

# 고고학 자료 통계분석

---

Week 2 : 기초 통계(1)

숭실대학교 사학과 석사과정 1학기  
주 찬 혁

# 계획

주차	제목	내용
1	Intro	소개, R 설치
2	기초 통계(1)	모집단과 표본, 기술통계량
3	기초 통계(2)	변수의 종류, 가설과 오류, 분석 절차
4	전처리	데이터 전처리
5	시각화	다양한 종류의 그래프
6	검정	t-검정, Median Polish
7	회귀분석	선형회귀, 다중선형회귀, 로지스틱회귀
8	군집분석	K-means,
9	판별분석	DA, MDA
10	주성분분석	PCA

# 복습

- 고고학 자료에 대한 통계 분석은 왜 필요한가?
- 본인의 컴퓨터에 R을 설치
- 본인의 컴퓨터에 R Studio를 설치
- 필요한 R Package를 설치

# 모집단과 표본

- 모집단(Population) : 연구 대상(고고학에서는 주로 과거의 실제 양상)
- 모수(Parameter) : 모집단의 통계값
  - 모평균( $\mu$ ), 모표준편차( $\sigma$ ), 모비율( $p$ )
- 표본(Sample)
- 통계량(Statistic) : 표본의 통계값
  - 표본 평균( $\bar{x}$ ), 표본 표준편차( $S$ )

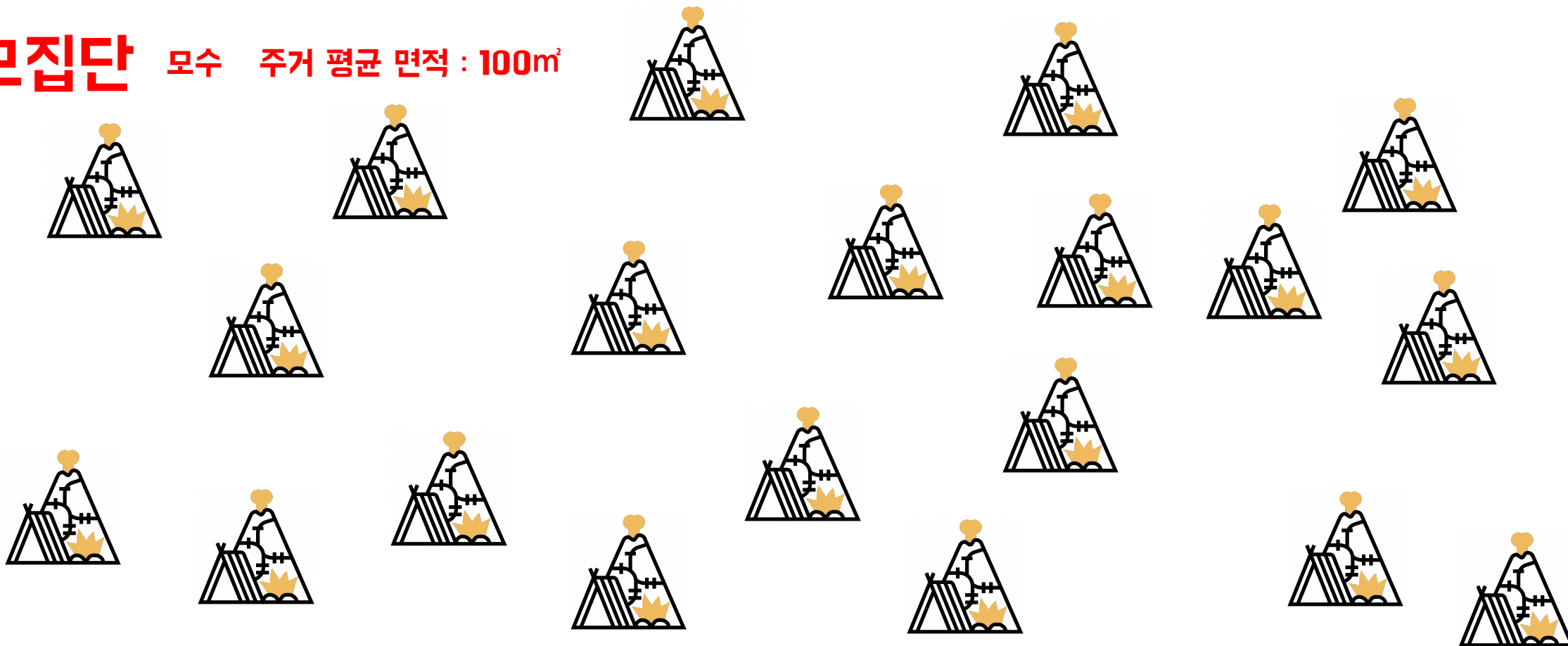
# 모집단과 표본

A유적의 주거 면적을 연구하는 중...

A유적의 실제 과거 모습 = 알고 싶은 대상

주거 : 20기

**모집단** 모수 주거 평균 면적 : 100m<sup>2</sup>



# 모집단과 표본

A유적의 주거 면적을 연구하는 중...

A유적의 현재 모습 = 모집단을 알아내기 위해 주어진 것

주거지 : 4기

표본

통계량    주거지 평균 면적 :  $89\text{m}^2$



# 표본조사의 이유

- 모집단을 알 수 없기 때문에(특히 고고학은...)
- 너무 많은 비용 발생
- 너무 많은 시간 소요
- 기타 등등...

