R Studio로 하는 통계

The Statics for Ubuntu(5월)

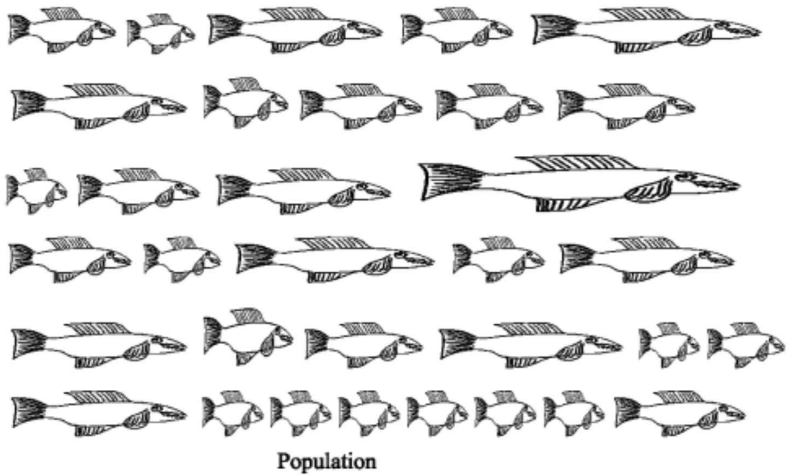
2020년 5월 6일 20:00~

오늘 함께 할 내용

통계의 핵심 - 유의 확률

통계 처리의 핵심툴 - R Studio

생물학 연구에서 통계의 필요성 살펴보기 1

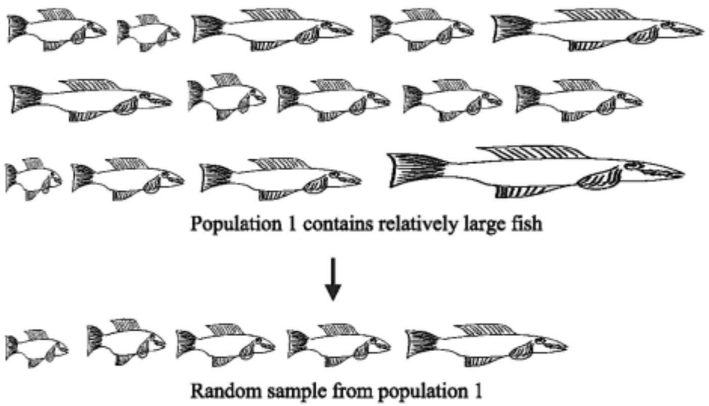


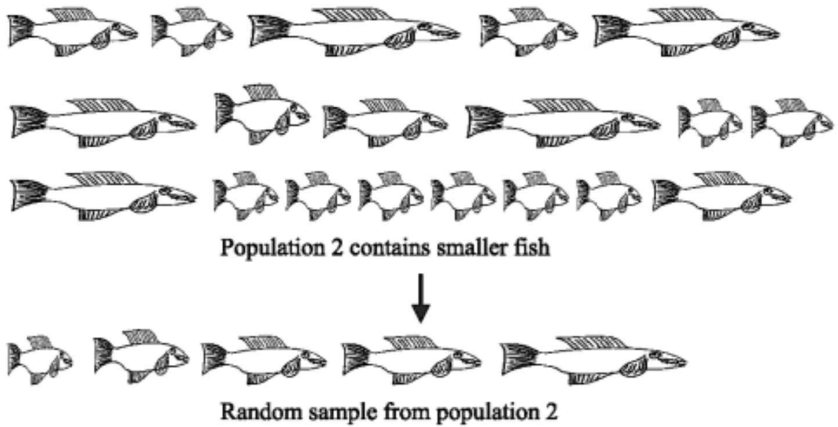
동일 집단에서 표집되었음에도 불구하고

우연히 다른 소집단이 표집됨

이미지: 박태성, 김희발(2011). 쉽게 풀어쓴 생물통계학. 교보문고

생물학 연구에서 통계의 필요성 살펴보기 2



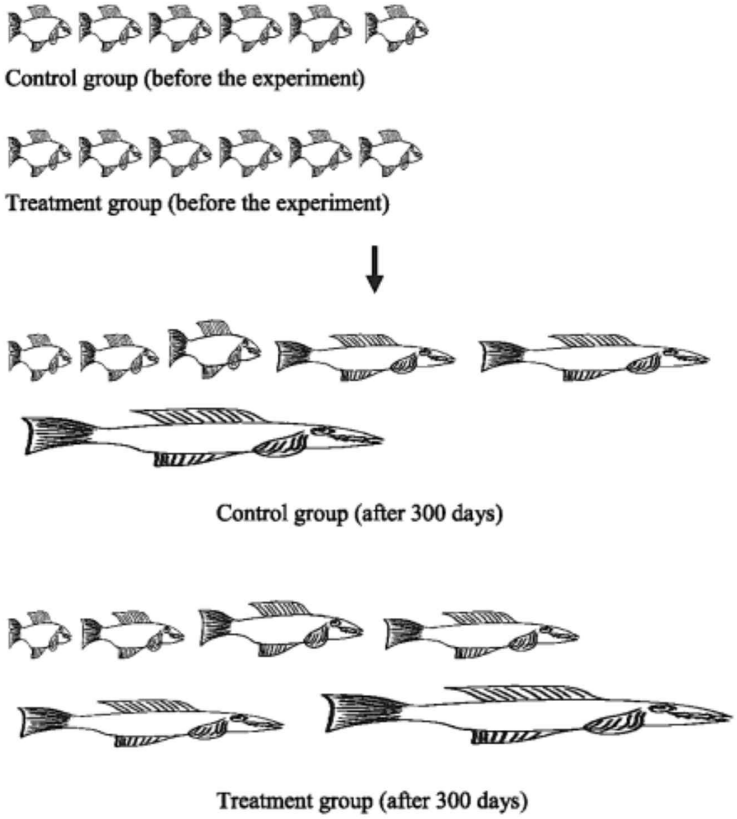


다른 집단에서 표집되었음에도 불구하고

우연히 비슷한 소집단이 표집될 수도 있음

이미지: 박태성, 김희발(2011). 쉽게 풀어쓴 생물통계학. 교보문고

생물학 연구에서 통계의 필요성 살펴보기 3



대조군 - 아무런 처리도 하지 않음
실험군 - 300일 동안 비타민 보충체 처리

대조군과 비교했을 때 어느 정도 성장한 효과가 있는 것으로 보이지만 성장 효과와 비교해 볼 때 상대적으로 미비해 보이는 경우

이미지: 박태성, 김희발(2011). 쉽게 풀어쓴 생물통계학. 교보문고

통계의 핵심

- 유의 확률

통계적 접근이란?

1. 예측은 기본적으로 불확실하다. -> 통계적 접근은 회의적, 비판적

1. 통계적 가설의 종류 : 귀무가설(또는 영가설; H_0), 대립가설(또는 실험 가설; H_1)
    a. 귀무가설 : 차이가 없다.
    b. 대립가설 : 차이가 있다.
    c. 생기는 의문? 어느 정도 차이면 인정해줄 수 있을까?

예화를 통해 익히기

1920년대 말 영국에서 실제 있었던 일이라고 전해지는 일화

어떤 한 사람이 밀크 홍차의 맛만 봐도 '차를 먼저 따랐는지 우유를 먼저 따랐는지'를 알 수 있다고 주장

그 자리에 현대 통계학의 아버지라고 불리는 한 남자(로널드 A. 피셔)가 검증 실험을 제안

순서를 다르게 한 차를 8잔(4잔 + 4잔) 준비하여 무작위로 맛보게 하였다.

결과) 연속해서 5잔의 배합 순서를 맞추었다.

질문) **맛을 통해 차와 우유의 배합 순서를 알아내는 능력을 가지고 있다고 인정할 수 있을까?**

예화를 통해 익히기 - 계속

> 귀무 가설 : 배합 순서를 알아내는 능력을 갖고 있지 않다.

