R통계분석-기초분석

R을 활용한 통계분석

김 덕 기 (toby123@cbnu.ac.kr)

공군사관학교

R – 특징

[R의 탄생]

S, S-PLUS의 환경을 기초로 만들어진 공개적인 통계적 도구.

S-1980년 AT&T 연구소에서 개발.

R-1995년 Auckland대학 통계학과 소속 Robert and Ross 개발.

R-코어팀에 의해 지속적으로 개발

R-주요 정보 : http://www.r-project.org

[R의 장-단점]

장점- 무료 Software, UNIX, Macintosh, Windows의 모든 환경에서 구동 됨.

R은 도움말 기능이 편리하고, 그래픽 성능이 우수, S-plus와 연계 가능.

배우기 쉬운 통계적 함수를 내장, 사용자 정의함수를 쉽게 작성 등.

단점- 제한적인 그래픽 인터페이스, 배우는 초기과정이 다소 힘듬.

상업적 지원이 없고, 명령어 자체가 프로그램밍 언어로 이해가 필요.

R – Program 설치방법

[R-Program Download]

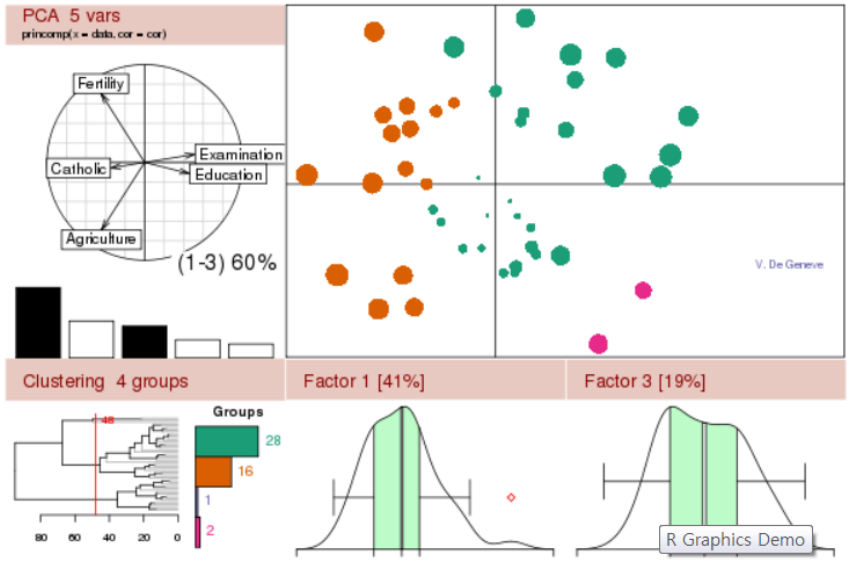
➔ Internet 접속 : http://www.r-project.org로 이동.

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

The R Project for Statistical Computing

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

Download, Packages
CRAN



(Continue…)

- Download R for Linux
- Download R for MacOS X
- Download R for Windows

Subdirectories:

base

contrib

old contrib

Rtools

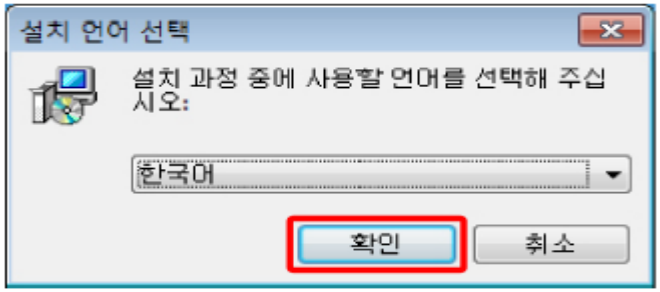
Download R 3.3.2 for Windows (62 megabytes, 32/64 bit)

Installation and other instructions

New features in this version

R – 설치하기

- 설치과정 – 다운로드 받은 파일을 실행하여 윈도우 설치과정을 진행합시다.

