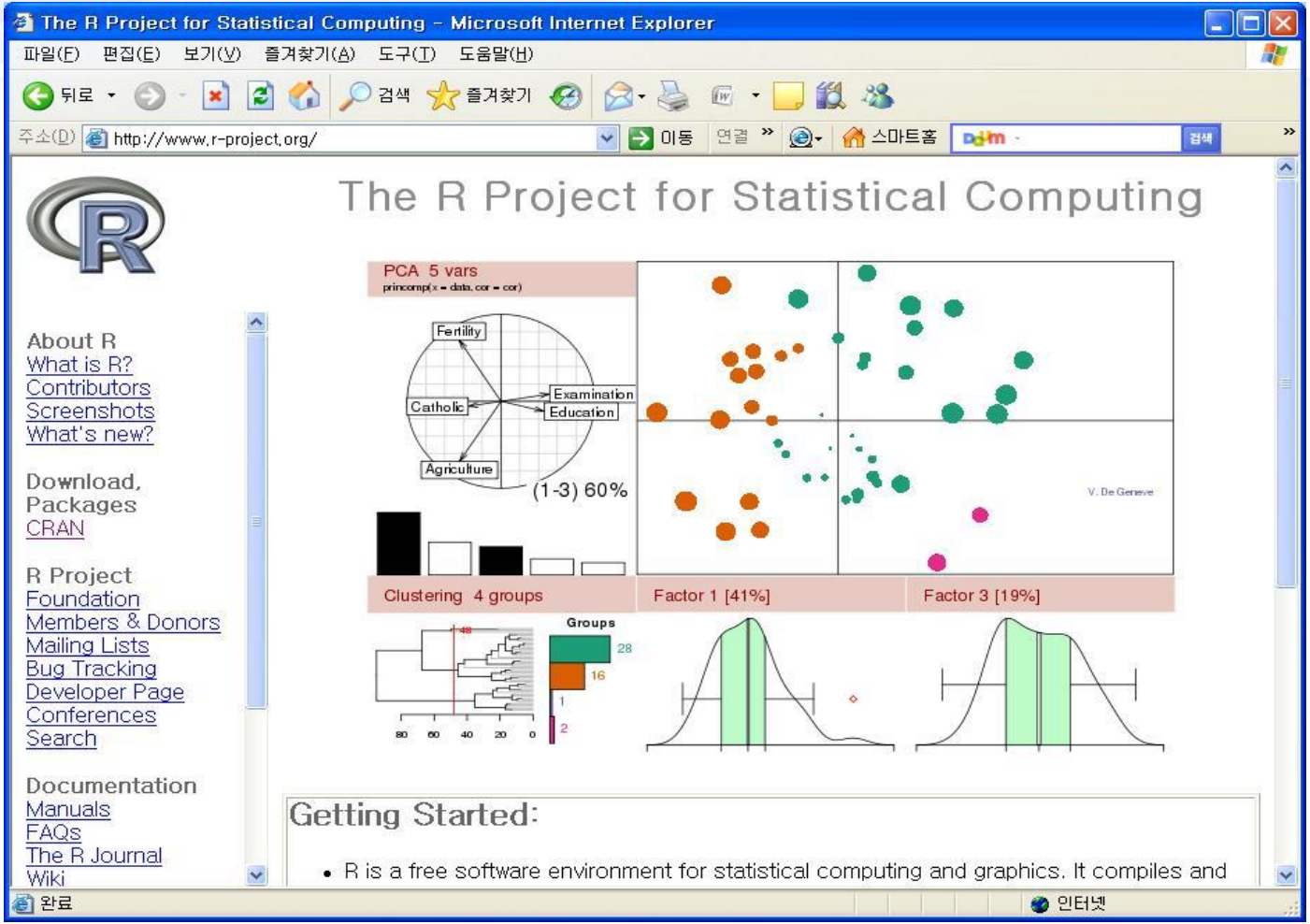
R 사용안내



이 화면을 R실습하기 타이틀 화면으로 넣어주셔요.

1. R 사용법

1 R의 소개

- R : 인터넷에서 자유롭게 다운받아 사용할 수 있는 통계시스템
- R은 프로그램의 기능을 가지고 있어 자체로서도 유용한 기능들을 구현할 수 있을 뿐만 아니라 자료 처리 및 그래픽스 분야에 탁월한 기능을 가지고 있음

R 사이트

- http://www.r-project.org
- 프로그램 및 매뉴얼 등을 무료로 다운

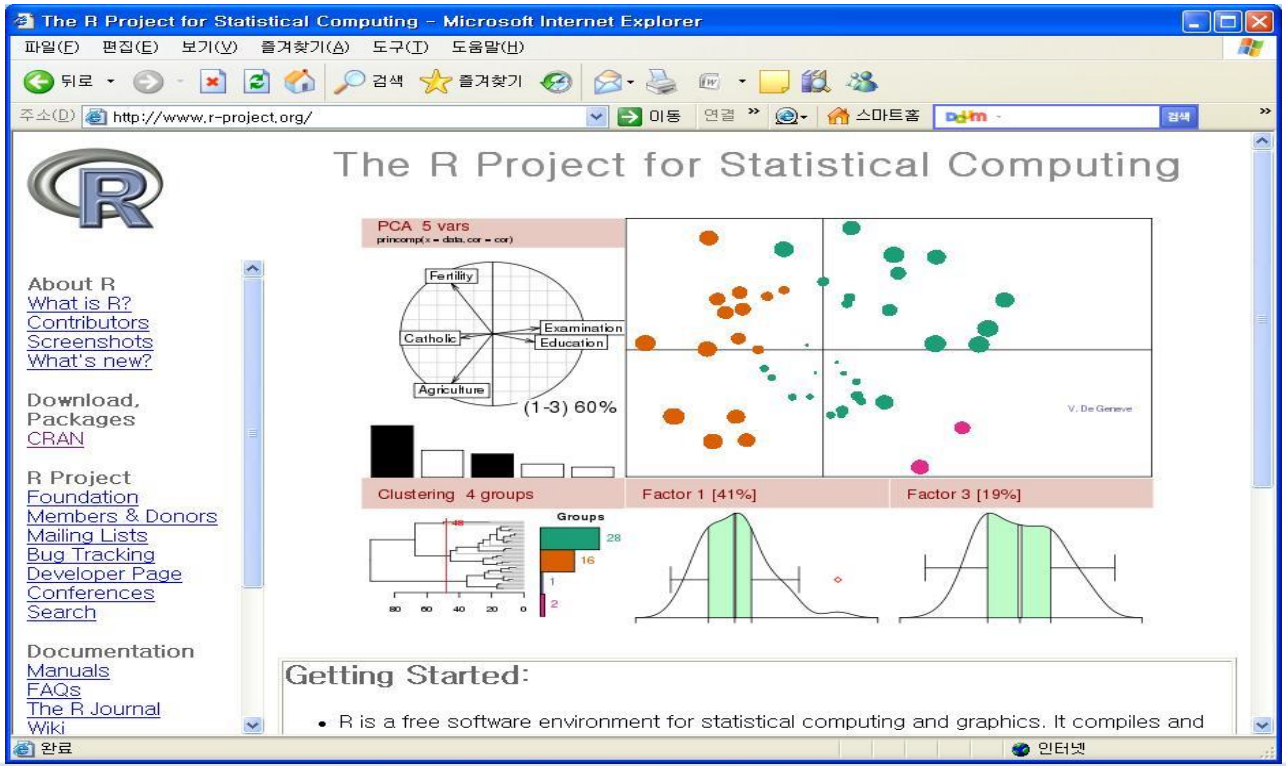
2 R의 기능

- 자료처리기능
  - 자료의 구성, 소팅, 결합 등이 프로그램 처리로 쉽게 이루어짐
- 자료분석 기능
  - 자료를 분석하기 위해 필요한 수치적 계산 및 분석 결과 제공 등이 다양
- 언어기능
  - 자 대화형으로 프로그래밍언어로서 함수문을 쉽게 작성할 수 있으며, C언어 및 FORTRAN 언어 등과 인터페이스 (Interface)가 가능
- 그래픽스
  - 대화형 그래픽스에 의한 자료분석의 기능 및 분석결과의 그래픽스 처리 기능이 뛰어남

