R 프로그래밍 #1

2019.03.05

한국생명공학연구원 김하성

2019학기 UST R 프로그래밍과 데이터 분석

- 장소: 한국생명공학연구원 연구동 세미나실 1213호 (매주수요일 13:00~16:00)
- 강사: 한국생명공학연구원 바이오합성연구센터 김하성
- 연락처: 042-860-4372, haseong@kribb.re.kr (생명연 연구동 1143)

강의 목표

- 이공계열 대학원생이 통계 분석용 범용 프로그램 언어인 R의 사용법을 익히고 이를 이용한 프로그래밍 기술의 습득을 주요 목적으로 함. 생물 데이터 위주의 분석 예제를 통해 R의 활용 기술을 배움.
- 이 강좌는 오프라인 강의를 기본으로 함. R 사용에 대한 기본 개념과 프로그래밍을 중심으로 필요시 기초 통계지식 강의. 실습 데이터로는 주로 생물데이터를 사용함. 모든 강의 시간 실습을 위한 개인 노트북 지참 필수. 강의 자료는 pdf로 제작 하여 수업 진행과 관련된 정보와 함께 웹사이트를 통해 배포됨. 상황에 따라 강의 일정이 조정될 수 있음.

참고 자료

- Using R for Introductory Statistics by John Verzani
 - o Free version of 1st Edition
 - https://cran.r-project.org/doc/contrib/Verzani-SimpleR.pdf
 - http://cbb.sjtu.edu.cn/~mywu/bi217/usingR.pdf
 - Second edition
 - https://www.crcpress.com/Using-R-for-Introductory-Statistics-Second-Edition/Verzani/p/book/9781466590731
- R for Data Science (https://r4ds.had.co.nz)
- Bioinformatics Data Skills by Vince Buffalo
- 일반통계학 (영지문화사, 김우철 외)

평가 세부 항목

• 출석 50% / 과제 50% / 수업참여 +alpha

https://github.com/greendaygh/2019-R-programming

R is a platform and language for statistical computing

Similar to the S language and environment which was developed at Bell Laboratories



R provides a wide variety of **statistical** (linear and nonlinear modelling, classical statistical tests, timeseries analysis, classification, clustering, ...) and **graphical** techniques, and is highly extensible.

Getting started with R installation

- R site https://www.r-project.org
- Windows, Mac and Linux versions available
- Executed using command line, or a graphical user interface (GUI)

The R Project for Statistical Computing

Getting Started

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To **download R**, please choose your preferred CRAN mirror.

If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.



Korea

Japan

https://ftp.harukasan.org/CRAN/

https://cran.seoul.go.kr/

http://healthstat.snu.ac.kr/CRAN/

https://cran.biodisk.org/

Malaysia

https://wbc.upm.edu.my/cran/http://wbc.upm.edu.my/cran/



CRAN
Mirrors
What's new?
Task Views
Search

About R R Homepage The R Journal

R Sources
R Binaries
Packages
Other

The Comprehensive R Archive Network

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
- Download R for (Mac) OS X

Download and Install R

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 (Friday 2017-06-30, Single Candle) R-3.4.1.tar.gz, read what's new in the latest version.



R for Windows

Subdirectories:

base

contrib

Binaries for base distribution (managed by Duncan Murdoch). This is what you want to <u>install</u> R for the first time.

