

2022년 한국생명공학연구원 연구데이터
분석과정 R

합성생물학전문연구소 김하성

2022-10-24

Contents

Chapter 1

Introduction

1.1 강의 개요

- 목표: 생물 데이터 분석을 위한 R 사용법과 (Rstudio, Tidyverse, Bioconductor 포함) 프로그래밍 기술을 습득함
- 장소: 코빅 3층 전산교육장(1304호)
- 강사: 한국생명공학연구원 합성생물학전문연구단 김하성
- 연락처: 042-860-4372, haseong@kribb.re.kr
- 강의자료: <https://greendaygh.github.io/kribbr2022/>
- 강의관련 게시판: <https://github.com/greendaygh/kribbr2022/issues>

1.2 강의 계획

1. (05.19, 13:30~17:30) [R/Rstudio] 사용법 및 데이터 구조
2. (05.26, 13:30~17:30) [R/Rstudio] 프로그래밍 기초
3. (06.09, 13:30~17:30) [R/Tidyverse] 데이터 분석 기초
4. (06.16, 13:30~17:30) [R/Tidyverse] 데이터 분석 중급
5. (07.07, 13:30~17:30) [R/Tidyverse] 데이터 가시화
6. (07.14, 13:30~17:30) [R/Tidyverse] 데이터 가시화 활용
7. (08.04, 13:30~17:30) [R/Bioconductor] 바이오 데이터 분석 기초
8. (08.11, 13:30~17:30) [R/Bioconductor] 서열 비교 및 계통 분석
9. (09.01, 13:30~17:30) [R/Bioconductor] 지놈 스케일 바이오 데이터 분석
10. (09.15, 13:30~17:30) [R/Bioconductor] NGS 데이터 소개 및 통계 기초
11. (10.06, 13:30~17:30) [R/Bioconductor] NGS 기반 Differentially Expressed Genes 분석
12. (10.20, 13:30~17:30) [R/Bioconductor] NGS 기반 Gene Set Enrichment Analysis

1.3 참고 자료

- R 홈페이지
- Rstudio 홈페이지
- Bioconductor
- R 기본 문서들
- R ebooks
- Cheat Sheets
- RStudio Webinars
- Shiny
- Hadley github
- R for Data Science
- Using R for Introductory Statistics by John Verzani
 - Free version of 1st Edition
 - Second edition
- Bioinformatics Data Skills by Vince Buffalo
- Introductory Statistics with R by Dalgaard
- Modern Statistics for Modern Biology
- bios221
- Annotation Workshop 2021
- CSAMA 2022
- RNA-Seq CSAMA 2022
- Annotation_Resources 2015
- Sequence analysis
- 일반통계학 (영지문화사, 김우철 외)

Chapter 2

R/Rstudio basics

2.1 What is R / Rstudio



R은 통계나 생물통계, 유전학을 연구하는 사람들 사이에서 널리 사용되는 오픈소스 프로그래밍 언어입니다. Bell Lab에서 개발한 S 언어에서 유래했으며 많은 라이브러리 (다른 사람들이 만들어 놓은 코드)가 있어서 쉽게 가져다 사용할 수 있습니다. R은 복잡한 수식이나 통계 알고리즘을 간단히 구현하고 사용할 수 있으며 C, C++, Python 등 다른 언어들과의 병행 사용도 가능합니다. R은 IEEE에서 조사하는 Top programming languages에서 2018년 7위, 2019년 5위, 2020년 6위, 2021년 7위로 꾸준히 높은 사용자를 확보하며 빅데이터, AI 시대의 주요한 프로그래밍 언어로 사용되고 있습니다.

R은 데이터를 통계분석에 널리 사용되는데 이는 데이터를 눈으로 확인하기 위한 visualization이나 벡터 연산 등의 강력한 기능 때문에 점점 더 많은 사람들이 사용하고 있습니다. 기존에는 속도나 확장성이 다른 언어들에 비해 단점으로 지적되었으나 R 언어의 계속적인 개발과 업데이트로 이러한 단점들이 빠르게 보완되고 있습니다. R 사용을 위해서는 R 언어의 코어 프로그램을 먼저 설치하고 그 다음 R 언어용 IDE(Integrated Development Environment)인 RStudio 설치가 필요합니다.



Rstudio는 R 언어를 위한 오픈소스 기반 통합개발환경(IDE)으로 R 프로그래밍을 위한 편리한 기능들을 제공해 줍니다. R언어가 주목을 받고 두터운 사용자 층을



Figure 2.1: <https://spectrum.ieee.org/top-programming-languages/>

확보할 수 있게된 핵심 동력이 Rstudio입니다. 자체적으로 최고수준의 오픈소스 개발팀이 있으며 tidyverse, 'shiny' 등의 데이터 분석 관련 주요 패키지를 개발하였고 정기적으로 conference 개최를 하면서 기술 보급의 핵심 역할을 하고 있습니다.