3 R 다운받기

◆ R 사이트 초기 화면에서 왼편 창의 CRAN을 선택

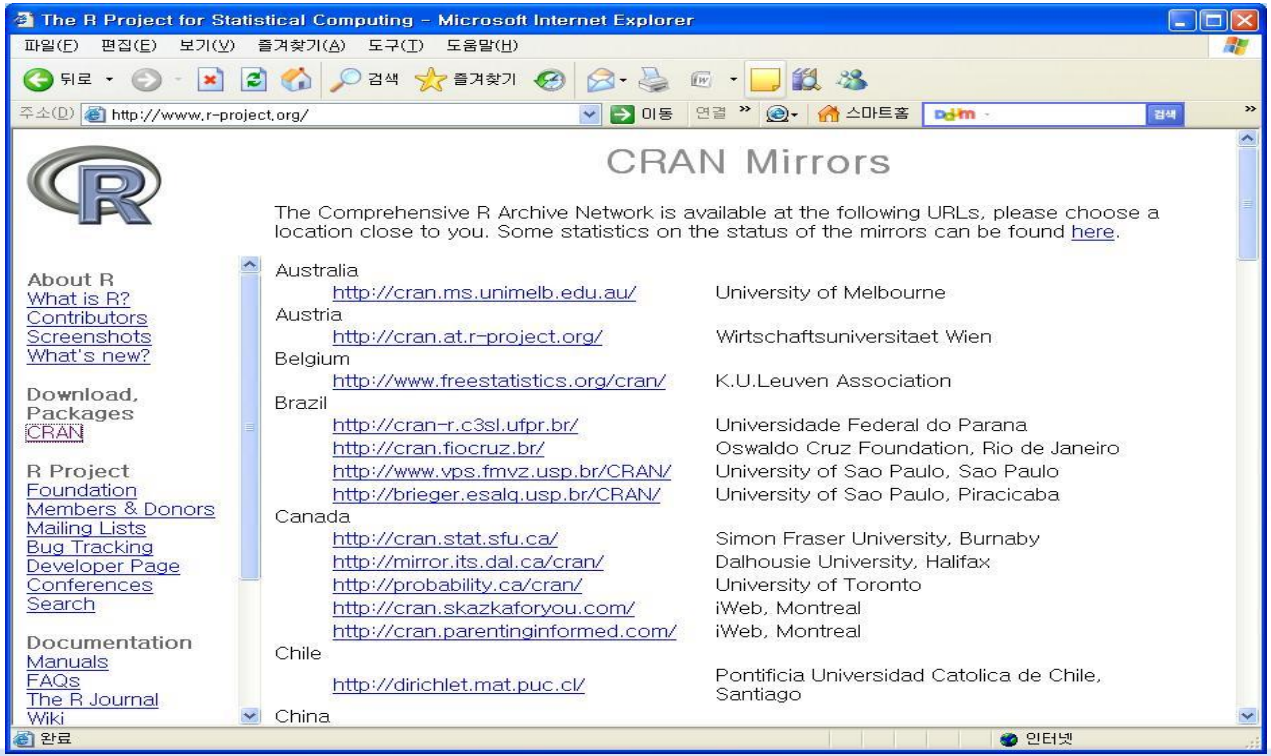
R 사이트(www.r-project.org)