> 대립 가설 : 배합 순서를 알아내는 능력을 가지고 있다.

판단 기준 : 연속해서 4잔의 배합 순서를 맞춘다. (1/16)

얻은 결과 : 연속해서 5잔의 배합 순서를 맞췄다. (1/32)

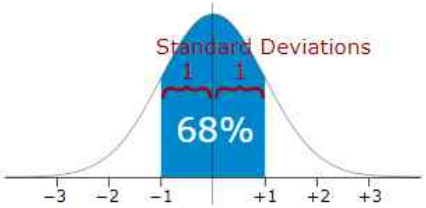
-> 1/32(0.03%)의 확률을 가지는 사건을 우연히 맞춘다는 것은 불가능에 가까운일.
그렇다면 우린 어떻게 받아들여야 할까? 능력이 있다는 것을 받아들이는 것이 타당하지 않을까?

예화를 통해 익히기

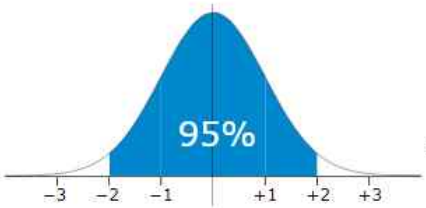
어떤 사람의 키를 측정하였더니 185cm라고 한다.
이 사람의 키가 크다고 할 수 있을까?

어느 정도의 확률을 기준으로 이야기할까?

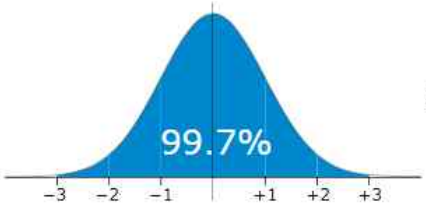
자연에서 얻어지는 데이터의 분포 형태에 대한 이해가 필요!



68% of values are within
1 standard deviation of the mean



95% of values are within
2 standard deviations of the mean



99.7% of values are within
3 standard deviations of the mean

유의 확률

> 어려운 표현 : 귀무 가설이 참인데도 귀무 가설이 거짓이라고 판단할 확률

-> 조금 쉬운 표현 : 차이가 없음에도 불구하고 차이가 있다고 판단할 가능성

-> 조금 더 쉬운 표현 : 잘못된 결론을 내릴 확률

- 통계에서는 이러한 경우를 1종 오류라고 부르지요.

유의 수준

> 어려운 표현 : 귀무 가설이 맞는데 귀무 가설이 거짓이라고 할 확률의 최대값

-> 조금 쉬운 표현 : 귀무 가설을 버리고 대립 가설을 수용하는 기준선

-> 더 쉬운 표현 : 틀린 정도를 용인할 수 있는 수준

정리해보기

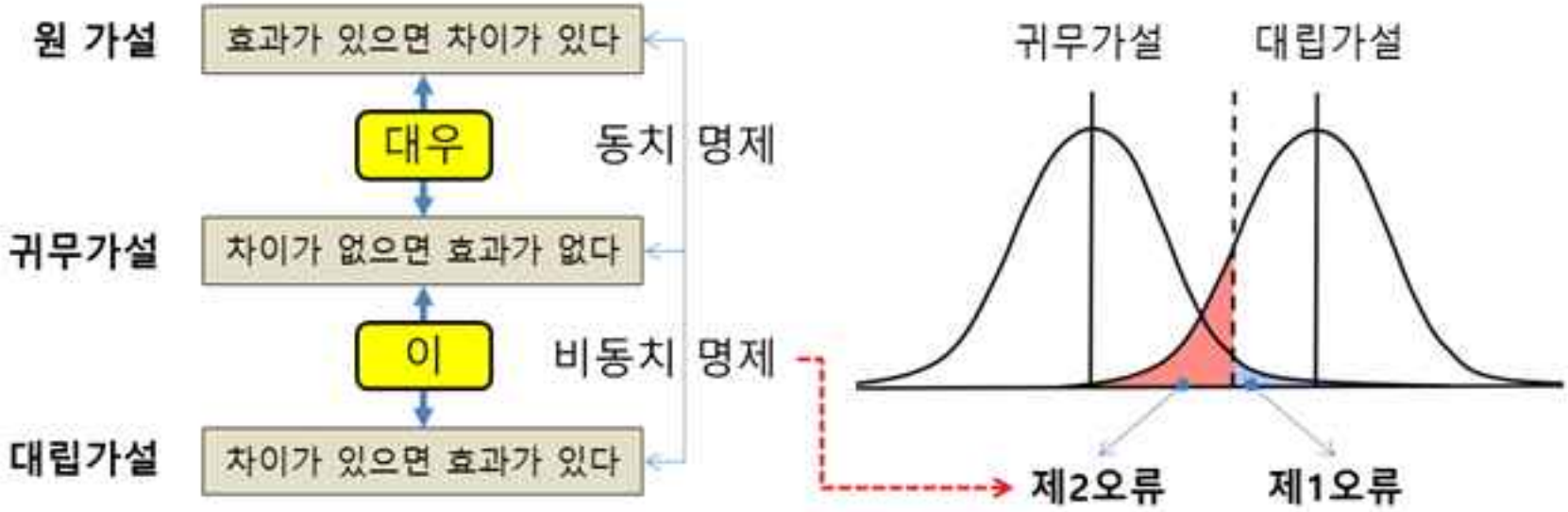
패널티 킥을 잘 찬다고 할 때 어느 정도를 잘 찬다고 할 수 있을까?

100번 중 98을 넣은 사람과 100번 중 90번을 넣은 사람이 있다고 할 때

누구를 킥커로 세울 것인가?

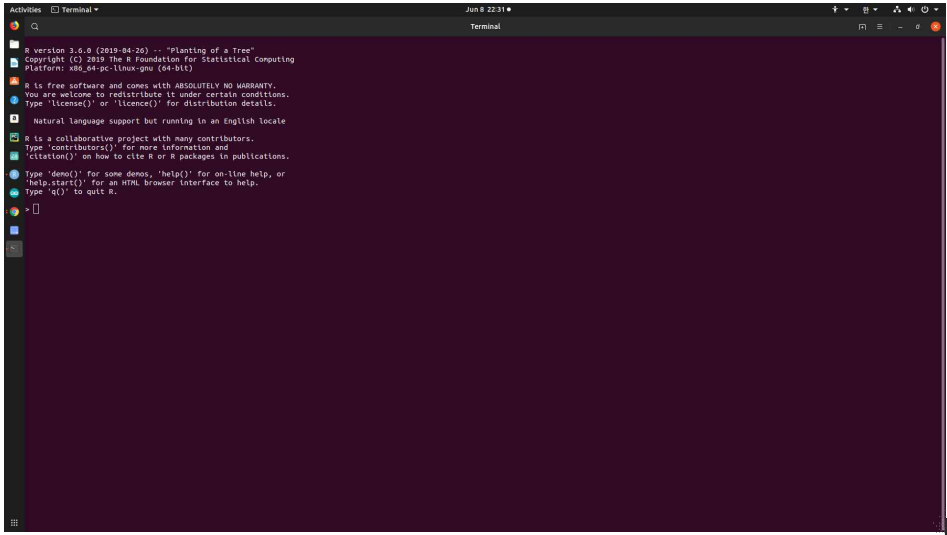
이를 위해 100번 중 95번을 성공한 사람을 기준으로 결정을 한다면?