2.2 R / Rstudio Installation

2.2.1 R 설치

- R 사이트에 접속 후 (<https://www.r-project.org/>) 좌측 메뉴 상단에 위치한 CRAN 클릭.
- 미러 사이트 목록에서 Korea의 아무 사이트나 들어감
- Download R for Windows를 클릭 후 base 링크 들어가서
- Download R x.x.x for Windows 링크 클릭으로 실행 프로그램 다운로드
- 로컬 컴퓨터에 Download 된 R-x.x.x-win.exe 를 실행 (2022.5 현재 R 버전은 4.2.0).
- 설치 프로그램의 지시에 따라 R 언어 소프트웨어 설치를 완료

2.2.2 Rstudio 설치

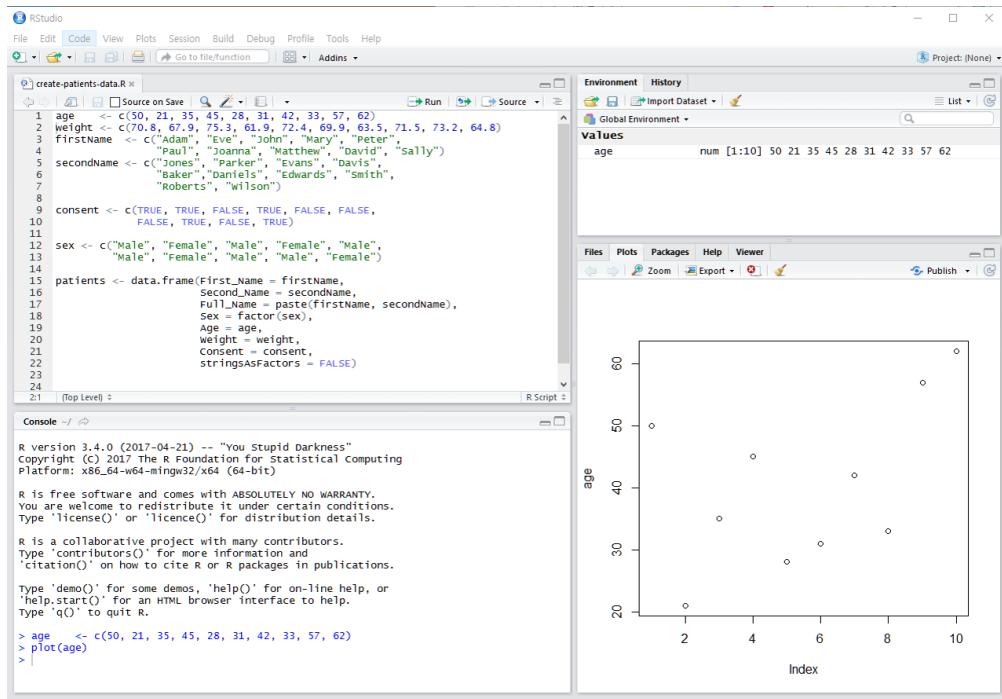
- 사이트에 접속 (<https://www.rstudio.com/>), 상단의 Products > RStudio 클릭
- RStudio Desktop 선택
- Download RStudio Desktop 클릭

Products	About RStudio		Additional Websites	
OPEN SOURCE	LEARNING	SUPPORT	ANALYSE & EXPLORE	CONNECT & INTEGRATE
RStudio Desktop	Education	Frequently Asked Questions	Tidyverse	Professional Drivers
RStudio Server	Videos & Webinars	RStudio Support	ggplot2	Launcher Plugin SDK
Shiny Server	Cheatsheets	RStudio Community	dplyr	Databases
R Packages	rstudio::conf	Certified Partners	tidy	Environments
HOSTED SERVICES	ABOUT US	Product Security	purrr	Sparklyr
RStudio Academy	About the Company	RStudio Documentation	COMMUNICATE & INTERACT	Plumber
RStudio Cloud	What Makes Us Different	Contact Us	Shiny	Reticulate
RStudio Public Package Manager	Analyst Reports	RStudio Legal Terms	rmarkdown	Ursa Labs
shinyapps.io	RStudio Swag	Email Subscription Management	flexdashboard	MODEL & PREDICT
PROFESSIONAL	Careers			Tensorflow
RStudio Team				Tidymodels
RStudio Workbench				Spark MLlib
RStudio Connect				
RStudio Package Manager				

Figure 2.2: <https://www.rstudio.com/>

- RStudio Desktop Free 버전의 Download를 선택하고
- Download RStudio for Windows 클릭, 다운로드
- 로컬 컴퓨터에 다운로드된 RStudio-x.x.x.exe 실행 (2022.5 현재 RStudio Desktop 2022.02.2+485)
- 설치 가이드에 따라 설치 완료

2.3 Rstudio interface



- 기본 화면에서 좌측 상단의 공간은 코드편집창, 좌측 하단은 콘솔창
- 각 위치를 기호에 따라서 바꿀 수 있음 (View -> Pane)