Binaries of contributed CRAN packages (for $R \ge 2.11.x$; managed by Uwe Ligges). There is also information on third party software available for CRAN Windows services and

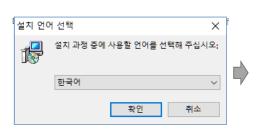


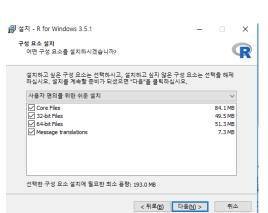
R-3.5.2 for Windows (32/64 bit)

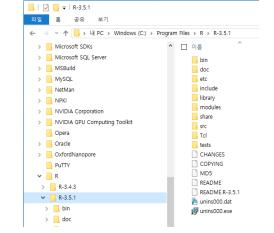
Download R 3.5.2 for Windows (79 megabytes, 32/64 bit)

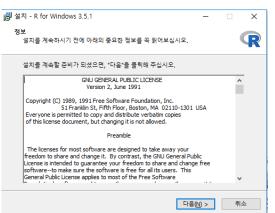
<u>Installation and other instructions</u> New features in this version

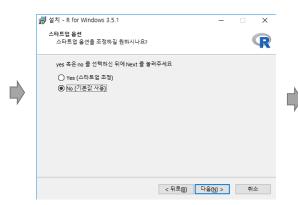
installation

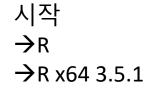


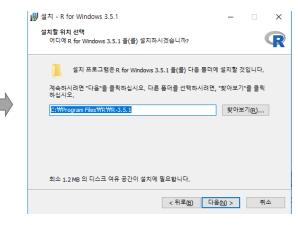




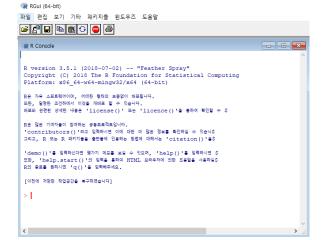








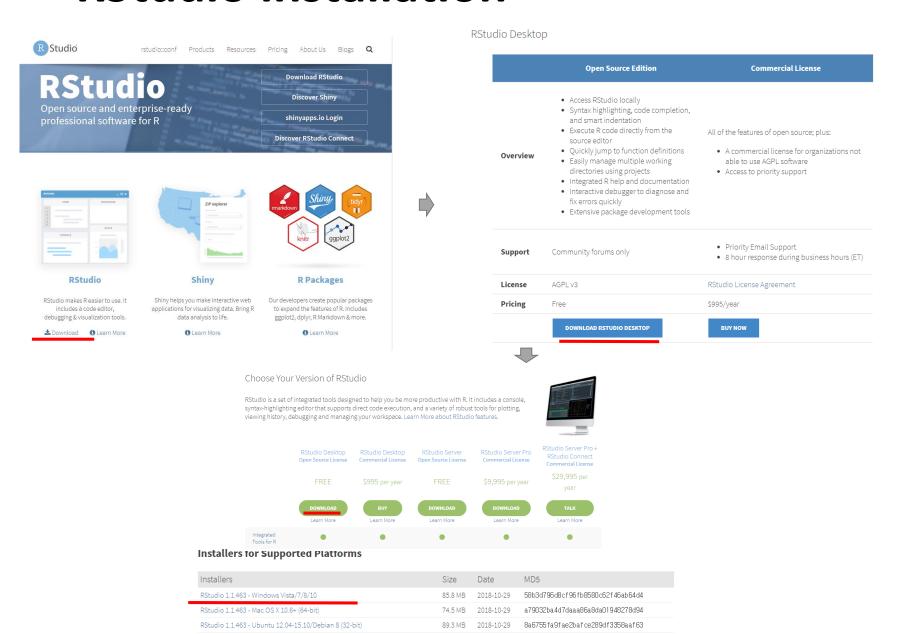




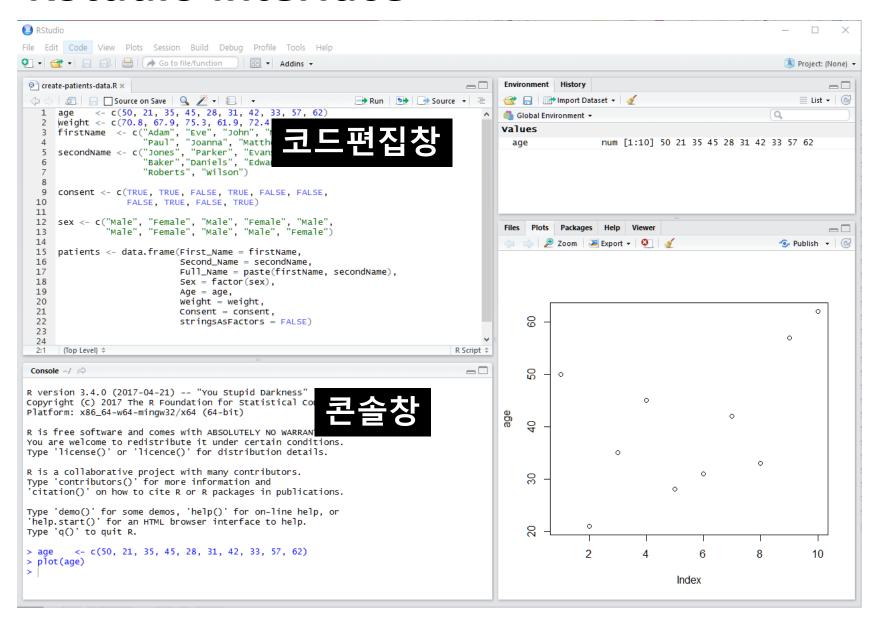


RStudio installation

https://www.rstudio.com/

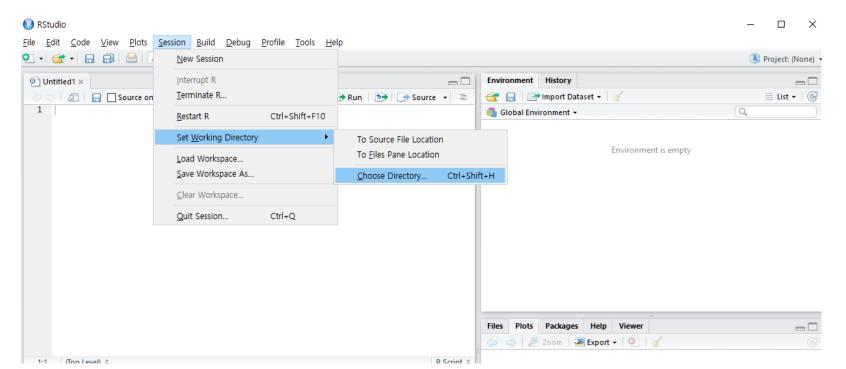


RStudio interface



Set working directory

```
File → New file → R script
getwd()
dir()
## make a directory of "Rwork"
setwd("C:\Users\haseong\Documents\Rwork")
getwd()
dir()
## make a new R file, "Rprog01.R"
```



R's command line (Console)

R can be used like a calculator

```
>2 + 2
>((2 - 1)^2 + (1 - 3)^2)^(1/2)
>1 / 200 * 30
>2 + 2; 2 - 2
```

- Combining commands with;
- Prompt
- Command history (up-arrow / down-arrow keys)
- Errors
 - 2^^2

Object

- R is a programming language for statistical computing
- Assignment operator (<-)

Valid object name <- value

- keyboard shortcut: Alt + (the minus sign)
- Variable

```
x <- 2
y <- x^2 - 2*x + 1
y
x <- "two"
some_data <- 9.8
X <- 2
pi</pre>
```

Variable type of (storage) mode

Туре	Explanation
285	Numeric (Integer)
34.67	
4.23E-4	
TRUE, T	Logical
FALSE, F	
'B'	Character
"Hello" or 'Hello'	
NULL	NULL

Name and Case sensitivity

- Characters (letters), numbers, "_", "."
- A and a are different symbols
- Names are effectively unlimited in length

```
i_use_snake_case
otherPeopleUseCamelCase
some.people.use.periods
And_aFew.People_RENOUNCEconvention
```