# 표집(Sampling)

- 모집단으로부터 표본을 추출하는 과정
- ~~비확률표집(Non-probability Sampling)~~
  - ~~눈덩이표집, 할당표집 등 ...~~
- 확률표집(Probability Sampling)
  - ~~유층표집, 체계적표집 등 ...~~
  - 임의표집(Random Sampling)
    - 복원추출 : 한 번 선택된 요소의 재추출 가능
    - 비복원추출 : 한 번 선택된 요소는 추출에서 제외



# 비복원추출(Sampling without replacement)

모집단



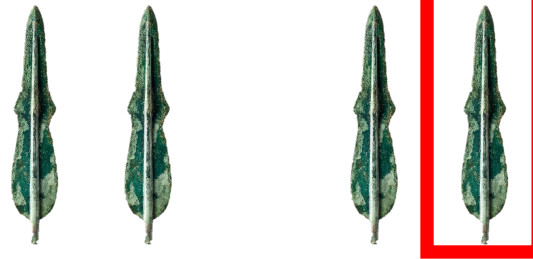
**X 1000**

추출 1회차



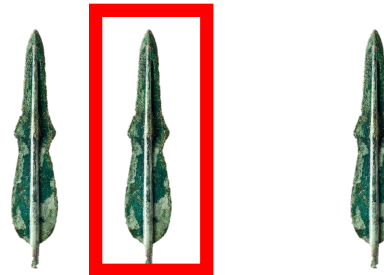
확률  
... **1/1000**

추출 2회차



... **1/999**

추출 3회차



... **1/998**

# 복원추출(Sampling with replacement)

모집단



**X 1000**

추출 1회차



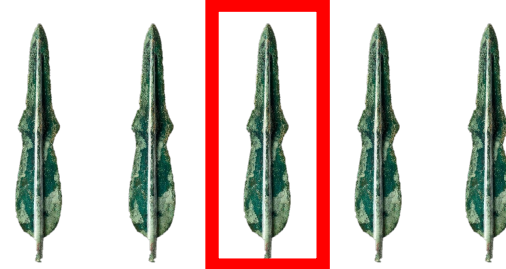
확률  
... 1/1000

추출 2회차



... 1/1000

추출 3회차



... 1/1000

# 표집(Sampling)

- 대부분의 고고학자료는 임의표집된 것이 아니므로 편향의 가능성이 있음
- 고고학적 맥락을 살펴 편향을 줄이거나 없애야함

# 기술통계량

- 대상에 대한 수치적 요약 -> **표본(데이터)의 전반적인 형태를 알 수 있음**
- 평균(Mean)
- 중앙값(Median)
- ~~최빈값(Mode)~~
- 표준편차(Standard Deviation)
- 사분위수(Quartile)
- 표준오차(Standard Error, SE)

# 평균(Mean)

- 모든 값을 더해서 값의 개수로 나눈 것
- 데이터의 분포가 대칭을 이루고 있는 경우 유용

$$\frac{x_1 + x_2 + x_3 \dots x_n}{n}$$

# 중앙값(Median)

- 값을 정렬했을 때 가장 중앙의 값
- 데이터의 분포가 한 쪽으로 치우친 경우 유용

$x_1$   
1

$x_2$   
2

$x_3$   
3

$x_4$   
4

$x_5$   
5

# 표준편차(Standard Deviation)

- 분산(Variance)
  - 편차 제곱의 평균(편차 : 평균과의 차이)
  - 자료가 퍼진 정도
  - '제곱' 하기 때문에 값이 크게 증가한다는 단점
- 표준편차(Standard Deviation)
  - 분산의 제곱근
  - 분산의 단점 해결

