

강의에 앞서..

통계·데이터과학과 박서영 교수

통계학의 기본 개념

데이터를 통계적으로 분석하고 해석하는 방법

R 소프트웨어를 활용한 실습

강의 목차

1강	데이터와 통계학	박서영
2강	데이터 요약 I	박서영
3강	데이터 요약 II	박서영
4강	확률	이기재
5강	확률변수	이기재
6강	확률분포	이기재
7강	표본분포	이기재

강의 목차

8강	추정 I	이금희
9강	추정 II	이금희
10강	가설검정 I	이금희
11강	가설검정 II	이금희
12강	통계적 비교 I	장영재
13강	통계적 비교 II	장영재
14강	회귀모형 I	장영재
15강	회귀모형 II	장영재

학습 방법

- ▶ 강의 시청, 교재 읽기
- ▶ 교재의 예제와 연습문제 풀어보기
- ▶ 워크북을 이용한 복습

1강

데이터와 통계학

통계·데이터과학과 박서영 교수

목차

- 1 통계학이란
- 2 통계학의 주요 개념
- 3 R과 RStudio 설치 및 시작
- 4 R의 데이터 형태와 연산

01

통계학이란

데이터: 세상을 이해하는 창

- ▶ 어떤 현상을 이해하기 위해 그 현상을 관찰하여 데이터를 수집
- ▶ 전통적인 데이터 수집 방법
 - 관찰, 설문조사, 실험 등



데이터 폭발(Data explosion)

▶ 컴퓨터와 정보통신 기술 발달로 매일 방대한 양의 데이터가 생산됨

- 뉴욕 타임즈가 하루에 실는 정보의 양은 17세기 영국의 평범한 한 사람이 평생 소비하는 정보의 양과 비슷하다

(Wurman, S.A. (1987) "Information Anxiety" New York: Doubleday , p.32)

- 페이스북에서는 하루에 4 페타바이트의 정보가 생성된다 (<https://kinsta.com/blog/facebook-statistics/>, Jan 3, 2021)

1 petabyte = 10^{15} bytes

- ▶ 데이터에서 쓸모있는 정보를 얻기 위한 별도의 과정이 필요
- ▶ **통계학**: 불확실한 현상을 이해하기 위해 데이터를 수집하고, 데이터 패턴을 요약, 분석하여 불확실한 현상에 대한 결론을 찾는 학문

통계학의 역할

- ▶ 데이터의 수집
- ▶ 데이터의 요약
- ▶ 추론

데이터의 수집

- 알고 싶은 현상을 왜곡되지 않게, 잘 반영하는 데이터를 수집하기 위해 통계적 원리를 사용

! 예제: 선거 여론조사

대통령 선거를 앞두고 유권자의 지지성향을 조사하여 선거전략을 세우고자 한다. 전체 유권자의 연령별, 성별 분포를 고려하여 전체를 대표할 수 있는 일부 유권자를 뽑아 조사한다.

데이터의 수집

- 알고 싶은 현상을 왜곡되지 않게, 잘 반영하는 데이터를 수집하기 위해 통계적 원리를 사용

! 예제: 임상시험

특정 감염병 예방을 위해 개발된 백신의 효과를 알아보기 위해, 3만명의 자원자를 모집한 후 랜덤으로 두 그룹으로 나누고, 한 그룹은 백신, 다른 그룹은 플라시보를 투여한다. 3개월 동안 추적 관찰하여 백신의 효과를 증명할 수 있는 데이터를 얻는다.

데이터의 요약

- ▶ 데이터가 가진 특징과 패턴을 정확하고 효과적으로 드러내기 위한 통계적 방법을 사용: **기술통계**

! 예제: 소아의 몸무게

소아의 몸무게를 조사하여 나이별로 몸무게의 평균, 중간값, 사분위수 등 요약통계량을 구한다. 나이에 따른 몸무게의 변화를 보여주기 위해 그래프를 작성한다.

데이터의 요약

- ▶ 데이터가 가진 특징과 패턴을 정확하고 효과적으로 드러내기 위한 통계적 방법을 사용: **기술통계**

- ! 예제: 미세먼지

지역별 미세먼지 농도를 수집하여 지도 위에 미세먼지 농도를 색깔로 표현한다.

- ▶ 데이터를 이용하여 우리의 관심 대상에 대해 추측하고 그 추측의 신뢰성을 계량화: 추측통계 (추론통계)

! 예제: 평균 연봉

대한민국 임금노동자의 평균 연봉을 알아내기 위해
서 랜덤 표집한 300명의 연봉을 조사하여 평균 연봉
추정치와 95% 신뢰구간을 구한다.

추론

- ▶ 데이터를 이용하여 우리의 관심 대상에 대해 추측하고 그 추측의 신뢰성을 계량화: 추측통계 (추론통계)

! 예제: 항암제 효과

새로 개발된 항암제의 효과를 알아보기 위하여 무작위 배정 임상시험에서 관측한 치료군과 대조군의 암 재발률을 비교한다.

