

통계학 특강

신지은 (서울시립대학교 통계학과)

Email: jieunstat@gmail.com



목차

1

통계와 R 프로그래밍

- 1) 통계, 통계학이란?
- 2) 통계와 프로그래밍
- 3) R과 R studio 설치

2

R 기초 연습하기

3

자료 정리

- 1) 모집단과 표본
- 2) 자료의 종류
- 3) 기술 통계량
- 4) 이변량 자료와 표본상관계수

4

그래프

- 1) 분할표
- 2) 상자그림
- 3) 도수분포표와 히스토그램
- 4) 줄기와 잎 그림

1. 통계와 R 프로그래밍

통계, 통계학이란?



**'통계 (Statistic)'란
무엇일까요?**

떠오르는 생각을 자유롭게 이야기 해볼까요?

통계, 통계학이란?

- 통계란 무엇일까요?

다음의 시나리오를 통해 '통계적 분석 절차'를 이해해 봅시다.

시나리오

- 우리 반과 옆 반이 1년 간 매 달 시험을 치렀다고 하자.
- 우리 반과 옆 반의 시험성적을 분석하고자 할 때, 어떤 절차로 해야 할까?

통계, 통계학이란?

- 통계란 무엇일까요?

다음의 시나리오를 통해 '통계적 분석 절차'를 이해해 봅시다.

시나리오

- 우리 반과 옆 반이 1년 간 매 달 시험을 치렀다고 하자.
- 우리 반과 옆 반의 시험성적을 분석하고자 할 때, 어떤 절차로 해야 할까?

우리 반과 옆 반의 1년 간
각 시험 점수 결과를
요약해보자.

우리 반과 옆 반의 시험 점수는
어떻게 변화하였나?

우리 반과 옆 반의 시험 점수가
차이가 있을까?

통계, 통계학이란?

- 통계란 무엇일까요?

다음의 시나리오를 통해 '통계적 분석 절차'를 이해해 봅시다.

시나리오

- 우리 반과 옆 반이 1년 간 매 달 시험을 치렀다고 하자.
- 우리 반과 옆 반의 시험성적을 분석하고자 할 때, 어떤 절차로 해야할까?

우리 반과 옆 반의 1년 간
각 시험 점수 결과를
요약해보자.

우리 반과 옆 반의 시험 점수는
어떻게 변화하였나?

우리 반과 옆 반의 시험 점수가
차이가 있을까?

통계, 통계학이란?

- '어떻게' 분석할 수 있을까?

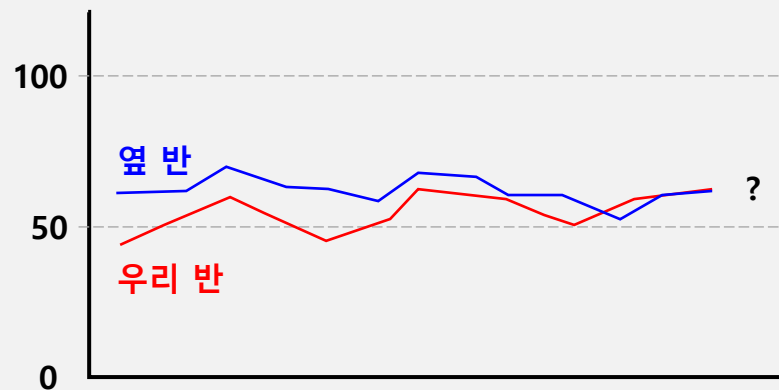
우리 반과 옆 반의 1년 간
각 시험 점수 결과를
요약해보자.

1월	평균	최솟값	최댓값
우리반	42	9	92
옆반	55	20	86

⋮

12월	평균	최솟값	최댓값
우리반	54	18	98
옆반	60	28	92

우리 반과 옆 반의 시험 점수는
어떻게 변화하였나?



우리 반과 옆 반의 시험 점수가
차이가 있을까?

그래프로 상으로는 차이가
줄어드는 것 같은데..

'수학적인 근거'가 있는 걸까?

통계적
검정

통계, 통계학이란?

- 통계(統計)의 사전적 정의

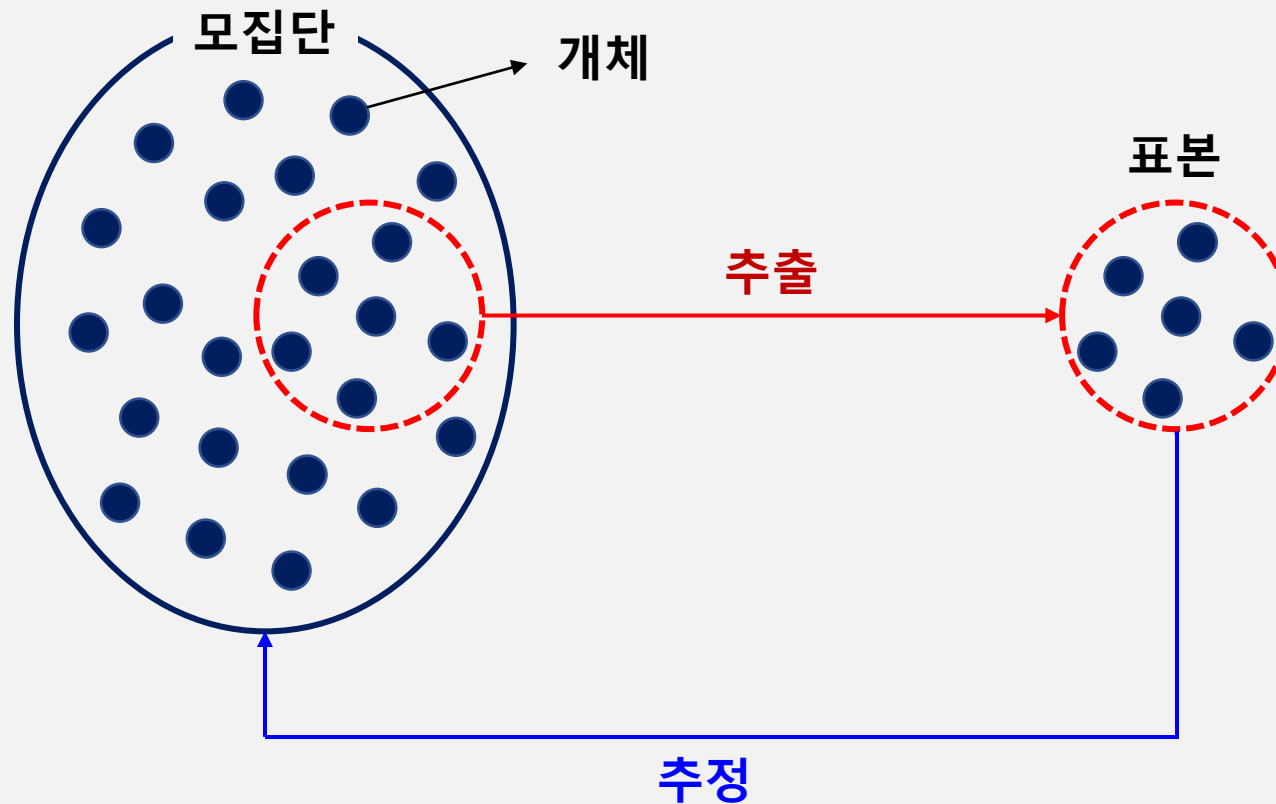
어떤 자료나 정보를 **분석·정리**하여 그 내용을 특징짓는 횟수·빈도·비율 등의 **수치를 산출**해 내는 일.
또는, 그 산출된 수치.

- 통계학 (統計學, statistics)

산술적 방법을 기초로 하여, 주로 다량의 데이터를 관찰하고 정리 및 분석하는 방법을 연구하는 수학의 한 분야

통계, 통계학이란?

통계의 분류



기술통계

수집한 데이터(표본)의 특성 분석

ex. 평균값(mean), 중위수(median), 최빈값(mode), 최댓값(maximum), 최솟값(minimum), 범위(range), 분산(variance), 표준편차(standard deviation)

추론통계

- 표본의 특성을 파악하여 전체 데이터(모집단)의 특성으로 일반화할 수 있는지 여부를 판단
- 모집단의 특성을 추정하는 것이 목적

ex. 선거 출구조사

통계와 프로그래밍

- 프로그래밍 (programming)

우리가 해결해야 할 문제를 컴퓨터가 처리할 수 있도록 문제해결 절차를 체계적으로 서술하는 과정

- 프로그래밍 언어 (programming language)

컴퓨터가 어떤 작업을 수행하기 위한 프로그램을 작성하는데 사용하는 언어

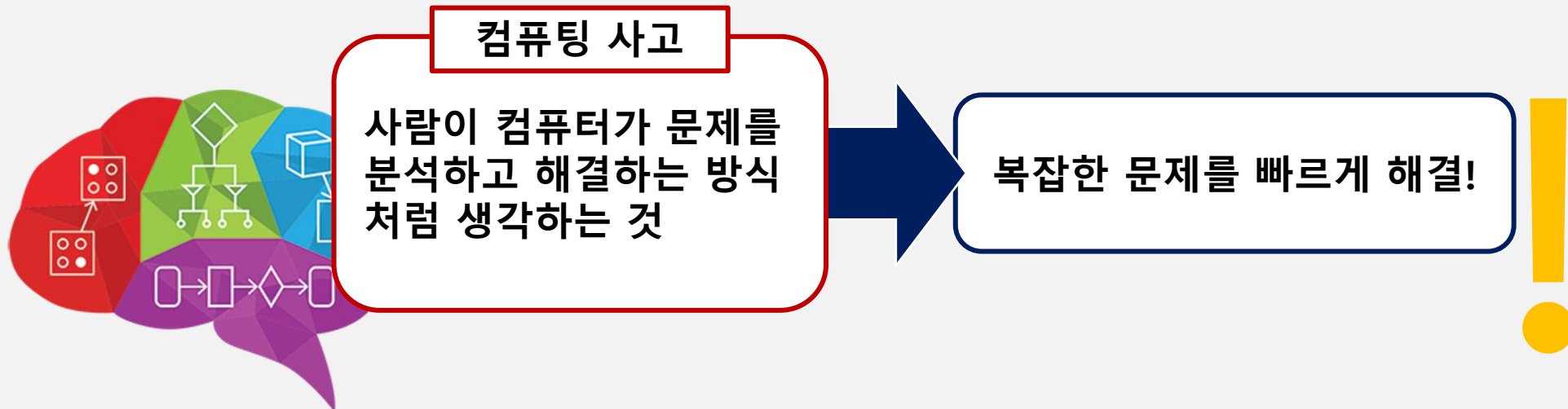
통계와 프로그래밍

- 프로그래밍 (programming)

우리가 해결해야 할 문제를 컴퓨터가 처리할 수 있도록 문제해결 절차를 체계적으로 서술하는 과정

- 프로그래밍 언어 (programming language)

컴퓨터가 어떤 작업을 수행하기 위한 프로그램을 작성하는데 사용하는 언어



통계와 프로그래밍

구현하고 싶은 것

1부터 5까지의 정수를 더하기

명령
(프로그램)

```
sum = 0
for (i in 1:5) {
  sum = sum + i
}

print(sum)
```

컴파일

기계어로 번역
(컴퓨터가 처리하는 과정)

```
1000 1011 0100 0101 1111 1000
1000 0011 1100 0100 0000 1100
0000 0011 0100 0101 1111 1100
```

실행

실행 결과

$1+2+3+4+5$
 $=15$

통계와 프로그래밍

- 우리가 실습할 통계 프로그램: R

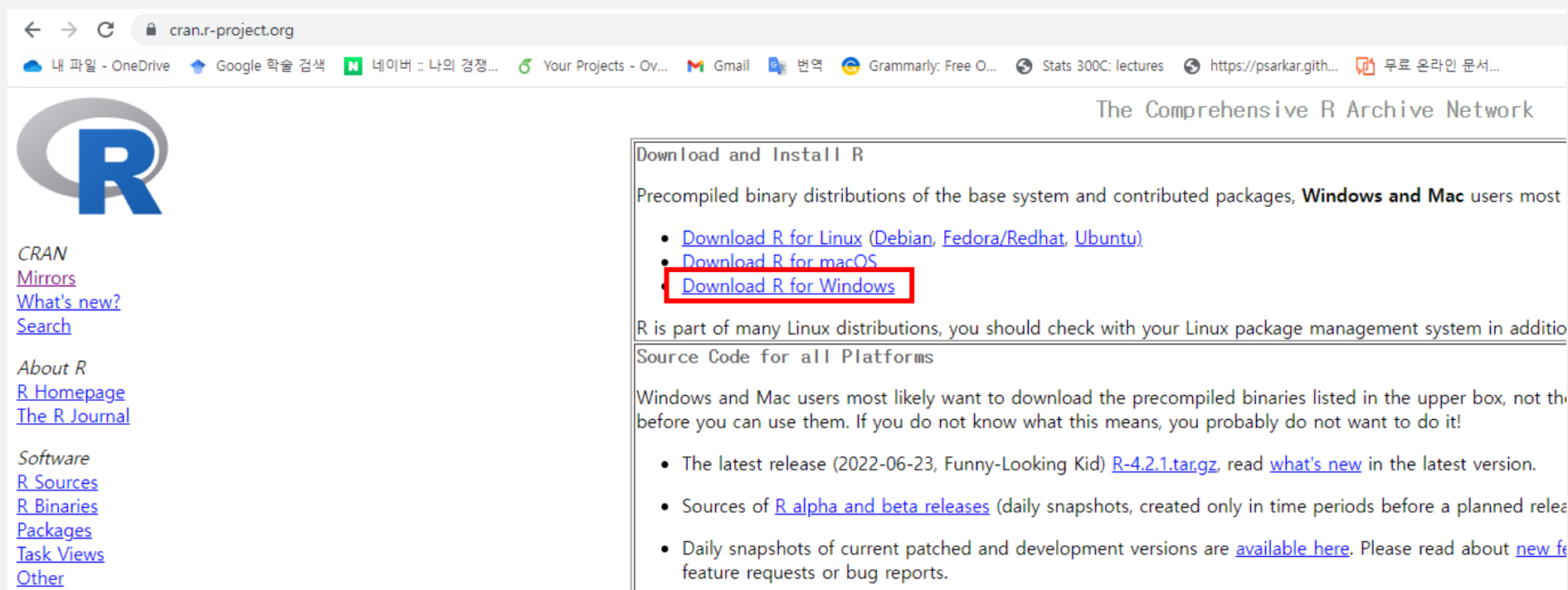


1. 데이터 분석에 특화된 언어
2. 배우기 쉬운 언어
3. 탄탄한 사용자 커뮤니티
4. 다양한 패키지 제공
5. 미적이고 기능적인 통계 그래프 제공
6. 편리한 프로그래밍 환경
7. 무료 사용

R과 R studio 설치

■ R 설치하기

1 <https://cran.r-project.org> 에 접속



The screenshot shows the CRAN website (cran.r-project.org) with the following content:

- CRAN logo
- CRAN Mirrors
- What's new?
- Search
- About R
 - R Homepage
 - The R Journal
- Software
 - R Sources
 - R Binaries
 - Packages
 - Task Views
 - Other

The main content area is titled "The Comprehensive R Archive Network" and "Download and Install R". It states: "Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most".