$$V = \frac{(x_1 - m)^2 + (x_2 - m)^2 + \dots + (x_n - m)^2}{n}$$

$$s = \sqrt{V}$$

# 표준오차(Standard Error, SE)

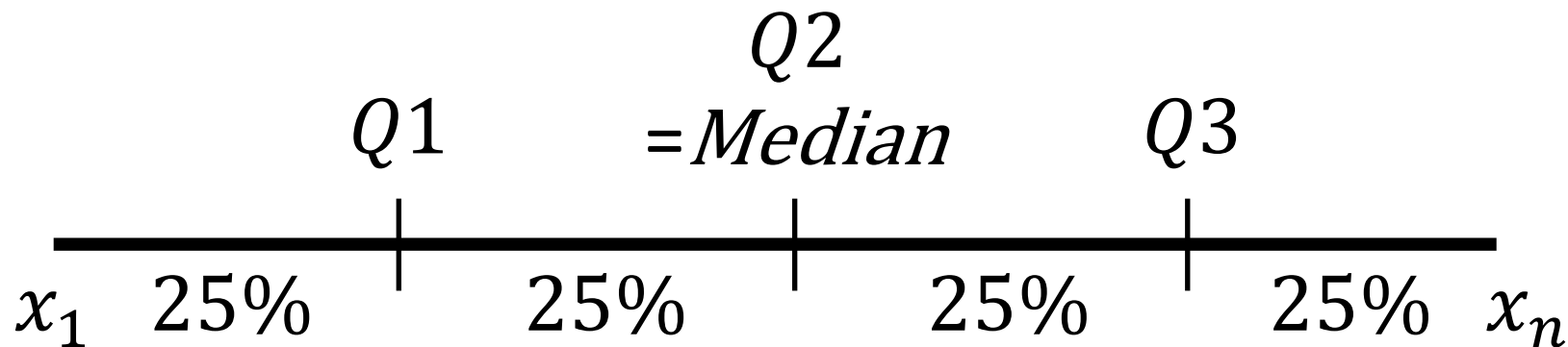
- 표본 평균과 모평균 사이의 차이
- 표본평균의 표준편차

$$SE = \frac{\sigma}{\sqrt{n}}$$



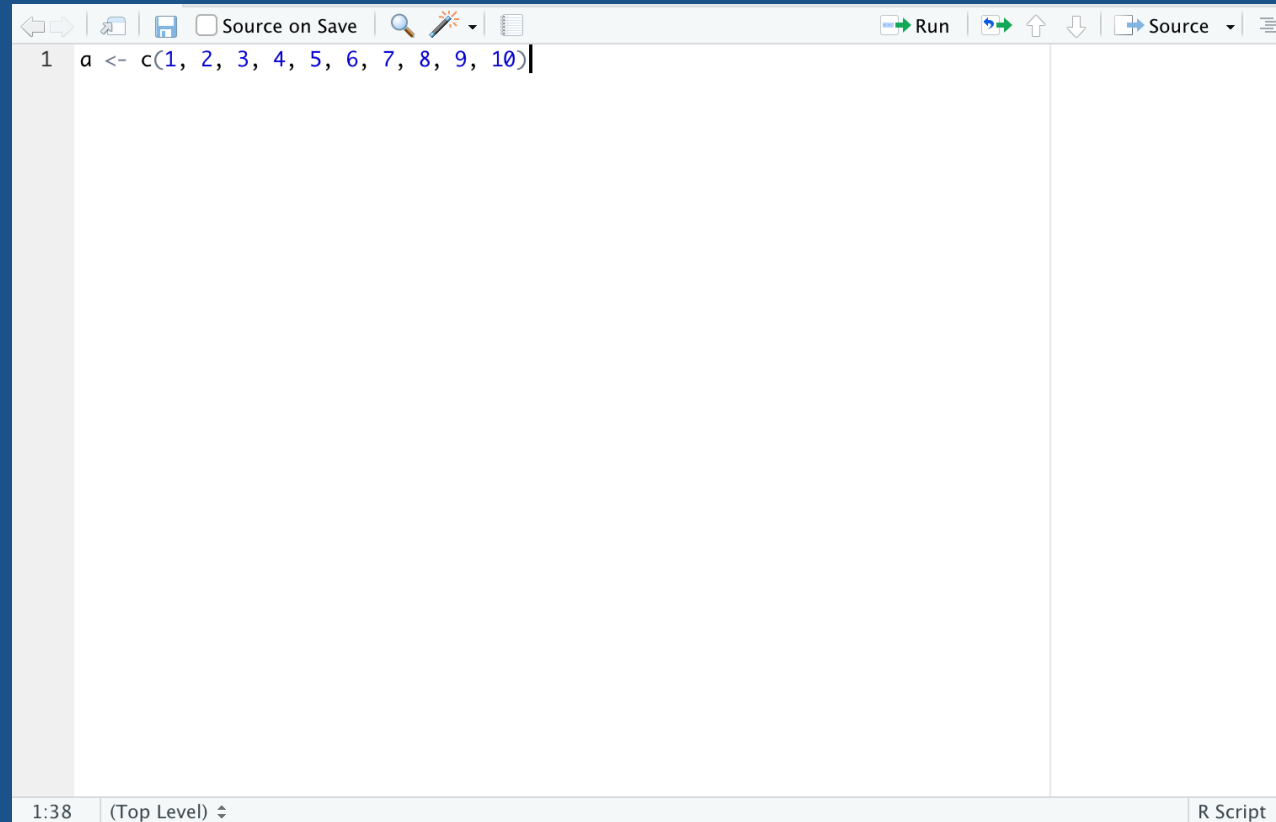
# 사분위수(Quartile)

- 값을 4개의 동일한 부분으로 나눈 값
- 중앙값이 사용될 때 활용하기 좋음
- 데이터의 분포가 한 쪽으로 치우친 경우 유용



# R 실습 - 데이터 준비

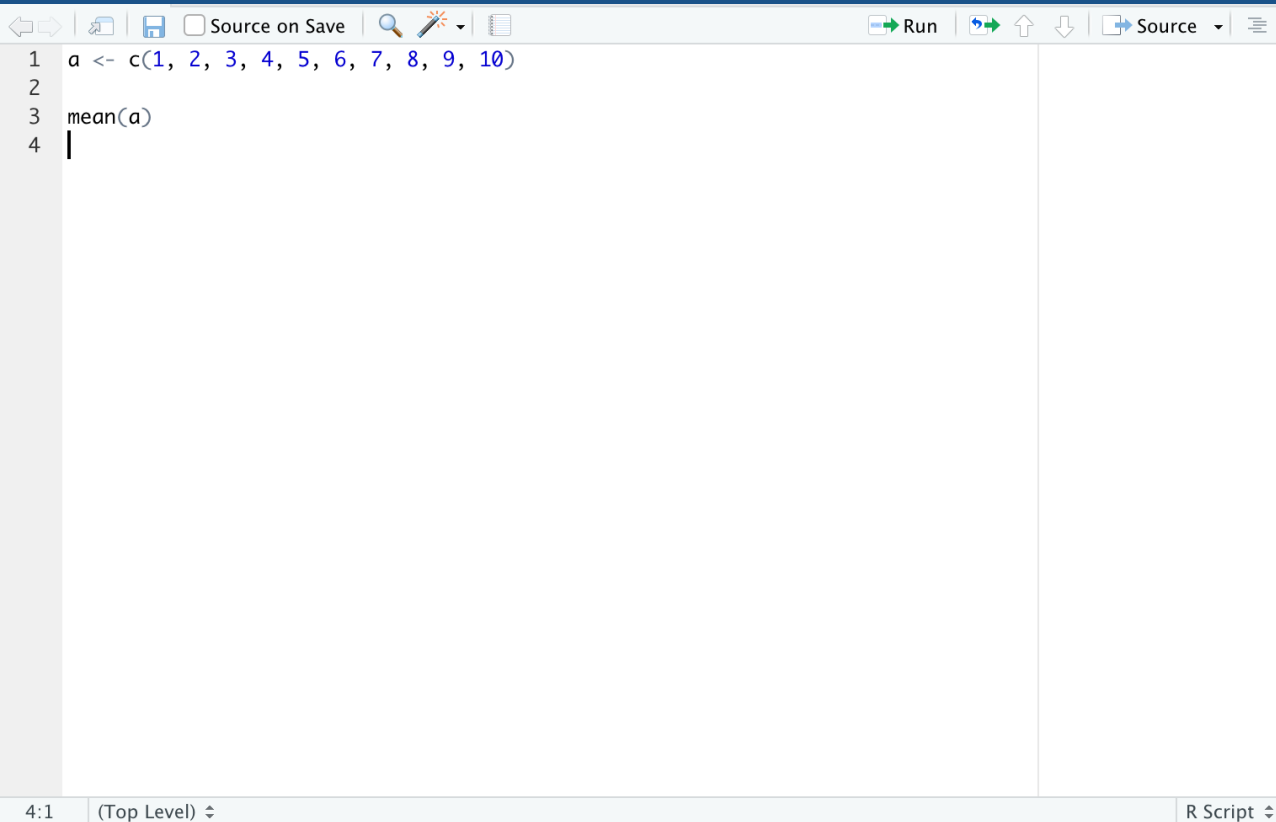
> 변수명 <- c(값1, 값2, 값3 ...)



```
1 a <- c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10)
```