데이터

▶ 데이터의 기본요소

- 단위(unit): 관측되는 개별 대상
- 변수(variable): 각 단위에 대해 관측되는 특성
- 관찰값(observation): 각 단위로부터 관측한 특성의 값

▶ 데이터: 하나 이상의 변수에 대한 관찰값의 모음

데이터

! 예제: 4명의 데이터

시연이는 여성이고 키 161cm, 몸무게 50kg이다.

이안이는 남성이고 키 175cm, 몸무게 73kg이다.

연하는 여성이고 키 163cm, 몸무게 55kg이다.

가현이는 여성이고 키 171cm, 몸무게 60kg이다.

➤ 단위: 시연이, 이안이, 연하, 가현이

➤ 변수: 성별, 키, 몸무게

데이터

통계학이란

예제: 4명의 데이터



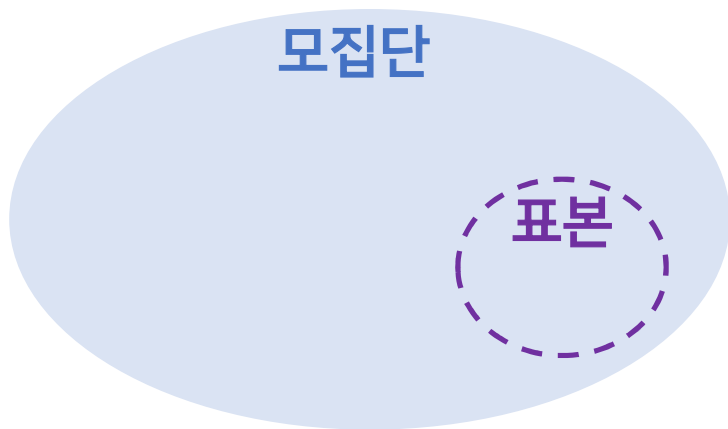
이름	성별	키(cm)	몸무게(kg)
시연이	여	161	50
이안이	남	175	73
연하	여	163	55
가현이	여	171	60

02

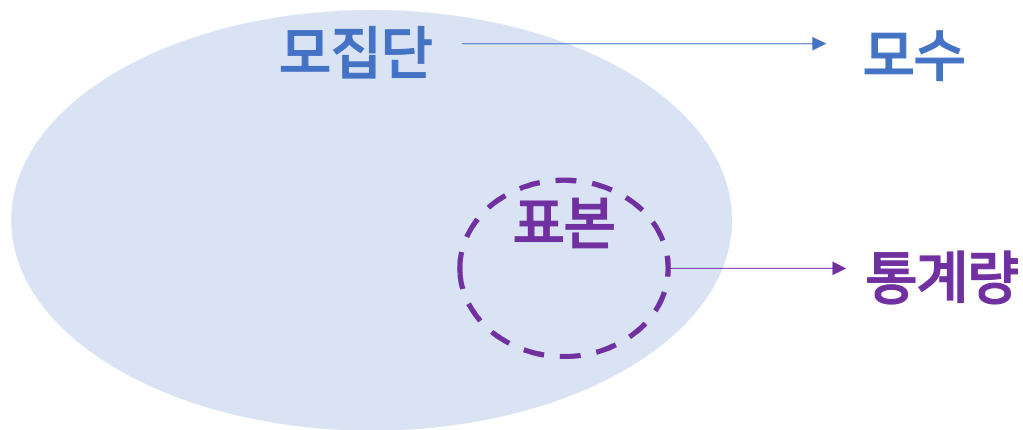
통계학의 주요 개념

모집단과 표본

- ▶ 모집단(population): 관심 대상이 되는 모든 개체의 모임
- ▶ 표본(sample): 모집단을 알기 위해 실제로 관측한 모집단의 일부



- ▶ 모집단(population): 관심 대상이 되는 모든 개체의 모임
- ▶ 표본(sample): 모집단을 알기 위해 실제로 관측한 모집단의 일부
- ▶ 모수(parameter): 모집단의 특성을 나타내는 대푯값
- ▶ 통계량(statistic): 표본의 특성을 나타내는 대푯값



예제: 주거비

▶ 대한민국의 1가구당 평균 주거비를 알아보려고 한다.
전국의 모든 가구의 주거비를 설문하는 것은 너무 많은
시간과 비용이 필요하므로, 랜덤으로 뽑은 1,000가구에
방문하여 주거비를 조사한다.

- 모집단: 대한민국의 모든 가구
- 표본: 랜덤으로 뽑은 1000가구
- 모수: 대한민국의 가구당 평균 주거비
- 통계량: 표본 1000가구의 평균 주거비



모집단과 모수

- ▶ 대부분의 경우 모집단은 너무 커서 모든 개체를 조사할 수 없다
- ▶ 모집단의 종류
 - 유한모집단: 개체 수가 유한개
 - 무한모집단: 개체 수가 무한개
- ▶ 모수:
 - 값이 고정되어있다
 - 대부분의 경우 값을 알 수 없다
 - 예외: 개체 수가 작은 유한모집단인 경우 모든 개체를 조사하면 모수를 알아낼 수 있다