- [Download R for Linux \(Debian, Fedora/Redhat, Ubuntu\)](#)
- [Download R for macOS](#)
- [Download R for Windows](#) (highlighted with a red box)

R is part of many Linux distributions, you should check with your Linux package management system in addition.

Source Code for all Platforms

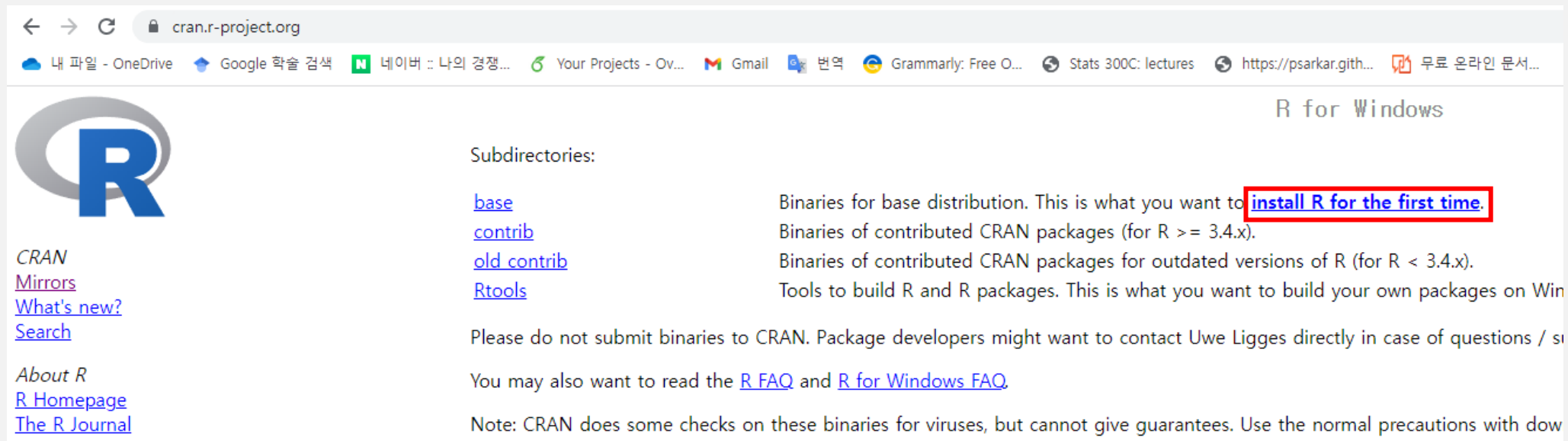
Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code, before you can use them. If you do not know what this means, you probably do not want to do it!

- The latest release (2022-06-23, Funny-Looking Kid) [R-4.2.1.tar.gz](#), read [what's new](#) in the latest version.
- Sources of [R alpha and beta releases](#) (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions are [available here](#). Please read about [new features](#), [feature requests](#) or [bug reports](#).

2 Download R for Windows 클릭


R과 R studio 설치

3 Install R for the first time 클릭



← → ↻ cran.r-project.org

내 파일 - OneDrive Google 학술 검색 네이버 :: 나의 경쟁... Your Projects - Ov... Gmail 번역 Grammarly: Free O... Stats 300C: lectures <https://psarkar.github.io> 무료 온라인 문서...

 R for Windows

Subdirectories:

base	Binaries for base distribution. This is what you want to install R for the first time.
contrib	Binaries of contributed CRAN packages (for R >= 3.4.x).
old contrib	Binaries of contributed CRAN packages for outdated versions of R (for R < 3.4.x).
Rtools	Tools to build R and R packages. This is what you want to build your own packages on Win

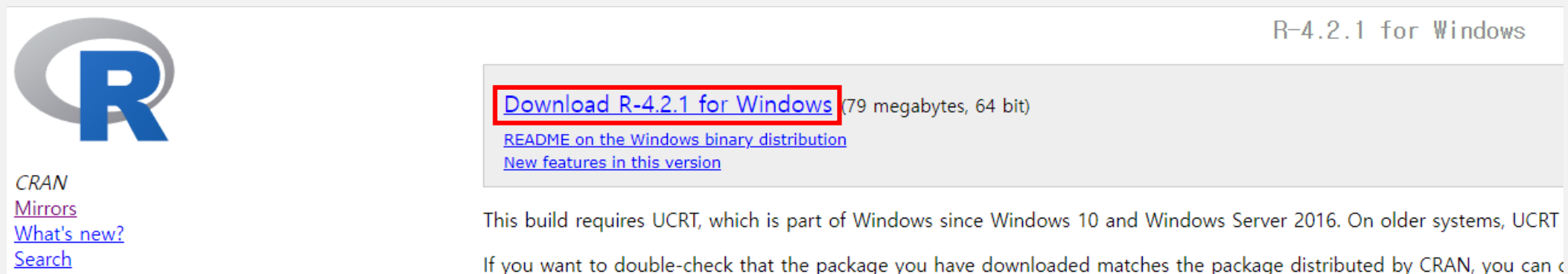
Please do not submit binaries to CRAN. Package developers might want to contact Uwe Ligges directly in case of questions / s


You may also want to read the [R FAQ](#) and [R for Windows FAQ](#).

Note: CRAN does some checks on these binaries for viruses, but cannot give guarantees. Use the normal precautions with dow

CRAN
[Mirrors](#)
[What's new?](#)
[Search](#)
 About R
[R Homepage](#)
[The R Journal](#)

4 Download R-4.2.1 for Windows 클릭



 R-4.2.1 for Windows

Download R-4.2.1 for Windows (79 megabytes, 64 bit)

[README on the Windows binary distribution](#)
[New features in this version](#)

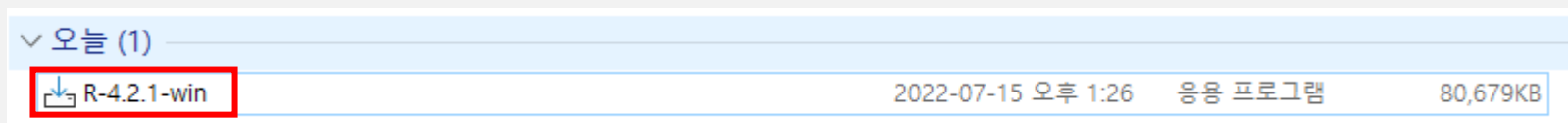
This build requires UCRT, which is part of Windows since Windows 10 and Windows Server 2016. On older systems, UCRT

If you want to double-check that the package you have downloaded matches the package distributed by CRAN, you can c

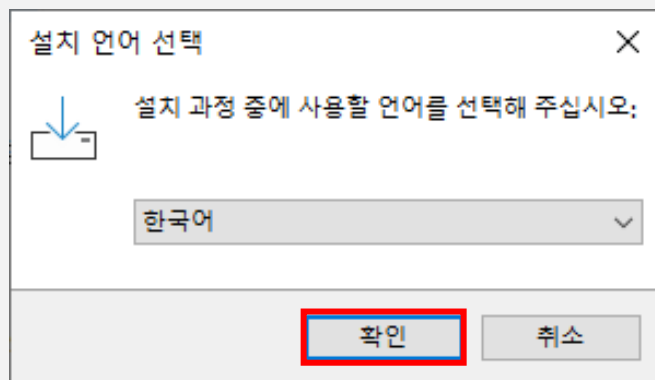
CRAN
[Mirrors](#)
[What's new?](#)
[Search](#)