3 R 다운받기

◆ R 다운받기 위한 Mirror사이트 선택

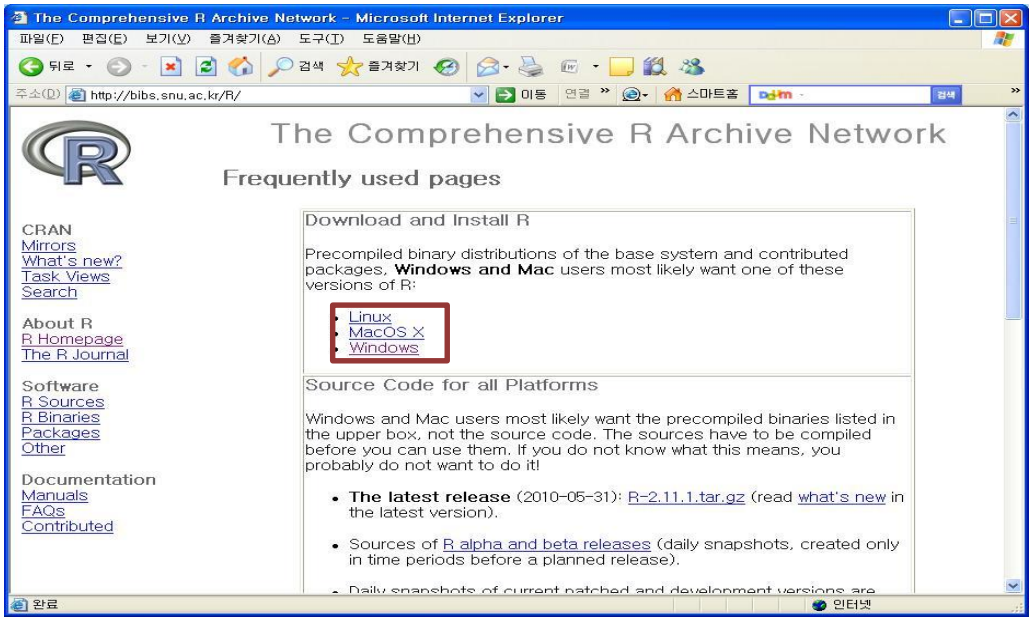
Mirror 사이트 선택



3 R 다운받기

◆ OS(Linux 또는 Window) 선택

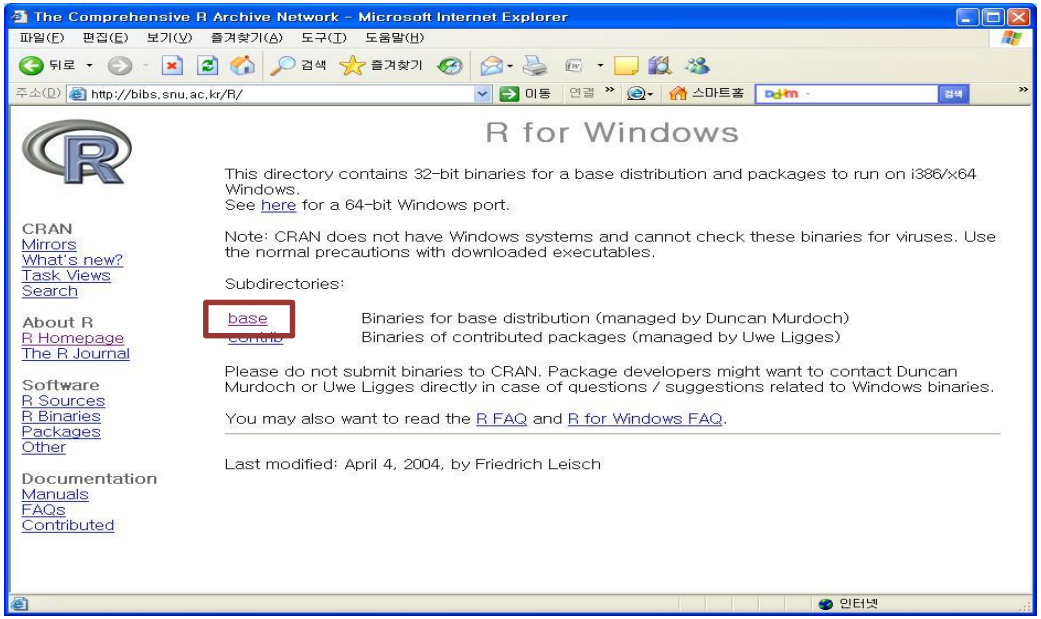
OS 선택



3 R 다운받기

◆ base 선택

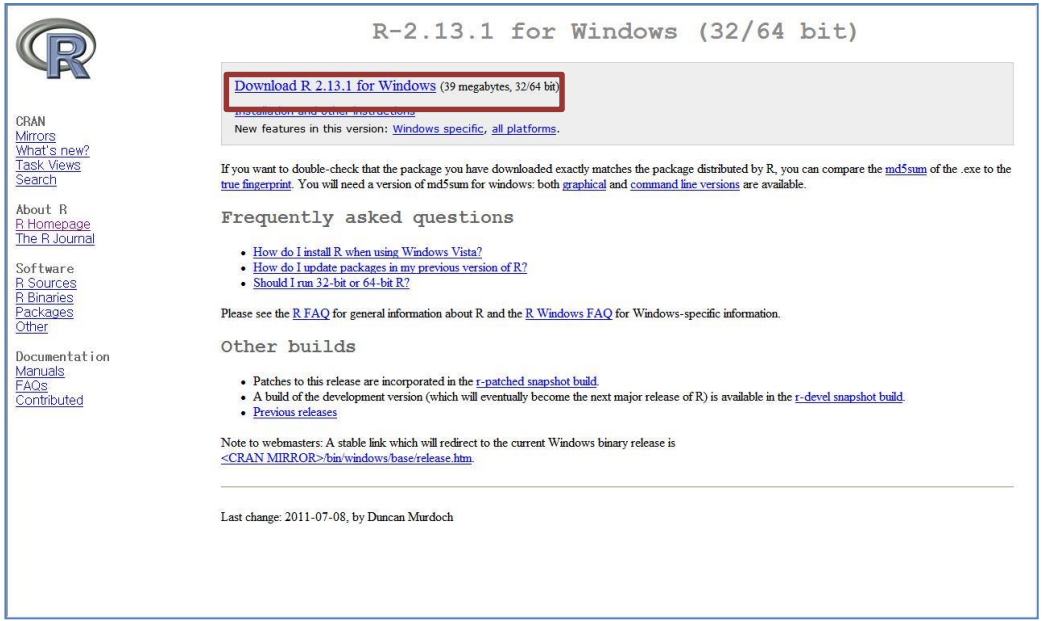
base 선택



3 R 다운받기

- Download R

Download R 선택



3 R 다운받기

R Contributors

About R
What is R?
Contributors
Screenshots
What's new?

Download, Packages
CRAN

R Project
Foundation
Members & Donors
Mailing Lists
Bug Tracking
Developer Page
Conferences
Search

Documentation
Manuals
FAQs
The R Journal
Wiki
Books
Certification
Other

Misc
Bioconductor
Related Projects
User Groups
Links

R Project Contributors

The current R is the result of a collaborative effort with contributions from all over the world. R was initially written by Robert Gentleman and Ross Ihaka—also known as "*R & R*" of the Statistics Department of the University of Auckland. Since mid-1997 there has been a core group with write access to the R source, currently consisting of

- Douglas Bates
- John Chambers
- Peter Dalgaard
- Seth Falcon
- Robert Gentleman
- Kurt Hornik
- Stefano Iacus
- Ross Ihaka
- Friedrich Leisch
- Uwe Ligges
- Thomas Lumley
- Martin Maechler
- Duncan Murdoch
- Paul Murrell
- Martyn Plummer
- Brian Ripley
- Deepayan Sarkar
- Duncan Temple Lang
- Luke Tierney
- Simon Urbanek

plus Heiner Schwarte up to October 1999 and Guido Masarotto up to June 2003.

R would not be what it is today without the invaluable help of these people, who contributed by donating code, bug fixes and documentation:

> Valerio Aimale, Thomas Baier, Roger Bivand, Ben Bolker, David Brahm, G□ran Brostr□m, Patrick Burns, Vince Carey, Saikat DebRoy, Brian D'Urso, Lyndon Drake, Dirk Eddelbuettel, John Fox, Paul Gilbert, Torsten Hothorn, Robert King, Kjetil Kjernsmo, Philippe Lambert, Jan de Leeuw, Jim Lindsey, Patrick Lindsey, Catherine Loader, Gordon Maclean, John Maindonald, David Meyer, Jens Oehlsch□gel, Steve Oncley, Richard O'Keefe, Hubert Palme, Jos□ C. Pinheiro, Anthony Rossini, Jonathan Rougier, G□nther Sawitzki, Bill Simpson, Gordon Smyth, Adrian Trapletti, Terry Therneau, Bill Venables, Gregory R. Warnes, Andreas Weingessel, Morten Welinder, Simon Wood, and Achim Zeileis,

We have probably omitted some important names here because of incomplete record keeping. If we have overlooked you, please let us know and we'll update the list. Many more, too numerous to mention here, have contributed by sending bug reports and suggesting various improvements. Simon Davies whilst at the University of Auckland wrote the original version of glm(). Julian Harris and Wing Kwong (Tiki) Wan whilst at the University of Auckland assisted Ross Ihaka with

3 R 다운받기

R manuals

About R
What is R?
Contributors
Screenshots
What's new?

Download, Packages
CRAN

R Project
Foundation
Members & Donors
Mailing Lists
Bug Tracking
Developer Page
Conferences
Search

Documentation
Manuals
FAQs
The R Journal
Wiki
Books
Certification
Other

Misc
Bioconductor
Related Projects
User Groups
Links

The R Manuals

edited by the R Development Core Team.

Current Version: 2.13.1 (July 2011)

The following manuals for R were created on Debian Linux and may differ from the manuals for Mac or Windows on platform-specific pages, but most parts will be identical for all platforms. The correct version of the manuals for each platform are part of the respective R installations. Here they can be downloaded as PDF files or directly browsed as HTML:

- **An Introduction to R** is based on the former "Notes on R", gives an introduction to the language and how to use R for doing statistical analysis and graphics. [browse HTML | download PDF]
- A draft of **The R language definition** documents the language *per se*. That is, the objects that it works on, and the details of the expression evaluation process, which are useful to know when programming R functions. [browse HTML | download PDF]
- **Writing R Extensions** covers how to create your own packages, write R help files, and the foreign language (C, C++, Fortran, ...) interfaces. [browse HTML | download PDF]
- **R Data Import/Export** describes the import and export facilities available either in R itself or via packages which are available from CRAN. [browse HTML | download PDF]
- **R Installation and Administration** [browse HTML | download PDF]
- **R Internals**: a guide to the internal structures of R and coding standards for the core team working on R itself. [browse HTML | download PDF]
- **The R Reference Index**: contains all help files of the R standard and recommended packages in printable form. [download PDF, 16MB, approx. 3100 pages]

Translations of manuals into other languages than English are available from the contributed documentation section (only a few translations are available).

The latex or texinfo sources of the latest version of these documents are contained in every R source distribution (in the subdirectory `doc/manual` of the extracted archive). Older versions of the manual can be found in the respective archives of the R sources. The HTML versions of the manuals are also part of most R installations (accessible using function `help.start()`).

3 R 다운받기

R Books

About R
What is R?
Contributors
Screenshots
What's new?

Download, Packages
CRAN

R Project
Foundation
Members & Donors
Mailing Lists
Bug Tracking
Developer Page
Conferences
Search

Documentation
Manuals
FAQs
The R Journal
Wiki
Books
Certification
Other

Misc
Bioconductor
Related Projects
User Groups
Links

[50] Roger S. Bivand, Edzer J. Pebesma, and Virgilio Gómez-Rubio. *Applied Spatial Data Analysis with R*. Springer, New York, 2008. ISBN 978-0-387-78170-. [bib | Discount Info | Publisher Info]

Applied Spatial Data Analysis with R is divided into two basic parts, the first presenting R packages, functions, classes and methods for handling spatial data. This part is of interest to users who need to access and visualise spatial data. Data import and export for many file formats for spatial data are covered in detail, as is the interface between R and the open source GRASS GIS. The second part showcases more specialised kinds of spatial data analysis, including spatial point pattern analysis, interpolation and geostatistics, areal data analysis and disease mapping. The coverage of methods of spatial data analysis ranges from standard techniques to new developments, and the examples used are largely taken from the spatial statistics literature. All the examples can be run using R contributed packages available from the CRAN website, with code and additional data sets from the book's own website. This book will be of interest to researchers who intend to use R to handle, visualise, and analyse spatial data. It will also be of interest to spatial data analysts who do not use R, but who are interested in practical aspects of implementing software for spatial data analysis. It is a suitable companion book for introductory spatial statistics courses and for applied methods courses in a wide range of subjects using spatial data, including human and physical geography, geographical information systems, the environmental sciences, ecology, public health and disease control, economics, public administration and political science. The book has a website where coloured figures, complete code examples, data sets, and other support material may be found: http://www.asdar-book.org.