# R 실습 - 평균

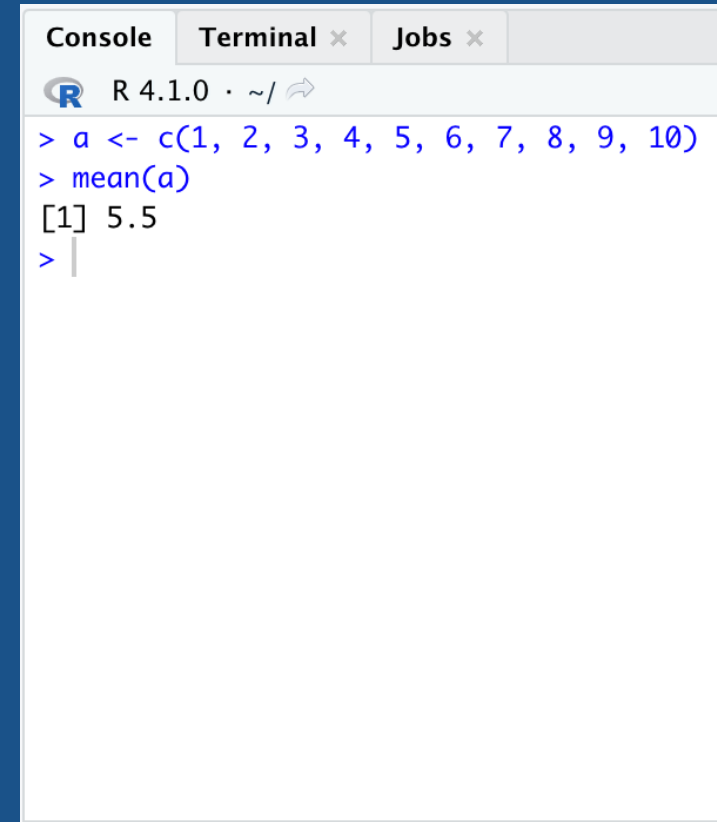
> mean(데이터)



The image shows the RStudio editor window. The source editor on the left contains the following R code:

```
1 a <- c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10)
2
3 mean(a)
4 |
```

The top toolbar includes icons for navigation, saving, and running code. The status bar at the bottom indicates the current position is 4:1 and the context is (Top Level).



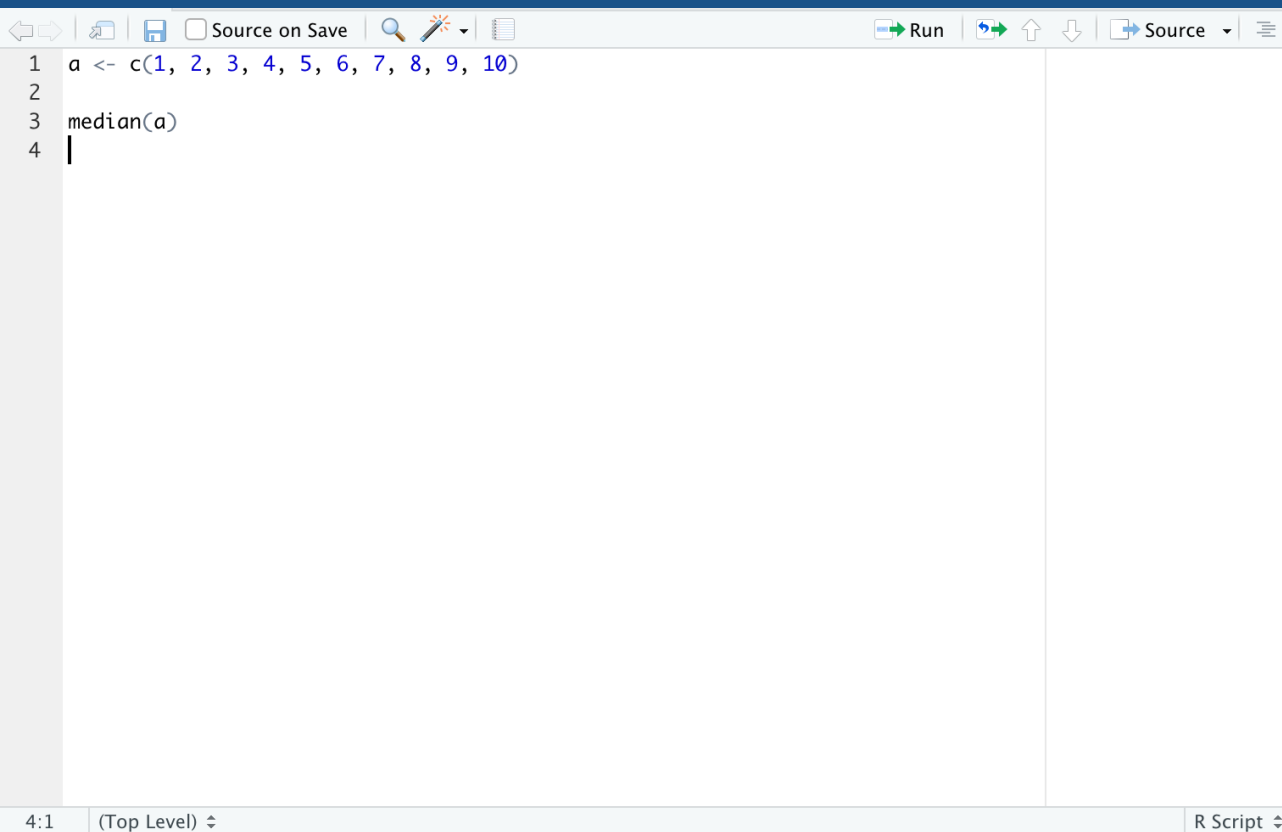
The image shows the R console window. The console output is as follows:

```
R 4.1.0 · ~/
> a <- c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10)
> mean(a)
[1] 5.5
> |
```