1종 오류와 2종 오류


원 가설
효과가 있으면 차이가 있다
대우
동치 명제
귀무가설
차이가 없으면 효과가 없다
이
비동치 명제
대립가설
차이가 있으면 효과가 있다
귀무가설
대립가설
제2오류
제1오류

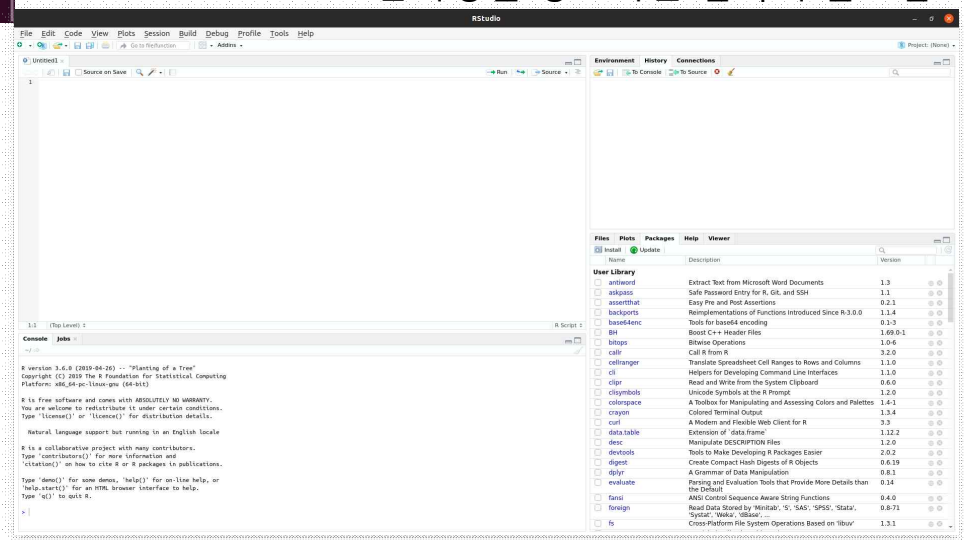
통계의 처리의 핵심 툴

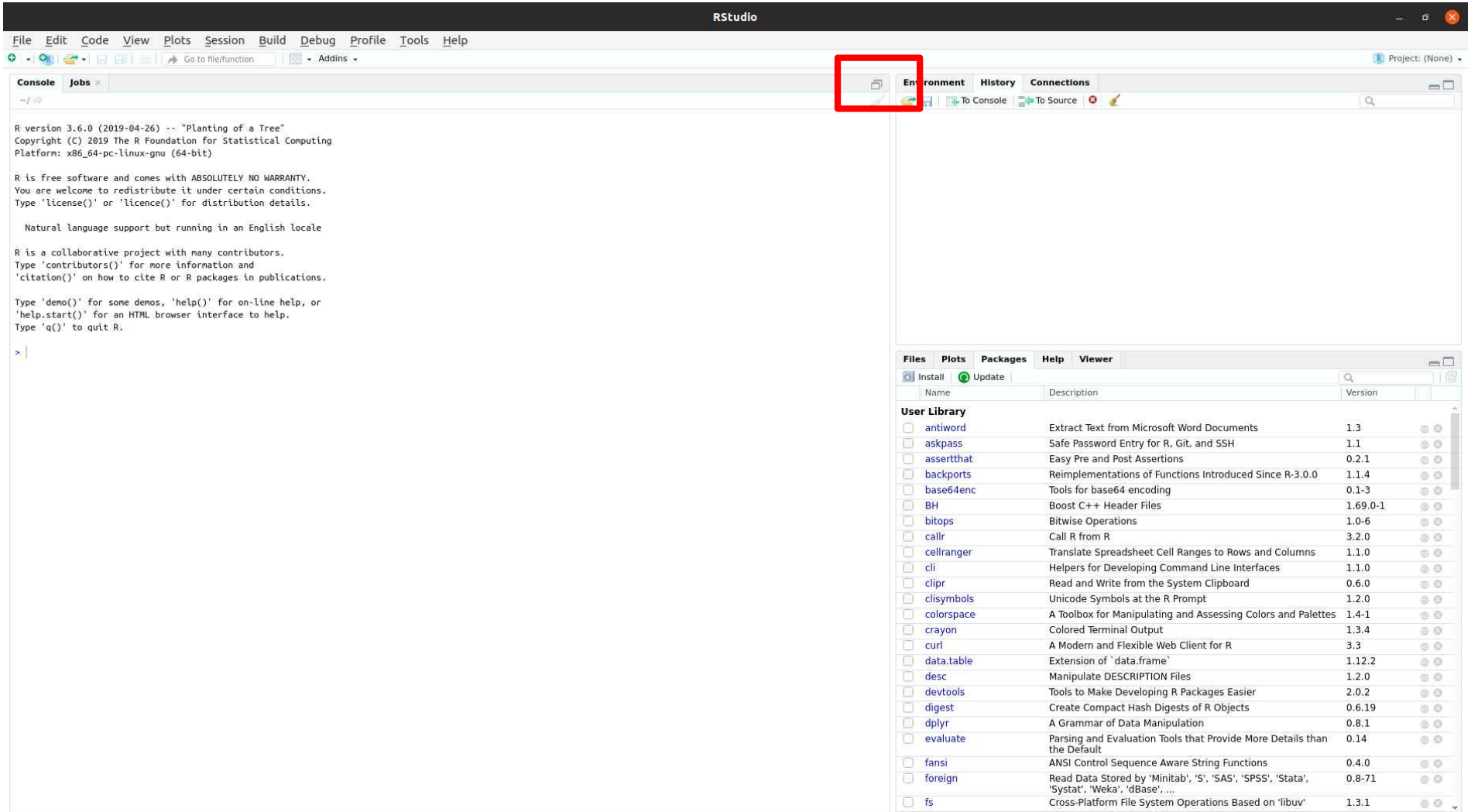
- **R Studio**

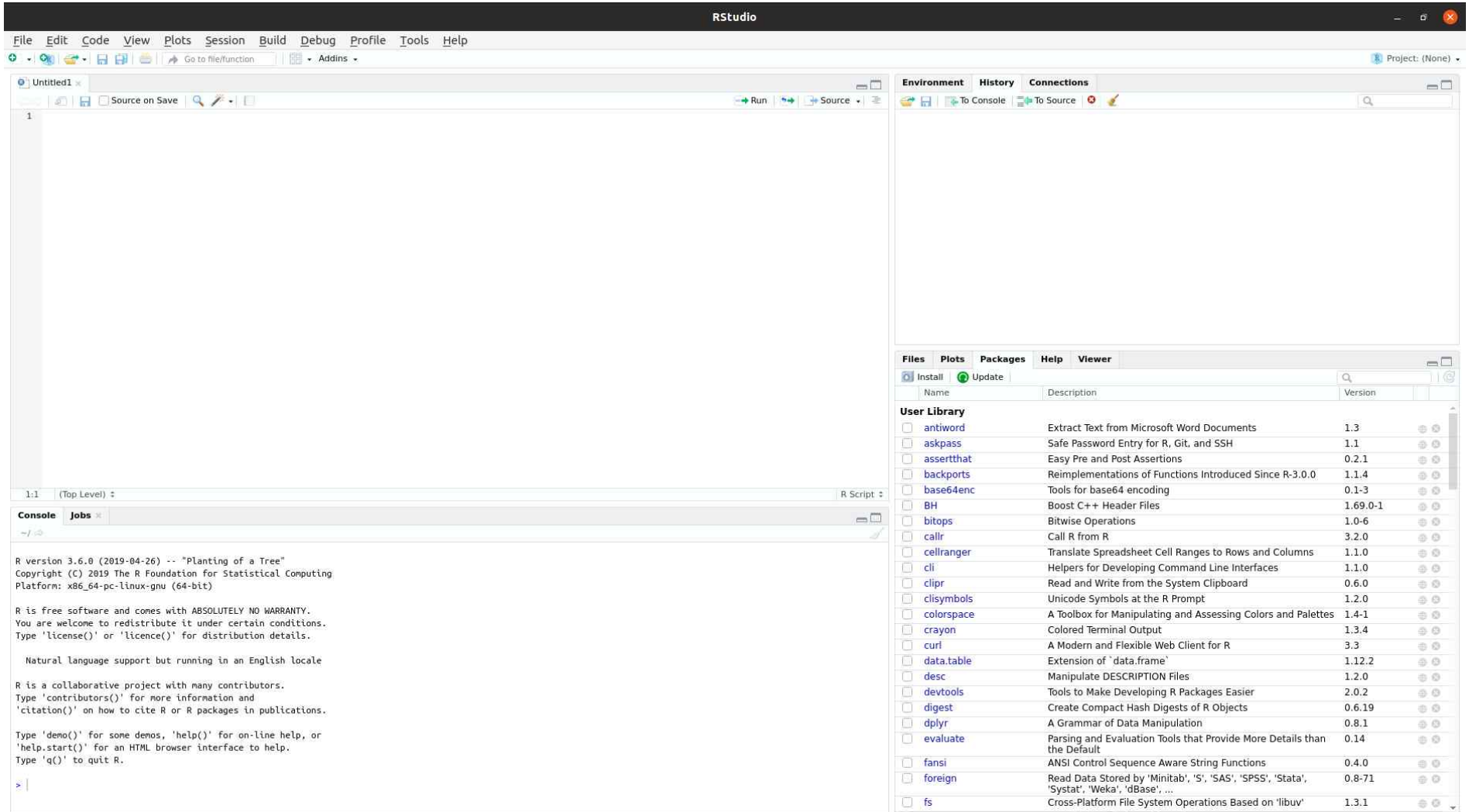


R을 실행시킨 모습 - 깔끔한 철학이 돋보임

R Studio를 실행시킨 모습
- 편리성을 강조하는 철학이 돋보임




RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Console Jobs
~/
R version 3.6.0 (2019-04-26) -- "Planting of a Tree"
Copyright (C) 2019 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
Natural language support but running in an English locale
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
>
Environment History Connections
To Console
To Source
Files Plots Packages Help Viewer
Install
Update
Name
Description
Version
User Library
antiword Extract Text from Microsoft Word Documents 1.3
askpass Safe Password Entry for R, Git, and SSH 1.1
assertthat Easy Pre and Post Assertions 0.2.1
backports Reimplementations of Functions Introduced Since R-3.0.0 1.1.4
base64enc Tools for base64 encoding 0.1-3
BH Boost C++ Header Files 1.69.0-1
bitops Bitwise Operations 1.0-6
callr Call R from R 3.2.0
cellranger Translate Spreadsheet Cell Ranges to Rows and Columns 1.1.0
cli Helpers for Developing Command Line Interfaces 1.1.0
clipr Read and Write from the System Clipboard 0.6.0
clisymbols Unicode Symbols at the R Prompt 1.2.0
colorspace A Toolbox for Manipulating and Assessing Colors and Palettes 1.4-1
crayon Colored Terminal Output 1.3.4
curl A Modern and Flexible Web Client for R 3.3
data.table Extension of `data.frame` 1.12.2
desc Manipulate DESCRIPTION Files 1.2.0
devtools Tools to Make Developing R Packages Easier 2.0.2
digest Create Compact Hash Digests of R Objects 0.6.19
dplyr A Grammar of Data Manipulation 0.8.1
evaluate Parsing and Evaluation Tools that Provide More Details than the Default 0.14
fansi ANSI Control Sequence Aware String Functions 0.4.0
foreign Read Data Stored by 'Minitab', 'S', 'SAS', 'SPSS', 'Stata', 'Systat', 'Weka', 'dBase', ... 0.8-71