[51] Roger D. Peng and Francesca Dominici. *Statistical Methods for Environmental Epidemiology with R: A Case Study in Air Pollution and Health*. Springer, New York, 2008. ISBN 978-0-387-78166-2. [bib | Discount Info | Publisher Info]

Advances in statistical methodology and computing have played an important role in allowing researchers to more accurately assess the health effects of ambient air pollution. The methods and software developed in this area are applicable to a wide array of problems in environmental epidemiology. This book provides an overview of the methods used for investigating the health effects of air pollution and gives examples and case studies in R which demonstrate the application of those methods to real data. The book will be useful to statisticians, epidemiologists, and graduate students working in the area of air pollution and health and others analyzing similar data. The authors describe the different existing approaches to statistical modeling and cover basic aspects of analyzing and understanding air pollution and health data. The case studies in each chapter demonstrate how to use R to apply and interpret different statistical models and to explore the effects of potential confounding factors. A working knowledge of R and regression modeling is assumed. In-depth knowledge of R programming is not required to understand and run the examples. Researchers in this area will find the book useful as a "live" reference. Software for all of the analyses in the book is downloadable from the web and is available under a Free Software license. The reader is free to run the examples in the book and modify the code to suit their needs. In addition to providing the software for developing the statistical models, the authors provide the entire database from the National Morbidity, Mortality, and Air Pollution Study (NMMAPS) in a convenient R package. With the database, readers can run the examples and experiment with their own methods and ideas.

[52] Robert Gentleman. *Bioinformatics with R*. Chapman & Hall/CRC, Boca Raton, FL, 2008. ISBN 1-420-06367-7. [bib | Discount Info | Publisher Info]

The Bioconductor project was initiated in 2001 to provide a resource of R packages that specifically address bioinformatics problems. Written by the leader of this project and the original developer of the R software, this book provides an overview of techniques to develop R programming skills for bioinformatics. The book presents comprehensive coverage of a broad range of key topics, including R language fundamentals, object-oriented programming in R, foreign language interfaces, building R packages, handling different data technologies, and debugging. It includes a number of detailed illustrative bioinformatics examples as well as exercises to demonstrate techniques.

[53] Robert Gentleman. *R Programming for Bioinformatics*. Computer Science & Data Analysis. Chapman & Hall/CRC, Boca Raton, FL, 2008. ISBN 978-1-420-06367-7. [bib | Discount Info | Publisher Info | http://www.bioconductor.org/pub/RBioinf/]

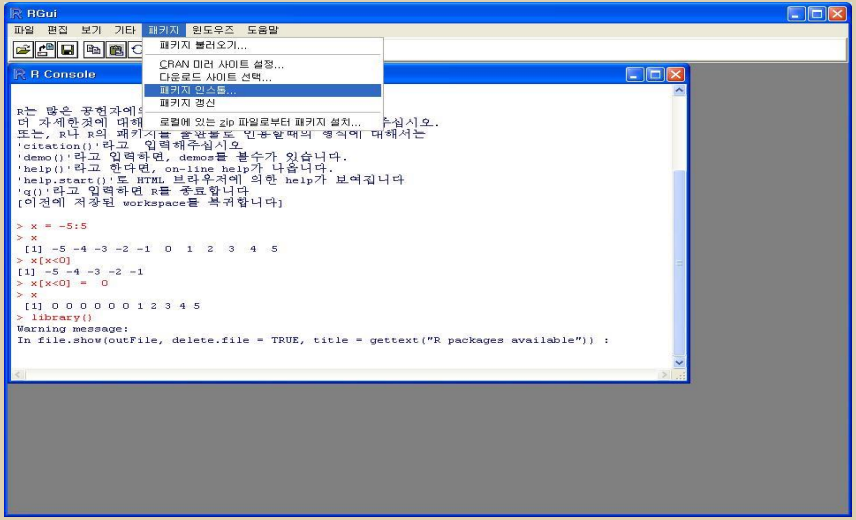
Thanks to its data handling and modeling capabilities and its flexibility, R is becoming the most widely used software in bioinformatics. R Programming for Bioinformatics builds the programming skills needed to use R for solving bioinformatics and computational biology problems. Drawing on the author's experiences as an R expert, the book begins with coverage on the general properties of the R language, several unique programming aspects of R, and object-oriented programming in R. It presents methods for data input and output as well as database interactions. The author also examines different facets of string handling and manipulations, discusses the interfacing of R with other languages, and describes how to write software packages. He concludes with a discussion

4 R 패키지

◆ 모든 R 함수와 다양한 데이터 들은 패키지(package)로 제공됨

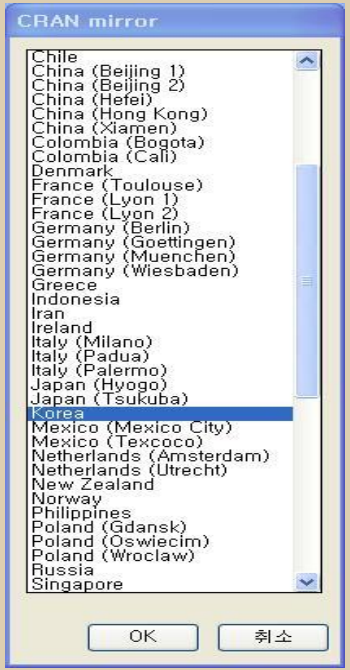
예) 군집분석 패키지 “cluster”를 설치하는 방법

① [패키지] - [package 인스톨]을 선택 또는
> install.packages(“cluster”)