Logical & Character vectors

The elements of a logical vector can have the values TRUE, FALSE, and NA (for "not available")

```
> temp <- 2 > 13

> x <- c(1:10)

> temp <- x > 5

> temp <- x > 2 & x < 7

> temp <- x > 2 | x < 7
```

Character strings with double (") or single (') quotes

```
x <- c("a", "b", "c")
?paste
paste("test", "string", sep="")
paste(c("a", "b"), 1, sep="-")</pre>
```

```
Q1: > x <- c("X1", "Y2", "X3", "Y4", "X5", "Y6", "X7", "Y8", "X9", "Y10")
> x
> "X1" "Y2" "X3" "Y4" "X5" "Y6" "X7" "Y8" "X9" "Y10"
```

Generate the same x using paste() and seq()

Object - Vectors

```
> x <- c(10.4, 5.6, 3.1, 6.4, 21.7)
```

This is an assignment statement using the function c() which in this context can take an arbitrary number of vector arguments and whose value is a vector got by concatenating its arguments end to end

```
> assign("x", c(10.4, 5.6, 3.1, 6.4, 21.7))
> y <- c(x, 0, x)
```

Vector arithmetic

```
> v <- 2*x + y + 1
> sum((x - mean(x))^2) / (length(x)-1)
```

Generating regular sequences

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

> x <- c(1:10)

> ?seq

> x <- seq(1, 10, 1)

> x <- seq(1, 10, 2)
```

```
> ?rep
> x <- rep(c(1,2,3), times=3)
> x <- rep(c(1,2,3), each=5)
> x <- rep(rep(1:3, each=5), times=3)
```

Index vectors & subsets of a vector

Subsets of the elements of a vector can be selected by appending to the name of the vector an *index vector* in square brackets.

```
> x <- c("X1", "Y2", "X3", "Y4", "X5", "Y6", "X7", "Y8", "X9", "Y10")
> x[1]
> "X1"
```

A vector of negative integral quantities

```
> x <- c("X1", "Y2", "X3", "Y4", "X5", "Y6", "X7", "Y8", "X9", "Y10")
> x[-1]
> x[-c(1:5)]
```

A vector of character strings

```
> fruit <- c(5, 10, 1, 20)
> names(fruit) <- c("orange", "banana", "apple", "peach")
> lunch <- fruit[c("apple", "orange")]</pre>
```

Logical vectors

```
> x <- c(1:10)
> temp <- x > 2 & x < 7
> x[temp]
```

Basic concepts in R - functions

Define a function

```
my_sine <- function(x){
     y <- sin(x)
     return(y)
}</pre>
```

Use

```
> my_sine(pi)
```

- This returns the sine of pi
 - One parameter: x
 - One argument: pi

Arguments

• Default arguments, named arguments

```
mean(x, trim=0.5)
mean(x)
```

Built-in Functions

- called by their names followed by ()
- arguments separated by commas
- Tab completion
- "combine function" creates a data set
- "mean" takes several numbers and summarizes with one

```
x <- pi
sin(x)
sqrt(x)
log(x)
log(x, 10)
x <- c(74, 122, 235, 111, 292)
mean(x)
sum(x)/length(x)</pre>
```

Generic functions

- Same name but different function
- Choice depends on the class of the first argument
- confusion at first but no need to remember all the names of functions

