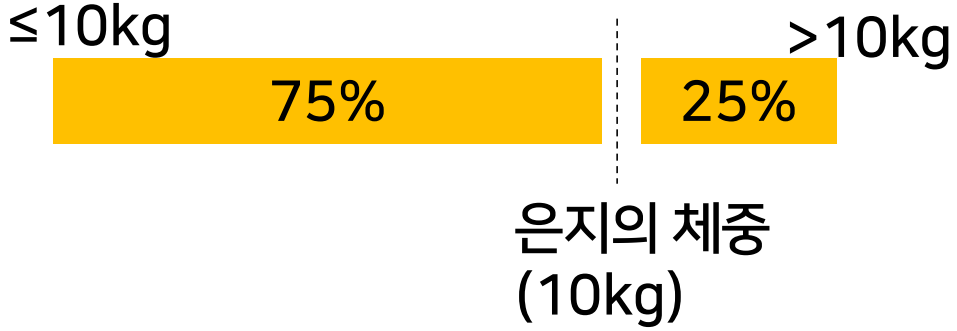
백분위수(percentile)

- p백분위수: 전체 데이터의 p%가 이 값보다 작거나 같은 값
- 1사분위수 = 25백분위수
- 2사분위수 = 50백분위수 = 중앙값
- 3사분위수 = 75백분위수



백분위수 예제

- 예제 2-14: 생후 12개월인 여아 은지의 체중은 10kg 이고 이것은 75백분위수에 해당된다고 한다



범위

- 관찰값의 최댓값 – 최솟값
- 데이터의 산포를 설명하는 가장 간단한 통계량
- 특이점의 영향을 심하게 받는다

다섯 수치요약과 상자그림

- 다섯 수치요약(five-number summary):
  - 최솟값, 1사분위수, 중앙값, 3사분위수, 최댓값
  - 데이터의 중심위치와 퍼진 정도를 모두 파악할 수 있다
- 상자그림(boxplot)
  - 다섯 수치요약을 나타낸 그래프

02

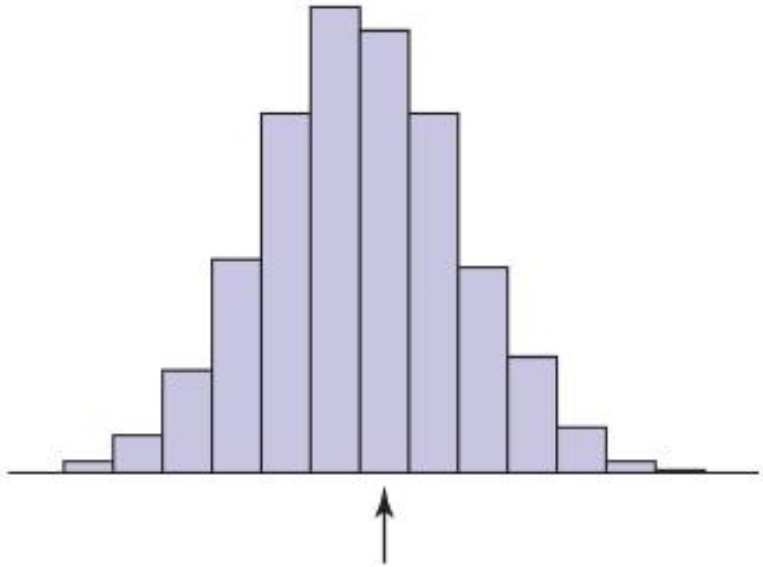
분포와 요약통계량

데이터의 중심위치

- 평균: 분포의 무게중심
  - 대칭적인 분포의 경우 데이터를 잘 대표한다
  - 분포가 기울어져 있거나 특이점이 있는 경우 데이터를 잘 대표하지 못한다
- 중앙값: 데이터를 크기 순으로 정렬했을 때 가장 가운데에 위치하는 값
  - 분포가 기울어져 있거나 특이점이 있는 경우 많이 쓰인다
- 최빈값: 빈도가 가장 높은 관찰값
  - 여러개 있을 수도, 하나도 없을 수도 있다
  - 분포의 중심위치에서 멀리 떨어져있을 수도 있다