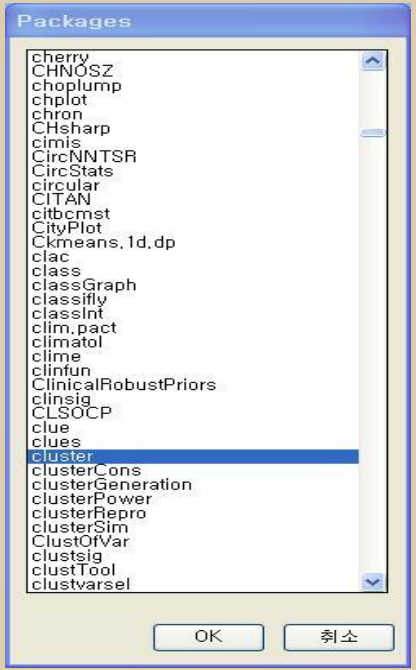


4 R 패키지

② Cran Mirror 선택

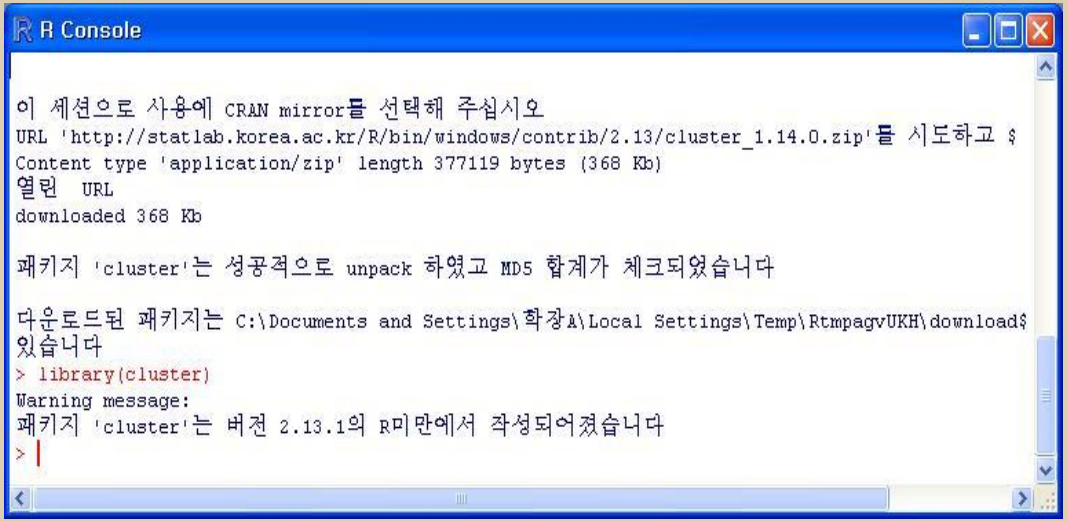


③ 패키지 cluster 선택



4 R 패키지

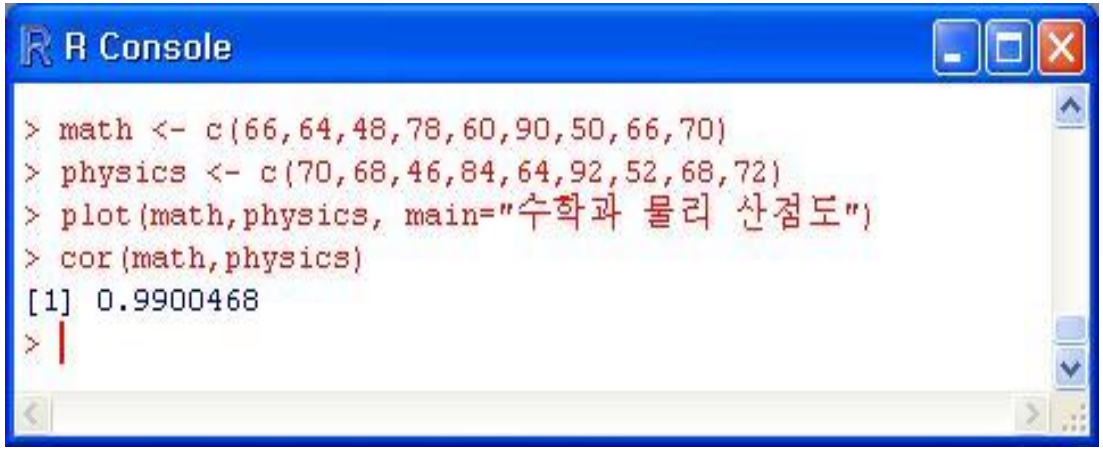
④ library(cluster) 입력

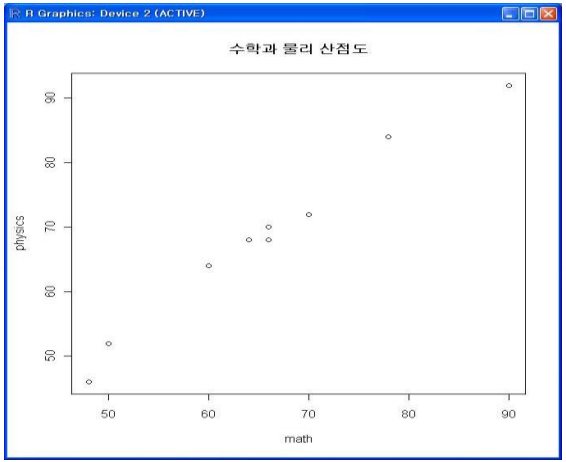


5 산점도 및 상관계수 예

- 어느 고등학교 학생 10명을 임의로 뽑아서 수학, 물리, 음악성적을 기록한 것이다. 두 변수간의 산점도를 그리고, 두 변수간의 상관계수를 알아보도록 하자.

학생번호	1	2	3	4	5	6	7	8	9	10
수학	68	64	48	46	78	60	90	50	66	70
물리	70	68	46	48	84	64	92	52	68	72





2. R 활용하기

1 R의 활용

◆ R을 활용하여 다음 [예제]를 구하여보자.

예제 1

▪ 어느 집단에서 표본을 10명 추출하여 다음과 같은 4개 문항에 대하여 설문조사를 실시하였다.

문항 1. 귀하의 성별은?

1) 남자 2) 여자

문항 2. 귀하의 나이는? (단위 : 세)

문항 3. 교육정도는?

1) 중졸이하 2) 고졸 3)대졸 및 그 이상

문항 4. 월수입(단위: 만원)

1 R의 활용

◆ R을 활용하여 다음 [예제]를 구하여보자.

자료

< 표본조사 자료 (10명) >

일련 번호	변수1 (성별)	변수2 (나이)	변수3 (교육정도)	변수4 (월수입)
1	1	21	2	150
2	2	22	1	100
3	1	33	2	200
4	2	33	3	220
5	1	28	2	170
6	1	41	3	300
7	2	39	2	290
8	1	32	3	220
9	2	44	1	370
10	1	55	3	410

1 R의 활용

◆ R을 활용하여 다음 [예제]를 구하여보자.

예제 1

1) 남자 여자의 수(이를 성별 도수분포표라 함)와 각 교육정도별 사람의 수(교육정도별 도수분포표)를 구하라.

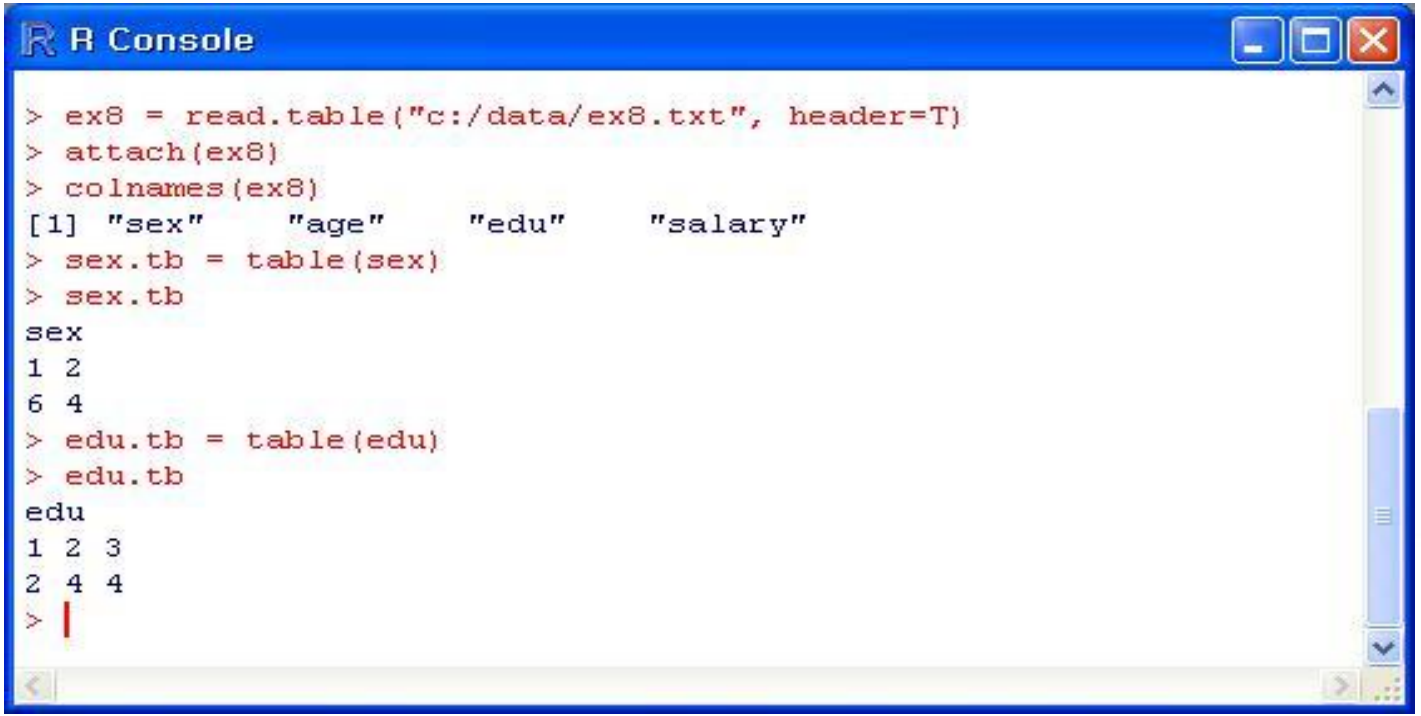
2) 교육정도별 도수분포를 나타내는 막대그림을 그려 어느 집단이 제일 많은지 관찰하라.

3) X축을 나이, Y축을 월수입으로 하는 산점도(scatter plot)를 성별로 구분되도록 그려 나이와 월수입의 관련성을 살펴보라.