표본과 통계량

- ▶ 모집단을 잘 반영하는 표본을 뽑는 것은 매우 중요하다
- ▶ **단순랜덤표집**(simple random sampling): 유한모집단에서 n 개의 개체로 이루어진 가능한 모든 부분집합이 표본으로 선택될 확률이 같도록 설계된 표본 표집 방법
- ▶ 통계량:
 - 모수를 추정하기 위해 표본에서 얻은 값
 - 표본을 새로 뽑으면 통계량의 값이 달라진다

03

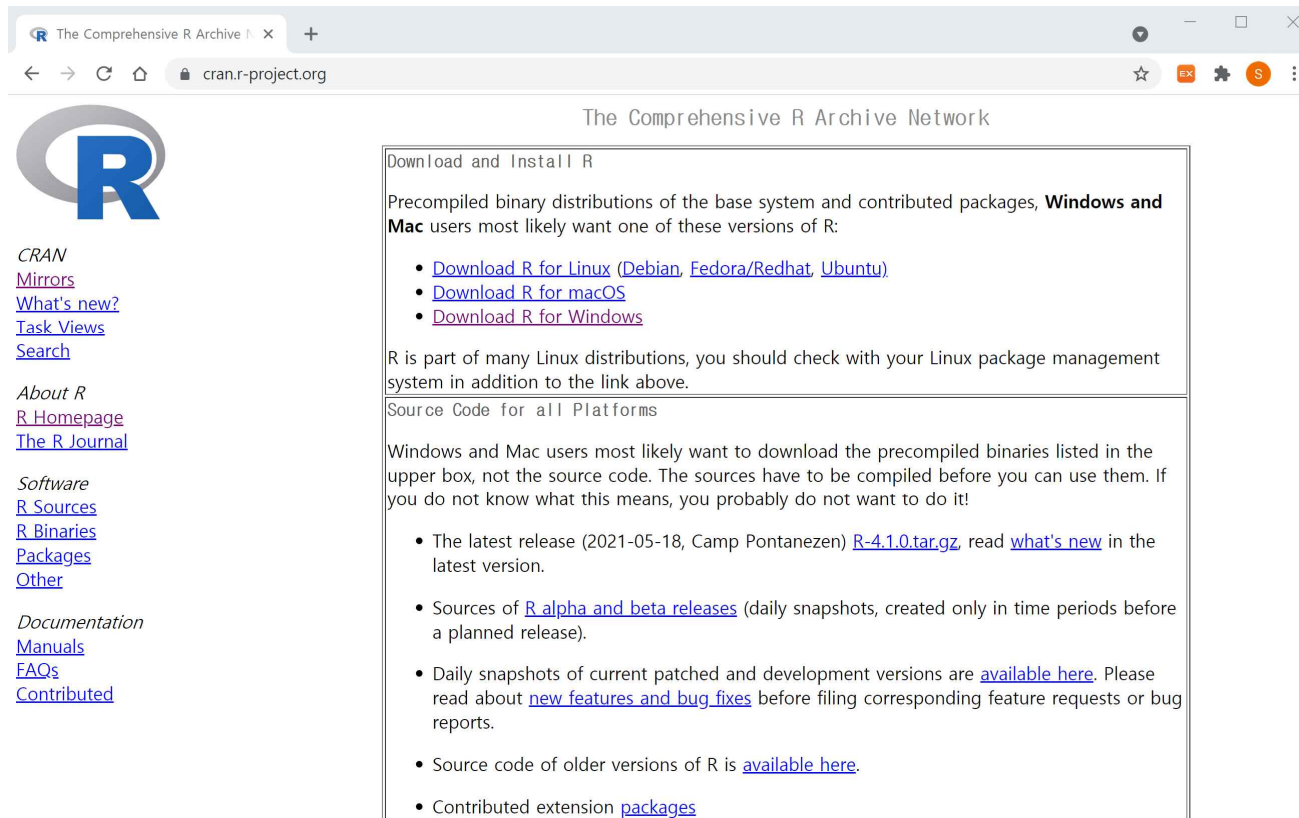
R과 RStudio 설치 및 시작

- ▶ 통계 분석과 그래프 작성에 쓰이는 무료 소프트웨어
- ▶ Windows, MacOS, Linux 등 다양한 컴퓨터 환경에 쉽게 설치, 사용 가능
- ▶ R development core team에 의하여 유지, 개선
- ▶ 누구나 새로운 함수를 개발하여 '패키지' 형태로 공유 가능 → 상업용 소프트웨어에 비해 다양한 분석 가능

R 설치

R과 RStudio 설치 및 시작

▶ CRAN (<https://cran.r-project.org>)에서 다운로드



The screenshot shows the CRAN website with the following content:

The Comprehensive R Archive Network

Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- [Download R for Linux](#) ([Debian](#), [Fedora/Redhat](#), [Ubuntu](#))
- [Download R for macOS](#)
- [Download R for Windows](#)