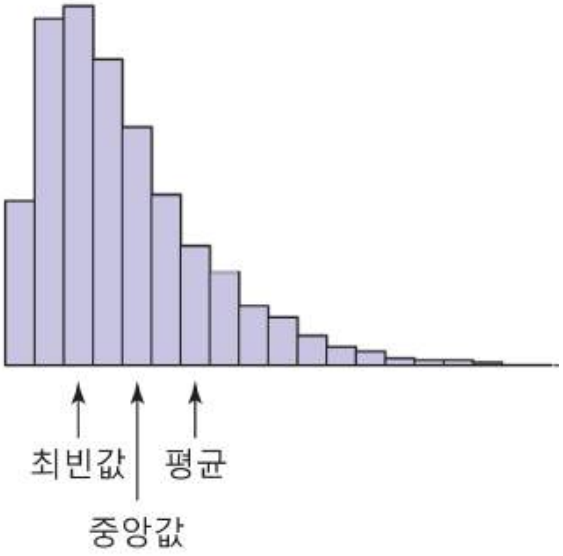
좌우 대칭인 종모양 분포

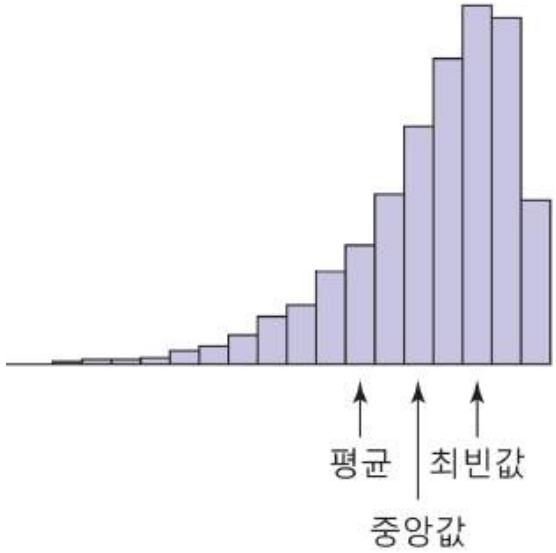
- 평균, 중앙값, 최빈값이 비슷하다



기울어진 분포

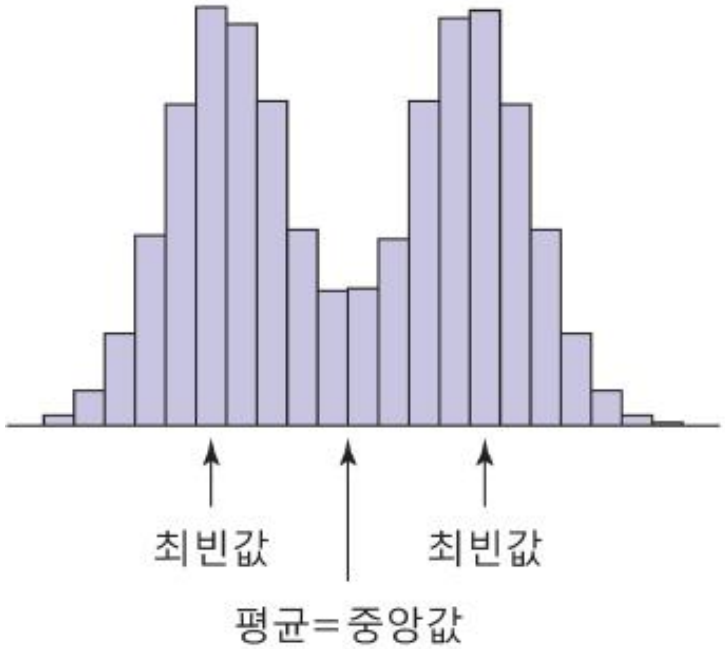
- 평균이 (중앙값에 비해) 긴 꼬리 쪽에 더 가깝게 된다





쌍봉우리형 분포

- 2개의 최빈값이 관찰될 수 있다



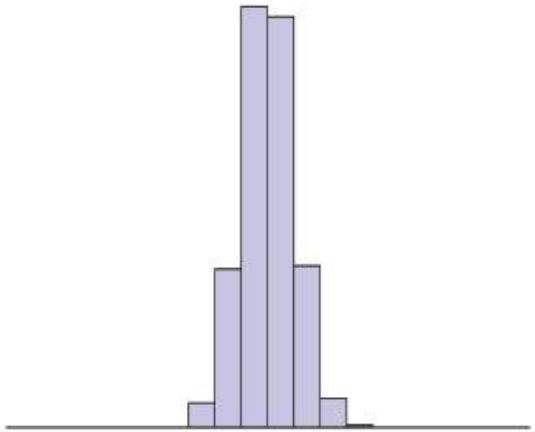
데이터의 산포

- 분산: 편차의 제곱의 평균
- 표준편차: 분산의 제곱근
- 범위: 최댓값 – 최솟값
- 사분위수 범위: 3사분위수-1사분위수

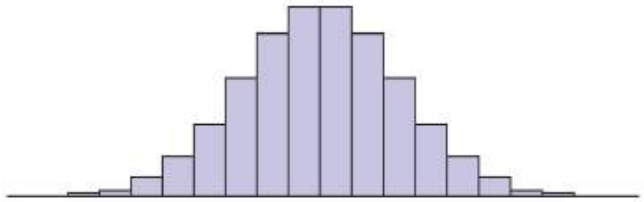
- 이 값들이 클 수록 데이터의 분포가 많이 퍼진 것이다
- 분산, 표준편차, 범위는 특이점의 영향을 크게 받는다

분산과 분포 형태

- 평균이 같고 분산이 다른 두 분포



분산이 작은 분포



분산이 큰 분포

03

R 실습

실습내용

- 그래프
  - 히스토그램
  - 점도표
  - 상자그림
- 요약통계량
  - 평균, 분산, 표준편차
  - 중앙값, 다섯수치요약
  - 사분위수 범위, 범위

히스토그램

- hist() 함수를 이용해서 그린다

```
hist(x, breaks, main, xlab, ylab, xlim, ylim, ...)
```