fs Cross-Platform File System Operations Based on 'libuv' 1.3.1

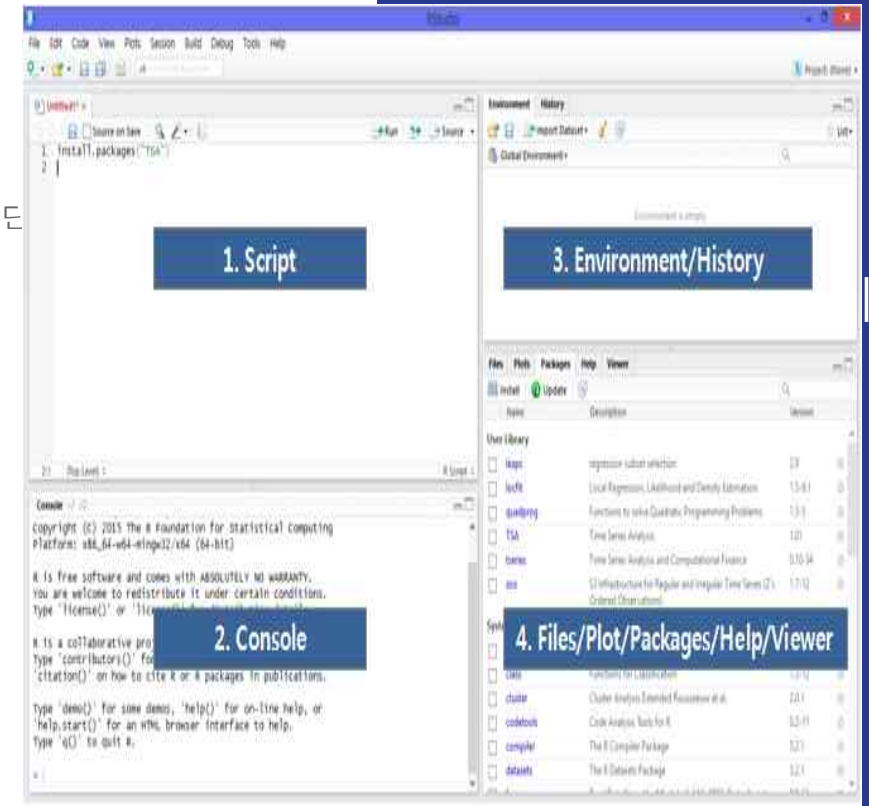

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Untitled1
Source on Save
Run
Source
1
1:1
(Top Level)
R Script
Console
Jobs
~/
R version 3.6.0 (2019-04-26) -- "Planting of a Tree"
Copyright (C) 2019 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
Natural language support but running in an English locale
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
>
Environment
History
Connections
To Console
To Source
Files
Plots
Packages
Help
Viewer
Install
Update
Name
Description
Version
User Library
antiword Extract Text from Microsoft Word Documents 1.3
askpass Safe Password Entry for R, Git, and SSH 1.1
assertthat Easy Pre and Post Assertions 0.2.1
backports Reimplementations of Functions Introduced Since R-3.0.0 1.1.4
base64enc Tools for base64 encoding 0.1-3
BH Boost C++ Header Files 1.69.0-1
bitops Bitwise Operations 1.0-6
callr Call R from R 3.2.0
cellranger Translate Spreadsheet Cell Ranges to Rows and Columns 1.1.0
cli Helpers for Developing Command Line Interfaces 1.1.0
clipr Read and Write from the System Clipboard 0.6.0
clisymbols Unicode Symbols at the R Prompt 1.2.0
colorspace A Toolbox for Manipulating and Assessing Colors and Palettes 1.4-1
crayon Colored Terminal Output 1.3.4
curl A Modern and Flexible Web Client for R 3.3
data.table Extension of `data.frame` 1.12.2
desc Manipulate DESCRIPTION Files 1.2.0
devtools Tools to Make Developing R Packages Easier 2.0.2
digest Create Compact Hash Digests of R Objects 0.6.19
dplyr A Grammar of Data Manipulation 0.8.1
evaluate Parsing and Evaluation Tools that Provide More Details than the Default 0.14
fansi ANSI Control Sequence Aware String Functions 0.4.0
foreign Read Data Stored by 'Minitab', 'S', 'SAS', 'SPSS', 'Stata', 'Systat', 'Weka', 'dBase', ... 0.8-71
fs Cross-Platform File System Operations Based on 'libuv' 1.3.1