2.3.1 Keyboard shortcuts

- 참고사이트
 - <https://support.rstudio.com/hc/en-us/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)
- 저장 (Ctrl + S)
- 주석처리 (Ctrl + Shift + C)
 - 또는 #으로 시작하는 라인
- 템 이동 (Ctrl + F11, Ctrl + F12)
- 코드편집창 확대 (Shift + Ctrl + 1) 콘솔창 확대 (Shift + Ctrl + 2)
- 커럼 편집 (Alt +)
- 자동 완성 기능 (Tab completion) in RStudio

Exercises

1. 코드편집창에서 다음을 입력/실행하고 단축키를 사용하여 주석을 넣으시오

```

1 x <- 10
2 y <- 20
3

```

- 단축키 Ctrl + enter로 코드 실행
- 단축키 Ctrl + 2로 커서 콘솔창으로 이동
- x값 x+y값 확인
- 단축키 Ctrl + 1로 코드편집창 이동
- 단축키 Ctrl + Shift + C 사용

```
# x <- 10
# y <- 20
```

2.3.2 Environment and Files

Values	x	10
y		20

Name	Size	Modified
..		
_book	361 B	May 19, 2022, 7:18 AM
_bookdown.yml	154 B	May 18, 2022, 5:47 PM
_common.R	274 B	May 18, 2022, 5:47 PM
_output.yml	40 B	May 19, 2022, 7:06 AM
.gitignore	8.1 KB	May 19, 2022, 7:33 AM
01-Rintro.Rmd	4.4 KB	May 19, 2022, 7:19 AM
02-Rprogramming.Rmd	525 B	May 18, 2022, 5:47 PM
03-parts.Rmd	1 KB	May 18, 2022, 5:47 PM
04-citations.Rmd	1.2 KB	May 18, 2022, 5:47 PM
05-blocks.Rmd	1.2 KB	May 18, 2022, 5:47 PM
06-share.Rmd		

2.4 Start a project

프로젝트를 만들어서 사용할 경우 파일이나 디렉토리, 내용 등을 쉽게 구분하여 사용 가능합니다. 아래와 같이 임의의 디렉토리에 `kribbR`이라는 디렉토리를 생성하고 `lecture1` 프로젝트를 만듭니다.

File > New Project > New Directory > New Project > “`kribbR`” > Create Project

시작할 때는 해당 디렉토리의 `xxx.Rproj` 파일을 클릭합니다. Rstudio 오른쪽 상단 프로젝트 선택을 통해서 빠르게 다른 프로젝트의 작업공간으로 이동할 수 있습니다.

2.4.1 Hello world

File > New File > R markdown > OK

```
mystring <- "Hello \n world!"  
cat(mystring)  
print(mystring)
```

2.5 Getting help

R은 방대한 양의 도움말 데이터를 제공하며 다음과 같은 명령어로 특정 함수의 도움말과 예제를 찾아볼 수 있습니다. `?` 명령을 사용하면 되며 구글이나 웹에서도 도움을 얻을 수 있습니다.

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

또한 <https://www.rstudio.com/resources/cheatsheets/> 에서는 다양한 R언어의 기능을 한 눈에 알아볼 수 있게 만든 cheatsheet 형태의 문서를 참고할 수 있습니다.

Base R Cheat Sheet

Getting Help

`?mean`
Get help of a particular function.
`help.search('weighted mean')`
Search the help files for a word or phrase.
`help(package = 'dplyr')`
Find help for a package.

[More about an object](#)

`str(iris)`
Get a summary of an object's structure.
`class(iris)`
Find the class an object belongs to.

Using Packages

`install.packages('dplyr')`
Download and install a package from CRAN.
`library(dplyr)`
Load the package into the session, making all its functions available to use.
`dplyr::select`
Use a particular function from a package.
`data(iris)`
Load a built-in dataset into the environment.

Working Directory

`getwd()`
Find the current working directory (where inputs are found and outputs are sent).
`setwd('C://file/path')`
Change the current working directory.

Use projects in RStudio to set the working directory to the folder you are working in.

Vectors		
Creating Vectors		
<code>c(2, 4, 6)</code>	2 4 6	Join elements into a vector
<code>2:6</code>	2 3 4 5 6	An integer sequence
<code>seq(2, 3, by=0.5)</code>	2.0 2.5 3.0	A complex sequence
<code>rep(1:2, times=3)</code>	1 2 1 2 1 2	Repeat a vector
<code>rep(1:2, each=3)</code>	1 1 1 2 2 2	Repeat elements of a vector

Vector Functions		
<code>sort(x)</code>	<code>rev(x)</code>	<code>unique(x)</code>
Return x sorted.	Return x reversed.	See unique values.
<code>table(x)</code>		See counts of values.

Selecting Vector Elements		
By Position		
<code>x[4]</code>	The fourth element.	
<code>x[-4]</code>	All but the fourth.	
<code>x[2:4]</code>	Elements two to four.	
<code>x[-(2:4)]</code>	All elements except two to four.	
<code>x[c(1, 5)]</code>	Elements one and five.	