The console window has tabs for Console, Terminal, and Jobs. The status bar at the bottom indicates the current position is 4:1 and the context is (Top Level).

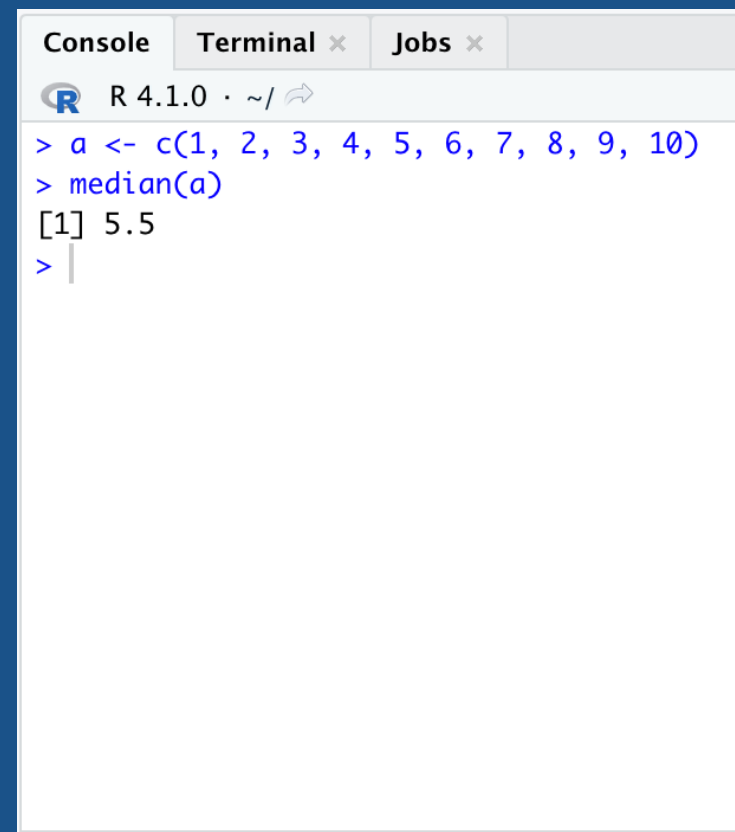
# R 실습 - 중앙값

> median(데이터)



```
1 a <- c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10)
2
3 median(a)
4 |
```

The image shows the RStudio editor interface. The source editor on the left contains four lines of R code: line 1 defines a vector 'a' with values 1 through 10; line 2 is empty; line 3 calls the 'median()' function on 'a'; line 4 has a cursor. The top toolbar includes icons for running and sourcing code. The bottom status bar shows '4:1' and '(Top Level)'.



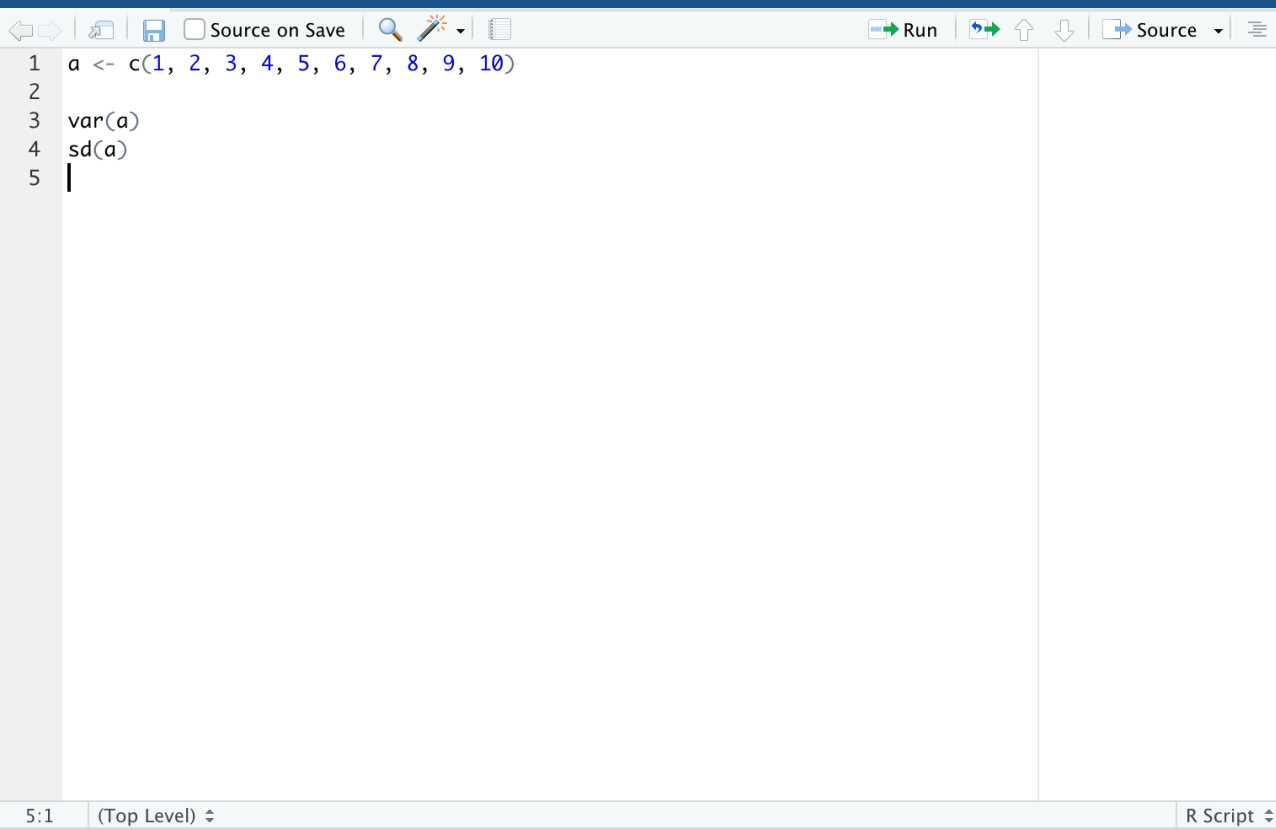
```
R 4.1.0 · ~/
> a <- c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10)
> median(a)
[1] 5.5
> |
```