R과 R studio 설치

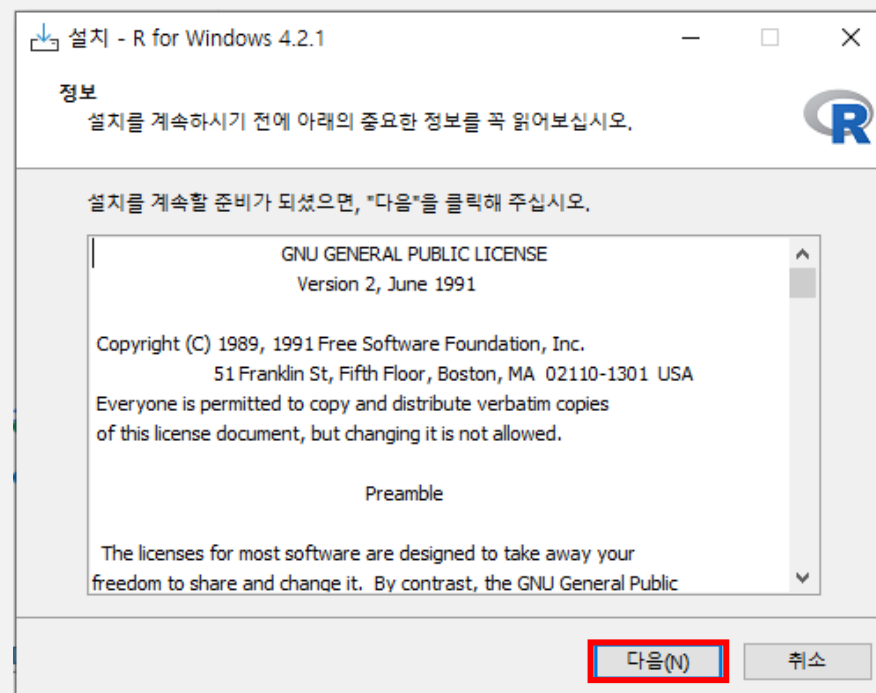
5 다운로드 – R-4.3.2-win 실행



6

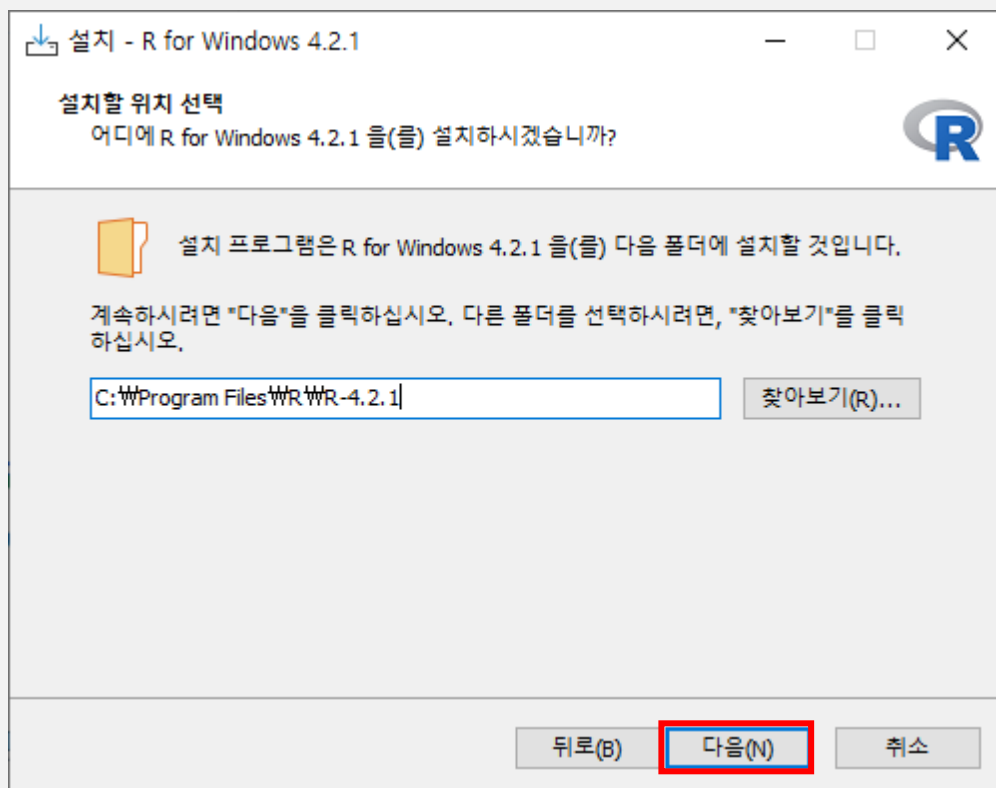


7

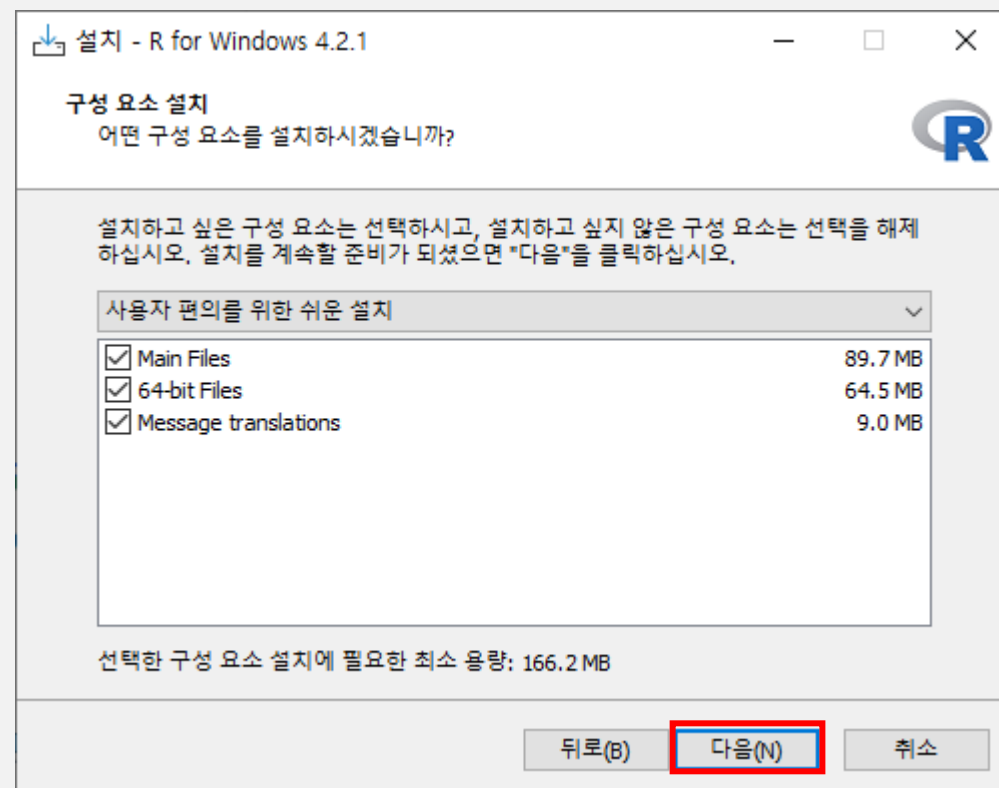


R과 R studio 설치

8

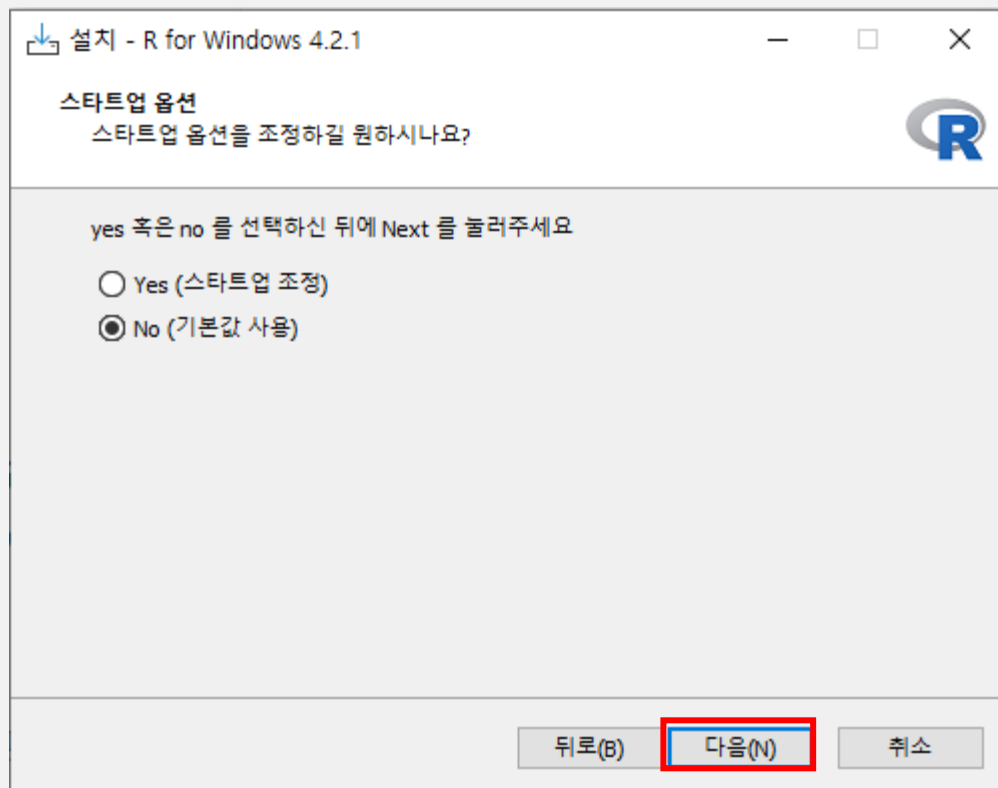


9

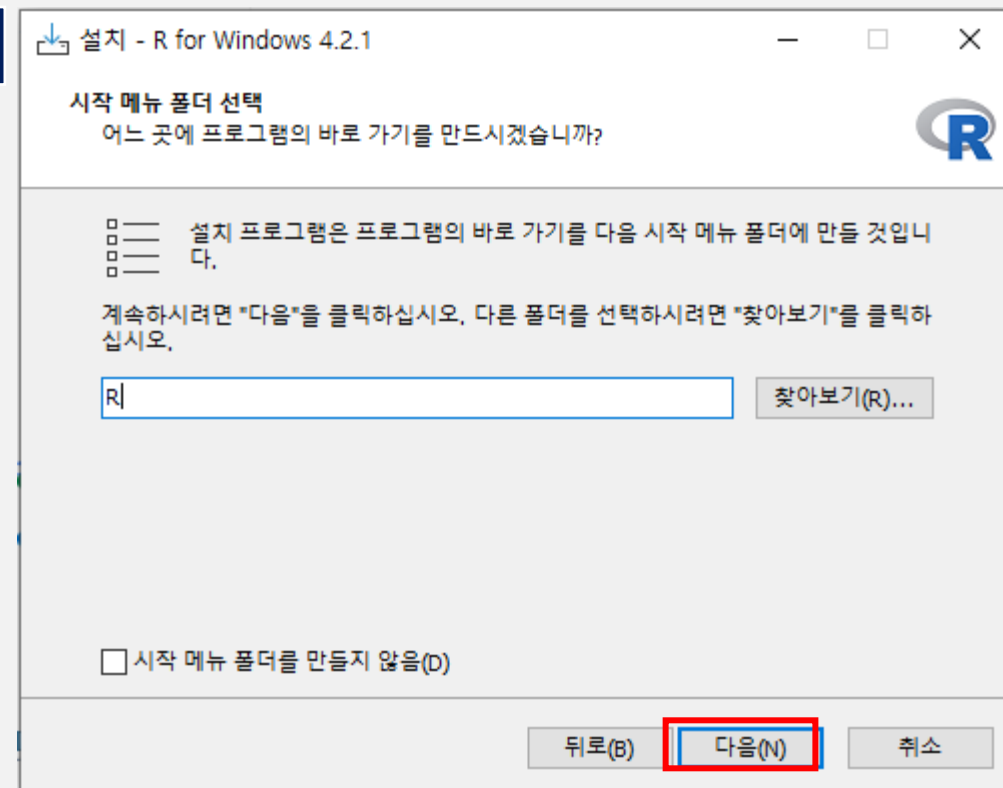


R과 R studio 설치

10

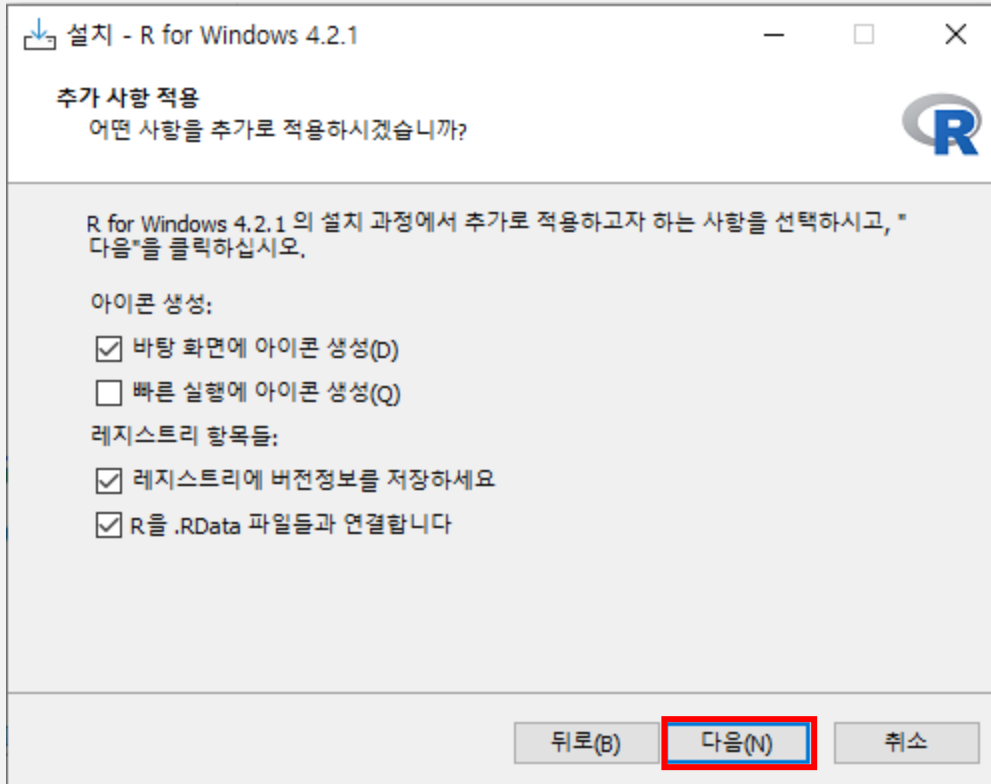


11

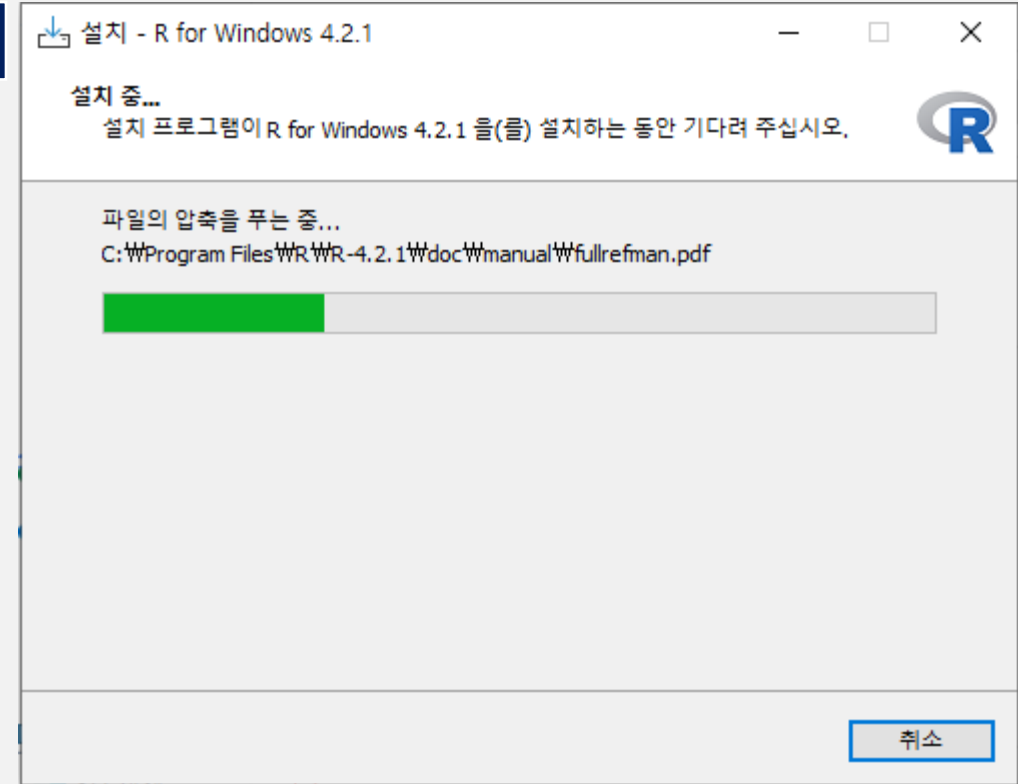


R과 R studio 설치

12



13



R과 R studio 설치

14



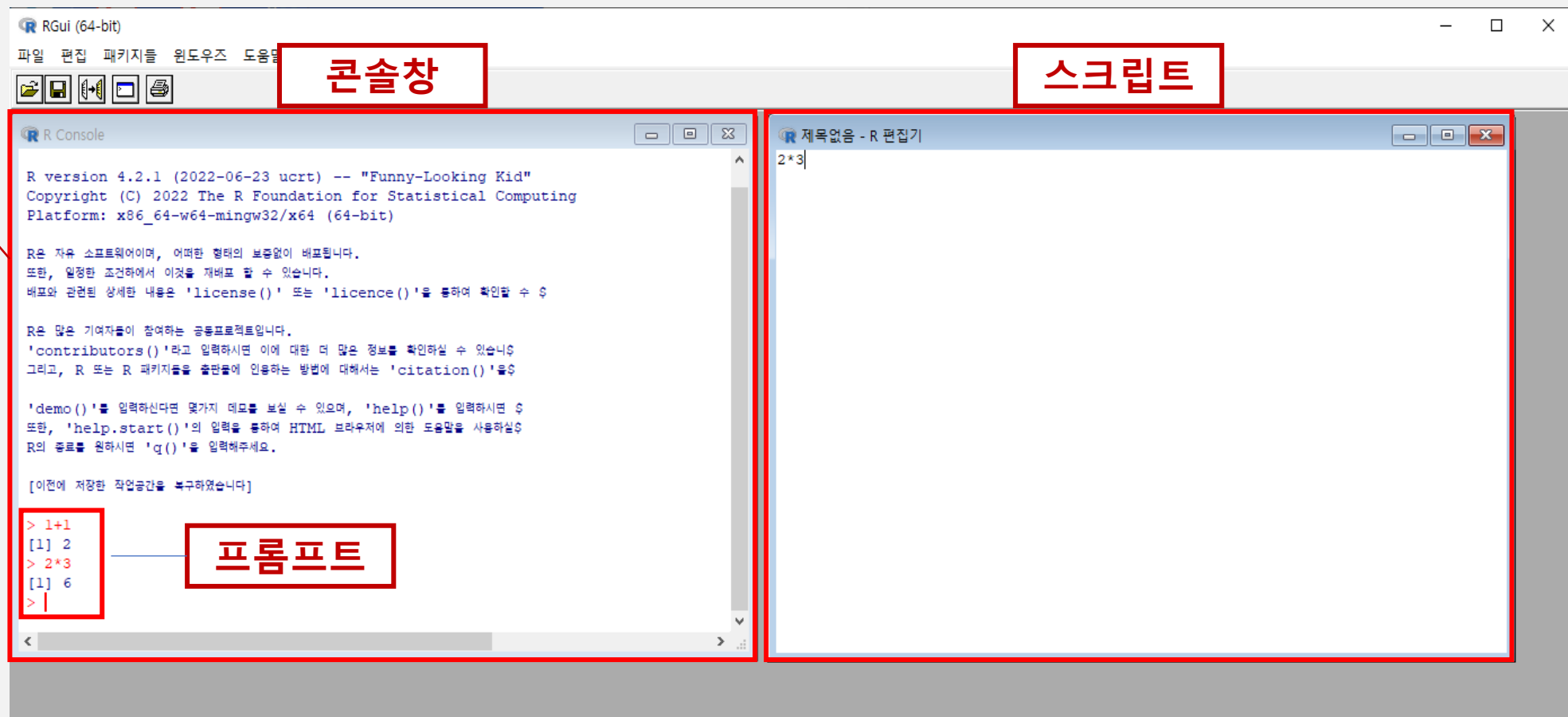
15

바탕화면에 아이콘 생성



R과 R studio 설치

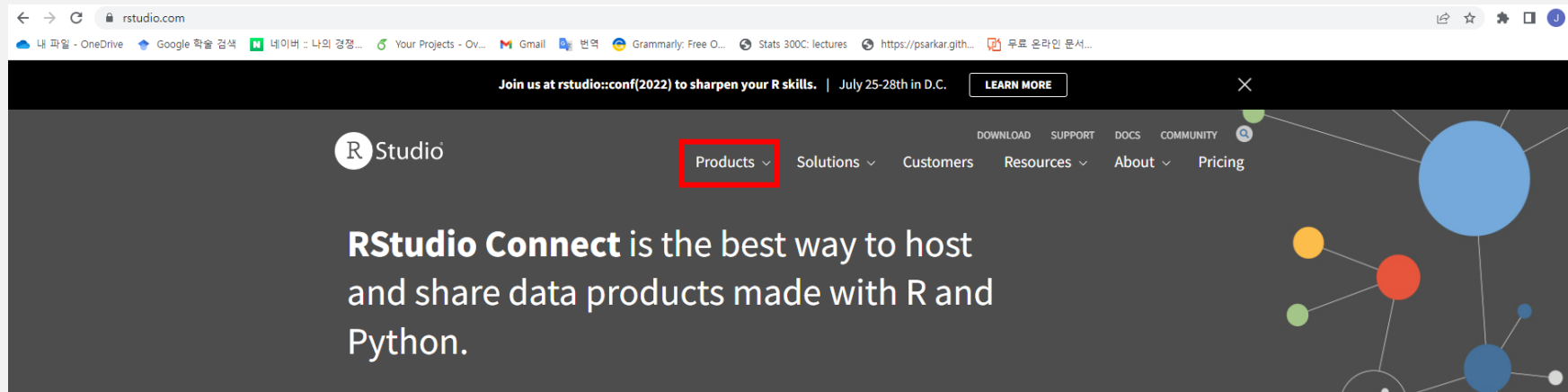
■ R 화면구성



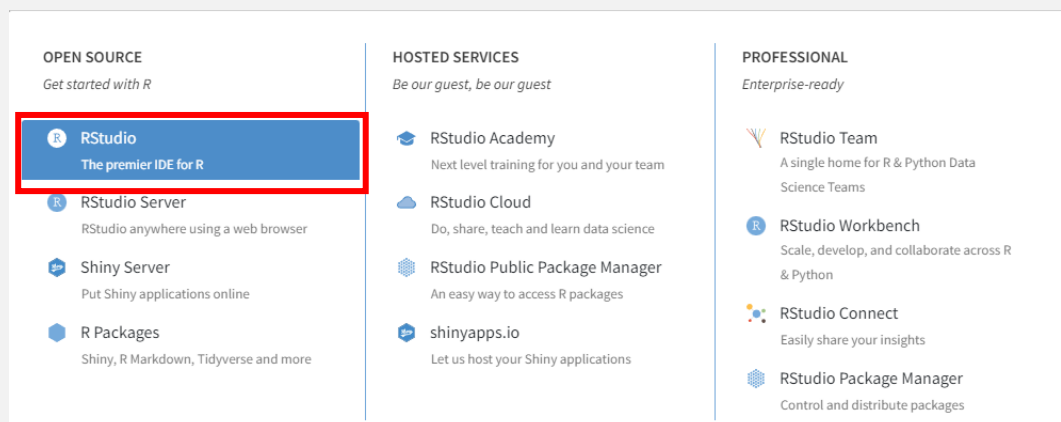
R과 R studio 설치

■ R studio 설치하기

1 <https://www.rstudio.com> 에 접속



2



R과 R studio 설치

3

R Studio Desktop

Open Source Edition

- Access RStudio locally
- Syntax highlighting, code completion, and smart indentation
- Execute R code directly from the source editor
- Quickly jump to function definitions
- View content changes in real-time with the Visual Markdown Editor
- Easily manage multiple working directories using projects
- Integrated R help and documentation
- Interactive debugger to diagnose and fix errors
- Extensive package development tools

Overview

RStudio Desktop Pro

All of the features of open source; plus:

- A commercial license for organizations not able to use AGPL software
- Access to priority support
- [RStudio Professional Drivers](#)
- Connect directly to your RStudio Workbench instance remotely

Support

Community forums only

- Priority Email Support
- 8 hour response during business hours (ET)

License

AGPL v3

[RStudio License Agreement](#)