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

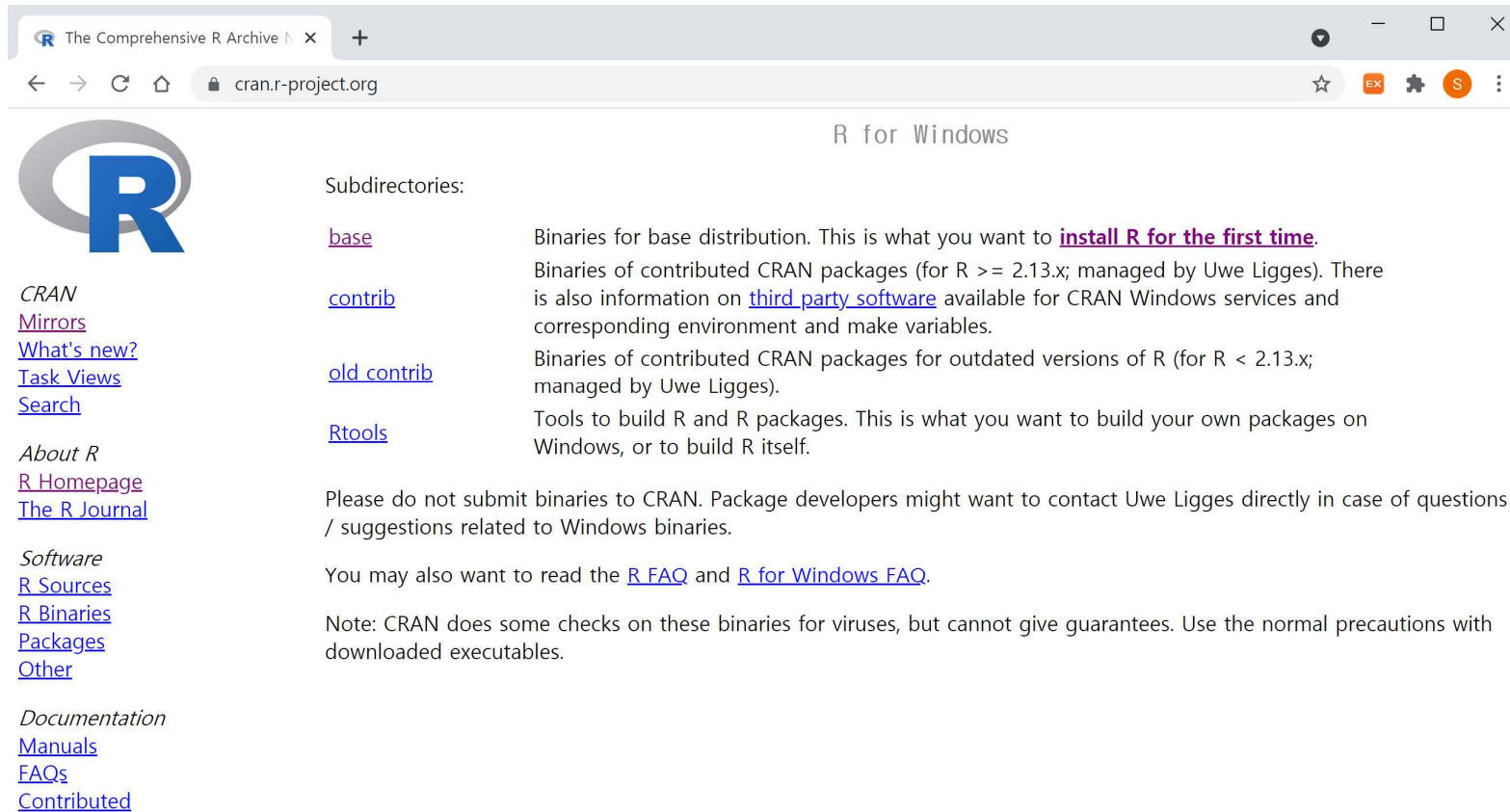
Source Code for all Platforms

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

- The latest release (2021-05-18, Camp Pontanezen) [R-4.1.0.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 and bug fixes](#) before filing corresponding feature requests or bug reports.
- Source code of older versions of R is [available here](#).
- Contributed extension [packages](#)

R 설치

R과 RStudio 설치 및 시작



The screenshot shows a web browser window with the address bar displaying 'cran.r-project.org'. The page title is 'The Comprehensive R Archive'. The main heading is 'R for Windows'. On the left side, there is a large blue 'R' logo and a list of links: 'CRAN', 'Mirrors', 'What's new?', 'Task Views', 'Search', 'About R', 'R Homepage', 'The R Journal', 'Software', 'R Sources', 'R Binaries', 'Packages', 'Other', 'Documentation', 'Manuals', 'FAQs', and 'Contributed'. The main content area is titled 'Subdirectories:' and lists four categories: 'base' (Binaries for base distribution), 'contrib' (Binaries of contributed CRAN packages), 'old contrib' (Binaries of contributed CRAN packages for outdated versions of R), and 'Rtools' (Tools to build R and R packages). Below these, there is a note about not submitting binaries to CRAN and a link to the 'R FAQ' and 'R for Windows FAQ'. A final note states that CRAN does some checks on these binaries for viruses but cannot give guarantees.

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 \geq 2.13.x; managed by Uwe Ligges). There is also information on [third party software](#) available for CRAN Windows services and corresponding environment and make variables.
- [old contrib](#): Binaries of contributed CRAN packages for outdated versions of R (for R $<$ 2.13.x; managed by Uwe Ligges).
- [Rtools](#): Tools to build R and R packages. This is what you want to build your own packages on Windows, or to build R itself.