By Value		
<code>x[x == 10]</code>	Elements which are equal to 10.	
<code>x[x < 0]</code>	All elements less than zero.	
<code>x[x %in% c(1, 2, 5)]</code>	Elements in the set 1, 2, 5.	

Named Vectors		
<code>x['apple']</code>	Element with name 'apple'.	

Programming		
For Loop		
<code>for (variable in sequence){ Do something }</code>		
Example		
<code>for (i in 1:4){ j <- i + 10 print(j) }</code>		

While Loop		
<code>while (condition){ Do something }</code>		
Example		
<code>while (i < 5){ print(i) i <- i + 1 }</code>		

If Statements		
<code>if (condition){ Do something } else { Do something different }</code>		
Example		
<code>if (i > 3){ print('Yes') } else { print('No') }</code>		

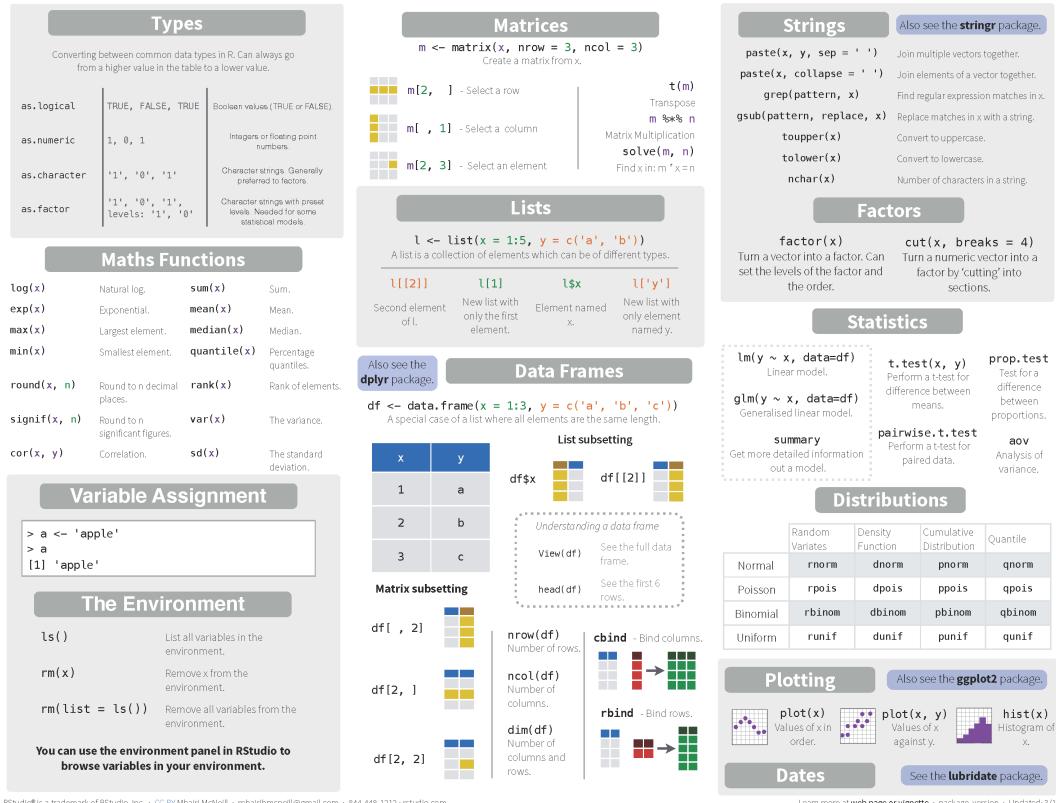
Functions		
<code>function_name <- function(var){ Do something return(new_variable) }</code>		
Example		
<code>square <- function(x){ squared <- x*x return(squared) }</code>		

Reading and Writing Data		
Also see the <code>readr</code> package.		
Input	Output	Description
<code>df <- read.table('file.txt')</code>	<code>write.table(df, 'file.txt')</code>	Read and write a delimited text file.
<code>df <- read.csv('file.csv')</code>	<code>write.csv(df, 'file.csv')</code>	Read and write a comma separated value file. This is a special case of <code>readable/writable</code> .
<code>load('file.RData')</code>	<code>save(df, file = 'file.Rdata')</code>	Read and write an R data file, a file type special for R.