Pricing

Free

\$995/year

[DOWNLOAD RSTUDIO DESKTOP](#)

[DOWNLOAD FREE RSTUDIO DESKTOP PRO TRIAL](#)

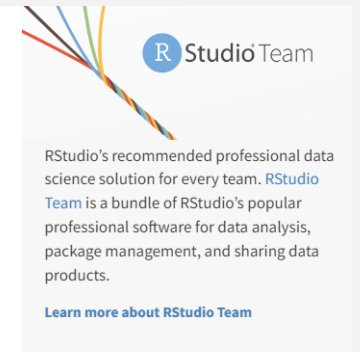
[Purchase](#) | [Contact Sales](#)

4

Choose Your Version

The RStudio IDE is a set of integrated tools designed to help you be more productive with R and Python. It includes a console, syntax-highlighting editor that supports direct code execution, and a variety of robust tools for plotting, viewing history, debugging and managing your workspace.

[LEARN MORE ABOUT THE RSTUDIO IDE](#)



RStudio Desktop

Open Source License

Free

[DOWNLOAD](#)

RStudio Desktop Pro

Commercial License

\$995

/year

[BUY](#)

RStudio Server

Open Source License

Free

[DOWNLOAD](#)

RStudio Workbench

Commercial License

\$4,975

/year

(5 Named Users)

[BUY](#)

R과 R studio 설치

5

RStudio Desktop 2022.07.0+548 - [Release Notes](#)

1. Install R. RStudio requires R 3.3.0+ [↗](#).

2. Download RStudio Desktop. Recommended for your system:



DOWNLOAD RSTUDIO FOR WINDOWS
2022.07.0+548 | 190.14MB

Requires Windows 10/11 (64-bit)



6

✓ 오늘 (2)



RStudio-2022.07.0-548

2022-07-15 오후 1:53

응용 프로그램

185,686KB



R-4.2.1-win

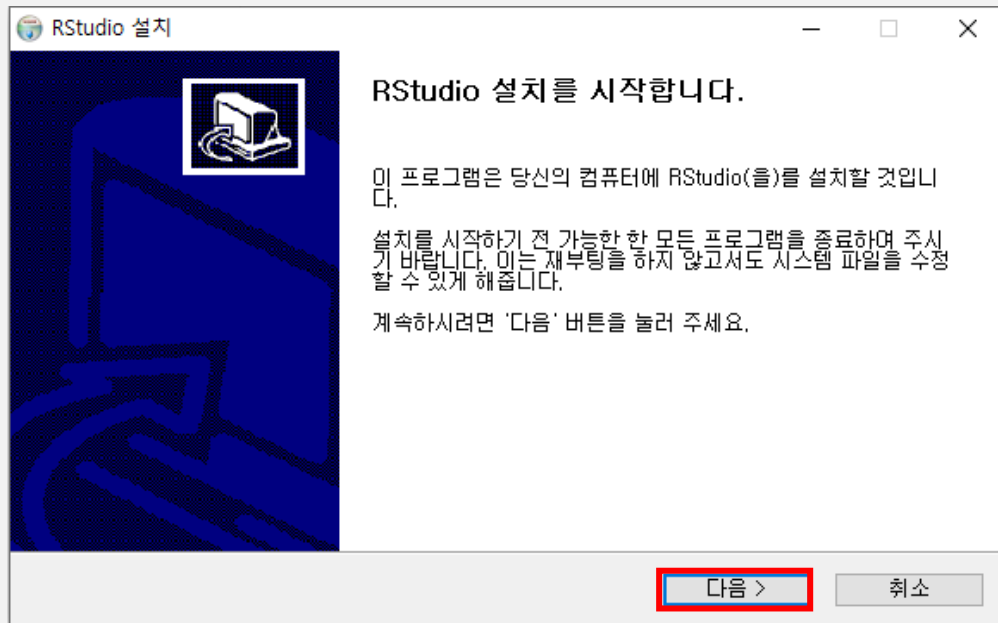
2022-07-15 오후 1:26

응용 프로그램

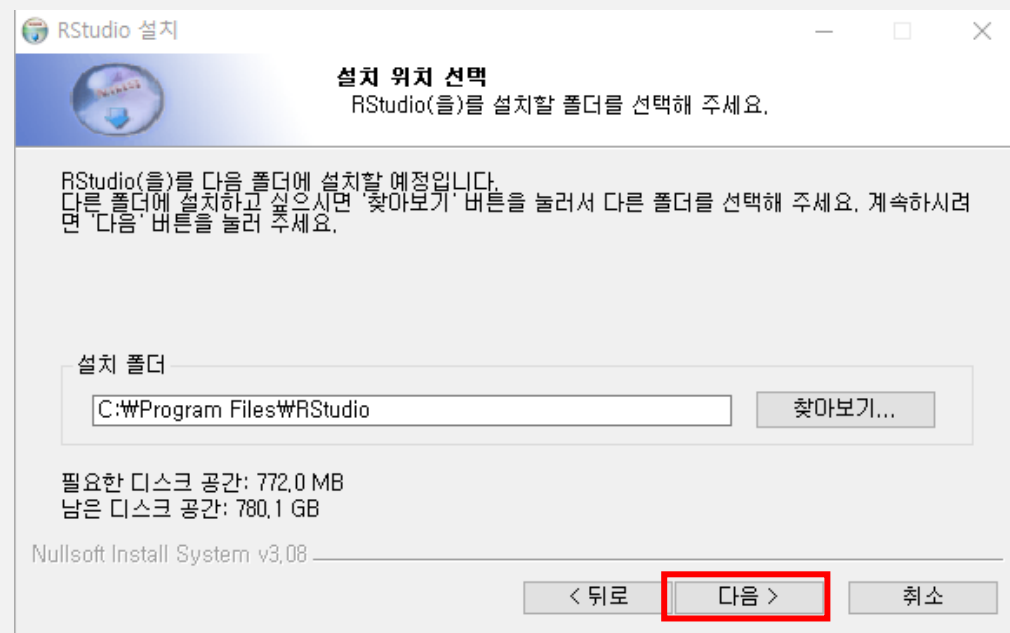
80,679KB

R과 R studio 설치

7

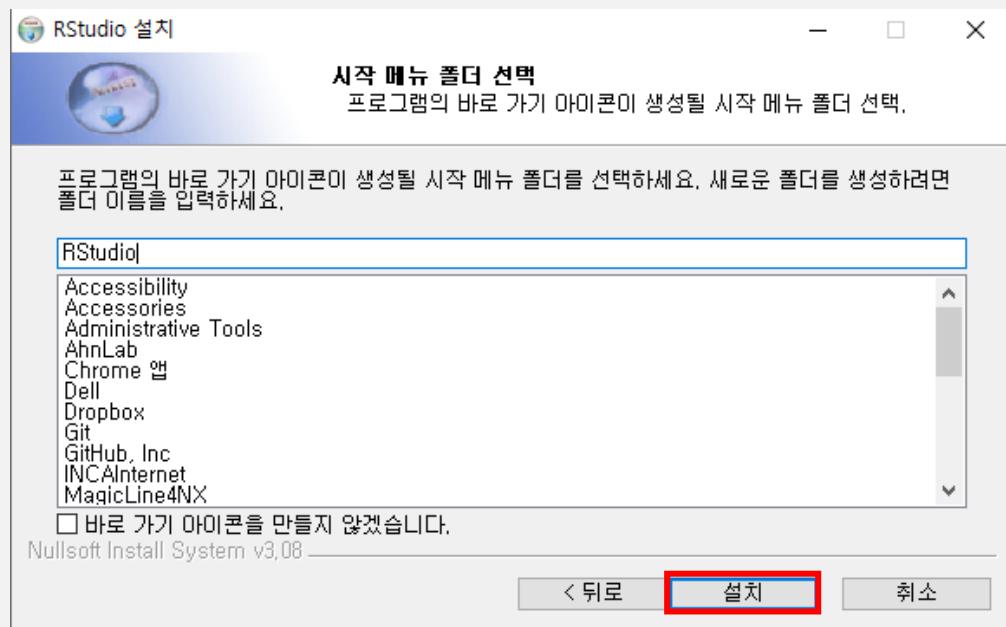


8

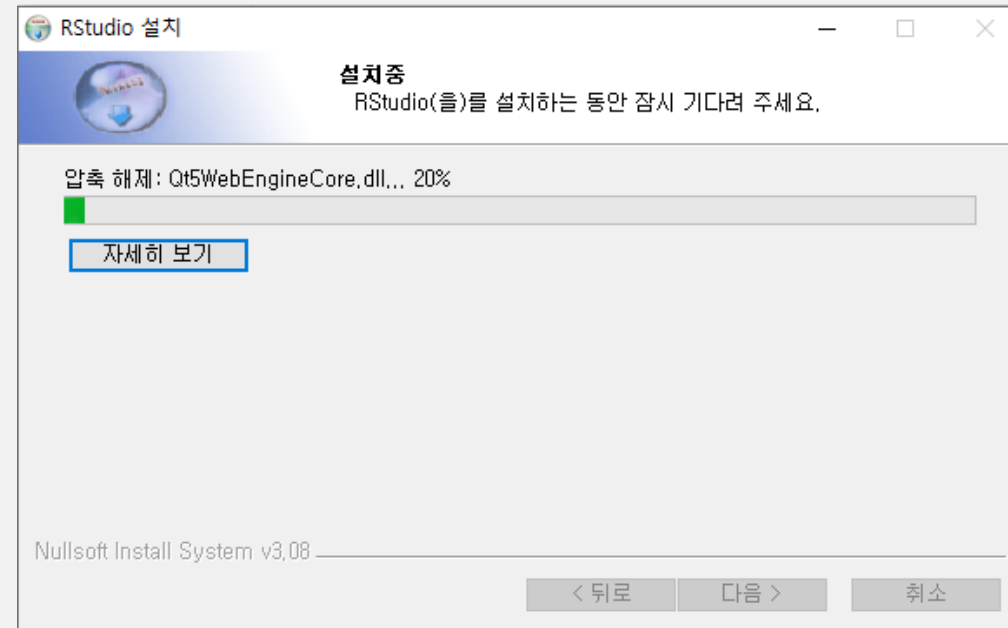


R과 R studio 설치

9



10

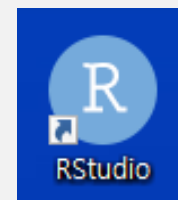


R과 R studio 설치

11

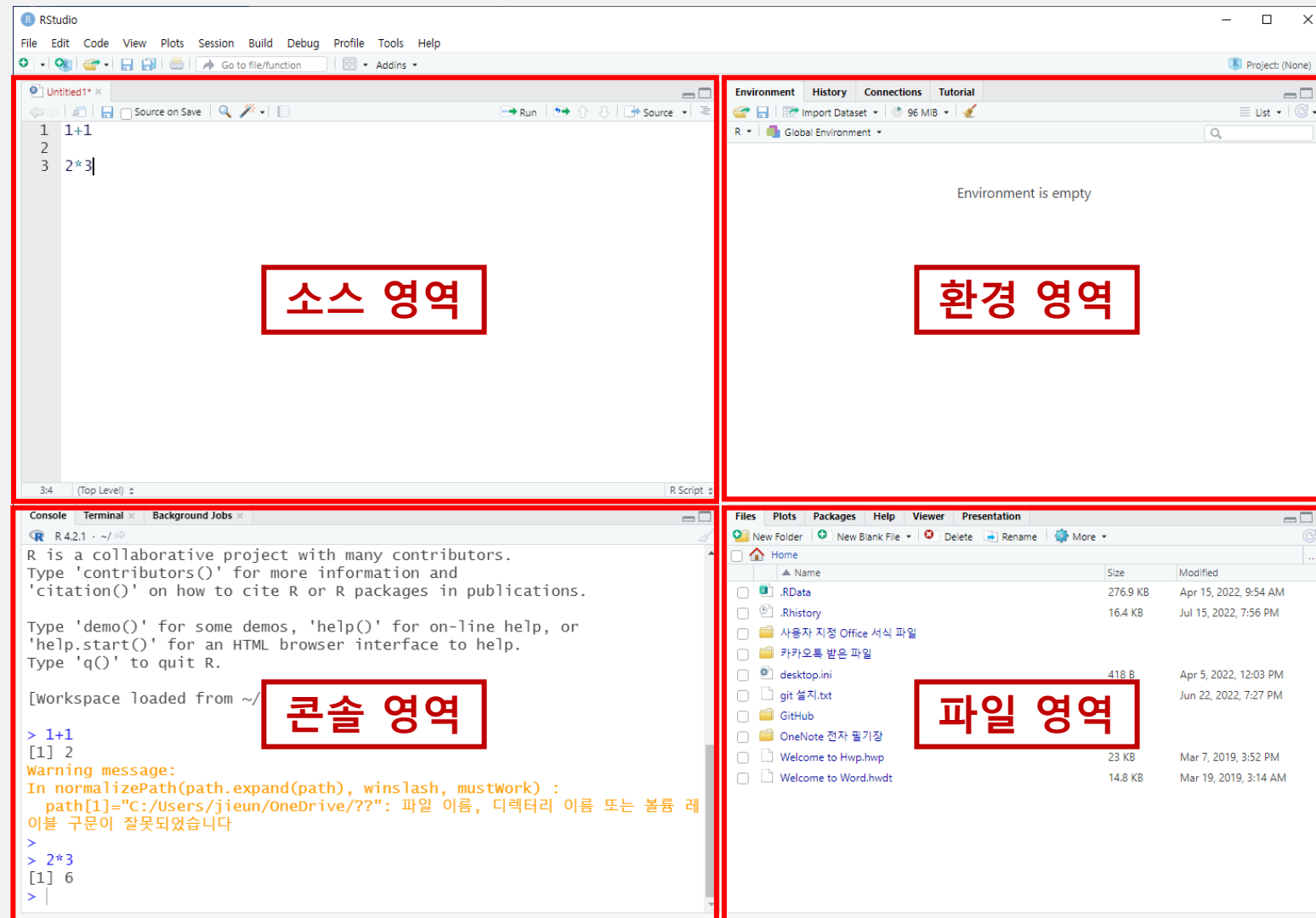


12 아이콘 클릭하여 실행



R과 R studio 설치

■ R studio 화면구성



R과 R studio 설치

- 실습해볼까요?

