R을 이용한 통계 기초와 데이터 분석

Lecture 1

남현진

한성대학교

2020

무제

Section 1

통계 이론

Section 2

R 실습

기본적인 R언어 문법

```
print("Hello, world!")
```

```
## [1] "Hello, world!"
```

기본적인 R언어 문법

```
2 + 2
```

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

- “[ ]” 는 그 줄의 첫 element의 위치 값을 보여준다.

```
1:10
```

```
##  [1]  1  2  3  4  5  6  7  8  9 10
```

- “:”는 앞부터 뒤까지를 포함하는 연속적인 숫자들을 출력한다.

기본적인 R언어 문법

```
x <- 2 + 2
y <- 3 * 6
```

- "<-" 는 해당 값을 부여하는 연산자이다.

```
y/x  #y divided by x
## [1] 4.5
```

- 저장된 값을 이용하여 원하는 계산을 실행할 수 있다.
- 실행하지 않는 코드는 "#"를 사용하여 적을 수 있다. 해당 코드를 설명하는 문장을 적을때 유용하게 쓸 수 있다.

데이터 구조의 이해

Data Structure

- 1차원 데이터: Vetor (동질성), List (이질성)
- 2차원 데이터: Matrix (동질성), Data frame (이질성)
- 3차원 데이터: Array

Vector

```
vec1 <- c(1, 2, 3)
vec1
## [1] 1 2 3
```

- R에서 벡터는 동질적인 값을 가지고 있는 숫자의 집합이다.
- 벡터를 만들기 위해서는 c()를 이용하여 해당 값을 부여한다.

```
vec2<- c("How", "are","you","?")
vec2
## [1] "How" "are" "you" "?"
```

- 따음표안에 숫자를 넣으면 R은 입력값을 문자열로 인식한다.

List

```
list1 <- list(vec1, vec2)
list1
## [[1]]
## [1] 1 2 3
##
## [[2]]
## [1] "How" "are" "you" "?"
```

- 리스트는 이질적인 변수들이 모여있는 데이터 형태이다.

Matrix

```
matrix1 <- matrix(1:12, nrow=2, ncol = 6)
matrix1
##      [,1] [,2] [,3] [,4] [,5] [,6]
## [1,]    1    3    5    7    9   11
## [2,]    2    4    6    8   10   12

matrix2 <- matrix(1:12, nrow=2, ncol = 6, byrow = TRUE)
matrix2
##      [,1] [,2] [,3] [,4] [,5] [,6]
## [1,]    1    2    3    4    5    6
## [2,]    7    8    9   10   11   12
```

Matrix

```
vec <- c(1,2,3)
cbind(vec,vec)
##      vec vec
## [1,]   1   1
## [2,]   2   2
## [3,]   3   3
rbind(vec,vec)
##     [,1] [,2] [,3]
## vec    1    2    3
## vec    1    2    3
```

- cbind를 사용하면 column(열)을 기준으로 두 벡터/행렬/데이터프래임이 결합한다.
- rbind를 사용하면 row를(행)기준으로 두 벡터/행렬/데이터프래임이 결합한다.

Data Frame

```
df1 <- data.frame(item = c('pencil', 'pen', 'eraser'),
                  stock = c(T,T,F),
                  price = c(1000,1300,500))
df1
##      item stock price
## 1 pencil  TRUE  1000
## 2    pen  TRUE  1300
## 3 eraser FALSE   500
```

- 데이터 프래임은 가장 많이 쓰이는 데이터 형식이다.

Data Frame

```
df2 <- as.data.frame(matrix2)

df2

##   V1 V2 V3 V4 V5 V6
## 1  1  2  3  4  5  6
## 2  7  8  9 10 11 12
```

- as.data.frame 함수를 사용하여 행렬을 데이터프레임 형식으로 바꿀 수 있다.

데이터 구조의 이해

Data Type

- Integer: 실수
- Numeric: 정수
- Character(string): 문자열
- Factor: 요인형
- Logical(boolean): 논리값

데이터 타입

```
head(sample_data)
```

```
## # A tibble: 6 x 5
##   name.first registered.age gender location.country marital.status
##   <chr>                <dbl> <fct>  <fct>            <lgl>
## 1 Janique                 21 male   Brazil           FALSE
## 2 Khaled                  24 male   Norway           FALSE
## 3 Maja                    30 female Denmark          FALSE
## 4 Latife                  28 female Turkey           FALSE
## 5 Sayenne                 32 female Netherlands      TRUE
## 6 Linda                   38 female Ireland          TRUE
```

```
str(sample_data)
```

```
## tibble [100 x 5] (S3: tbl_df/tbl/data.frame)
```

평균과 분산 구하기

1부터 1000까지의 100개의 난수를 생성한 후 평균과 분산을 구하기.

```
set.seed(1)
x <- sample(1:1000, 100, replace=T)
mean(x)
## [1] 534.38
sd(x)
## [1] 289.6076
var(x)
## [1] 83872.54
```