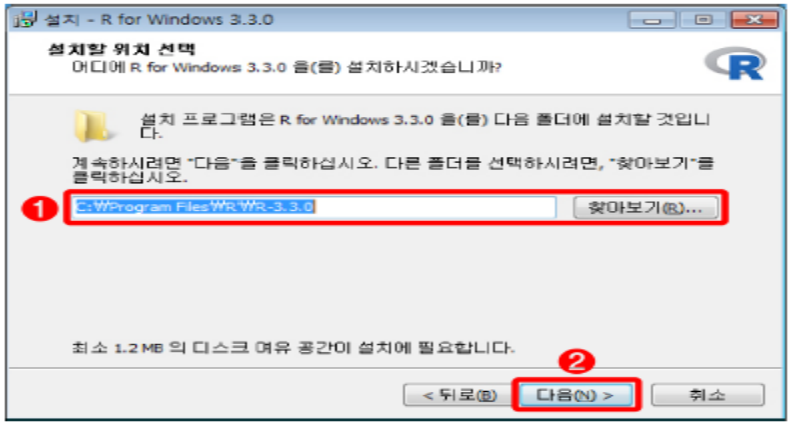


[그림 A–6] 설치 시 사용할 언어 선택

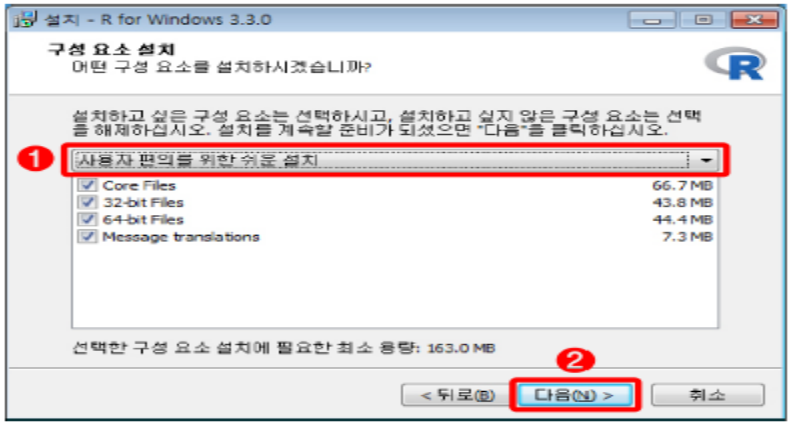
[그림 A–7] 설치 시작

[그림 A–8] 라이선스 확인

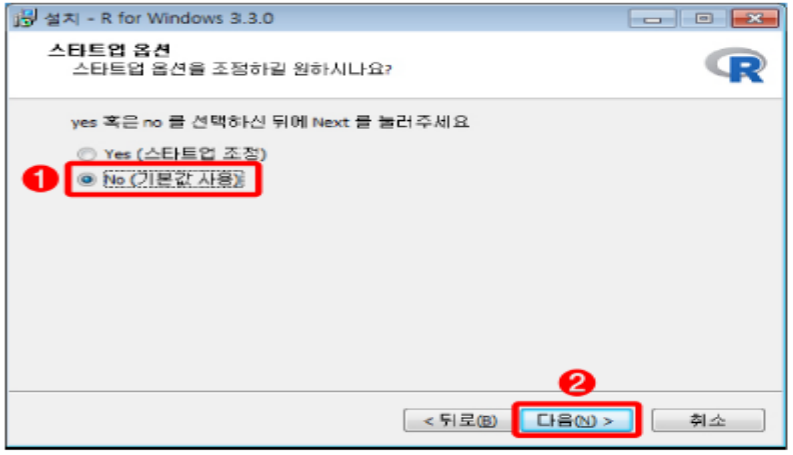
R – 설치하기



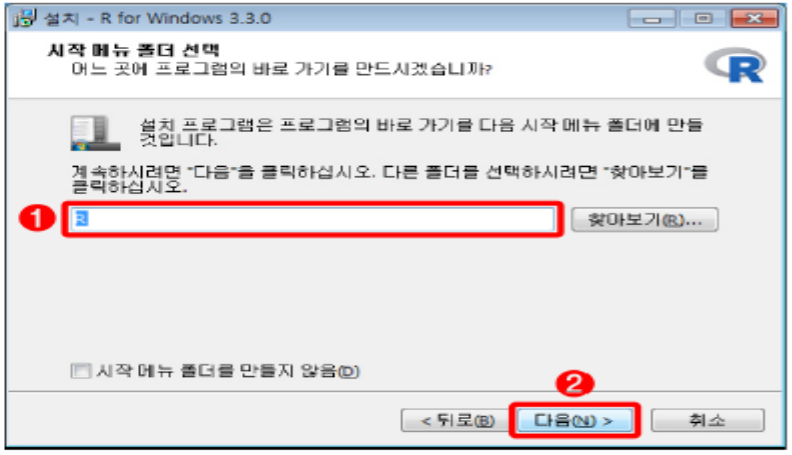
[그림 A–9] 설치 위치 확인



[그림 A–10] 설치 내용 확인

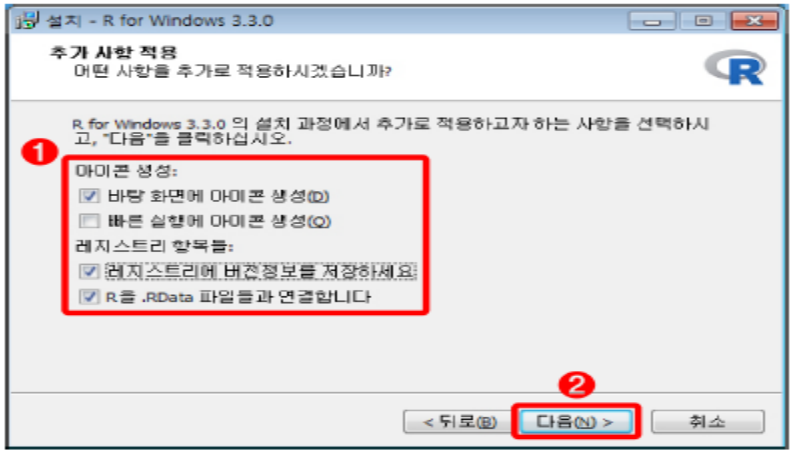


[그림 A–11] 기본 시작 옵션으로 설치

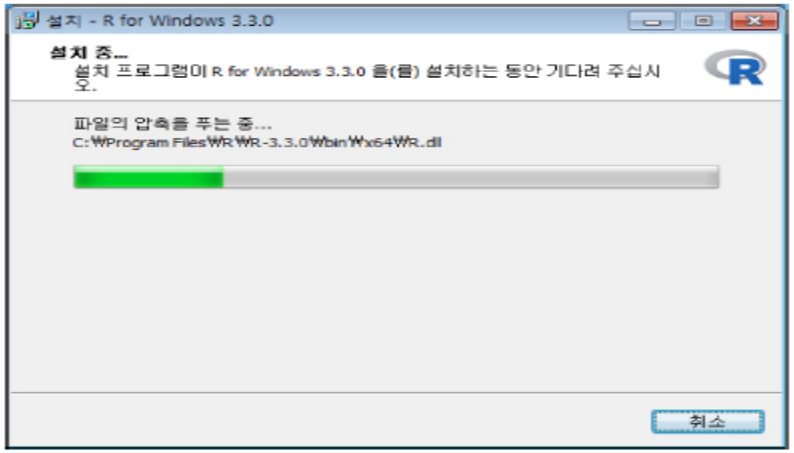


[그림 A–12] Windows의 프로그램 그룹 등록

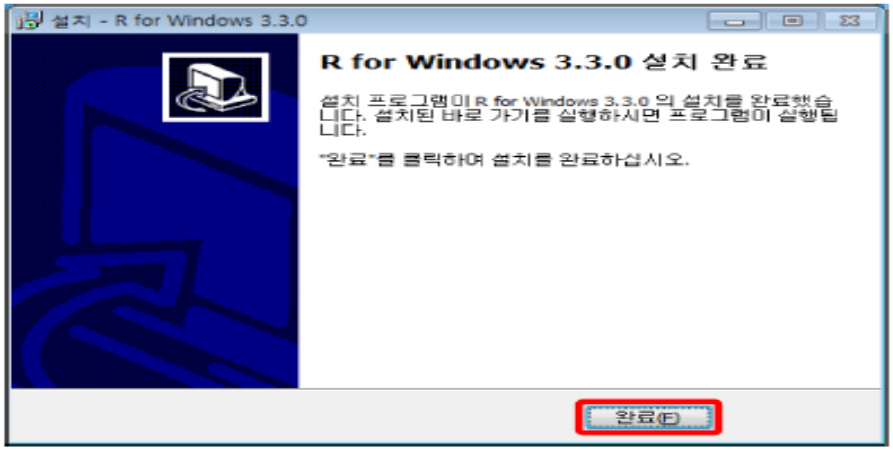
R – 설치하기



[그림 A–13] 추가 적용사항 확인



[그림 A–14] 설치 진행