The image shows the R console output. It displays the R version (4.1.0) and the current directory (~). The commands entered are 'a <- c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10)' and 'median(a)'. The output for 'median(a)' is '[1] 5.5'. The console also shows the prompt '>' and a cursor.

# R 실습 - 분산 및 표준편차

> var(데이터)

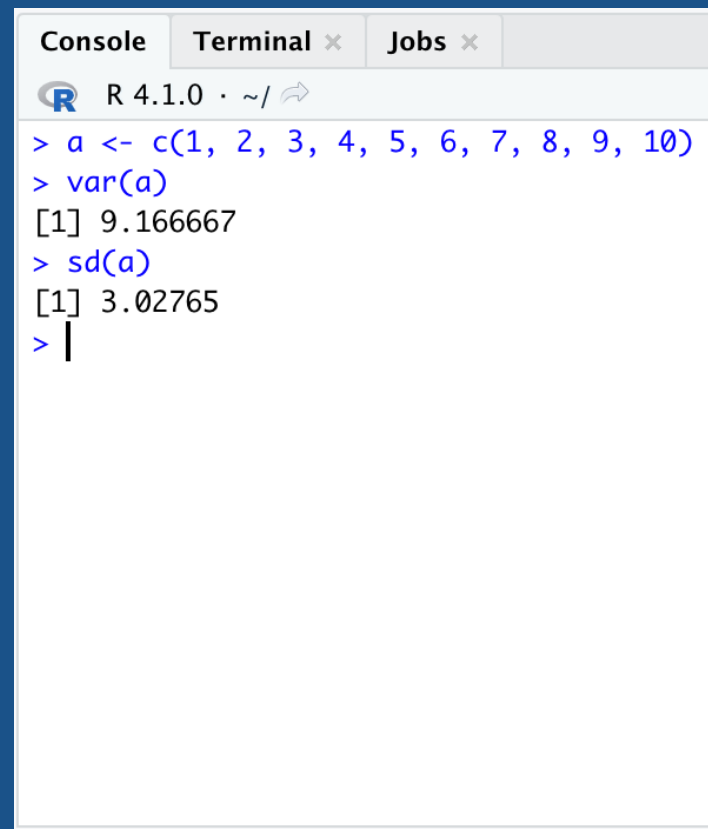
> sd(데이터)



The image shows the RStudio editor window. The source editor on the left contains the following R code:

```
1 a <- c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10)
2
3 var(a)
4 sd(a)
5 |
```

The top toolbar includes icons for navigation, saving, and running code. The status bar at the bottom indicates the cursor is at line 5, column 1, and the file is an R Script.



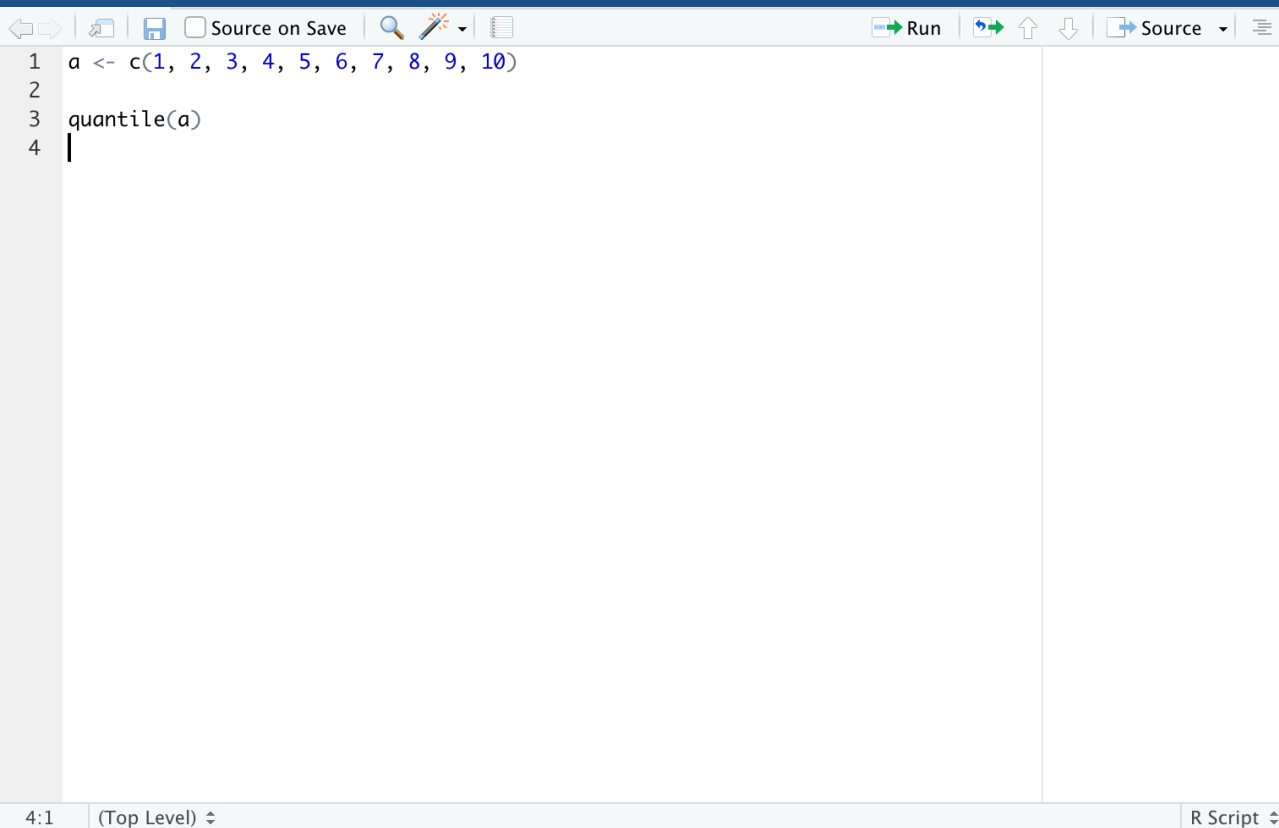
The image shows the RStudio console window. The console output is as follows:

```
R 4.1.0 · ~/
> a <- c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10)
> var(a)
[1] 9.166667
> sd(a)
[1] 3.02765
> |
```