1. 스크립트 창

R 코드를 작성하고 실행할 수 있는 창

(1) Ctrl + R or
(2) Ctrl + Enter or
(3) 블럭설정 후 RStudio Script 창의 상단

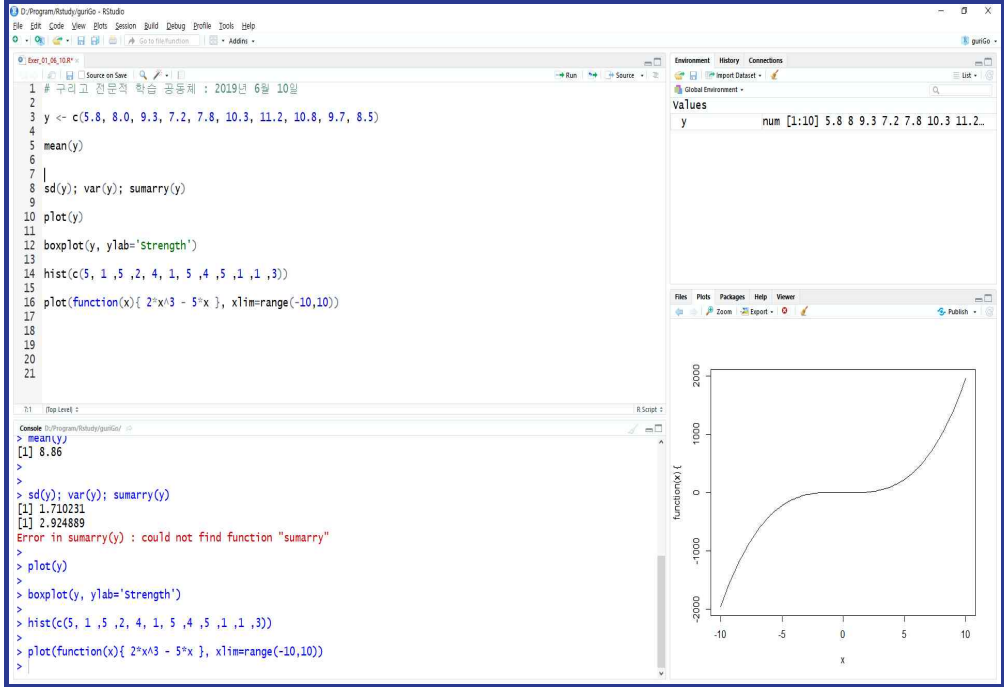
클릭



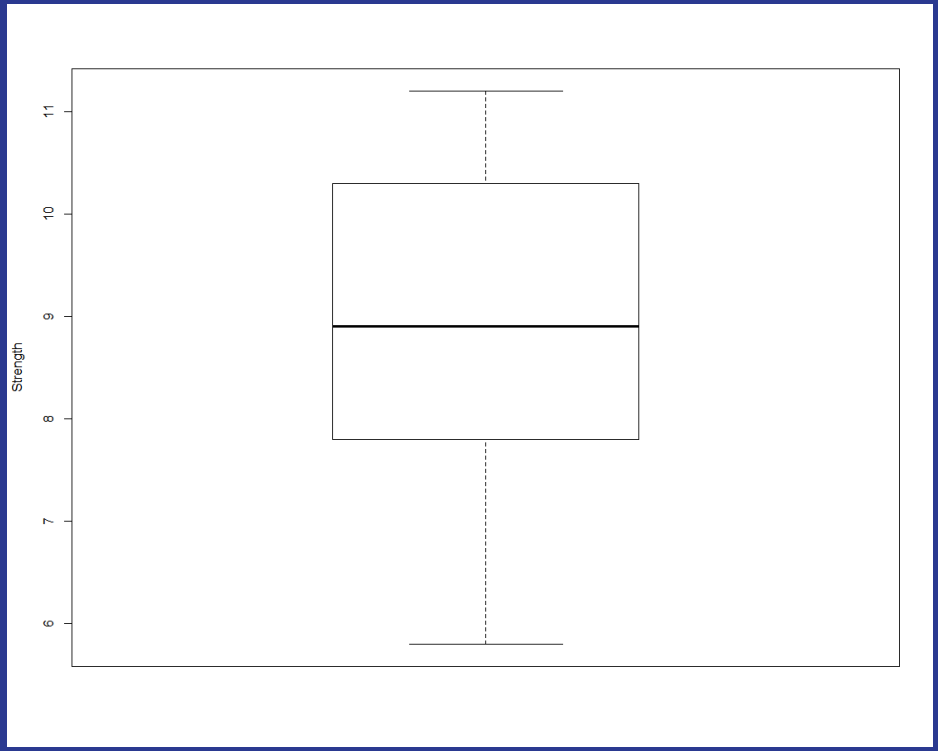
오

사용해보기

```
y <- c(5.8, 8.0, 9.3, 7.2, 7.8, 10.3, 11.2, 10.8, 9.7, 8.5)

mean(y)


sd(y); var(y); sumarry(y)

plot(y)

boxplot(y, ylab='Strength')

hist(c(5, 1 ,5 ,2, 4, 1, 5 ,4 ,5 ,1 ,1 ,3))

plot(function(x){ 2*x^3 - 5*x }, xlim=range(-10,10))
```



```
#====================================
# t 검정하기
# 얻어진 값을 평균 7과 비교하기
#====================================
data <- c(5.8, 8.0, 9.3, 7.2, 7.8, 10.3, 11.2, 10.8, 9.7, 8.5)
summary(data)
boxplot(data, ylab='Strength')
t.test(data, mu=7.0, alt='two.sided') # 양측(같다고 할 수 있는가?)
t.test(data, mu=7.0, alt='greater') # 단측(크다고 할 수 있는가?)
t.test(data, mu=7.0, alt='less') # 단측(작다고 할 수 있는가?)
```



생물학적 맥락에서 이해하기

Panama Gold의 경우 수확을 위해 요구되는 최적의 수분량은 평균 수분함유량은 50g/kg이라고 알려져 있다.

> 수확 여부를 결정하기 위해서는 농작물의 수분 함유량이 최적의 수분 함유량과 비교하여야 한다.

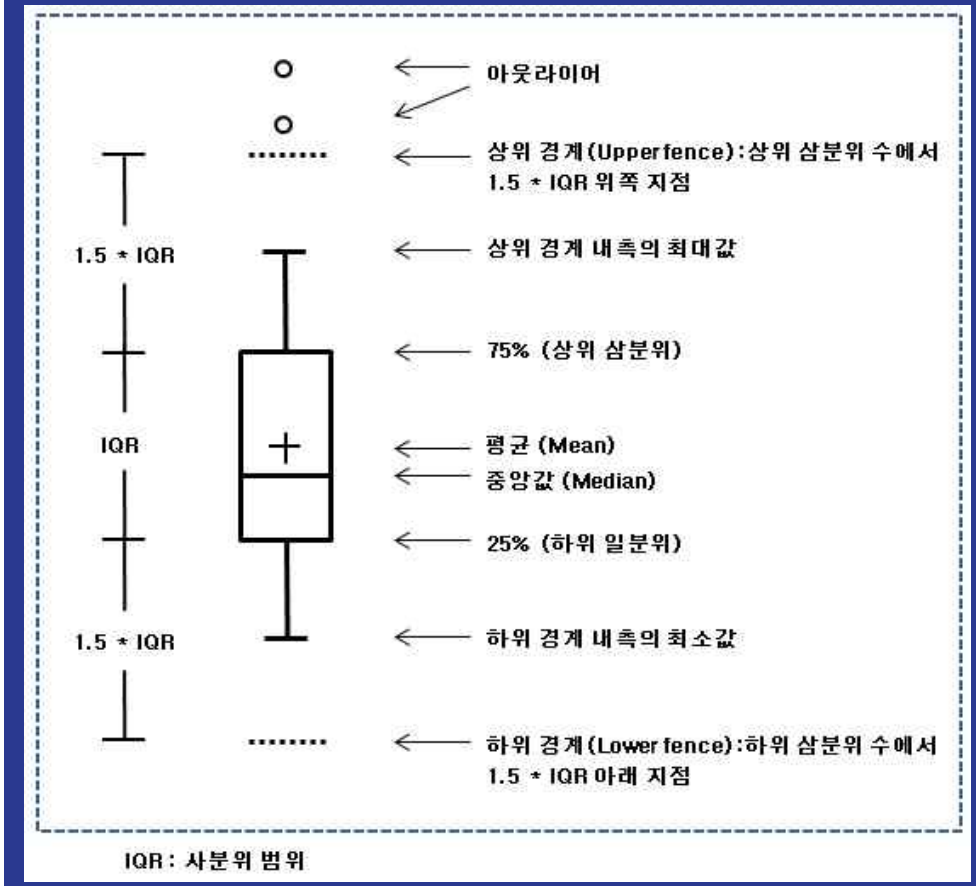
> 이를 위해서 임의로 1.0kg을 반복하여 아홉번씩 수확한 후 수분 함유량을 측정하였다.
(얻어진 데이터 44, 42, 43, 49, 43, 47, 45, 46, 43g/kg)