Conditions		a == b	Are equal	a > b	Greater than	a >= b	Greater than or equal to	is.na(a)	Is missing
		a != b	Not equal	a < b	Less than	a <= b	Less than or equal to	is.null(a)	Is null

RStudio® is a trademark of RStudio, Inc. • CC BY Mhairi McNeill • mhairimcnell@gmail.com

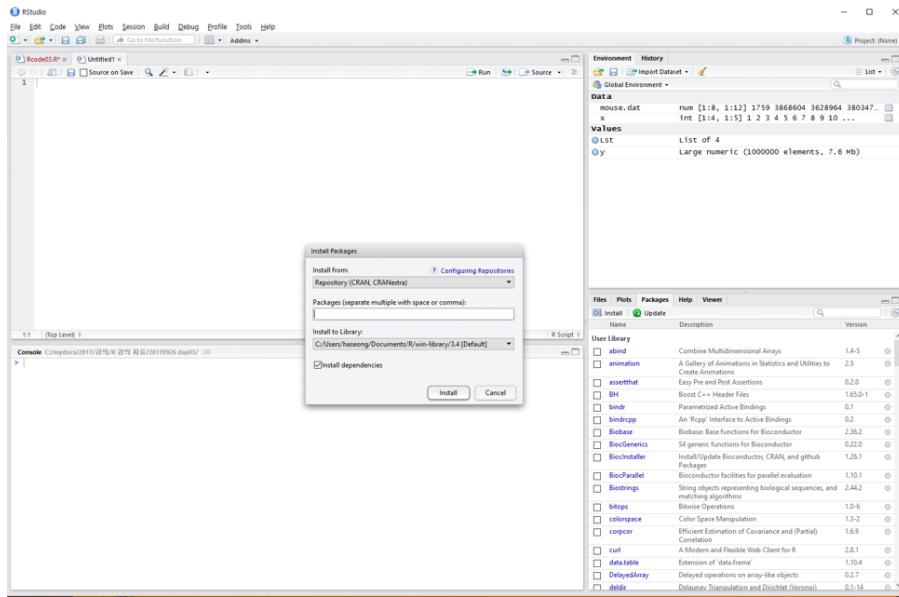
Learn more at [web page](#) or [vignette](#) • package version • Updated: 3/15



2.6 R packages and Dataset

R 패키지는 함수와 데이터셋의 묶음으로 다른 사람들이 만들어 놓은 코드나 기능을 가져와서 사용하므로써 코드 작성의 수고로움을 줄이고 편리하고 검증된 함수(기능)를 빠르게 도입하여 사용할 수 있다는 장점이 있습니다. 예를 들어 `sd()` 함수는 stats package에서 제공하는 함수로써 표준편차 계산을 위한 별도의 함수를 만들어서 사용할 필요가 없이 바로 (stats 패키지는 R 기본 패키지로) 별도 설치 없이 바로 사용 가능합니다.

이러한 패키지는 인터넷의 repository에서 구할 수 있으며 대표적인 repository는 The Comprehensive R Archive Network (CRAN) (<http://cran.r-project.org/web/views/>) 와 생물학자를 위한 Bioconductor (<http://www.bioconductor.org/>) 가 있습니다. 이러한 패키지의 설치는 아래와 같이 RStudio를 이용하거나 콘솔창에서 `install.packages()` 함수를 이용할 수 있습니다.



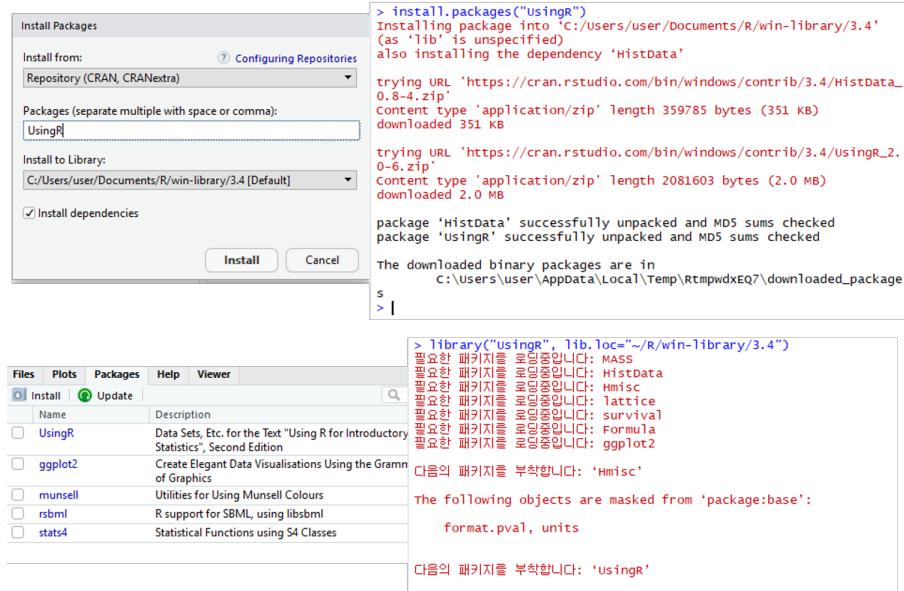
패키지를 설치하고 사용하기 위해서는 `library()` 함수를 사용해서 관련 명령어를 사용하기 전에 미리 `loading` 해 두어야 합니다. 한 번 로딩으로 작업 세션이 끝날때까지 관련된 함수를 사용할 수 있으나 R 세션이나 RStudio를 재시작 할 경우 다시 로딩해야 사용할 수 있습니다.

```
library(UsingR)
```

- R 설치 디렉토리
- R 패키지 설치 디렉토리

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

Packages → Install



일반적으로 패키지 안에 관련된 데이터도 같이 저장되어 있으며 `data()` 함수를 이용해서 패키지 데이터를 사용자 작업공간에 복사해서 사용 가능합니다.

```
head(rivers)
length(rivers)
class(rivers)
data(rivers)
data(package="UsingR")
library(HistData)
head(Cavendish)
str(Cavendish)
head(Cavendish$density2)
```

이 저작물은 크리에이티브 커먼즈 저작자표시-비영리-변경금지 4.0 국제 라이선스에 따라 이용할 수 있습니다.