- 데이터 파일 구성 : C:₩data₩ex8.txt

	sex	age	edu	salary
1	1	21	2	150
2	2	22	1	100
3	1	33	2	200
4	2	33	3	220
5	1	28	2	170
6	1	41	3	300
7	2	39	2	290
8	1	32	3	220
9	2	44	1	370
10	1	55	3	410

- 성별과 교육 정도의 도수분포표 구하기

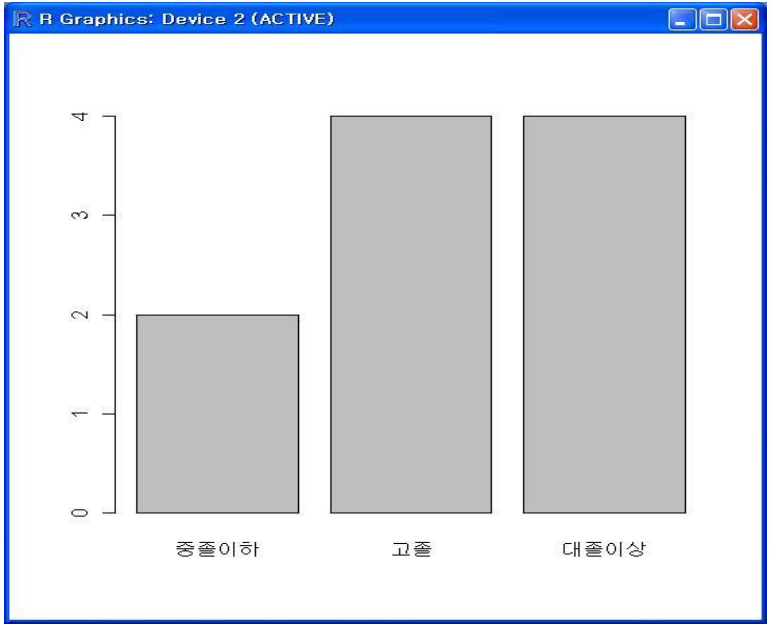


1 R의 활용

- 교육 정도 별 도수분포표를 나타내는 막대그림

R Console

```
> rownames(edu.tb) = c("중졸이하","고졸","대졸이상")
> edu.tb
edu
중졸이하     고졸 대졸이상
       2        4        4
> barplot(edu.tb)
>
```

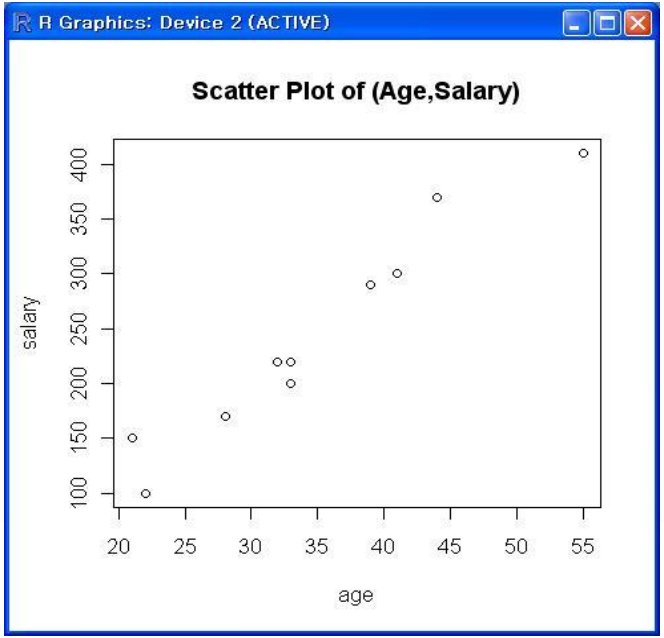


1 R의 활용

- [나이, 월수입] 산점도(scatter plot) 그리기

R Console

```
> ex8
   sex age edu salary
1    1  21   2    150
2    2  22   1    100
3    1  33   2    200
4    2  33   3    220
5    1  28   2    170
6    1  41   3    300
7    2  39   2    290
8    1  32   3    220
9    2  44   1    370
10   1  55   3    410
> plot(age, salary)
> title("Scatter Plot of (Age,Salary)")
>
```



2 R을 이용하여 이산형 그래프 그리기

◆ R을 활용하여 다음 [예제]를 구하여보자.

예제 2

▪ 세 그룹(C1,C2,C3)이 다섯 자선단체(T1 T5) 에 기부하는 가상자료를 예를 들어 막대그림, 원그림을 그리는 프로그램을 작성하여 보자.

< perc.txt >

	C1	C2	C3
T1	5.4	3.1	3.5
T2	5.7	8.6	25.0
T3	20.4	26.0	22.0
T4	36.3	34.1	28.0
T5	14.4	11.4	4.5

참고: 위와 같이 행렬 형태의 자료를 data.frame 이라고 한다.
read.table 함수를 이용하여 읽으면 된다.

2 R을 이용하여 이산형 그래프 그리기

- 데이터 읽기

```
> percData <- read.table("c:/mydata/perc.txt", header=T)
> percData <- as.matrix(percData)
> var.name  <- colnames(percData)
> case.name <- rownames(percData)
```

: 데이터의 변수이름과 행의 이름을 각각 var.name, case.name 에 저장

- 막대그림 그리기

```
> # barplot
> barplot(percData, names=var.name)
> legend(locator(1), case.name)
> title("Barplot")
```

: 막대그림함수(barplot),

: 마우스로 선택한 임의의 위치에(locator(1)) 범례(legend)를 나타냄

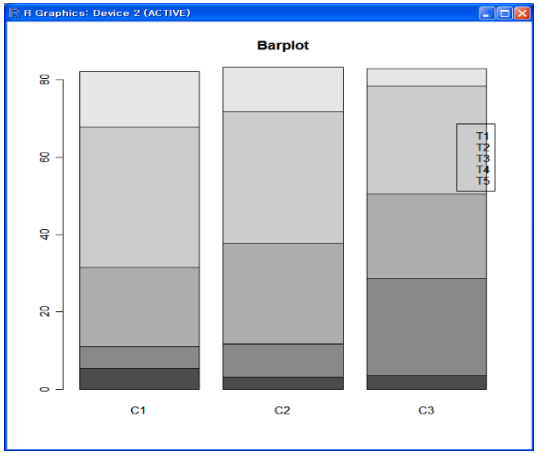
2 R을 이용하여 이산형 그래프 그리기

막대그림 그리기

```
> # barplot
> barplot(percData, names=var.name)
> legend(locator(1), case.name)
> title("Barplot")
```

: 막대그림함수(barplot),

: 마우스로 선택한 임의의 위치에(locator(1))
범례(legend)를 나타냄

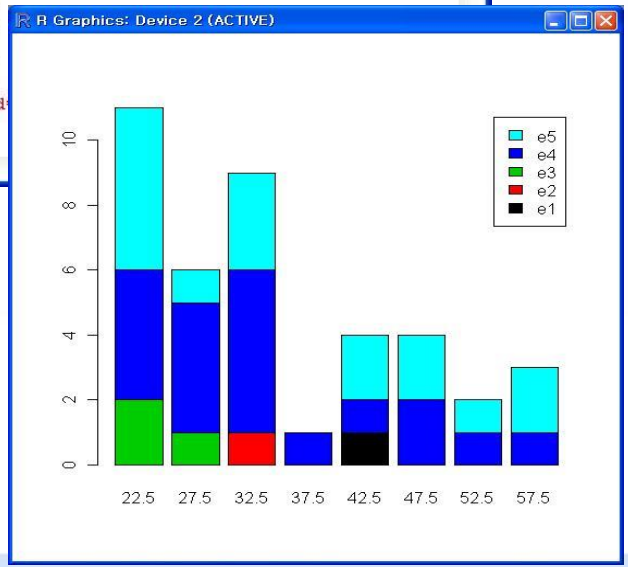


3 R을 이용하여 연속형 그래프 그리기

◆ [예제3] 자료에서 교육 정도 별 나이에 대한 히스토그램 그리기

R Console

```
> age.edu.count = matrix(0,ncol=8, nrow=5)
> colnames(age.edu.count) = bmid
> rownames(age.edu.count) = c("e1","e2","e3","e4","e5")
> for(i in 1:5) { age.e = age[edu==i]
+      age.hist = hist(age.e, breaks=bpoint)
+      age.edu.count[i,] = age.hist$counts }
> age.edu.count
   22.5 27.5 32.5 37.5 42.5 47.5 52.5 57.5
e1    0    0    0    0    1    0    0    0
e2    0    0    1    0    0    0    0    0
e3    2    1    0    0    0    0    0    0
e4    4    4    5    1    1    2    1    1
e5    5    1    3    0    2    2    1    2
> barplot(age.edu.count, col=c(1:5), legend
>
```