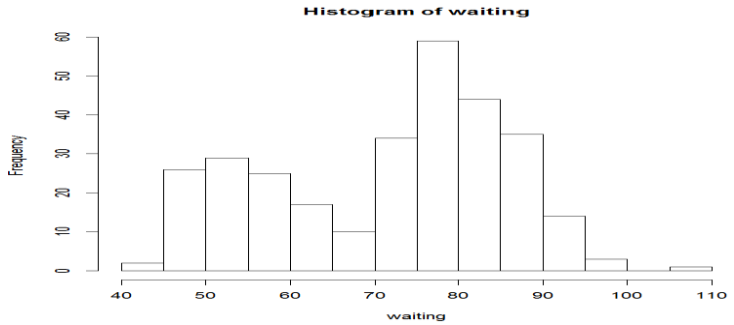


[그림 A–15] 설치 완료

외부데이터 불러오기 및 기초 사용법-1

```
geyser <- read.table(file="D:₩₩2019-통계교육₩₩R-data₩₩geyser299.txt",header=T)
attach(geyser)  # detach(geyser)
geyser
geyser[1:5,]
hist(waiting) # hist(geyser$waiting)  # quit()
```

	waiting	duration
1	80	4.02
2	71	2.15
3	57	4.00
4	80	4.00
5	75	4.00



=> 테이블 형태의 Text 데이터를 불어오기 위해서는 "read.table" 명령어를 이용. 뒤에 "header=T"는 첫 번째 행이 변수로 입력되어 있는 경우 지정.

다음으로 해 주어야 할 것이 test데이터속에 변수를 도수분포표 또는 다른 그래프로 표현하기 위해서는 attach()를 이용하여 test데이터안에 속해있는 변수들을 따로 이용할 수 있도록 해 주어야만 한다.