  - x: 데이터 벡터
  - breaks: 계급에 대한 정보
    - 계급의 개수
    - 계급을 나누는 값들의 벡터
  - main: 그래프의 제목
  - xlab: x축 제목
  - ylab: y축 제목
  - xlim: x축의 범위
  - ylim: y축의 범위

교재 예제 2-6의 히스토그램

```
score<-c(93, 83, 91, 68, 75, 87, 89, 96, 97, 67, 83, 81, 87, 80, 64,
         83, 88, 76, 91, 78, 72, 80, 69, 80, 84, 71, 91, 81, 88, 73)

hist(score)

hist(score, main="")
```

교재 예제 2-7의 히스토그램


```
rv<-c(0.8, 0.8, 0.8,  0.9, 0.9, 0.9, 0.9, 0.9,
      1, 1, 1.8,
      2, 2.1, 2.3, 2.4, 2.8, 2.9,
      3, 3.2, 3.3, 3.5, 3.8, 3.8, 3.9,
      4, 4.2, 4.4, 4.5,
      5.1, 5.3, 5.3, 5.4,
      14, 17, 18, 19,
      21, 21, 23, 25, 27, 28, 32, 34, 36,
      41, 42, 44, 48, 49,
      51, 54, 59, 60, 61, 62, 80,
      240)

hist(rv)
hist(rv, main="", xlab="CRP", breaks=20)
hist(rv, main="", xlab="CRP", breaks=seq(0, 240, 20))
```

교재 예제 2-8의 히스토그램


```
set.seed(2021)
rn<-c(rnorm(100, 5, 2), rnorm(100, 10, 2))

hist(rn)
hist(rn, breaks=20, main="", xlab="value")
hist(rn, breaks=5, main="", xlab="value")
```

점도표

- stripchart() 함수에서 method="stack" 옵션을 사용해서 그린다.

```
stripchart(x, method, pch, offset, cex, axes, ...)
```

  - x: 데이터 벡터
  - method:
    - "stack": 같은 관찰값이 있을 경우 위에 쌓아올린다
    - "overplot": 같은 관찰값이 있을 경우 겹쳐 그린다
    - "jitter": 관찰값에 약간의 노이즈를 더한다
  - pch: 점의 종류
  - cex: 점의 크기
  - offset: "stack" 방법을 쓸 경우, 점 사이의 간격
- width=500, height=200 으로 크기를 지정해서 extract한다

교재 예제 2-9의 점도표


```
age<-c(57, 61, 47, 57, 48, 58, 57, 61, 54, 50, 68, 51)
m.age<-mean(age)

stripchart(age)

stripchart(age, method="stack", pch=19)

stripchart(age, method="stack", pch=19, offset=5, cex=1.5)

stripchart(age, method="stack", pch=19, offset=5, cex=1.5, axes=F)
axis(1, at=seq(45, 70, 5))

stripchart(age, method="stack", pch=19, offset=5, cex=1.5, axes=F, xlim=c(45, 70))
axis(1, at=seq(45, 70, 5))

par(xpd=TRUE )
stripchart(age, method="stack", pch=19, offset=5, cex=1.5, axes=F, xlim=c(45, 70))
axis(1, at=seq(45, 70, 5))
points(m.age, -5, pch=17, cex=2, col="red")

#### extract with width=500, height=200
```

상자그림

- boxplot() 함수를 이용해서 그린다

```
boxplot(x,...)
```

  - x: 데이터 벡터

교재 예제 2-15,16의 상자그림


```r
### 2-15

age<-c(57, 61, 47, 57, 48, 58, 57, 61, 54, 50, 68, 51)
boxplot(age, ylab="Age")
```

```r
### 2-16

member<-c(92, 107, 180, 90, 78, 91, 102, 88, 106, 125, 95, 102, 162)
boxplot(member, ylab="Number of board members")
```

요약통계량

x가 데이터 벡터일 때

- 평균: mean(x)
- 분산: var(x)
- 표준편차: sd(x)
- 중앙값: median(x)
- 다섯수치 요약: fivenum(x)
- 사분위수 범위: IQR(x)
- 범위: range(x)

교재 예제 2-15 데이터의 요약통계량


```
member<-c(92, 107, 180, 90, 78, 91, 102, 88, 106, 125, 95, 102, 162)

mean(member)
var(member)
sd(member)
median(member)
fivenum(member)
IQR(member)
range(member)
```

정리하기

- 중앙값은 데이터를 크기 순서대로 늘어놓았을 때 정확히 중앙에 위치하는 값을 말한다.
- 중앙값은 특이점의 영향을 거의 받지 않는다.
- 사분위수는 크기 순서대로 늘어놓은 데이터를 4등분하는 값이다.
- 사분위수 범위는 3사분위수 - 1사분위수이다.
- 다섯 수치요약이란 양적 데이터의 분포를 최솟값, 1사분위수, 중앙값, 3사분위수, 최댓값으로 정리한 것이다
- 상자그림은 다섯 수치요약을 나타낸 그래프이다.

다음시간 안내

4강

확률

한국방송통신대학교