새 파일 열기

[File]-[New File]-[R Script] 또는 [Ctrl]+[Shift] + [N]

파일 저장하기

[File]-[Save As]

2. R 기초 연습하기

R 기초 연습하기

실습 내용 (tutorial.R)

1. 단순계산
2. 객체에 값 할당하기
3. 모든 객체 보기
4. 예약어 NaN, Inf, NA
5. R의 객체
6. Help 기능
7. 패키지 사용하기
8. 자료 입력하기
9. 데이터 프레임 다루기

R 기초 연습하기

- 연습해볼까요?

Q1. 45, 55, 51, 63, 55를 원소로 갖는 `wright` 벡터를 생성하고, 각각의 원소에 Choi, Cho, Shin, Moon, Chae의 이름을 할당하시오.

Q2. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12의 자료를 두 가지 방법으로 3행과 4열의 행렬을 만들고 행 (row)과 열 (col)의 이름을 지정하시오.

3. 자료 정리

모집단과 표본

▪ 모집단 (population)

: 흥미가 있는 대상의 전체

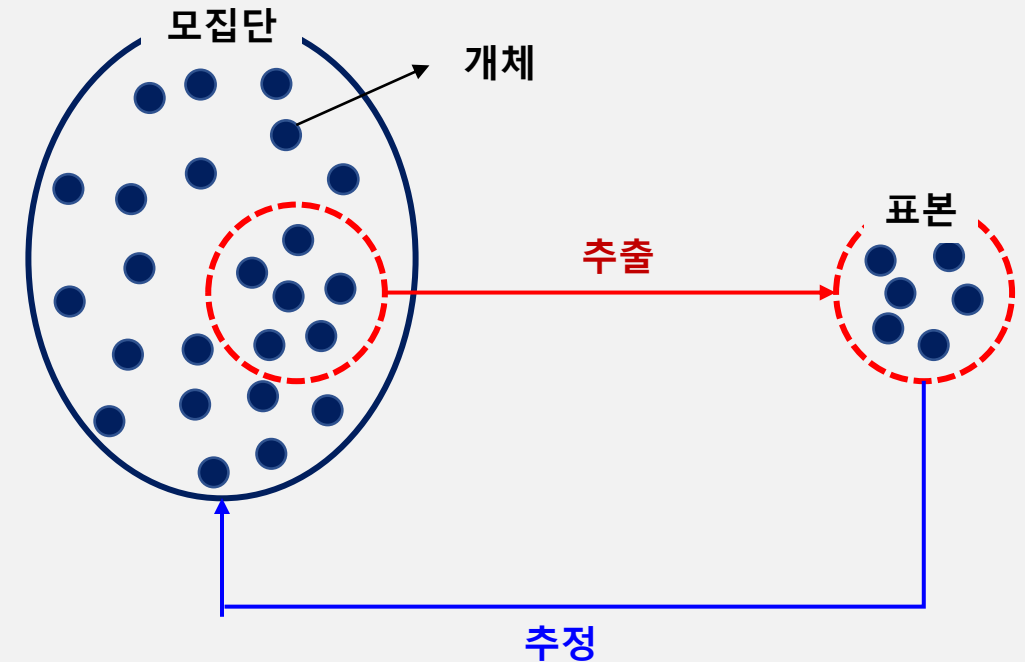
ex) 한국인 20대 남성의 신장,

유권자의 정당에 대한 지지율,

공장에서 생산되는 제품 (전구 등)의 수명,

어떤 농산물에 대한 전국의 농가 1호당의 수확량,

새로 개발된 의약품의 어떤 환자에 대한 효과 등



모집단과 표본

- 모집단 (population)

: 흥미가 있는 대상의 전체

ex) 한국인 20대 남성의 신장,

유권자의 정당에 대한 지지율,

공장에서 생산되는 제품 (전구 등)의 수명,

어떤 농산물에 대한 전국의 농가 1호당의 수확량,

새로 개발된 의약품의 어떤 환자에 대한 효과 등

20대 한국남성 전체

유권자 전체

생산품 전체

전국의 농가 전체

환자 전체

- 문제의 모형화

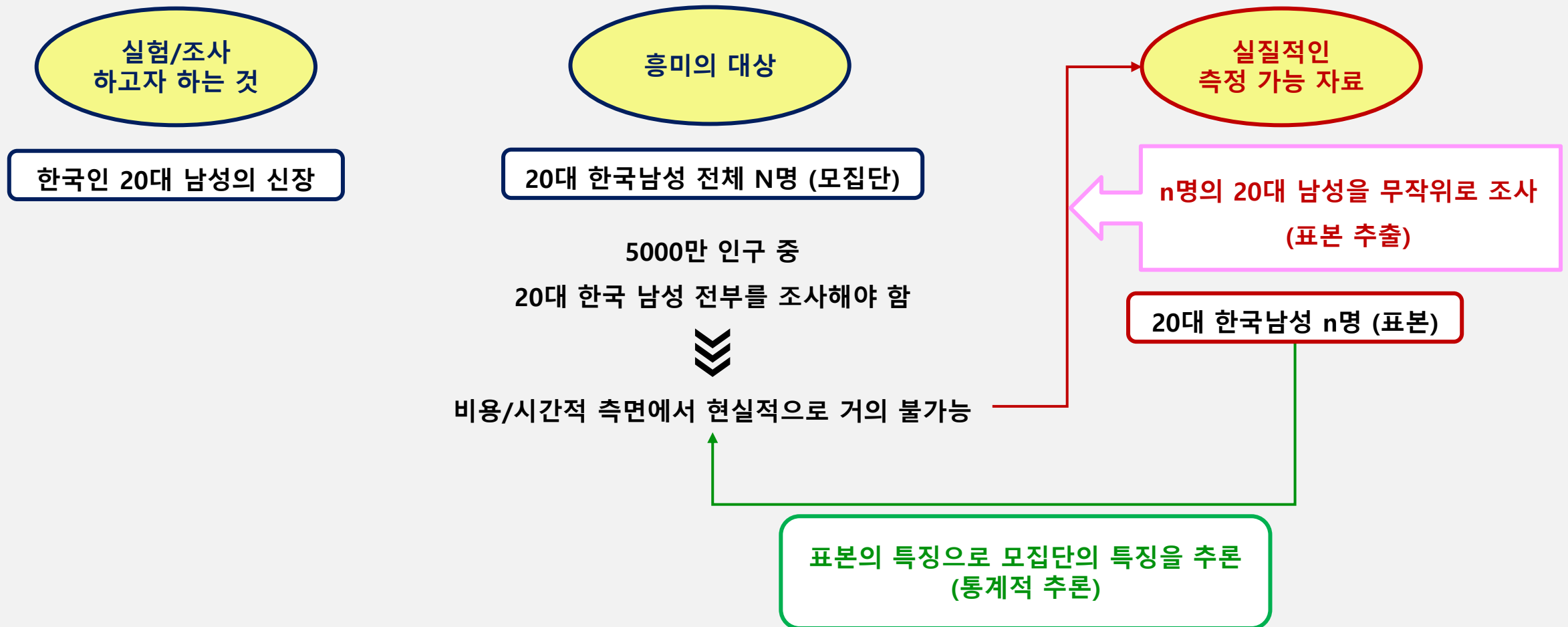
실험/조사
하고자 하는 것



흥미의 대상

모집단과 표본

- 통계적 추론 (statistical inference): 모집단을 조사하는 것이 가능한가?



모집단과 표본

- 통계적 추론 (statistical inference): 모집단을 조사하는 것이 가능한가?

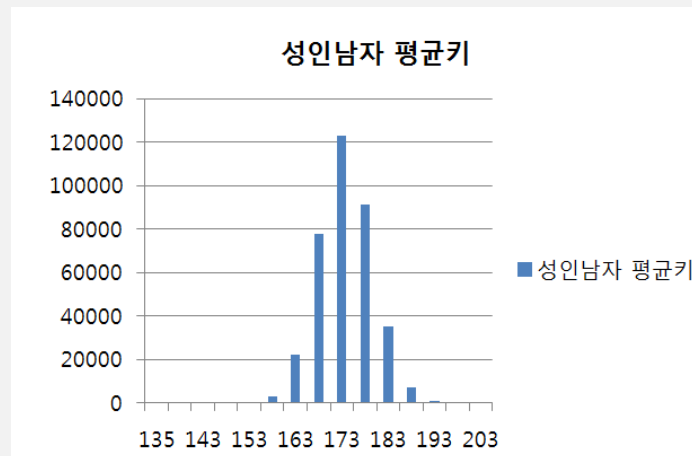
: 몇 개의 실험대상을 무작위(random)하게 추출하여 모은 자료로 모집단의 특징을 추론하는 것.

이렇게 모여진 자료를 표본(sample)이라고 하고, 모집단에서 표본을 취한 것을 표본추출(sampling)이라고 한다.

- 무작위 추출(random sampling)

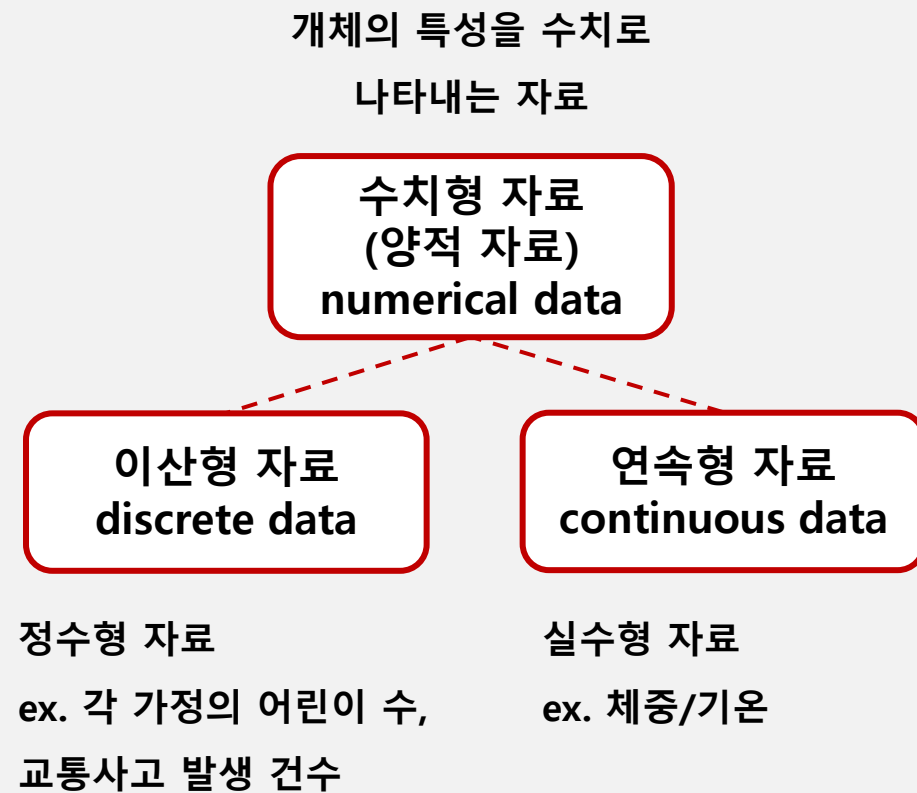
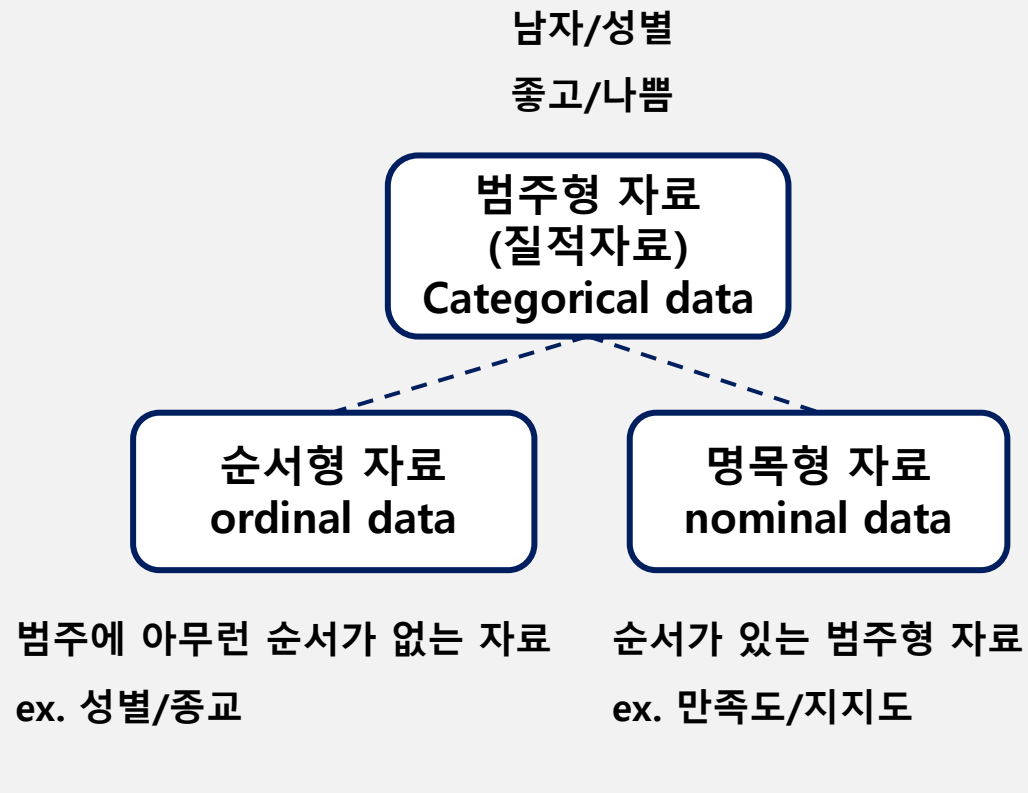
: 표본 추출은 무작위 추출이어야 한다. 이는 모집단에 포함되는 모든 원소가 같은 가능성으로 추출되어야 함을 의미한다.

ex. 20대 남성의 신장을 조사할 때, 키가 큰 집단의 평균 신장을 조사한 뒤 이를 한국인의 평균 신장이라고 하면 안된다.