코딩방법 및 변경 : 단일응답

(설문1) 성별을 답해 주십시오 ( )
(설문2) 혈액형을 답해 주십시오 ( )형
(설문3) 가장 좋아하는 색을 답해 주십시오 ( )색
(설문4) 귀하의 연령을 답해 주십시오 ( )세
(설문5) 이 상품을 처음으로 알게 된 계기는 다음 중 어느 것입니까?
1.TV의 광고 2.라디오의 광고 3.신문의 광고
4.저널광고 5.아는 사람의 소개 6.기타 ( )
(설문6) 이 상품의 만족도는?
1.대단히 불만 2.불만 3.약간 불만 4.약간 만족 5.만족
6.대단히 만족

(설문1) 남 ➔ 1, 여 ➔ 2로 바꾸어 입력.
(설문2) A ➔ 1, B ➔ 2, AB ➔ 3, O ➔ 4로 바꾸어 입력.
(설문3) 응답 자체를 입력 후 필요한 시점에 코딩.
(설문4) 응답 자체를 입력 후 필요한 시점에 코딩.
(설문5) 선택된 숫자 그대로 입력.
(설문6) 선택된 숫자 그대로 입력.

* 무응답의 경우 : 설문1-2(숫자)는 99, 설문3-4(문자)는 공란.