The console window has tabs for Console, Terminal, and Jobs. The status bar at the bottom indicates the cursor is at line 5, column 1, and the file is an R Script.

# R 실습 - 사분위수

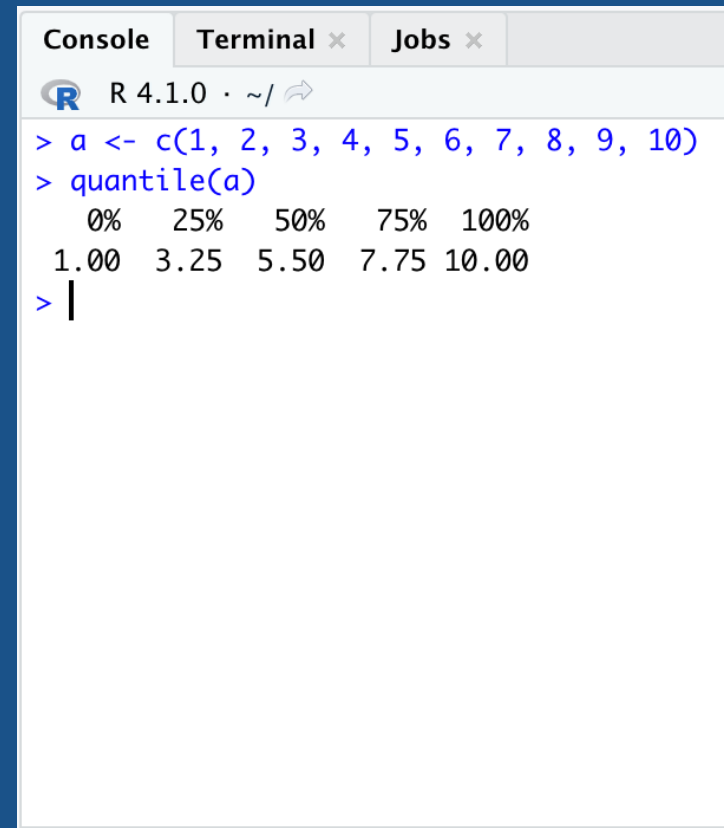
> quantile(데이터)



The image shows the RStudio editor window. The source editor on the left contains the following R code:

```
1 a <- c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10)
2
3 quantile(a)
4 |
```

The top toolbar includes icons for navigation, saving, and running code. The bottom status bar shows the cursor position at line 4, column 1, and the file type as 'R Script'.