```
x <- c(74, 122, 235, 111, 292)
y <- c(T, F, T, T)
summary(x)
summary(y)</pre>
```

Vectorized functions

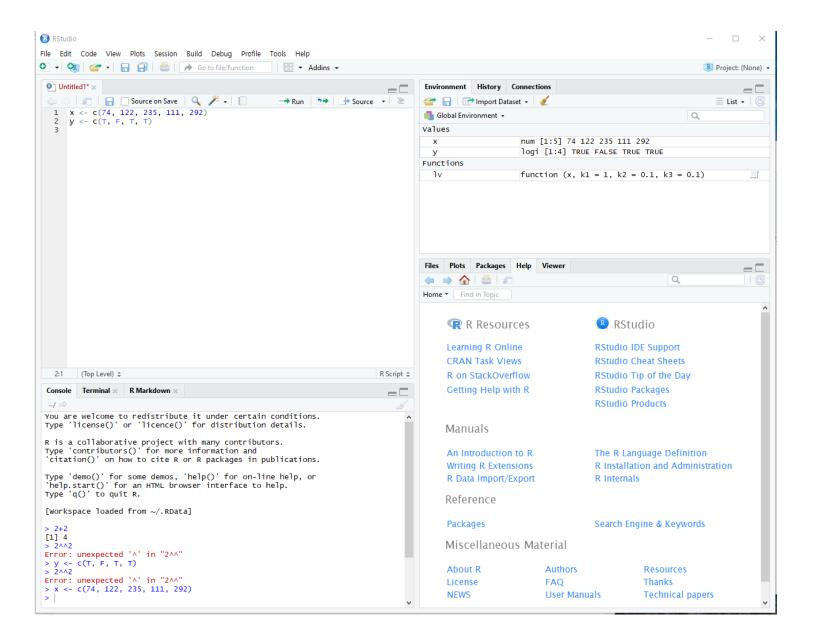
Do the same thing for each number of a vector

```
x <- seq(1, 10)
x + x
sqrt(x)
x - mean(x)</pre>
```

Keyboard shortcut

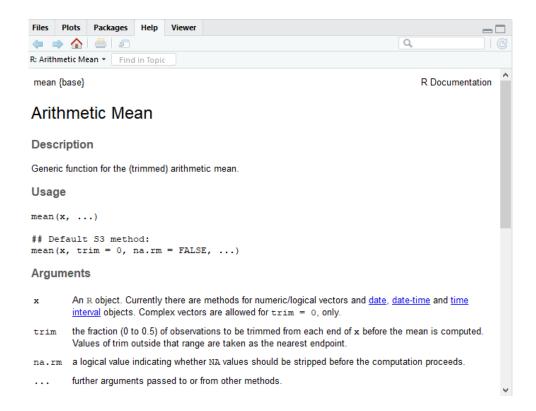
- https://support.rstudio.com/hc/enus/articles/200711853-Keyboard-Shortcuts
- Tools → Keyboard shortcut Quick Reference (Alt + Shift + K)
- 코드편집창 이동 (Ctrl+1) 콘솔창 이동(Ctrl+2)
- 한 줄 실행 (Ctrl+Enter)
- 주석처리 (Ctrl + Shift + C)
 - Starting with a hashmark ('#'), everything to the end of the line is a comment
 - > # this is a comment line
 - > this is a comment line

Environment



Help