자료의 종류

■ 자료의 종류



기술 통계량

▪ 기술통계량 1. 중심척도

① 평균 (mean)

- 중심위치에 대한 가장 대표적인 척도로 n 개의 자료값 x_1, x_2, \dots, x_n 이 있을 때

$$\bar{x} = (x_1 + \dots + x_n)/n$$

으로 계산된다.

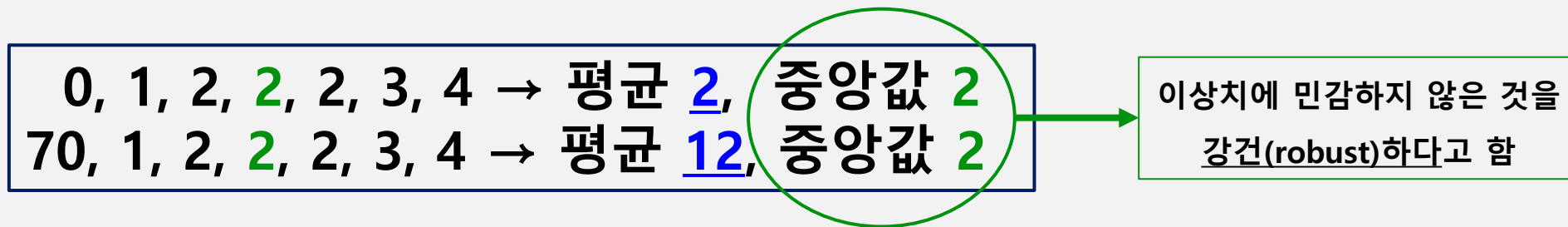
- 모든 자료에 똑같이 $1/n$ 씩의 가중치를 주어 중심을 구한 것과 같다.
- 평균은 자료의 분포가 한 쪽으로 치우치지 않고 하나의 축을 중심으로 좌우 대칭으로 흩어진 형태의 자료의 특성을 표현하기에 적합하다.
- 그러나 아주 큰 값 또는 아주 작은 값 등의 이상치 (outlier)가 있을 때에는 이상치의 영향을 많이 받아 평균을 사용하는 것이 부적절하다.

기술 통계량

기술통계량 1. 중심척도

② 중앙값 (median, 혹은 중위수)

- 자료들을 작은 값부터 큰 값까지 순서대로 배열하였을 때 가운데에 위치하는 값.
 - n 개의 자료가 있을 때 n 이 홀수라면 $(n+1)/2$ 번째 값이 되고, 짝수라면 $(n/2)$ 번째 값과 $(n/2) + 1$ 번째 값의 평균이 중앙값이 된다.
 - 양 끝에 아주 큰 값 또는 아주 작은 값이 있더라도 중앙값에는 영향을 미치지 않게 된다는 이점이 있다.
 - 따라서 분포의 형태가 좌우대칭이 아니고 어느 한쪽으로 치우쳐 있을 때 중심위치를 나타내는 척도로서 유용하게 쓸 수 있다.
- ex. 두 가지 자료 (0, 1, 2, 2, 2, 3, 4)와 (70, 1, 2, 2, 2, 3, 4)의 평균, 중앙값을 비교해보자

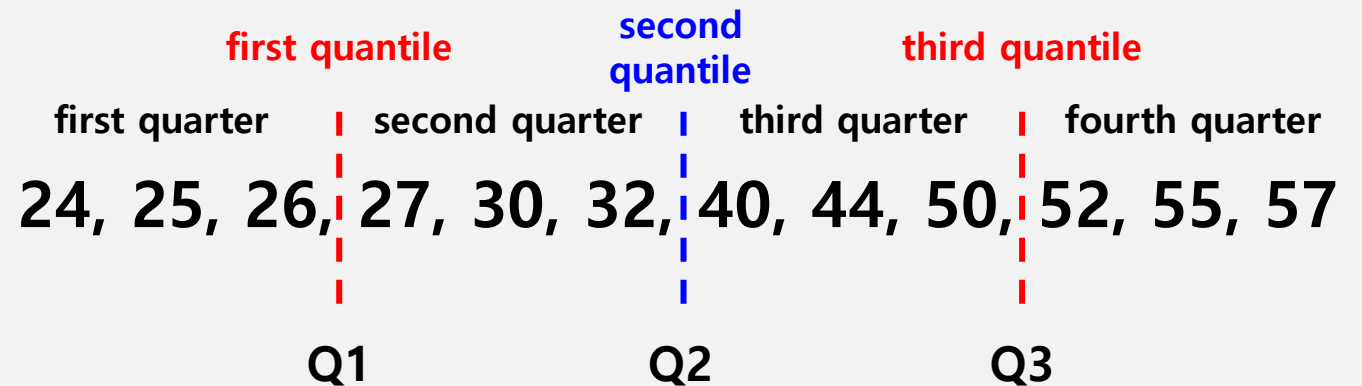


기술 통계량

■ 기술통계량 1. 중심척도

② 중앙값 (median, 혹은 중위수)

- 사분위수 (quartile)
: 자료를 크기 순으로 배열하였을 때 4등분하는 위치에 오는 값
- 제 1사분위수 (Q1): 하위 25%에 해당하는 값
- 제 2사분위수 (Q2): 50%에 해당하는 값 (=중앙값)
- 제 3사분위수 (Q3): 상위 75%에 해당하는 값
- 100등분 한 백분위수 (percentile)로 확장할 수 있음.

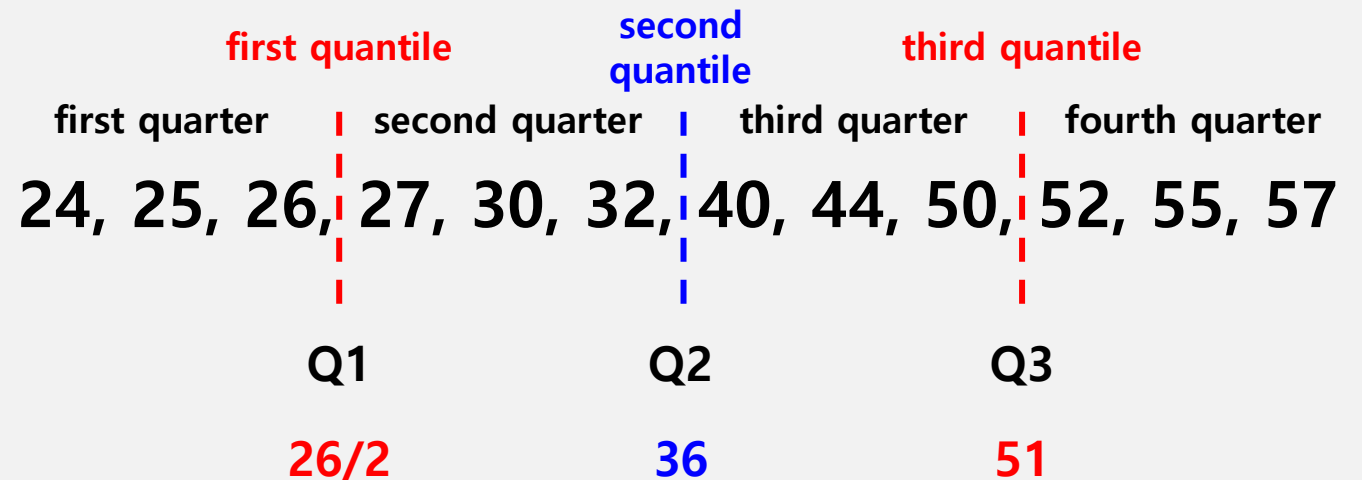


기술 통계량

기술통계량 1. 중심척도

② 중앙값 (median, 혹은 중위수)

- 사분위수 (quartile)
: 자료를 크기 순으로 배열하였을 때 4등분하는 위치에 오는 값
- 제 1사분위수 (Q1): 하위 25%에 해당하는 값
- 제 2사분위수 (Q2): 50%에 해당하는 값 (=중앙값)
- 제 3사분위수 (Q3): 상위 75%에 해당하는 값
- 100등분 한 백분위수 (percentile)로 확장할 수 있음.



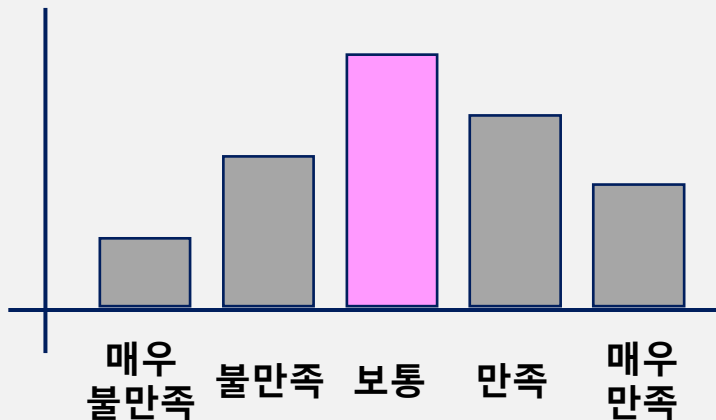
기술 통계량

■ 기술통계량 1. 중심척도

③ 최빈값 (mode)

- 가장 빈번히 나타난 자료값.
- 양적 자료보다는 질적 자료, **명목형 자료**에서 주로 사용된다.
- 분포가 하나의 봉우리를 갖는 형태가 아니고 두 개 (또는 그 이상의) 봉우리 모양으로 흩어진 경우 (이봉분포)에 유용하게 쓰일 수 있다.

[범주형 자료의 예]



[연속형 자료의 예]

0, 1, 2, 12, 12, 14, 18, 21, 21, 23, 24, 25,
28, 29, 30, 30, 30, 33, 36, 44, 45, 47, 51

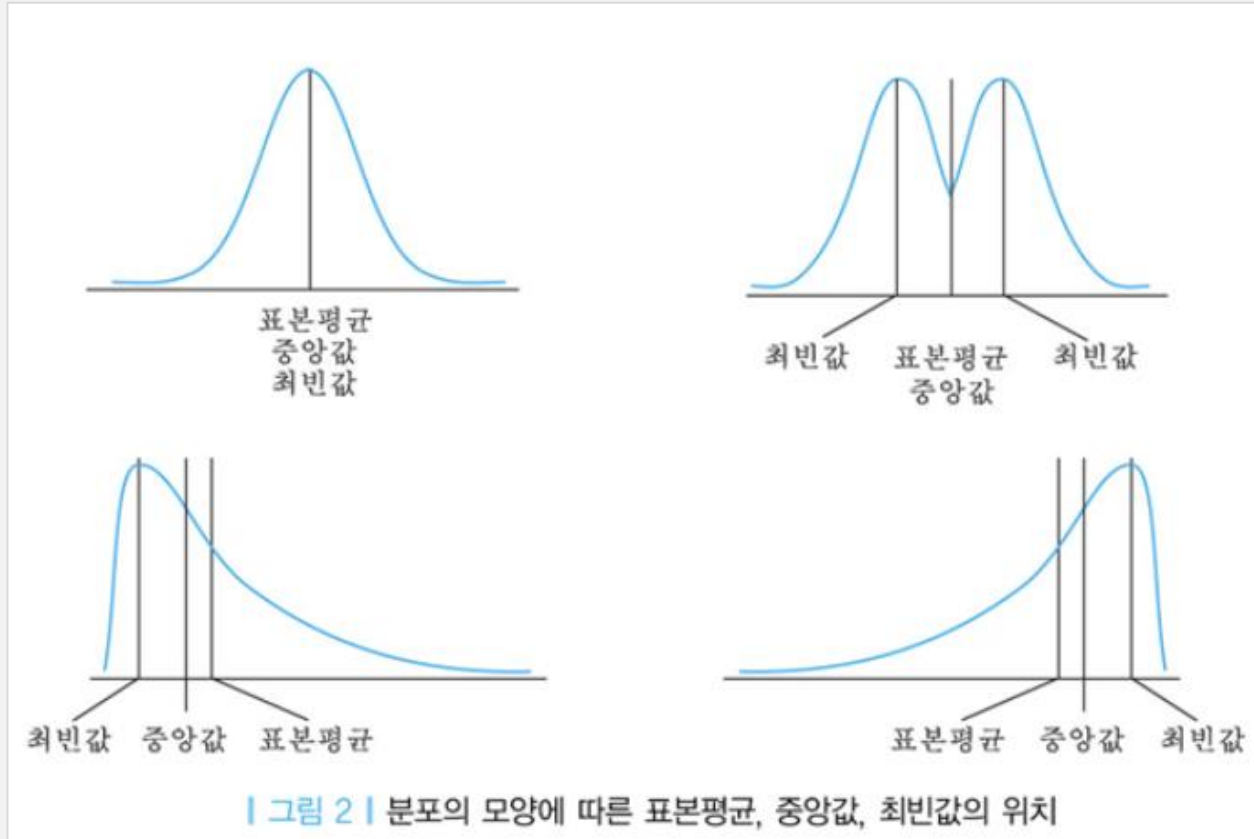


최빈값: 30

기술 통계량

■ 기술통계량 1. 중심척도

- 평균, 중앙값, 최빈값의 비교



기술 통계량

▪ 기술통계량 1. 중심척도

- 예제

컴퓨터 실험실의 23대 소형 컴퓨터의 지난 달 각각 발생한 총 중단시간 (단위: 분)이 아래와 같이 관찰되었다.

0, 1, 2, 12, 12, 14, 18, 21, 21, 23, 24, 25,
28, 29, 30, 30, 30, 33, 36, 44, 45, 47, 51

위의 자료를 R을 이용하여 다음 물음에 답해보자.

Q1. 컴퓨터의 평균 중단시간은 몇 분인가?

Q2. 컴퓨터의 중단시간의 중위수는 몇 분인가?

Q3. 컴퓨터의 중단시간의 최빈값은 몇 분인가?

기술 통계량

▪ 기술통계량 2. 산포의 척도

① 분산과 표준편차

- 분산 (variance)

: 각 자료값들과 평균과의 차이 $x_i - \bar{x}$ 로 산포를 나타낸다. 즉, 평균으로부터 멀리 떨어져 있을수록 $x_i - \bar{x}$ 의 절댓값이 커짐.

표본분산 s^2 은 다음과 같은 식으로 구한다.