4 원 그래프

- 자료파일

education.txt - 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

Grade	Y1970	Y1990
초등졸	28.3	21.1
중졸	24.7	20.2
고졸	36.0	43.6
대졸	11.0	15.1

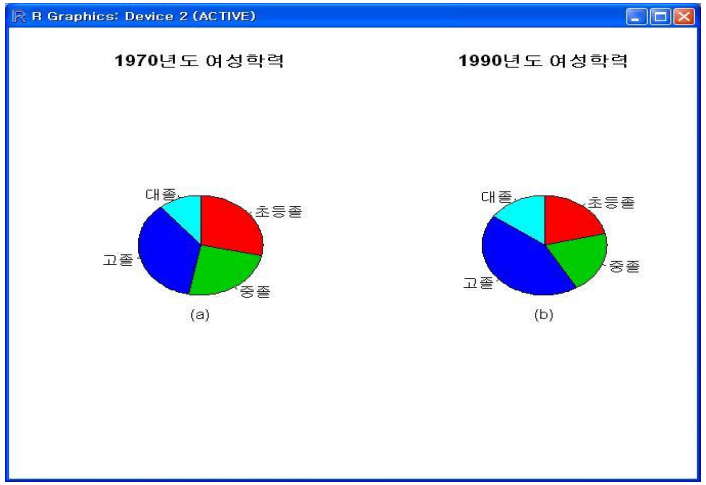
- R 명령어 실행화면

R Console

```
> edu = read.table("c:/data/education.txt", header=T)
> edu
   Grade Y1970 Y1990
1 초등졸  28.3  21.1
2   중졸  24.7  20.2
3   고졸  36.0  43.6
4   대졸  11.0  15.1
> name = edu[,1]
> grade70 = edu[,2]
> grade90 = edu[,3]
> par(mfrow=c(1,2))
> pie(grade70, label=name, clockwise=T, col=c(2:5), main="1970년도 여성학력")
> mtext(outer=F, "(a)", side=1, line=-6)
> pie(grade90, label=name, clockwise=T, col=c(2:5), main="1990년도 여성학력")
> mtext(outer=F, "(b)", side=1, line=-6)
>
```

4 원 그래프

풀이



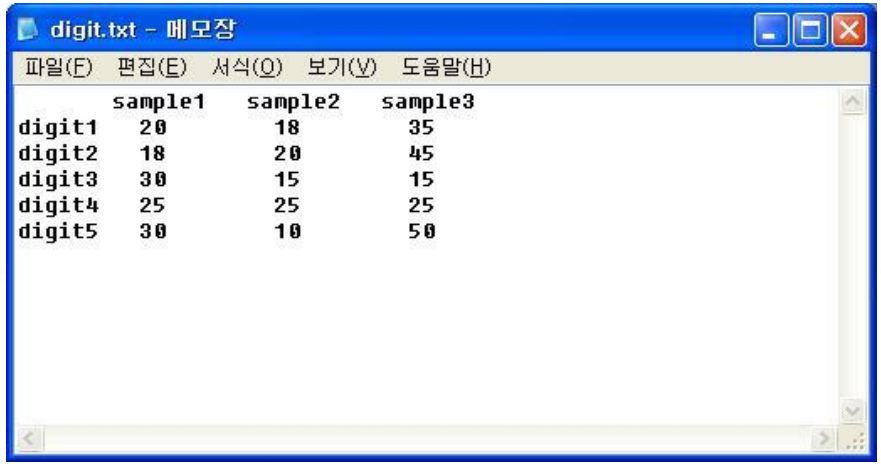
- 두 그래프를 보면 그 구성비의 변화를 표를 이용하여 비교하는 것보다 쉽게 비교
- 원그래프에서 시각적으로 보면 1970년도 고졸의 구성비보다 1990년도의 구성비가 증가했음을 쉽게 알 수 있음

5 막대 그래프(Bar chart)

◆ 막대그래프 : 각 범주에 속한 도수나 비율을 하나의 막대로 나타낸 그래프

- 예제) 5×3 행렬로 구성된 digit 자료를 이용하여
  막대그래프를 그리고 각 sample에 따른 구성비를
  설명하여라.

<digit.txt 자료>

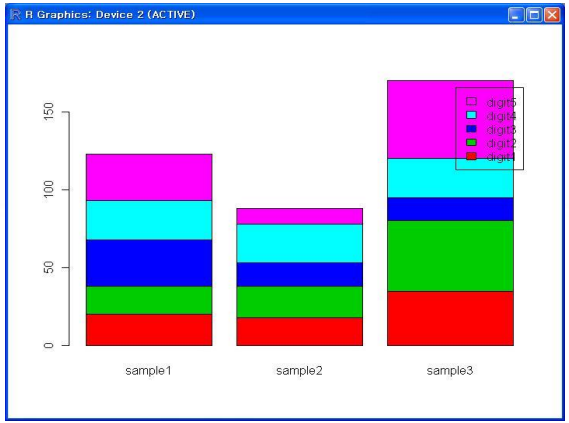


5 막대 그래프(Bar chart)

풀이

R Console

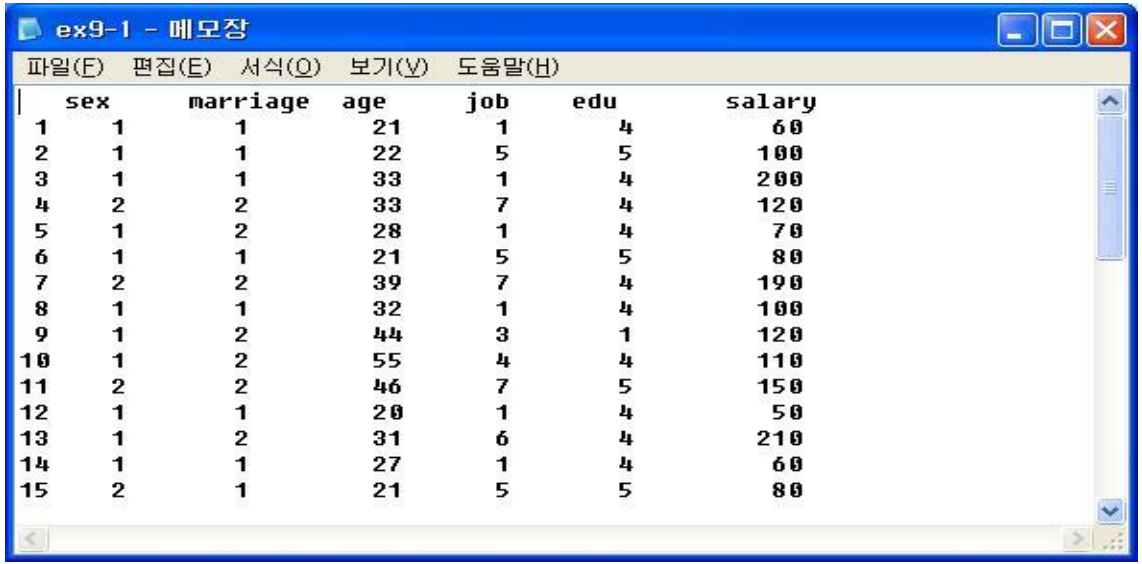
```
> digit = read.table("c:/data/digit.txt", header=T)
> digit = as.matrix(digit)
> par(mfrow=c(1,1))
> barplot(digit, legend.text=rownames(digit), col=c(2:6))
>
```