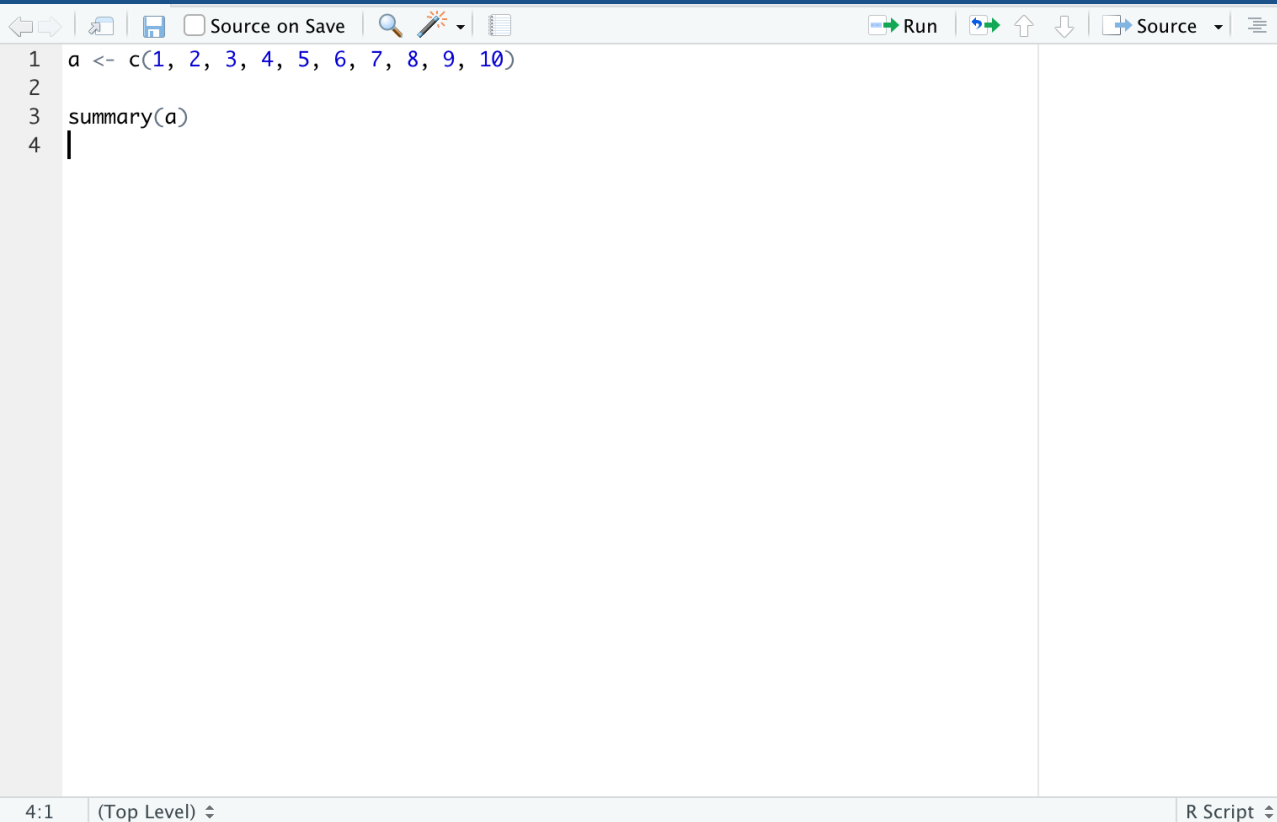
The image shows the RStudio console window. The console output is as follows:

```
R 4.1.0 · ~/
> a <- c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10)
> quantile(a)
      0%    25%    50%    75%   100%
 1.00  3.25  5.50  7.75 10.00
> |
```

The console window has tabs for 'Console', 'Terminal', and 'Jobs'. The output shows the quantiles of the vector 'a' at 0%, 25%, 50%, 75%, and 100% percentiles.

# R 실습 - Tip

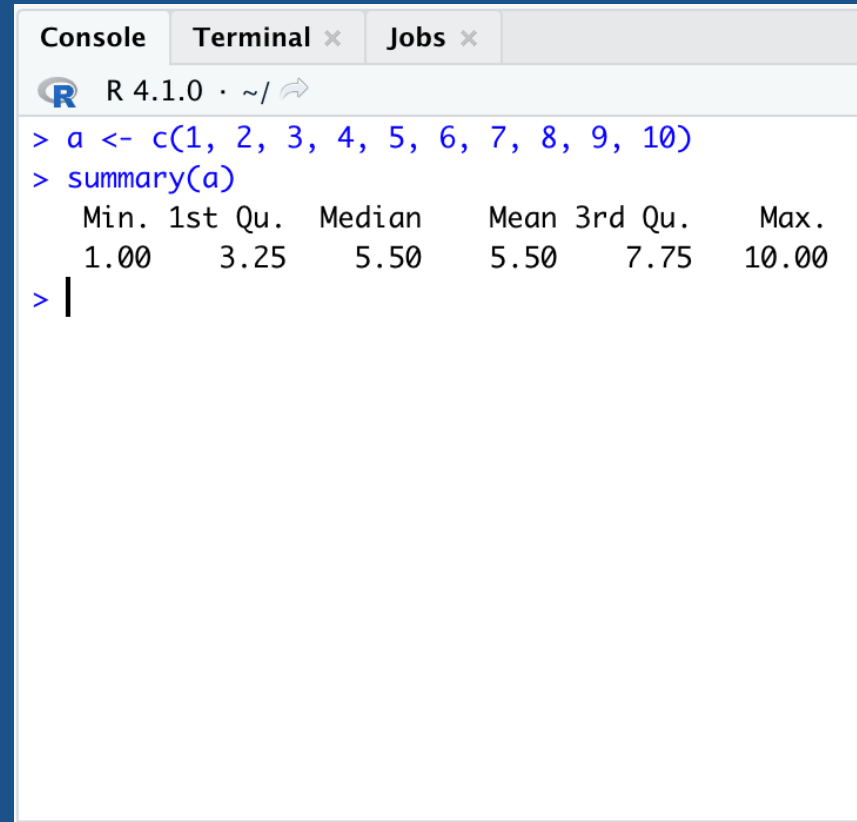
> summary(**데이터**)



The image shows the RStudio editor window. The source editor on the left contains the following R code:

```
1 a <- c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10)
2
3 summary(a)
4 |
```

The top toolbar includes icons for navigation, saving, and running code. The status bar at the bottom indicates the current position is 4:1 and the file is at the Top Level.



The image shows the RStudio console window. The console output displays the result of the `summary(a)` command, showing the distribution of the data vector `a`.

```
R 4.1.0 · ~/
> a <- c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10)
> summary(a)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 1.00   3.25   5.50   5.50   7.75  10.00
> |
```

The console window has tabs for Console, Terminal, and Jobs. The status bar at the bottom indicates the current position is 4:1 and the file is at the Top Level.

# R 실습 - 표준오차

plotrix 패키지로 간단히 계산 가능

> plotrix::std.error(데이터)

```
1 library(plotrix)
2
3 a <- c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10)
4
5 plotrix::std.error(a)
6
```

6:1 (Top Level) R Script

```
Console Terminal x Jobs x
R 4.1.0 · ~/
> library(plotrix)
> a <- c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10)
> plotrix::std.error(a)
[1] 0.9574271
>
```