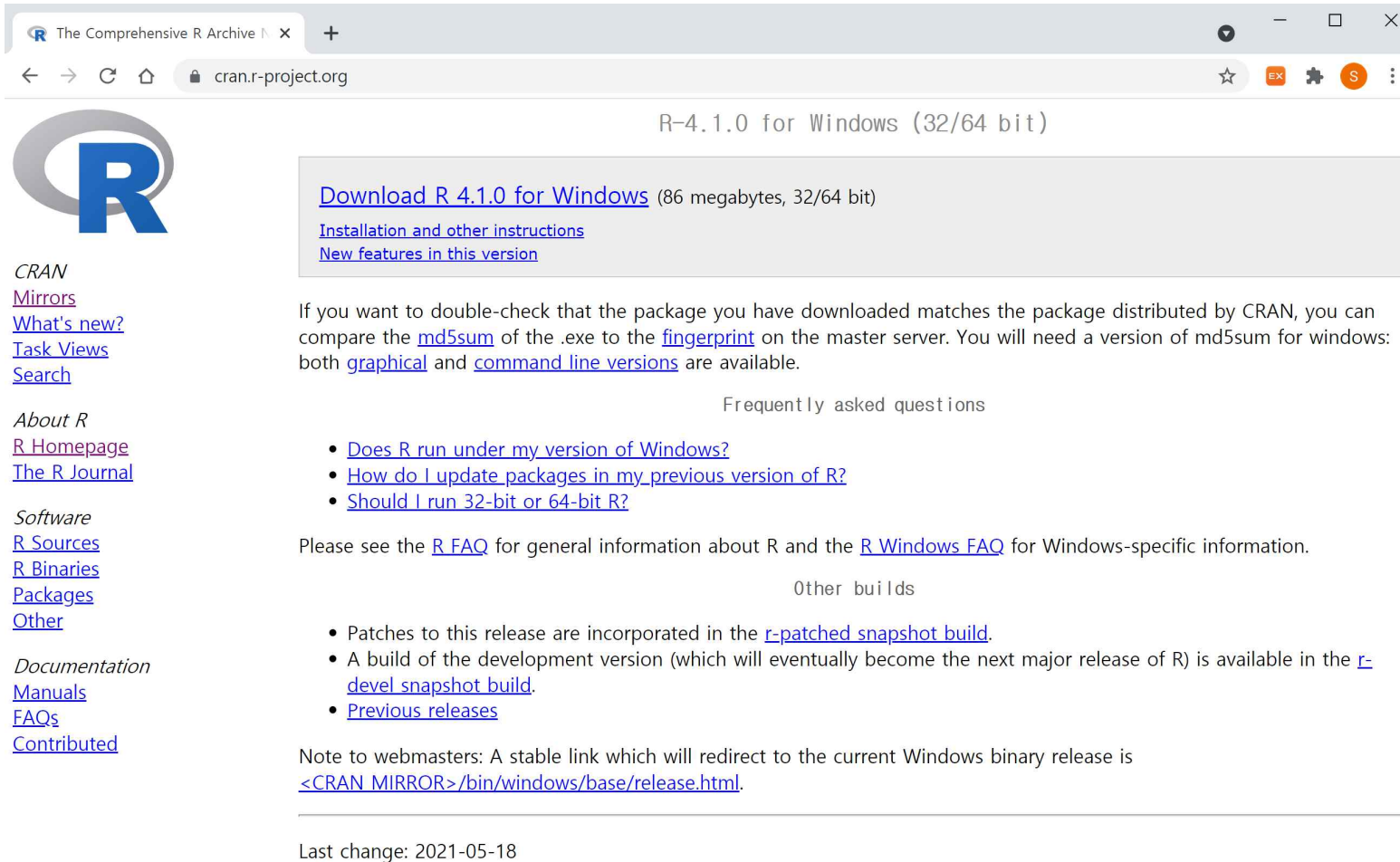
Please do not submit binaries to CRAN. Package developers might want to contact Uwe Ligges directly in case of questions / suggestions related to Windows binaries.

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 downloaded executables.

R 설치

R과 RStudio 설치 및 시작



The screenshot shows the CRAN website for R 4.1.0 for Windows (32/64 bit). The page includes the R logo, navigation links (CRAN, Mirrors, What's new?, Task Views, Search, About R, R Homepage, The R Journal, Software, R Sources, R Binaries, Packages, Other, Documentation, Manuals, FAQs, Contributed), and a section for downloading R 4.1.0 for Windows. The download section provides a link to the Windows installer (86 megabytes, 32/64 bit) and links to installation instructions and new features. Below this, there is a section for frequently asked questions and a note to webmasters about a stable link to the current Windows binary release.

R-4.1.0 for Windows (32/64 bit)

[Download R 4.1.0 for Windows](#) (86 megabytes, 32/64 bit)
[Installation and other instructions](#)
[New features in this version](#)

If you want to double-check that the package you have downloaded matches the package distributed by CRAN, you can compare the [md5sum](#) of the .exe to the [fingerprint](#) on the master server. You will need a version of md5sum for windows: both [graphical](#) and [command line versions](#) are available.