코딩방법 및 변경 : 예제

[예제 1] 다음은 위 설문을 20명에게 실시하여 얻은 응답 결과이다. 변수를 정의하고 **(설문1~설문6)** 코딩변경으로 설문1(1➔남자, 2➔여자), 설문2(1➔A, …,4➔O)를 수행하고, 설문4는 새로운 변수 연령대(20미만이면 1, 20이상 30미만이면 2, 30이상 40미만이면 3, 40이상은 4로 값을 부여)를 만들어 보자. (파일 : 예제1.csv)

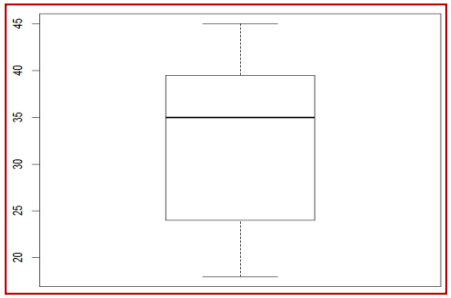
응답자	설문1	설문2	설문3	설문4	설문5	설문6
1	1	1	적	32	3	3
2	1	3	황	20	5	4
3	1	2	적	25	6	5
4	2	4	청	38	1	6
5	2	4	백	45	1	3
6	2	3	흑	36	2	4
7	1	2	백	39	4	5
8	1	1	적	23	3	1
9	1	1	녹	22	2	2
10	1	1	백	40	3	3
11	2	1	백	35	3	1
12	1	1	적	32	5	2
13	2	2	적	40	6	3
14	1	3	적	38	3	4
15	2	4	녹	40	3	3
16	1	1	녹	35	2	5
17	2	3	녹	42	1	3
18	1	2	백	35	1	5
19	2	1	백	20	2	5
20	2	3	적	18	3	4

외부데이터 불러오기 및 기초 사용법-2

```
# 작업디렉토리 변경 : 다운로드 받은 자료의 위치를 사전에 지정, file> new script 이용.

ex1 = read.csv("예제1.csv", header=T)
str(ex1)
attach(ex1)
stem(설문4); boxplot(설문4);
table(설문3); barplot(table(설문3))
length(설문4); mean(설문4); median(설문4); var(설문4); sd(설문4); quantile(설문4)
summary(설문4)
```

```
'data.frame':   20 obs. of  7 variables:
 $ 사람 : int  1 2 3 4 5 6 7 8 9 10 ...
 $ 설문1: int  1 1 1 2 2 2 1 1 1 1 ...
 $ 설문2: int  1 3 2 4 4 3 2 1 1 1 ...
 $ 설문3: Factor w/ 6 levels "녹","백","적",..: 3 5 3 4 2 6 2 3 1 2 ...
 $ 설문4: int  32 20 25 38 45 36 39 23 22 40 ...
 $ 설문5: int  3 5 6 1 1 2 4 3 2 3 ...
 $ 설문6: int  3 4 5 6 3 4 5 1 2 3 ...
```




```
[1] 20
[1] 32.75
[1] 69.88158
[1] 8.35952
   0%   25%   50%   75%  100%
18.00 24.50 35.00 39.25 45.00
```

```
The decimal point is 1 digit(s) to the right of the |

1 | 8
2 | 00235
3 | 225556889
4 | 00025
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
18.00	24.50	35.00	32.75	39.25	45.00

R - 실습문제

실습 1 : geyser299.txt

➔ waiting, duration에 대한 히스토그램을 그려 각 자료에 대한 분포를 해석하고, 두 자료 간의 연관성을 해석하기 위해 산점도를 그려 pattern을 해석하시오.

➔ waiting, duration 자료에 대해 기초통계량을 구해 자료의 특징을 해석하시오.

➔ 추가적으로 두 자료의 상자그림을 그려 자료의 특징을 해석하시오.

실습 2: 예제.csv

➔ (p.10)의 형태로 자료의 코딩변경을 수행하라.

➔ 범주형과 연속형을 나누어 적절한 그래프로 정리하고 해석하라.

➔ 성별에 따라 만족도에 대한 상자그림을 그려 해석하라.

➔ 연속형자료에 대해 수치(통계량-중심, 산포, 분포)들을 계산하여 결과를 해석하라.