6. 데이터 프레임

- 자료를 Spreadsheet와 유사한 표의 형식으로 표현함

- 열(column)들은 여러 변수들, 행(row)은 같은 사람이나 사물에 대한 다른 측정값 또는 변수 값을 표현함

- R 내부에서 만들기도 하고, 다른 프로그램에서 만든 자료를 불러오기도 함

- 하나의 list로서 원소(변수)들은 동일한 길이의 열벡터들이 됨

7. data.frame 함수

> 학번=c(1601,1602,1603,1604,1605,1606,1607) // 수치형 벡터

> 성별=c("m","f","m","m","f","m","f") // 문자형 벡터

> 점수=c(70,80,NA,89,75,50,55)

> exam=data.frame(학번,성별,점수)

> exam1=data.frame(code=학번,sex=성별,score=점수)

|  |  |
| --- | --- |
| > exam  학번 성별 점수  1 1601 m 70  2 1602 f 80  3 1603 m NA  4 1604 m 89  5 1605 f 75  6 1606 m 50  7 1607 f 55 | > exam1  code sex score  1 1601 m 70  2 1602 f 80  3 1603 m NA  4 1604 m 89  5 1605 f 75  6 1606 m 50  7 1607 f 55 |

|  |  |
| --- | --- |
| > head(exam1) // 앞 6행까지만 출력  code sex score  1 1601 m 70  2 1602 f 80  3 1603 m NA  4 1604 m 89  5 1605 f 75  6 1606 m 50 | > tail(exam1) // 뒤 6행까지만 출력  code sex score  2 1602 f 80  3 1603 m NA  4 1604 m 89  5 1605 f 75  6 1606 m 50  7 1607 f 55 |

8. data.frame안에 있는 자료접근

> exam1$code // code열의 요소들을 나타냄

[1] 1601 1602 1603 1604 1605 1606 1607

> exam1$sex

[1] "m" "f" "m" "m" "f" "m" "f"

> mean(exam1$score) // NA가 있어서 평균이 구해지지않음

[1] NA

> mean(exam1$score, na.rm=T) // mean: 평균을 구함, na.rm=T: NA를 삭제함

[1] 69.83333

> exam1[ ,2] // 앞은 행 뒤는 열 Ex) 공백, 2는 모든행에 대한 2열을 나타냄

[1] "m" "f" "m" "m" "f" "m" "f"

> exam1[5,2]

[1] "f"

> exam1[2:4,1:2] // 2~4행, 1~2열

code sex

2 1602 f

3 1603 m

4 1604 m

> exam2=exam1[-3, ] // exam1에서 2행을 제거 후 exam2에 저장함

> exam2

code sex score

1 1601 m 70

2 1602 f 80

4 1604 m 89

5 1605 f 75

6 1606 m 50

7 1607 f 55

Ex 7) exam1의 세 번째, 네 번째 행을 삭제하라

> exam2=exam1[c(-3,-4), ]

> exam2

code sex score

1 1601 m 70

2 1602 f 80

5 1605 f 75

6 1606 m 50

7 1607 f 55

Ex 8) exam1의 두 번째 열을 삭제하라

> exam2=exam1[ ,-2]

> exam2

code score

1 1601 70

2 1602 80

3 1603 NA

4 1604 89

5 1605 75

6 1606 50

7 1607 55

Ex 9) exam1의 세 번째 행과, 두 번째 열을 삭제하여 새로운 데이터 프레임 hw에 저장하라 / Ex 10) hw 내용을 출력하라

> hw=exam1[-3,-2]

> hw

code score

1 1601 70

2 1602 80

4 1604 89

5 1605 75

6 1606 50

7 1607 55

9. 데이터 편집기

> test=edit(data.frame()) // var1, var2…를 클릭후 자신이 원하는 변수명으로 편집

|  |  |  |
| --- | --- | --- |
| Code | Score | 성별 |
| 12 | 90 | 여 |
| 34 | 90 | 남 |

편집후 나가고

> test

Code score 성별

1 12 90 여

2 34 90 남

> newtest=edit(test)

|  |  |  |
| --- | --- | --- |
| Code | Score | 성별 |
| 12 | 90 | 여 |
| 34 | 90 | 남 |
| 1 | 60 | 여 |

> newtest

code score 성별

1 12 90 여

2 34 90 남

3 1 60 여

|  |  |  |
| --- | --- | --- |
| Code | Score | 성별 |
| 12 | 90 | 여 |
| 34 | 90 | 남 |
|  |  |  |

code score 성별

1 12 90 여

1. 34 90 남
2. NA NA <NA>=

> newtest[-3, ]

code score 성별

1 12 90 여

2 34 90 남

10. 참고

- 데이터 편집기는 작은 규모의 자료 입력시 사용하지만 대량의 자료를 입력하기에는 불편하다

- 우리는 엑셀에서 자료를 입력한 후에 R로 불러 사용할 것이다

Ex) > read.csv(“test.csv”)

11. 엑셀 자료 파일을 R로 불러오는법

- 엑셀 자료 파일 → 다른 이름으로 저장 → 확장자 csv로 (utf-8 X) → R Console에서 작업 디렉토리 변경 (csv파일을 저장한 경로로) → > 하고싶은 함수명=read.csv("csv파일의 이름.csv") → 하고싶은 함수명