```
>?mean
>help("mean")
>example("mean")
>help.search("mean")
>help(package="MASS")
```



R packages

R comes ready loaded with various libraries of functions called packages ex) sum() is in the "base" package and sd() in the "stats" package

The packages can be found in numerous server locations on the web called **repositories**

The Comprehensive R Archive Network (CRAN)

http://cran.r-project.org/web/views/

Bioconductor specialised in genomics http://www.bioconductor.org/

To find functionality

Metacran http://www.r-pkg.org/

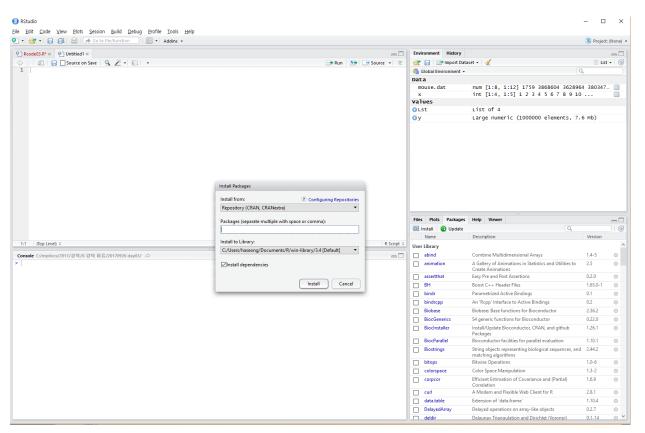
Installing packages

The Comprehensive R Archive Network (CRAN)

install.packages(name.of.package)

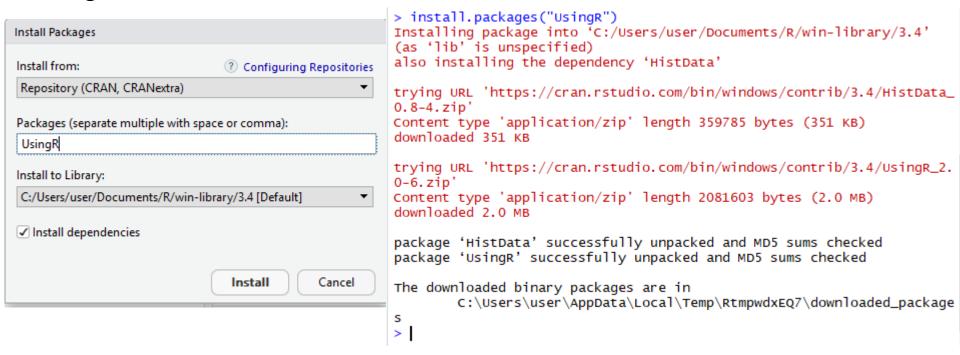
Bioconductor

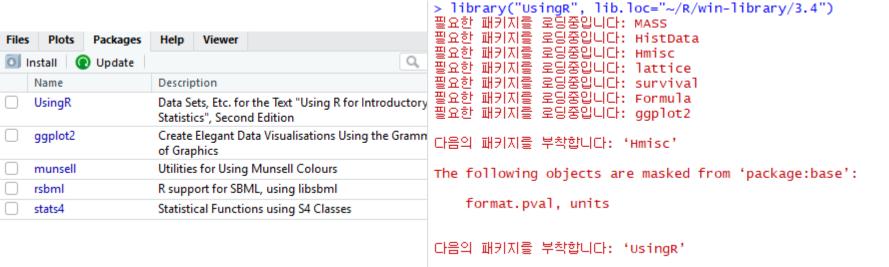
if (!requireNamespace("BiocManager"))
 install.packages("BiocManager")
BiocManager::install()



Toggling checkbox Load / Unload

Packages → Install





Packages loading

library(readxl)

```
> library(UsingR)

다음의 패키지를 부착합니다: 'UsingR'

The following object is masked from 'package:survival':