Frequently asked questions

- [Does R run under my version of Windows?](#)
- [How do I update packages in my previous version of R?](#)
- [Should I run 32-bit or 64-bit R?](#)

Please see the [R FAQ](#) for general information about R and the [R Windows FAQ](#) for Windows-specific information.

Other builds

- Patches to this release are incorporated in the [r-patched snapshot build](#).
- A build of the development version (which will eventually become the next major release of R) is available in the [r-devel snapshot build](#).
- [Previous releases](#)

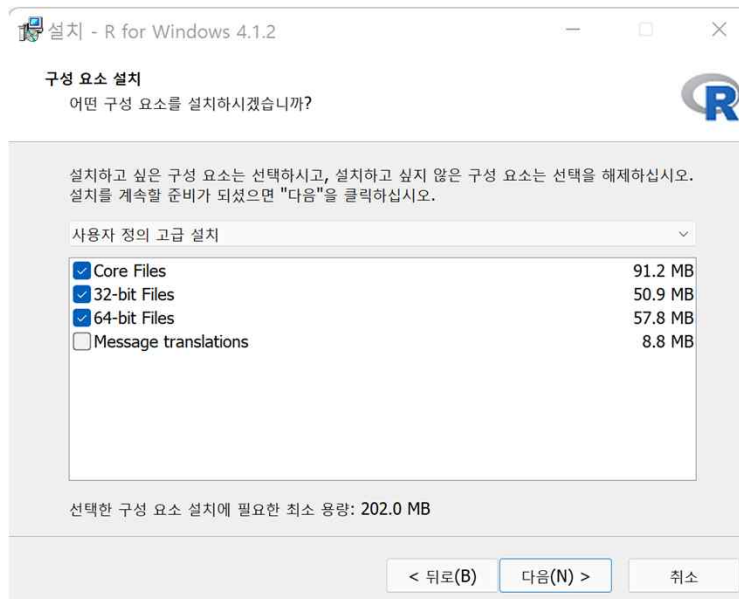
Note to webmasters: A stable link which will redirect to the current Windows binary release is [<CRAN MIRROR>/bin/windows/base/release.html](#).

Last change: 2021-05-18

R 설치

R과 RStudio 설치 및 시작

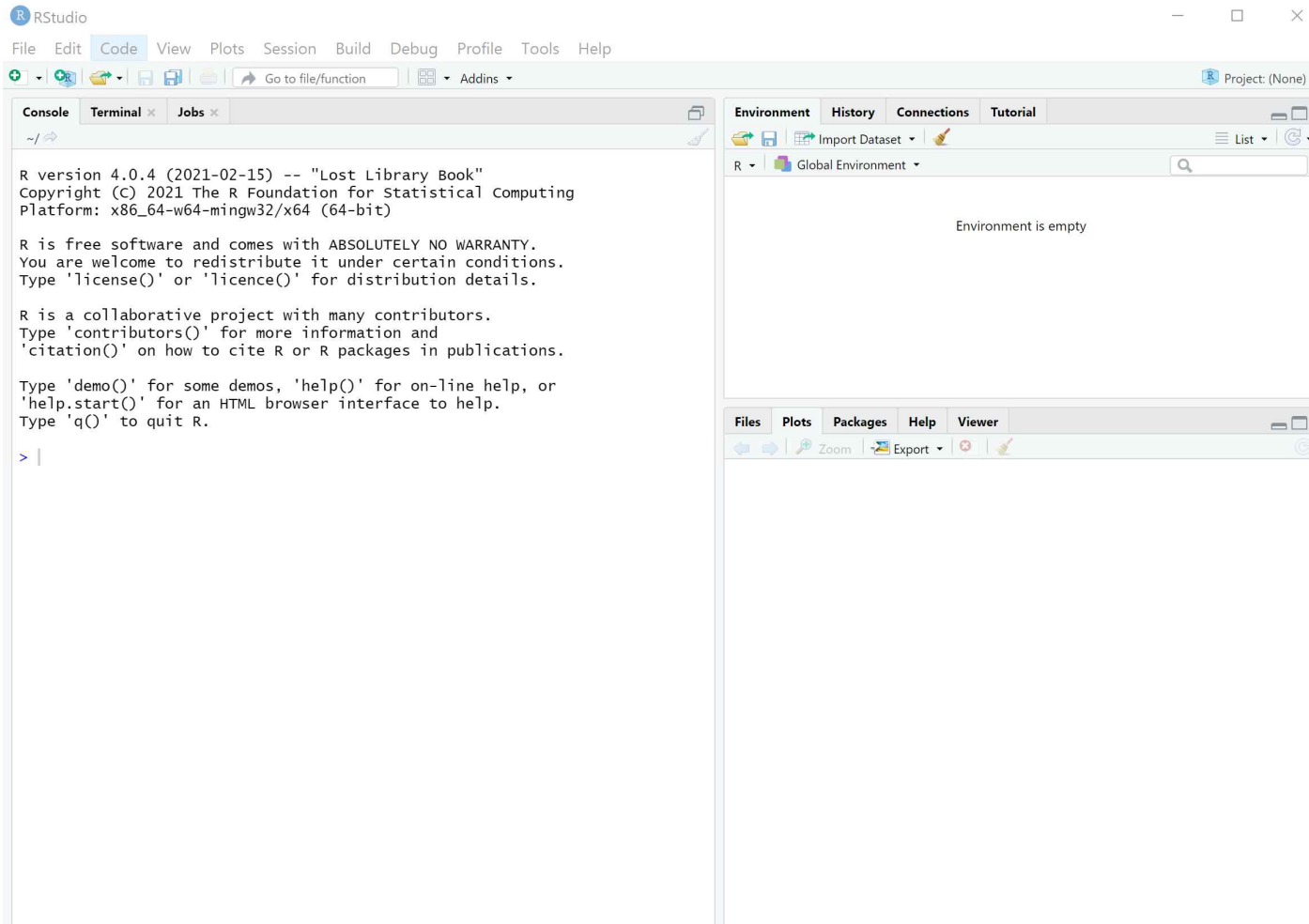
- 다운로드 받은 파일을 실행하여 설치 마법사 시작
- '구성 요소 설치' 단계에서 'Message translations' 체크박스를 해제하는 것을 추천
 - R의 출력언어가 한글 대신 영어가 된다
 - 오류메시지가 영어로 되어있으면 검색을 통해 해결책을 찾기가 훨씬 쉽다