6 R을 이용한 기술통계량 구하기

◆ 데이터 읽기

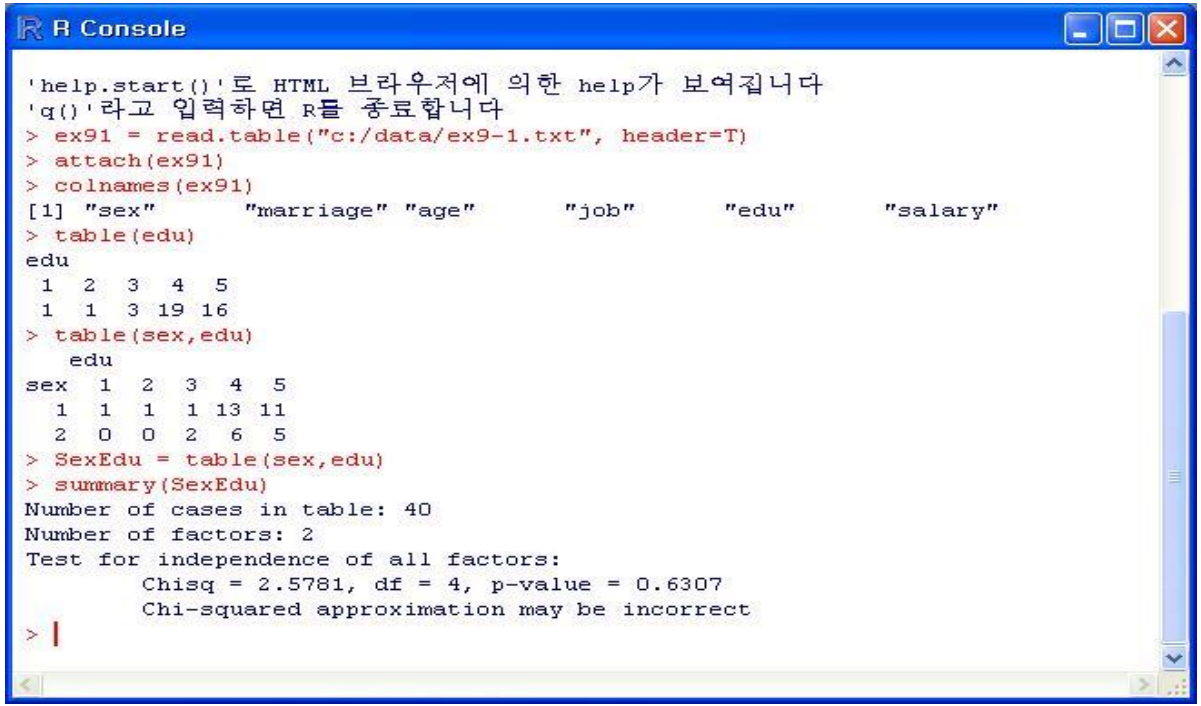
> “ex91=read.table(“c:/data/ex9-1.txt”, header=T)



	sex	marriage	age	job	edu	salary
1	1	1	21	1	4	60
2	1	1	22	5	5	100
3	1	1	33	1	4	200
4	2	2	33	7	4	120
5	1	2	28	1	4	70
6	1	1	21	5	5	80
7	2	2	39	7	4	190
8	1	1	32	1	4	100
9	1	2	44	3	1	120
10	1	2	55	4	4	110
11	2	2	46	7	5	150
12	1	1	20	1	4	50
13	1	2	31	6	4	210
14	1	1	27	1	4	60
15	2	1	21	5	5	80

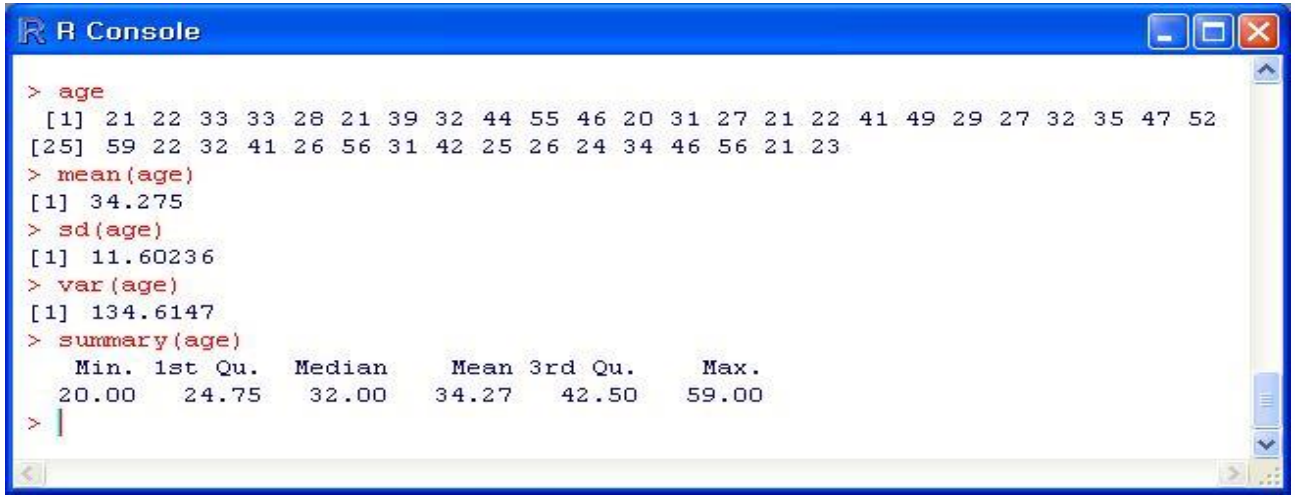
6 R을 이용한 기술통계량 구하기

- R을 이용한 빈도표 및 분할표, 독립성검정결과



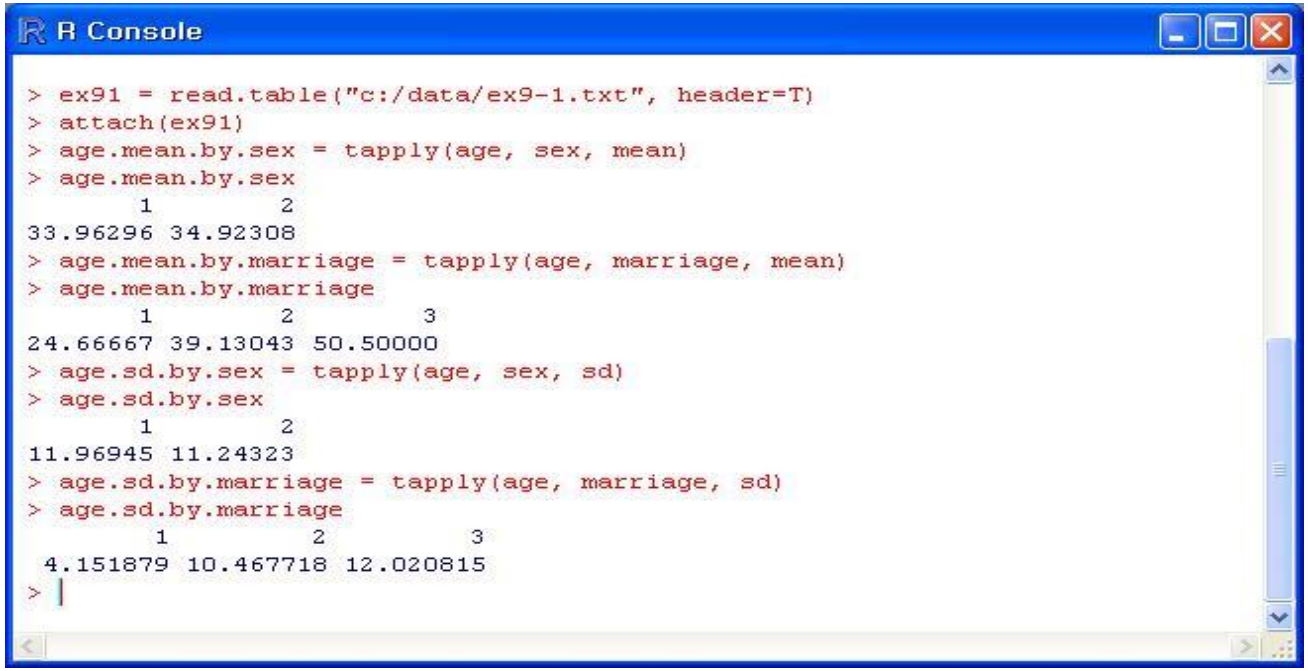
6 R을 이용한 기술통계량 구하기

- 나이 변수에 대한 기술통계량



6 R을 이용한 기술통계량 구하기

- 나이에 대한 성별, 결혼상태 별 평균 및 표준편차



6 R을 이용한 기술통계량 구하기

- 나이에 대한 (성별x결혼상태) 별 평균 및 표준편차