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

- 표준편차 (s.d., standard deviation)

: 분산의 제곱근. 분산을 구할 때 제곱을 취함으로써 원래 자료값의 단위가 달라진 것을 복구한 것이다.

표본표준편차 s 은 다음과 같은 식으로 구한다.

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

기술 통계량

기술통계량 2. 산포의 척도

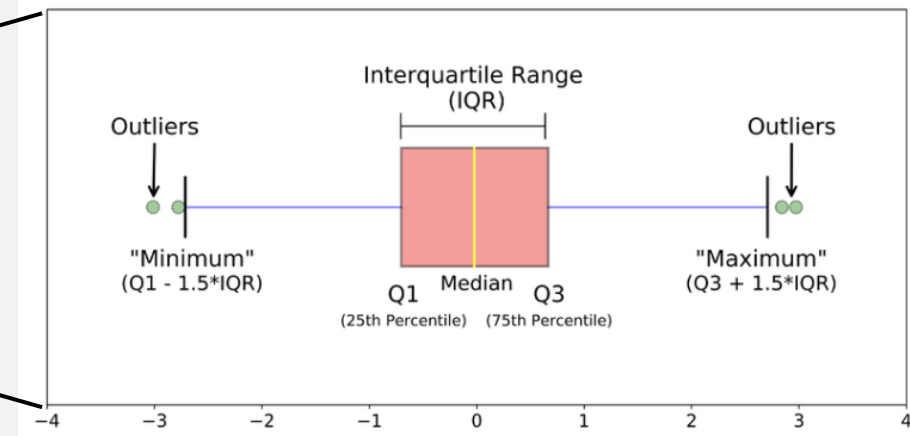
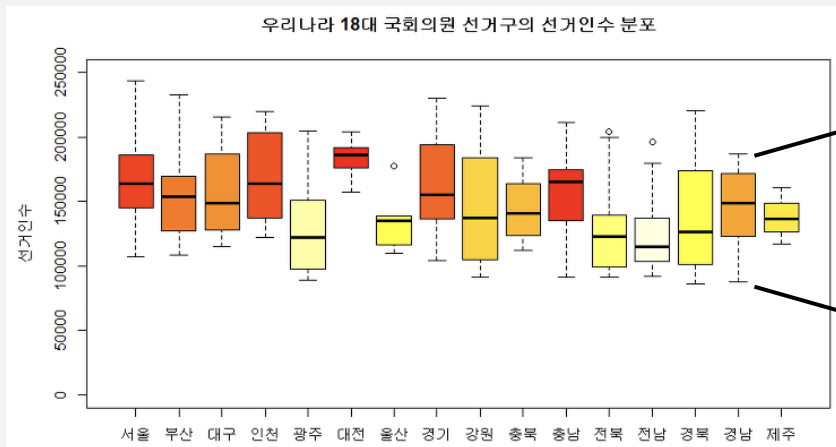
② 범위 (range)

: (최댓값 - 최솟값)의 차이로 간편하게 산포를 계산할 수 있다. 단, 이상치에 대한 영향을 많이 받는다.

- 사분위수 범위 (IQR, interquartile range)

: ($Q3 - Q1$) 즉, 제 3사분위수와 제 1사분위수의 차이로 계산되며, 범위의 극단값의 영향을 받지 않아 많이 사용된다.

상자그림 (box plot) 을 통해 중앙값, 사분위수, 범위를 이용하여 자료를 표현할 수 있다.



기술 통계량

■ 기술통계량 2. 산포의 척도

- 예제

프로그래밍 실습수업을 듣는 40명의 학생 중 여학생은 23명이고 27명은 남학생이다. 첫 학기의 프로그래밍의 시험점수는 다음과 같다.

Female	7, 59, 78, 79, 60, 65, 68, 71, 75, 48, 51, 55, 56, 41, 43, 44, 75, 78, 80, 81, 83, 83, 85
Male	48, 49, 49, 30, 30, 31, 32, 35, 37, 41, 86, 42, 51, 53, 56, 42, 44, 50, 51, 65, 67, 51, 56, 58, 64, 64, 75

전체 점수를 'marks'라는 변수에 저장하고 다음 물음에 답해보자.

Q1. 전체 점수에 대한 **분산**을 구해보자.

Q2. 전체 점수에 대한 **표준편차**를 구해보자.

Q3. 전체 점수에 대한 **범위**를 구해보자.

기술 통계량

▪ 기술통계량 2. 산포의 척도

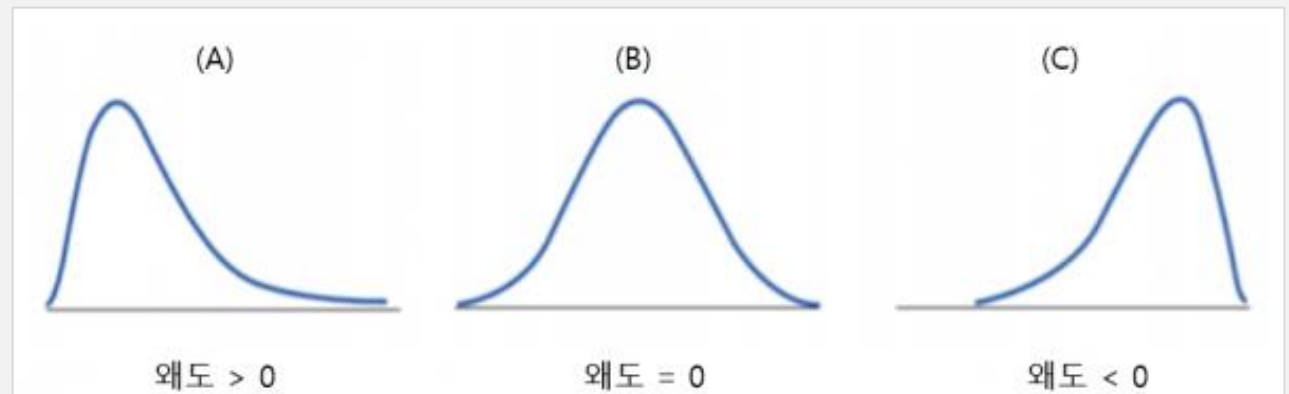
③ 왜도와 첨도

: 자료의 **분포 형태**에 대한 정보를 보여 주는 통계량들.

왜도 (skewness)

- 자료의 분포 형태가 **기울어진** 정도.
- 분포가 좌우 대칭이면 왜도=0, 오른쪽으로 긴 꼬리를 가지면 왜도 >0, 왼쪽으로 긴 꼬리를 가지면 왜도<0.

- 왜도 식 :
$$\frac{\sqrt{n} \sum_{i=1}^n (x_i - \bar{x})^3}{(\sum_{i=1}^n (x_i - \bar{x})^2)^{3/2}}$$



기술 통계량

기술통계량 2. 산포의 척도

③ 왜도와 첨도

: 자료의 **분포 형태**에 대한 정보를 보여 주는 통계량들.

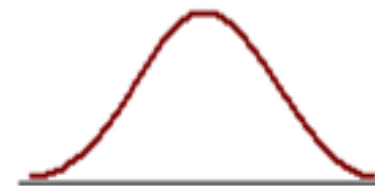
첨도 (kurtosis)

- 분포가 평균 주변에 **몰려 있는** 형태인지 멀리 **퍼져 있는** 형태인지 그 뾰족한 정도.
- 표준정규분포 (첨도=0)를 기준으로, 첨도>0이면 더 뾰족하게 몰려 있고 첨도<0이면 넓게 퍼져 있는 형태.

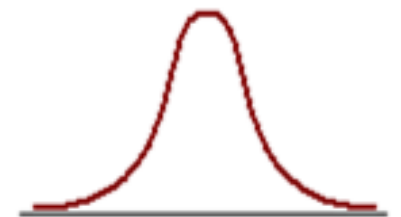
$$\text{첨도 식} : \frac{\sum_{i=1}^n (x_i - \bar{x})^4 / n}{(\sum_{i=1}^n (x_i - \bar{x})^2 / n)^2} - 3$$



첨도가 음수일때



첨도가 0일때



첨도가 양수일때

이변량 자료와 표본상관계수

- 이변량 자료

: 실험이나 관찰의 결과로서 하나의 개체에 대해서 복수 개의 수치가 표본으로 얻어질 수 있다.

이변량 자료란 2개의 수치가 조를 이루어 얻어진 자료를 의미한다.

ex. (엄마의 체중, 신생아의 체중)

(최고혈압, 최저혈압)

(어떤 기간의 강수량, 어떤 작물의 수확고)

(100m달리기 기록, 1500m달리기 기록)

- 3개의 수치가 조를 이루는 자료를 3변량 자료라고 하며,

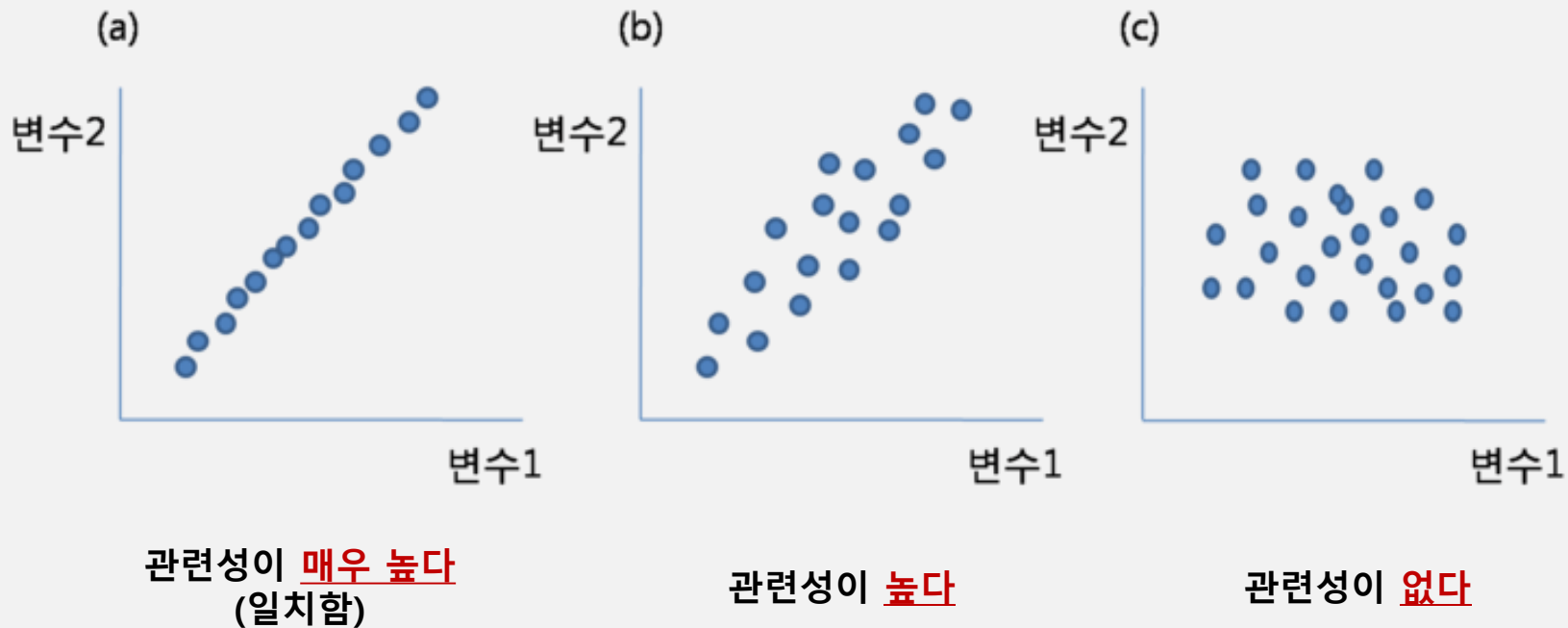
3변량 자료 이상부터는 다변량 자료라고 부른다.

ex. (국어, 영어, 수학) , 사람의 (신장, 체중, 앞은 키, 가슴둘레) , 야구선수의 (타율, 도루, 타점)

이변량 자료와 표본상관계수

- 이변량 자료의 관심사

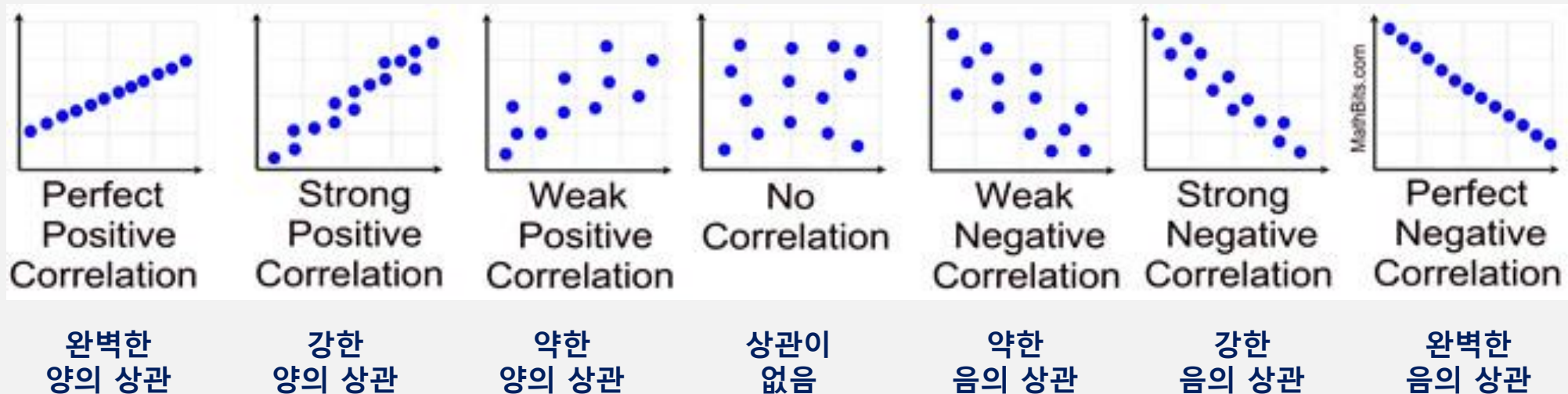
두 개 변량 간 관련성이 있는가? ⇒ 산점도를 그려보자



이변량 자료와 표본상관계수

이변량 자료의 관심사

두 개 변량 간 관련성이 있는가? \Rightarrow 산점도를 그려보자



이변량 자료와 표본상관계수

- 이변량 자료의 관심사: 두 개 변량 간 **관련성**이 있는가?

상관계수 (correlation coefficient)

- 변량 간의 관계의 강함을 보는 척도
- $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ 을 얻어진 표본 (2변량 자료)이라 하자. \bar{x} 와 \bar{y} 를 각각 x 와 y 의 표본평균으로 하였을 때

$$S_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

x 와 y **표본 공분산(sample covariance)**이라고 한다.

- 또 s_x^2 와 s_y^2 을 각각 x 와 y 의 표본분산이라고 하면

$$r = \frac{S_{xy}}{\sqrt{S_x^2 S_y^2}}$$

를 **표본상관계수 (sample correlation coefficient)**라고 한다.