귀무 가설: 50g/kg과 차이가 없다.
대립 가설:차이가 있다.

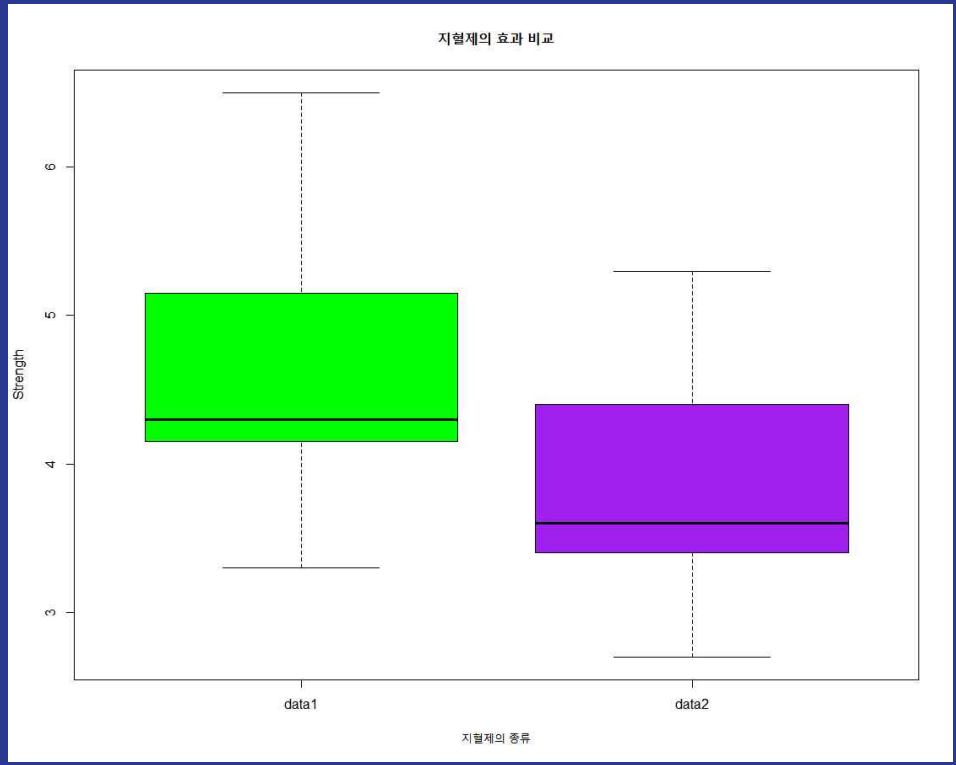
https://www.piqsels.com/en/search?q=straw&page=76

BoxPlot(상자도표)

