

2장. 데이터 준비하기

R 데이터 객체의 소개

◎ 통계 데이터의 유형

- 양적 데이터 (숫자형 데이터)
- 질적 데이터 (범주형 데이터)
 1. 명목형 데이터
 2. 순서형 데이터

◎ 통계 data set: 데이터가 행과 열의 2차원 형태로 배열된 상태

- 열: 변수, 하나의 열에는 같은 유형의 데이터만이 올 수 있음
- 행: 동일한 대상에 대한 여러 변수의 관찰값

R 데이터

◎ R 데이터 유형

- 숫자형(numeric), 문자형(character), 논리형(logical) 등등

◎ 다양한 구조의 데이터 객체

- 벡터: 1차원 구조
- 요인: 범주형 데이터를 표현하는 구조. 1차원 구조
- 행렬: 2차원 구조. 구성요소는 모두 동일한 유형의 데이터
- 배열: 2차원 이상의 구조. 동일 유형의 데이터로 구성
- 데이터 프레임: 2차원 구조. 여러 유형의 데이터로 구성
통계 데이터 세트에 가장 적합한 구조
- 리스트: 가장 포괄적인 구조

R 벡터

◎ 1차원으로 배열된 구조

- 유형: 숫자형 (정수형, 실수형), 문자형, 논리형

◎ 벡터의 생성: 함수 c()

```
> X <- c(FALSE, TRUE, FALSE)
> y1 <- c(2L, 4L, 8L)
> y2 <- c(2.2, 3.5, 11.4)
> z <- c("one", "two", "three")
```

◎ 벡터의 구성요소: 모두 같은 유형의 데이터

```
> typeof(x)
[1] "logical"
> typeof(y1)
[1] "integer"
> typeof(y2)
[1] "double"
> typeof(z)
[1] "character"
```

R 벡터의 길이

◎ 벡터를 구성하고 있는 요소 개수

- 확인하기: `length()` 함수를 사용

```
> y1
[1] 2 4 6
>length(y1)
[1] 3
```

◎ 스칼라(길이가 1인 벡터)의 생성

- Combine을 사용하지 않아도 된다 `c()`

```
> a<-1
a
[1] 1
```

R 데이터를 섞어서 입력했다면?

◎ 실습 1

```
> z1 <-c(TRUE,2L,1.1,4.5)
```

결과 값은???

◎ 실습2

- 숫자형과 논리형 데이터가 Combine된다면?

```
> z2 <- c(3, TRUE, FALSE)
```

결과 값은???

R 벡터의 구성요소에 이름 붙이기

◎ 처음 입력시 이름을 붙이는 방법:

```
> c(Seoul=9930, Busan=3497, Incheon=2944, Suwon=1194)
```

```
Seoul  Busan  Incheon  Suwon
9930   3497   2944    1194
```

◎ 데이터를 먼저 할당한 후 이름 붙이기:

```
> pop <- c(9930,3497,2944,1194)
```

```
> names(pop) <- c("Seoul", "Busan", "Incheon", "Suwon")
```

```
> pop
```

```
Seoul  Busan  Incheon  Suwon
9930   3497   2944    1194
```

```
> names(pop)
```

```
[1] "Seoul" "Busan" "Incheon" "Suwon"
```

R 벡터의 인덱싱(Indexing)

◎ 벡터의 일부분을 선택하여 가져오는 방법:

- 벡터의 일부분을 선택하는 작업.
- $x[a]$ 의 형태: 벡터 a 는 정수형, 논리형, 문자형(구성요소에 이름이 있는 경우)
- 정수형 벡터에 의한 인덱싱
 - 모두 양수: 지정된 위치의 자료 선택
 - 모두 음수: 지정된 위치의 자료 제외

```
> y <- c(2,4,6,8,10)
```

```
> y[c(1,3,5)]
```

```
[1] 2 6 10
```

```
> y[c(-2,-4)]
```

```
[1] 2 6 10
```

```
> y[c(2,2,2)]    # 같은 위치 반복 지정 가능
```

```
[1] 4 4 4
```

```
> y[6]
```

```
[1] NA
```

지정한 위치가 벡터 길이보다 큰 경우

R 벡터의 인덱싱(Indexing)

- 논리형 벡터에 의한 인덱싱
 - TRUE가 있는 위치의 자료막 선택
 - 벡터의 비교에 의한 자료 선택에서 유용하게 사용됨

```
> y
[1] 2 4 6 8 10

> y[c(FALSE, TRUE, FALSE, TRUE, FALSE)]
[1] 4 8

> y>5
[1] FALSE FALSE TRUE TRUE TRUE

> y[y>5] # 같은 위치 반복 지정 가능
[1] 6 8 10
```

R 벡터의 인덱싱(Indexing)

실습) 벡터 y의 자료 중 평균보다 큰 값을 인덱싱 하라!

```
> y [y > mean(y)]
```

```
[1] 8 10
```

- 문자형 벡터에 의한 인덱싱
 - 벡터의 구성요소에 이름이 있는 경우에만 적용 가능

```
> pop
```

Seoul	Busan	Inchon	Suwon
9930	3497	2944	1194

```
> pop[c("Seoul", "Suwon")]
```

Seoul	Suwon
9930	1194

R 막간의 퀴즈타임!

1. 다음의 데이터를 벡터 x에 입력하라.

17 16 20 24 22 15 21 18

1) 벡터 x에 입력된 데이터의 개수를 확인하라.

2) 벡터 x의 마지막 데이터의 값을 출력하라.

(단, x[8]과 같이 자료의 위치를 직접 숫자로 지정하지 않는다.)

2. 다음의 데이터를 벡터 y에 입력하라.

10.4 5.6 3.1 6.4 9.6 7.8 12.1

1) 벡터 y에 입력된 데이터의 개수를 확인하라.

2) 벡터 y의 마지막에서 3번째 데이터의 값을 출력하라.

(단, x[5]과 같이 자료의 위치를 직접 숫자로 지정하지 않는다.)



주요 이력

現) (주)RTMC 전략기획실장
前) (주)B사 웹로그분석 및 DP사업 完
前) (주)H금속사 회계팀 선물환 및 자금관리
前) (주)B건설사 회계팀 주식 공시
前) K문고 CRM VIP 군집전략 CRM프로젝트 보조연구원
前) L백화점 CRM Alert 전략 CRM프로젝트 보조연구원

BSL(스위스 로잔 비즈니스 스쿨) MBA
ASSIST 빅데이터경영통계 MBA

국가공인 ADSP(빅데이터 준전문가)

現) 코리아IT아카데미 빅데이터 R 강사
現) 코리아IT아카데미 빅데이터 기초 파이썬 강사
現) 코리아IT아카데미 빅데이터 기초 ML 강사
現) 코리아IT아카데미 빅데이터 기초통계 전담강사
現) 코리아IT아카데미 빅데이터 취업 강사

“자료는 대가이신 박동련 교수님께 도움을 받았음을 밝힙니다.”

[박영식] [완성에 이르기까지](#)