Chapter 3

Rmarkdown

Rmarkdown은 데이터를 분석하는 코드와 리포트를 동시에 수행할 수 있는 일종의 통합 문서입니다. 워드나 아래한글에서 프로그래밍과 데이터분석을 위한 코드를 작성할 수 있는 경우라고 생각해도 됩니다. Plain-text 기반의 markdown 문법을 사용하며 Rmarkdown으로 작성된 문서는 HTML, PDF, MS word, Beamer, HTML5 slides, books, website 등 다양한 포맷의 출력물로 변환할 수 있습니다.



Figure 3.1: Image from rmarkdown.rstudio.com

Rmarkdown 웹사이트에 Rmarkdown 소개 동영상과 Rmarkdown 공식 사이트 메뉴얼 관련 서적 Rmarkdown: The Definitive Guide를 참고하세요. 또한 Rmarkdown을 사용할 때 cheatsheet를 옆에 두고 수시로 보면서 사용하시면 많은 도움이 될 수 있습니다.

3.1 Rmarkdown의 기본 작동 원리

Rmarkdown은 plain text 기반으로 작성되며 Rmd라는 확장자를 갖는 파일로 저장됩니다. 다음과 같은 텍스트 파일이 Rmd 파일의 전형적인 예입니다.

```

---  

title: "Viridis Demo"  

output: html_document  

---  

```{r}  

library(viridis)

```  

10 The code below demonstrates two color palettes in the  

[viridis](https://github.com/sjmarnier/viridis) package. Each  

plot displays a contour map of the Maunga Whau volcano in  

Auckland, New Zealand.  

11 ## Viridis colors  

12  

13 ```{r}  

14 image(volcano, col = viridis(200))  

15 ````  

16 ## Magma colors  

17  

18 ```{r}  

19 image(volcano, col = viridis(200, option = "A"))  

20 ````  

21  

22 ````  

23

```

위 예제에서 네 가지 다른 종류의 컨텐츠를 볼 수 있습니다. 하나는 --- 으로 둘러쌓인 내용으로 YAML이라고 하며 JSON과 같은 데이터 직렬화를 수행하는 하나의 데이터 저장 포맷입니다. 백틱(`)으로 둘러쌓인 코드 청그(Code Chunks)라고 하는 부분에는 R이나 python 등의 다양한 코드(실제 작동하는)를 넣어서 사용합니다. 그리고 ### 으로 표시된 글은 제목 글을 나타내며 나머지는 일반적인 텍스트를 나타냅니다.

```

---  

title: "Lecture3"  

output:  

  html_document:  

    toc: yes  

    toc_float: yes  

    toc_depth: 2  

    number_sections: yes  

---

```

이러한 Rmarkdown 파일은 render라는 명령어로 원하는 포맷의 문서로 변환할 수 있습니다. 다음 예의 파일을 pdf 형식으로 rendering하기 위해서는 YAML에 pdf 임을 명시하고 아래와 같이 render함수를 사용하면 됩니다. 또는 Rstudio 코드 입력창 상단의 Knit 버튼으로 pdf나 html 문서를 생성할 수 있습니다.

```

1---  

2 title: "My R markdown example"  

3 output:  

4   pdf_document: default  

5---  

6  

7```{r setup, include=FALSE}  

8 library(tidyverse)  

9```  

10  

11## R Markdown  

12  

13 This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML,  

  PDF, and MS Word documents. When you click the **Knit** button a document will be generated  

  that includes both content as well as the output of any embedded R code chunks within the  

  document. You can embed an R code chunk like this:  

14  

15```{r}  

16 cars %>% head  

17 cars %>%  

18   ggplot(aes(x=speed, y=dist)) +  

19   geom_point()  

20```  

21  

  render("examples/test.Rmd", output_format = "pdf_document")