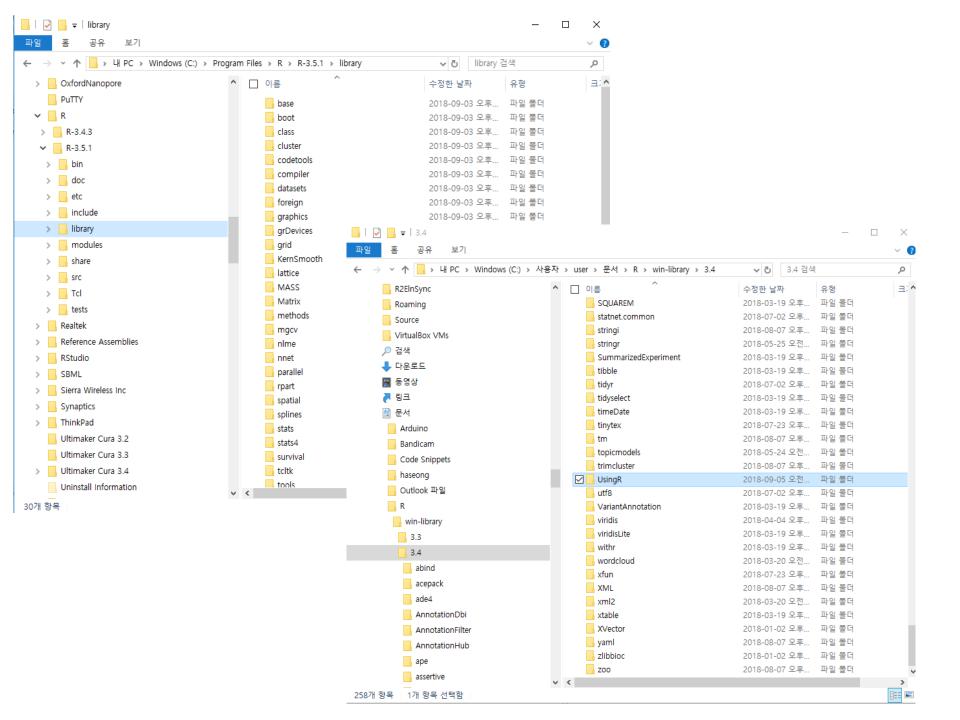
cancer

Warning message:
패키지 'UsingR'는 R 버전 3.4.4에서 작성되었습니다
> |
```

```
R 설치 디렉토리
R 패키지 설치 디렉토리
```

```
.libPaths()
path.package()
```

```
> path.package()
[1] "C:/Program Files/R/R-3.4.0/library/stats"
                                                   "C:/Program Files/R/R-3.4.0/library/graphics"
[3] "C:/Program Files/R/R-3.4.0/library/grDevices" "C:/Program Files/R/R-3.4.0/library/utils"
[5] "C:/Program Files/R/R-3.4.0/library/datasets"
                                                   "C:/Program Files/R/R-3.4.0/library/methods"
[7] "C:/PROGRA~1/R/R-34~1.0/library/base"
> library(readx1)
Warning message:
패키지 'readx1'는 R 버전 3.4.1에서 작성되었습니다
> path.package()
[1] "C:/Users/haseong/Documents/R/win-library/3.4/readxl" "C:/Program Files/R/R-3.4.0/library/stats"
[3] "C:/Program Files/R/R-3.4.0/library/graphics"
                                                          "C:/Program Files/R/R-3.4.0/library/grDevices"
[5] "C:/Program Files/R/R-3.4.0/library/utils"
                                                          "C:/Program Files/R/R-3.4.0/library/datasets"
[7] "C:/Program Files/R/R-3.4.0/library/methods"
                                                          "C:/PROGRA~1/R/R-34~1.0/library/base"
```



교재 및 참고자료

- R 기본 문서들 (소개, 사용, 설치, 운영)
 - https://cran.r-project.org/doc/manuals/r-release/R-intro.html
 - https://cran.r-project.org/doc/manuals/r-release/R-data.html
 - https://cran.r-project.org/doc/manuals/r-release/R-admin.html

• 참고 사이트

- R home (https://www.r-project.org/)
- Packages for biologists (http://www.bioconductor.org/)
- R studio (<u>https://www.rstudio.com/</u>)
- R-bloggers (https://www.r-bloggers.com/)
- Revolutionanalytics (http://blog.revolutionanalytics.com/)
- R 레퍼런스카드 (https://cran.r-project.org/doc/contrib/Short-refcard.pdf)

• 튜토리얼

- https://www.codeschool.com/courses/try-r
- https://www.youtube.com/playlist?list=PLFe6bHk0eK-AaHbHnbwLhrhE1Gyc5UOkE
- https://www.bioconductor.org/help/course-materials/2017/OMRF/B2_Common_Operations.html
- http://r4ds.had.co.nz/

- http://tutorials.iq.harvard.edu/R/Rintro/Rintro.html
- https://web.stanford.edu/class/bios221/labs/
- https://www.stat.berkeley.edu/~s133/

Next

- Get KRIBB email ID / Password
- R basics I
 - Matrix, Data.frame, List
 - Building functions