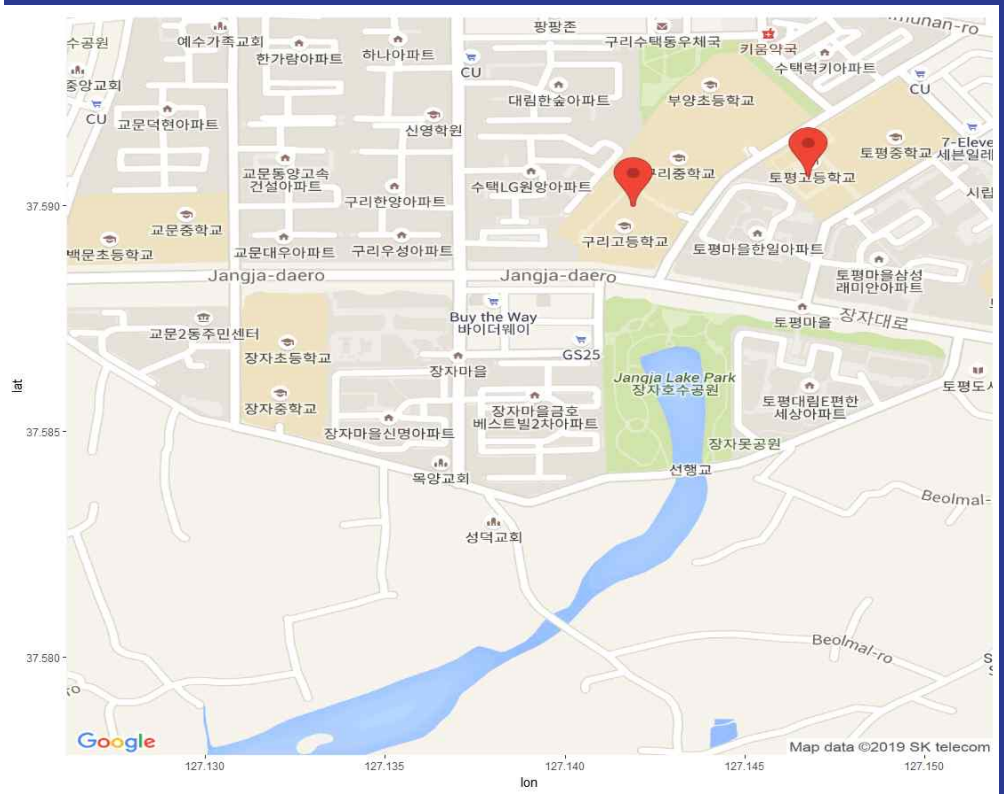
부제목을 입력하십시오



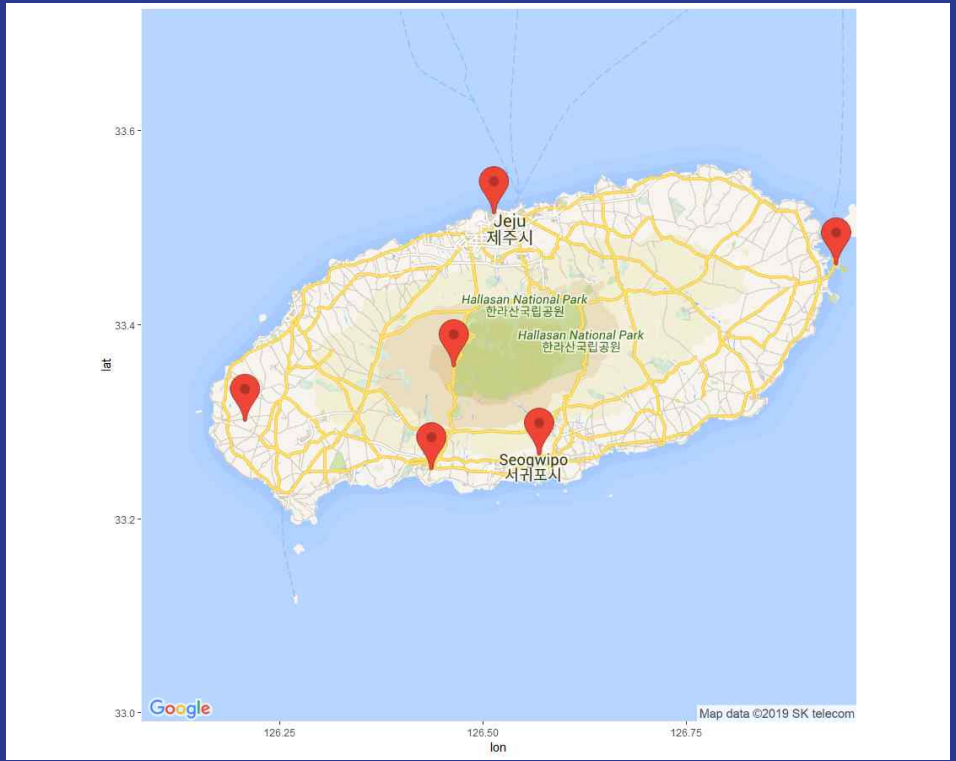
```
#====================================
# t 검정하기
# 얻어진 값을 비교하기
#====================================

data1 <- c(4.1, 3.3, 4.3, 4.2, 5.3, 6.5, 5.0)
data2 <- c(3.3, 4.3, 3.5, 3.6, 5.3, 2.7, 4.5)

ex_data <- data.frame(data1, data2)
summary(ex_data)

boxplot(ex_data, xlab='지혈제의 종류', ylab='Strength',
main = '지혈제의 효과 비교')
boxplot(ex_data, xlab='지혈제의 종류', ylab='Strength',
main = '지혈제의 효과 비교', col = c('green','purple'))

t.test(data1, data2, var.equal=TRUE,
alternative='two.sided')
```



```
#====================================
# 구글 지도를 이용해보기
#====================================

library(ggmap)
register_google(key='AIzaSyC2T3t4WwrJI7AWt1TWNZkyiHRKjSbkbJ4')

gc <- geocode(enc2utf8("경기도 구리시 수택동 785"))

cen <- as.numeric(gc) # in numbers

map <- get_googlemap(center=cen, zoom=16,
                     size=c(1200,960), maptype="roadmap")


ggmap(map) # on the map library(ggmap)
```

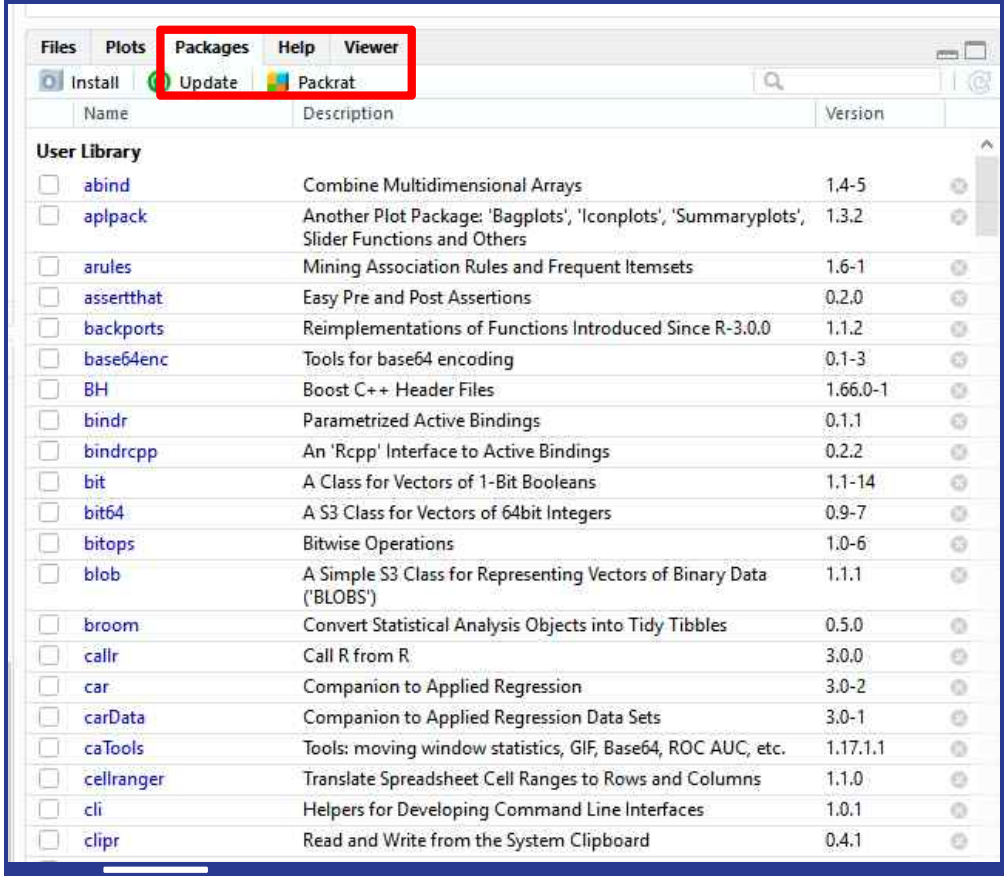


```
#====================================
# 구글 지도를 이용해보기
#====================================
register_google(key='AIzaSyC2T3t4WwrJI7AWt1TWNZkyiHRKjSbkbJ4') # 부여받은 키 등록

names <- c("용두암","성산일출봉","정방폭포", "중문관광단지","한라산1100고지","차귀도")

addr <- c("제주시 용두암길 15", "서귀포시 성산읍 성산리 174-10", "서귀포시 동홍동 299-3",
          "서귀포시 중문동 2624-1", "서귀포시 색달동 산1-2", "제주시 한경면 고산리 125")

gc <- geocode(enc2utf8(addr))

gc

df <- data.frame(name=names, lon=gc$lon, lat=gc$lat)
cen <- c(mean(df$lon),mean(df$lat))
map <- get_googlemap(center=cen,
maptype="roadmap", zoom=10, size=c(640,640),
                     marker=gc)
ggmap(map)
```