```

My R markdown example

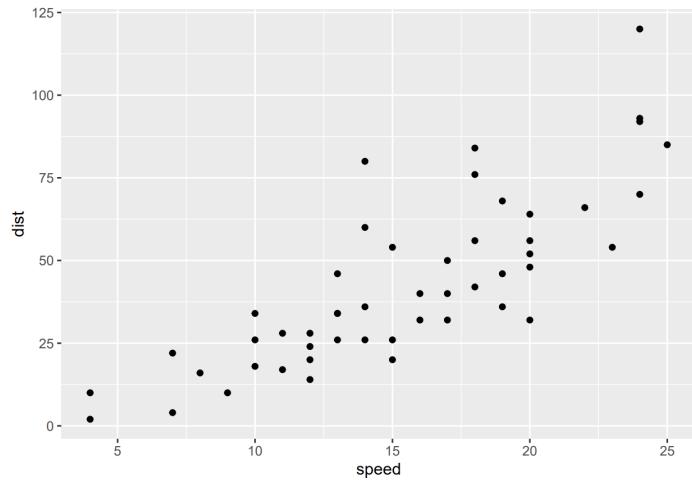
R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```

cars %>% head
cars %>%
  ggplot(aes(x=speed, y=dist)) +
  geom_point()

```



RMarkdown의 작동 원리는 Rmd 파일을 만든 후 render 함수를 부르면 knitr 소프트웨어가 R 코드를 실행시킨 후 markdown (.md) 파일을 생성합니다. 이후 .md 파일을 pandoc이라는 문서변환기가 원하는 문서 형태로 전환해 줍니다.

3.2 코드 입력

Rmarkdown에서 사용하는 코드첨크는 **CTRL+ALT+i** 단축키를 사용해서 넣을 수 있으며 다음과 같은 몇 가지 옵션으로 코드 스니펫들의 실행/숨김 여부를 결정할 수 있습니다.

- `include = FALSE` : 코드는 실행되지만 보고서에 결과와 코드가 보여지지 않음
- `echo = FALSE` : 코드는 실행되고 보고서에 결과가 포함되지만 코드는 보여지지 않음
- `eval = FALSE` : 코드가 실행되지 않지만 보고서에 코드는 보여짐
- `message = FALSE, warning=FALSE, error=FALSE` : 코드에 의해서 발생되는 메세지/경고/에러가 보고서에 보여지지 않음
- `fig.cap = "..."` : 코드로 그려지는 그래프에 캡션을 붙일 수 있음

```

43 ````{r}
44 # default
45 n <- c(1, 2, 3)
46 mean(n)
47 ````

48 ````{r, eval=F}
49 # eval=FALSE
50 n <- c(1, 2, 3)
51 mean(n)
52 ````

53 ````

54 ````{r, echo=F}
55 # echo=FALSE
56 n <- c(1, 2, 3)
57 mean(n)
58 ````

59 ````
```

Figure 3.2: 코드첨크 옵션 예시

실행 결과는 아래와 같습니다.

```

# default
n <- c(1, 2, 3)
mean(n)
#> [1] 2

# eval=FALSE
n <- c(1, 2, 3)
mean(n)

#> [1] 2
```

Rmarkdown에서는 ‘`r`’을 사용해서 코드첨크가 아닌 곳에 R 코드를 넣을 수 있습니다. 또한 R 언어 외에도 Python, SQL, Bash, Rcpp, Stan, JavaScript, CSS 등의 다양한 프로그래밍 언어에 대해서도 코드첨크 기능을 사용할 수 있습니다. 그런데 이러한 언어들이 사용 가능해지기 위해서는 해당 언어들을 실행해주는

엔진이 있어야 하며 python의 경우 `reticulate`라는 패키지가 이러한 기능을 담당합니다. 이 패키지를 설치할 경우 miniconda라는 가상환경 및 데이터 분석을 위한 오픈소스 패키지가 자동으로 설치됩니다.

```
library(reticulate)

x = "hello, python in R"
print(x.split(' '))
```

아래는 위에 해당하는 소스코드입니다.

```
```{r, eval=F}
library(reticulate)
```

```{python, eval=F}
x = "hello, python in R"
print(x.split(' '))
```

['hello', 'python', 'in', 'R']
```

3.3 Markdown 문법

마크다운은 plain text 기반의 마크업 언어로서 마크업 언어는 태그 등을 이용해서 문서의 데이터 구조를 명시하는데 이러한 태그를 사용하는 방법 체계를 마크업 언어라고 합니다. 가장 대표적으로 html 이 있습니다.

```
<html>
  <head>
    <title> Hello HTML </title>
  </head>
  <body>
    Hello markup world!
  </body>
</html>
```

마크다운도 몇 가지 태그를 이용해서 문서의 구조를 정의하고 있으며 상세한 내용은 Pandoc 마크다운 문서를 참고하시기 바랍니다. 마크다운언어의 철학은 쉽게 읽고 쓸 수 있는 문서입니다. plain text 기반으로 작성되어 쓰기 쉬우며 (아직도 사람들이 메모장 많이 사용하는 이유와 같습니다) 태그가 포함되어 있어도 읽는데 어려움이 없습니다. 위 html 언어와 아래 markdown 파일의 예들을 보시면 그 차이를 어렵지 않게 이해할 수 있습니다.

마크다운에서는 Enter를 한 번 입력해서 줄바꿈이 되지 않습니다.
 또는 문장 마지막에 공백을 두 개 입력하면 되겠습니다.

이 문장은 줄바꿈이 되지 않습니다

이 문장은 줄바꿈이 됩니다

마크다운 테그를 몇 가지 살펴보면 먼저 # 을 붙여서 만드는 header 가 있습니다.