- R을 편리하게 이용할 수 있게 해주는 편집기
- R을 설치한 후 <https://www.rstudio.com> 에서
Products>RStudio>RStudio Desktop 선택
- 다운로드 받은 파일을 실행하여 설치
- R을 실제로 이용할 때는 RStudio를 열면 된다

RStudio

R과 RStudio 설치 및 시작



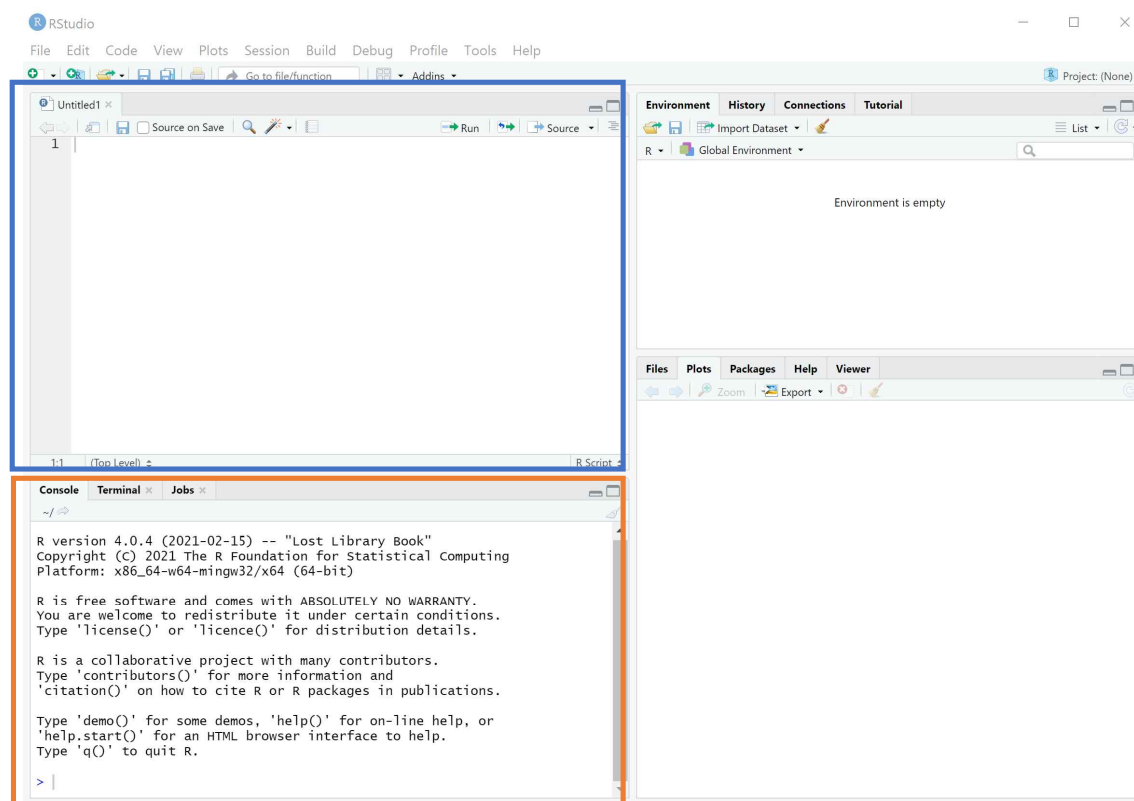
RStudio

R과 RStudio 설치 및 시작

- File>New File>R Script를 클릭해서 스크립트 창을 열어 사용하는 것이 편리하다

스크립트 창

콘솔 창



▶ 명령어를 실행하는 법

■ 콘솔 창을 이용하는 방법:

- 콘솔 창에 직접 입력하고 Enter를 누른다

■ 스크립트 창을 이용하는 방법:

- 1줄만 실행할 경우: 실행하려는 명령어에 커서를 놓은 후 후 Run 버튼을 클릭하거나 단축키 Ctrl+Enter를 누른다
- 여러줄 또는 1줄 전체가 아닌 일부분만 실행할 경우: 실행하려는 명령어를 마우스로 선택한 후 Run 버튼을 클릭하거나 단축키 Ctrl+Enter를 누른다

작업 디렉토리

- ▶ 작업 디렉토리 지정: 데이터 파일을 읽어들이거나 내보낼 때, 일일이 위치를 지정하지 않아도 되는 디폴트 위치를 설정하는 것
- ▶ `setwd()` 함수 안에 큰따옴표를 씌운 경로를 넣는다
- ▶ 경로의 각 단계를 구분할 때 일반적으로 쓰는 역슬래시(\) 대신 슬래시(/) 또는 두개의 역슬래시(\\)를 써야한다
- ▶ 예:

```
setwd("C:\\Users\\KNOU_stat\\R_exercise")  
setwd("C:/Users/KNOU_stat/R_exercise")
```
- ▶ 작업 디렉토리는 RStudio를 종료하면 해제된다

04

R의 데이터 형태와 연산

객체의 생성과 저장

객체 이름 <- 저장하고 싶은 값

```
a<-1  
b<-1  
c<-a+b  
c
```

벡터(vector)

- ▶ 벡터: 어떤 요소(값)들이 일렬로 늘어선 것
- ▶ 벡터를 만드는 법
 - `c()` 함수 안에 벡터의 각 요소를 쉼표로 구분하여 넣는다
 - `seq()` 등의 함수를 이용한다

```
height <- c(165, 151, 162, 160, 151, 152, 159, 163, 143, 161)

d<-1:3
e<-seq(1, 9, 2)
f<-rep(10, 5)
g<-c(d, f)
h<-c(4:1, seq(0, 9, 3))
```

벡터의 연산

▶ 벡터들 간 사칙연산 가능

- 벡터의 길이가 같은 경우: 각 벡터에서 같은 위치에 있는 숫자끼리 연산
- 벡터의 길이가 다른 경우: 길이가 짧은 벡터의 각 요소를 앞에서부터 재활용하면서 연산 (경고 메시지 출력)

```
e+f  
e-f  
e*f  
e/f  
d+f
```

```
## Warning in d + f: longer object length is not a  
## multiple of shorter object length
```

데이터형

R의 데이터 형태와 연산

- 숫자형: 사칙연산 가능
- 범주형: factor() 또는 as.factor() 이용하여 생성
- 문자형: as.character() 이용하여 생성
- 논리형: TRUE 또는 FALSE 값을 가진다

```
i<-1:4
j<-as.factor(1:4)
i+1
j+1
## Warning in Ops.factor(j, 1): '+' not meaningful for factors
k<-as.character(1:4)
l<-c("K", "N", "O", "U")
m<-i>2
```

행렬(matrix)

- ▶ 벡터 여러개의 모임
- ▶ 행렬의 요소들은 데이터형이 모두 같아야한다
- ▶ cbind(), rbind(), matrix() 함수 이용해서 생성

```
n<-rep(10, 5)
o<-1:5
p<-cbind(n, o)
q<-rbind(n, o)
r<-matrix(1:4, 2, 2)
s<-matrix(c(1, 4, 2, 7), 2, 2)
r+s
r %*% s
solve(s)
s[1,2]
s[1,]
s[,2]
```


데이터 프레임(data frame)

- ▶ 행렬과 비슷하나 데이터형이 다른 벡터들도 하나의 데이터 프레임에 저장 가능
- ▶ data.frame() 함수 이용해서 생성

```
name<-c("Kim", "Lee", "Park", "Choi")  
age<-c(20, 32, 17, 51)  
sex<-as.factor(c("Male", "Female", "Female", "Female"))
```

```
dat<-data.frame(name, age, sex)
```

```
dat$age  
dat$name  
dat$sex
```

정리하기

- 통계학이란 불확실한 현상을 이해하기 위해 데이터를 수집하고, 데이터 패턴을 요약, 분석하여 불확실한 현상에 대한 결론을 찾는 학문이다.
- 통계학의 역할에는 데이터의 수집, 데이터의 요약, 추론이 있다.
- 데이터는 하나 이상의 변수에 대한 관찰값의 모음이다. 데이터에서 관측되는 개별 대상을 단위라 하고, 각 단위에 대해 관측되는 특성은 변수라고 한다.
- 관심 대상이 되는 모든 개체의 모임을 모집단이라 하고, 모집단을 알기 위해 실제로 관측한 모집단의 일부를 표본이라고 한다. 모집단을 잘 대표하는 표본을 표집하는 방법 중 가장 기본이 되는 방법은 단순랜덤표집이다.
- 모수는 우리가 알고 싶은 모집단의 특성을 나타내는 대푯값이고, 모수를 알기 위해 표집한 표본의 특성을 나타내는 대푯값을 통계량이라고 한다.

2강

다음시간안내

데이터 요약 I