이변량 자료와 표본상관계수

- 이변량 자료의 관심사: 두 개 변량 간 **관련성**이 있는가?

상관계수 (correlation coefficient)

- 표본상관계수 r 는 항상 -1과 1 사이의 값, 즉 $-1 \leq r \leq 1$ 이고, r 이 -1 또는 1에 가까울수록 x 와 y 의 직선관계가 강하다는 것을 나타낸다.
- 또, 한 쪽의 변량이 증가할 때 다른 쪽의 변량도 **증가**하는 것 같은 경향이 있으면 표본상관계수는 **양의 값**을 갖고, 한 쪽이 증가할 때 다른 쪽이 **감소**하는 것 같은 경향이 있으면 표본 상관계수는 **음의 값**을 갖는다.



이변량 자료와 표본상관계수

■ 상관계수 (correlation coefficient)

예제

- 고등학교 A에 대해서 중간시험을 행하였다. x 를 중간시험의 성적, y 를 기말시험의 성적으로 한 것이 아래의 표이다.
다음 물음에 답해보자.

Q1. 중간시험과 기말시험의 산점도를 그려보고

중간시험과 기말시험의 연관성이 있는지 확인해보자.

Q2. 중간시험과 기말시험의 표본상관계수를 계산하고

중간시험과 기말시험의 연관성을 설명해보자.

중간	기말	x	y	x	y	x	y	x	y
49	50	45	11	33	78	40	50	49	31
59	61	52	49	40	52	50	71	46	23
32	50	76	76	78	70	61	61	53	32
66	61	47	36	51	23	47	20	68	47
59	69	52	48	66	62	55	40	48	37
59	38	39	29	59	45	72	59	25	48
68	47	45	56	68	64	42	33	21	10
68	64	63	60	56	72	62	28	47	14
60	82	75	61	31	46	40	28	58	54
68	78	60	61	57	38	66	50	38	52
54	47	75	55	53	33	23	26	53	48
40	47	76	58	59	59	51	56	51	31

4. 그래프

분할표

▪ 분할표 (contingency table)

- 교육수준이 결혼생활에 영향을 미치는지 알아보기 위해 1000명을 조사하였다.
- 각 응답자는 (교육수준, 결혼생활) 쌍의 한 범주를 응답하였다.
- 이와 같이 통계표 형태로 정리된 자료를 분할표라고 한다.

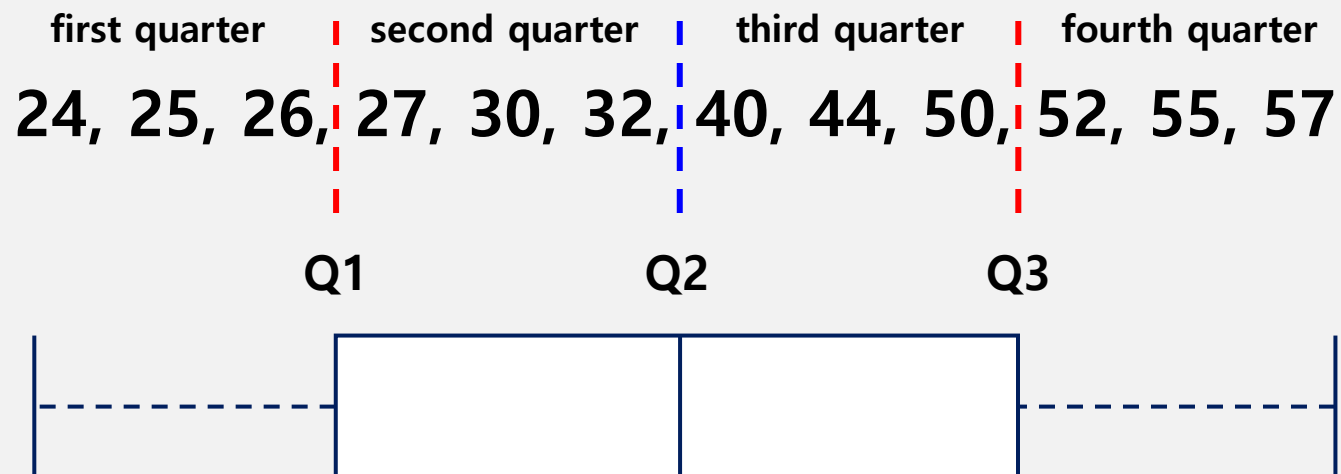
교육수준	결혼생활		
	빈약	원만	대단히 양호
대학	72	112	245
고등학교	65	90	120
중학교	95	103	98

[표] 교육수준과 결혼생활

상자그림

상자그림 (box plot)

- 자료의 분포에 대한 정보를 **사분위수**를 중심으로 나타내는 그림.
- 상자의 밑변과 윗변은 각각 제1사분위수 (Q1)와 제 3사분위수 (Q3)를 나타내고, 중간에 위치한 수평선은 중앙값 (Q2)을 나타낸다.
- 사분위 범위를 벗어난 최댓값 및 최솟값까지 수염 (whisker)라고 불리는 수직선을 점선으로 긋는다.
- 따라서 자료의 25%씩이 4개의 구간 사이에 위치함을 보여준다.



도수분포표와 히스토그램

▪ 도수분포표 (frequency distribution table)

- 아래의 프로그래밍 점수를 도수분포표로 나타내보자

Female	Male
7, 59, 78, 79, 60, 65, 68, 71, 75, 48, 51, 55, 56, 41, 43, 44, 75, 78, 80, 81, 83, 83, 85	48, 49, 49, 30, 30, 31, 32, 35, 37, 41, 86, 42, 51, 53, 56, 42, 44, 50, 51, 65, 67, 51, 56, 58, 64, 64, 75

프로그래밍 점수



1. 관측치의 최댓값과 최솟값의 차이, 즉 범위를 구한다.

$$\Rightarrow 85 - 7 = 78$$

2. 구간을 몇 개로 나눌 것인가?

$$\Rightarrow 10\text{개}$$

3. 구간 폭을 정하자

$$\Rightarrow \text{구간 폭} = (\text{최댓값} - \text{최솟값}) / \text{구간수} = 78 / 10 = 7.8$$

4. 도수와 상대도수, 누적도수, 누적상대도수 등을 산출한다.

도수분포표와 히스토그램

■ 도수분포표 (frequency distribution table)

- 아래의 프로그래밍 점수를 도수분포표로 나타내보자

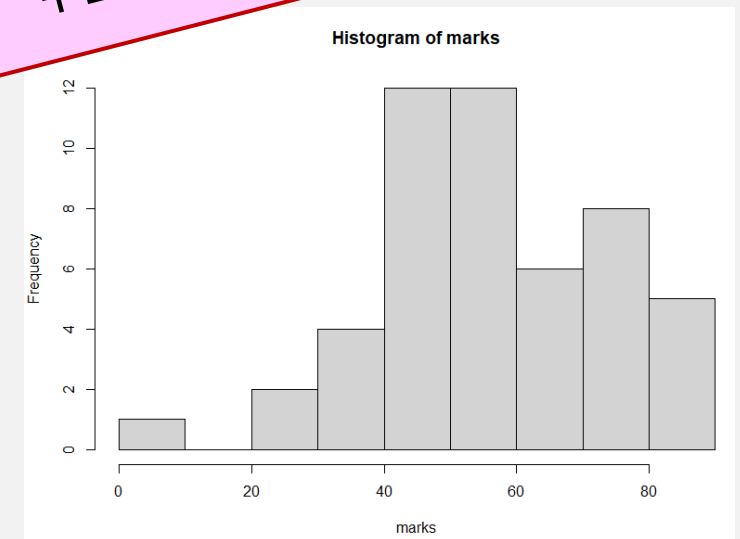
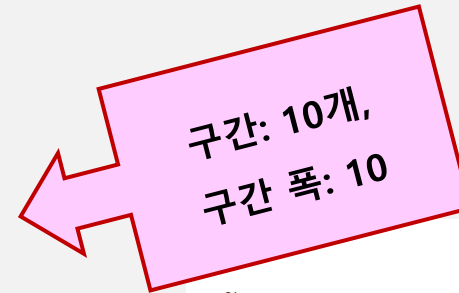
Female	Male
7, 59, 78, 79, 60, 65, 68, 71, 75, 48, 51, 55, 56, 41, 43, 44, 75, 78, 80, 81, 83, 83, 85	48, 49, 49, 30, 30, 31, 32, 35, 37, 41, 86, 42, 51, 53, 56, 42, 44, 50, 51, 65, 67, 51, 56, 58, 64, 64, 75

프로그래밍 점수



점수	학생 수 (명)
(0, 10]	1
(10, 20]	0
(20, 30]	2
(30, 40]	4
(40, 50]	12
(50, 60]	12
(60, 70]	7
(70, 80]	9
(80, 90]	5
(90, 100]	0
계	50

도수분포표



히스토그램

도수분포표와 히스토그램

▪ 도수분포표 (frequency distribution table)

- 많은 관측 값들이 있을 때 그들을 일정한 구간 (계급구간)으로 나누어 각 구간에 속한 자료의 수를 세어 표로 요약한 것.

구간 수의 선정

- 구간 수가 너무 적으면 각 구간에 속하는 도수가 서로 비슷하게 나타날 수 있어 분포상의 특징을 알아내기 어렵다.
- 구간 수가 너무 많으면 한 구간에 포함되는 자료가 하나도 없는 경우가 다수 발생할 수 있다.
- 구간 수 선정에 통일된 기준이 있는 것은 아니다. 주어진 자료에 대해 적절히 선정하면 된다.

자료의 개수	적절한 구간 수
40~100	5~9
100~200	8~12
200 이상	10~16

도수분포표과 히스토그램

- **도수분포표 (frequency distribution table)**
 - 상대도수 (relative frequency): 각 구간의 도수를 자료의 총 수로 나눈 값
 - 누적상대도수: 상대도수의 누적 값

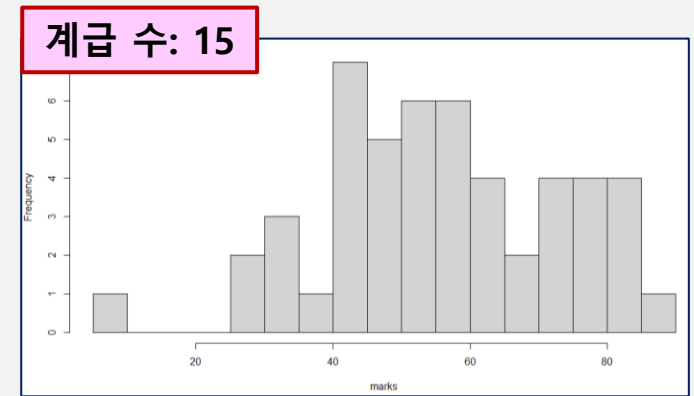
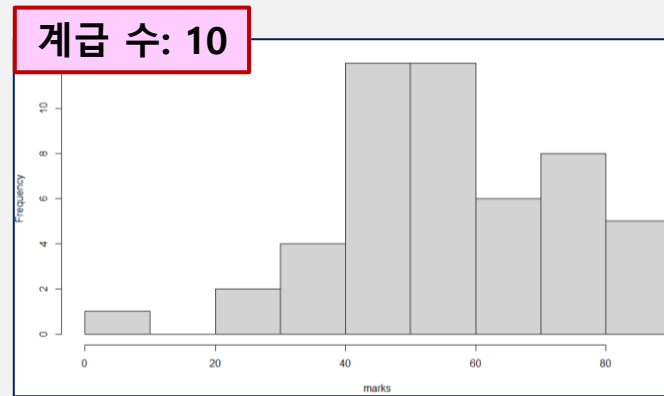
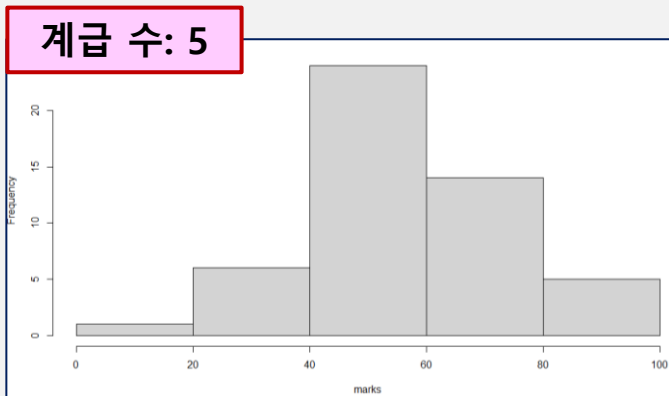
점수	학생 수 (명)	상대도수	상대누적도수
(0, 10]	1	0.02	0.02
(10, 20]	0	0.00	0.02
(20, 30]	2	0.04	0.06
(30, 40]	4	0.08	0.14
(40, 50]	12	0.24	0.38
(50, 60]	12	0.24	0.62
(60, 70]	7	0.12	0.74
(70, 80]	9	0.16	0.90
(80, 90]	5	0.10	1.00
(90, 100]	0	0.00	1.00
계	50	1.00	1.00

도수분포표와 히스토그램

■ 히스토그램 (histogram)

: 도수분포표를 그린 그림

점수	학생 수 (명)	상대도수	상대누적도수
(0, 10]	1	0.02	0.02
(10, 20]	0	0.00	0.02
(20, 30]	2	0.04	0.06
(30, 40]	4	0.08	0.14
(40, 50]	12	0.24	0.38
(50, 60]	12	0.24	0.62
(60, 70]	7	0.12	0.74
(70, 80]	9	0.16	0.90
(80, 90]	5	0.10	1.00
(90, 100]	0	0.00	1.00
계	50	1.00	1.00



도수분포표과 히스토그램

예제

- 남학생과 여학생의 히스토그램을 각각 그리고 점수 분포에 차이가 있는지 확인해보자.

Female	Male
7, 59, 78, 79, 60, 65, 68, 71, 75, 48, 51, 55, 56, 41, 43, 44, 75, 78, 80, 81, 83, 83, 85	48, 49, 49, 30, 30, 31, 32, 35, 37, 41, 86, 42, 51, 53, 56, 42, 44, 50, 51, 65, 67, 51, 56, 58, 64, 64, 75

줄기와 잎 그림

■ 줄기와 잎 그림 (stem-and-leaf plot)

- 히스토그램과 비슷하지만 조금 더 많은 정보를 주는 그림
1. 범위와 구간을 정한다.
 2. 적절히 나눈 구간의 단위를 줄기로 삼고, 구체적인 수치 값을 잎으로 삼아서 줄기에 해당하는 잎을 달아준다.
 3. 각 줄기 내에서 크기순으로 정렬한다.

Female	Male
7, 59, 78, 79, 60, 65, 68, 71, 75, 48, 51, 55, 56, 41, 43, 44, 75, 78, 80, 81, 83, 83, 85	48, 49, 49, 30, 30, 31, 32, 35, 37, 41, 86, 42, 51, 53, 56, 42, 44, 50, 51, 65, 67, 51, 56, 58, 64, 64, 75



줄기	잎
0	7
1	
2	
3	001257
4	11223448899
5	011113566689
6	0445578
7	1555889
8	013356