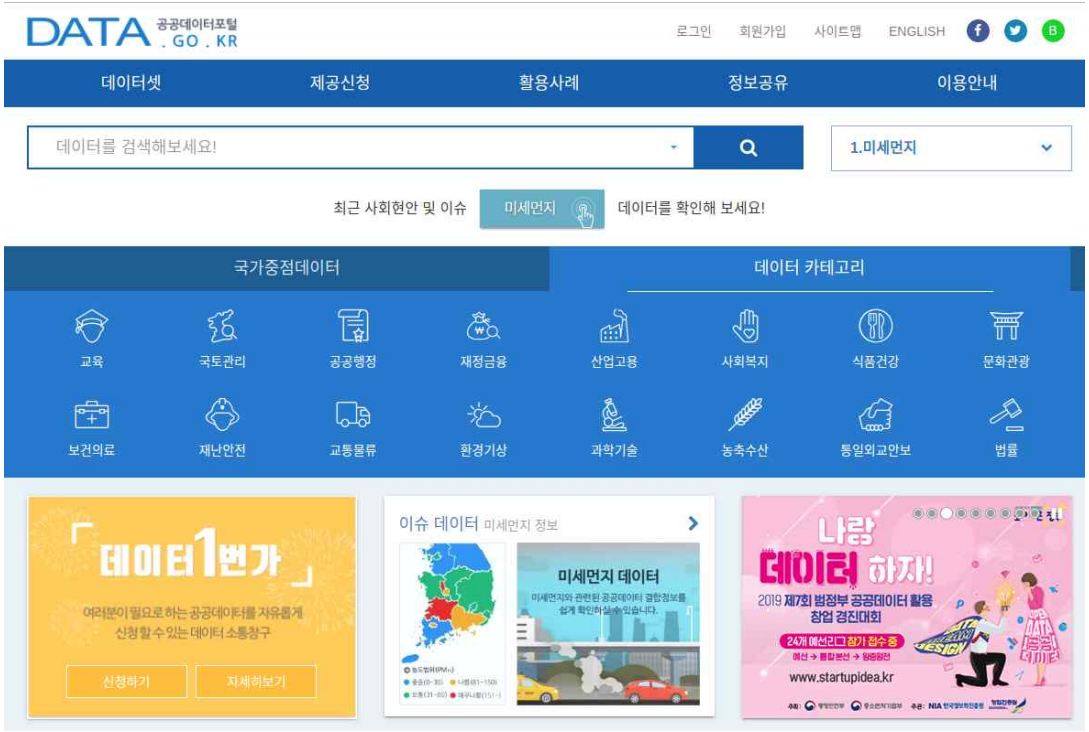


사용해보기 2 - 패키지 설치

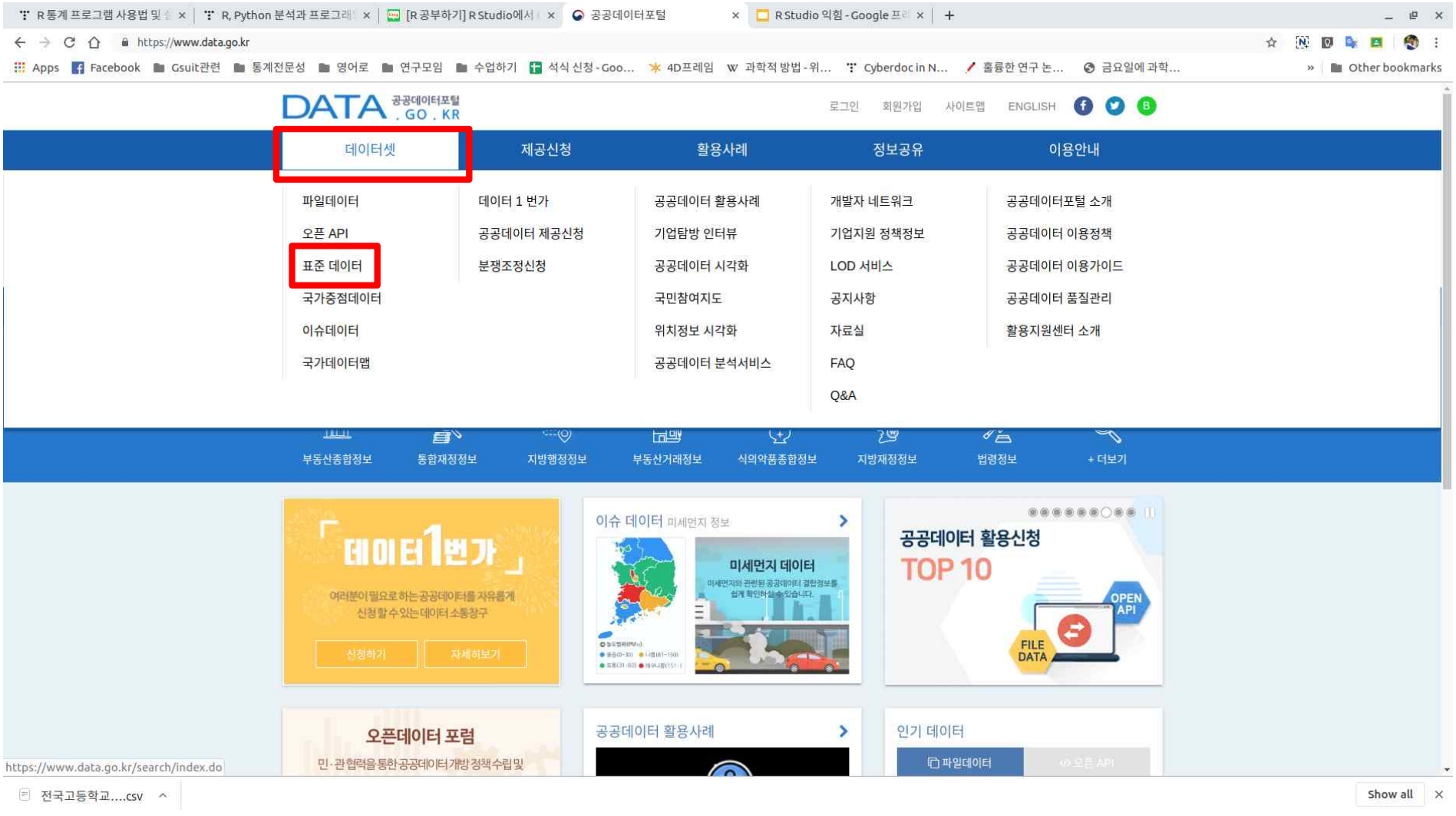
오른쪽 아래



공공데이터 포털에서 자료 가지고 오기



- 행정안전부에서 운영하는 공공데이터 통합제공 시스템
- 대한민국 정부가 보유한 다양한 공공데이터를 개방하여 누구나 편리하고 손쉽게 활용할 수 있게 하는 것을 목적으로 함


R통계 프로그램 사용법 및
R, Python 분석과 프로그래
[R 공부하기] R Studio에서
공공데이터포털
R Studio 익힘 - Google 프
https://www.data.go.kr
Apps
Facebook
Gsuit관련
통계전문성
영어로
연구모임
수업하기
석식 신청 - Goo...
4D프레임
과학적 방법 - 위...
Cyberdoc in N...
훌륭한 연구 논...
금요일에 과학...
Other bookmarks
DATA 공공데이터포털 .GO.KR
로그인
회원가입
사이트맵
ENGLISH
데이터셋
제공신청
활용사례
정보공유
이용안내
파일데이터
오픈 API
표준 데이터
국가중점데이터
이슈데이터
국가데이터맵
데이터 1 번가
공공데이터 제공신청
분쟁조정신청
공공데이터 활용사례
기업탐방 인터뷰
공공데이터 시각화
국민참여지도
위치정보 시각화
공공데이터 분석서비스
개발자 네트워크
기업지원 정책정보
LOD 서비스
공지사항
자료실
FAQ
Q&A
공공데이터포털 소개
공공데이터 이용정책
공공데이터 이용가이드
공공데이터 품질관리
활용지원센터 소개
부동산종합정보
통합재정정보
지방행정정보
부동산거래정보
식의약품종합정보
지방재정정보
법령정보
+ 더보기
데이터1번가
여러분이 필요로 하는 공공데이터를 자유롭게
신청할 수 있는 데이터 소통창구
신청하기
자세히보기
이슈 데이터 미세먼지 정보
미세먼지 데이터
공공데이터 활용신청
TOP 10
OPEN API
FILE DATA
오픈데이터 포럼
민·관 협력을 통한 공공데이터 개방 정책 수립 및
공공데이터 활용사례
인기 데이터
파일데이터
https://www.data.go.kr/search/index.do
전국고등학교....csv
Show all

통계의 처리 연습하기

- EXCEL

가설 - 예측

> 가설

관찰 또는 표본 추출로부터 얻어진 정보를 이용하여 시스템이 어떤 기능을 하는지에 대해 제시한 직관적이고 논리적인 추측(가추 또는 귀추적 사고, 퍼스)

> 예측

가설로부터 논리적으로 이끌어진 추측(prediction)


Activities
Jun 8 22:32
Type to search...
LibreOffice W...
Livepatch
Lucidchart Dia...
Lucidchart Dia...
Mahjongg
Mines
ONLYOFFICE D...
Passwords an...
Power Statistics
PyCharm Com...
R
Remmina
Rhythmbox
RStudio
Settings
Shotwell
Simple Scan
SimpleScreen...
Software & Up...
Software Upd...
Startup Applic...
Startup Disk C...
Sudoku
Sundry
Frequent
All