```
# A level-one header
## A level-two header
### A level-three header

# A level-one header {#l1-1}
## A level-two header {#l2-1}
### A level-three header {#l3-1}

# A level-one header {#l1-2}
## A level-two header {#l2-2}
### A level-three header {#l3-2}
```

Block quotations

This is block quote. This paragraph has two lines

> This is block quote. This paragraph has two lines

This is a block quote.

A block quote within a block quote.

```
> This is a block quote.
>
> > A block quote within a block quote.
```

Italic

Italic

Bold

Bold

Naver link

[Naver link] (<https://www.naver.com/>)

이미지를 직접 삽입하고 가운데 정렬합니다.

```
<center>
! [ ](images/rmarkdown/000002.png){width="200"}
</center>
```

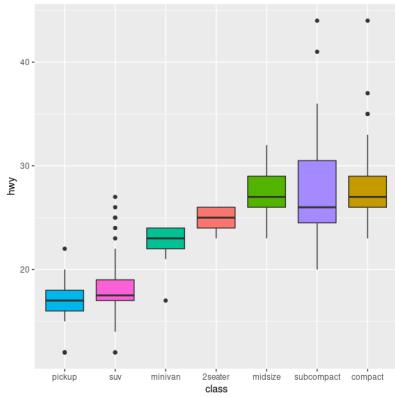


Figure 3.3: 자동차 모델에 따른 고속도로 연비 분포

1. 첫 번째
2. 두 번째
3. 세 번째

- 아이템 1
- 아이템 2
- 아이템 3
 - 아이템 3-1
 - 아이템 3-2

- 1.
 - 2.
 - 3.
- 1
- 2
- 3
 - 3-1
 - 3-2

참고로 소스코드 그대로 표현하기 위해서는 ~~~ 를 사용합니다.

3.4 스타일

아래와 같이 코드 청크를 이용해서 css 코드를 삽입하고 해당되는 class 또는 id에 해당하는 내용에 스타일을 적용할 수 있습니다.

Table 3.1: A knitr kable.

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360	175	3.15	3.440	17.02	0	0	3	2

```
```{css, echo=F}
#header1 {
 color: red;
}
```

### 소스코드

```
<div id='header1'>

</div>
```

## 3.5 테이블

kable 함수를 이용하여 Rmarkdown 문서에 포함되는 표를 원하는 방향으로 작성할 수 있습니다. mtcars는 데이터프레임 형식의 데이터입니다.

```
knitr:::kable(
 mtcars[1:5,],
 caption = "A knitr kable."
)
```

## 3.6 YAML 헤더

Rmarkdown 파일에서 YAML의 가장 중요한 기능은 output 포맷을 지정하는 것이며 title, author, date, 등을 설정할수도 있습니다.

```

layout: page
title: "R"
subtitle: "Rmarkdown"
output:
 html_document:
 css: style.css
```

```
includes:
 in_header: header.html
 after_body: footer.html
theme: default
toc: yes
toc_float: true
highlight: tango
code_folding: show
number_sections: TRUE
mainfont: NanumGothic

```

## 3.7 Output format

주요 문서 포맷으로 다음과 같은 몇 가지가 있습니다. 상세한 내용은 Rmarkdown output format을 참고하시기 바랍니다.

- html\_document - HTML document w/ Bootstrap CSS
- pdf\_document - PDF document (via LaTeX template)
- word\_document - Microsoft Word document (docx)
- ioslides\_presentation - HTML presentation with ioslides
- beamer\_presentation - PDF presentation with LaTeX Beamer
- powerpoint\_presentation: PowerPoint presentation

### Exercises

“KRIBBR2022-Lecture2”라는 이름의 다음과 같은 형태의 Rmarkdown 문서를 만들고 이 번 강의의 실습 코드 및 설명, 질문, 코멘트 등을 적어 보시기 바랍니다.

---

이 저작물은 크리에이티브 커먼즈 저작자표시-비영리-변경금지 4.0 국제 라이선스에 따라 이용할 수 있습니다.



# Chapter 4

## R programming

### 4.1 Console calculator

콘솔에서 바로 계산을 수행할 수 있습니다. 참고로 이전에 수행한 명령은 콘솔에 커서가 있는 상태에서 위 아래 화살표를 누르면 볼 수 있고 엔터를 눌러 재사용 할 수 있습니다. ;을 사용하면 두 개의 명령을 동시에 수행할 수 있습니다.

$$2 + 2$$

$$((2 - 1)^2 + (1 - 3)^2)^{1/2}$$

```
2 + 2
((2 - 1)^2 + (1 - 3)^2)^(1/2)
2 + 2; 2 - 2
```

#### Exercises

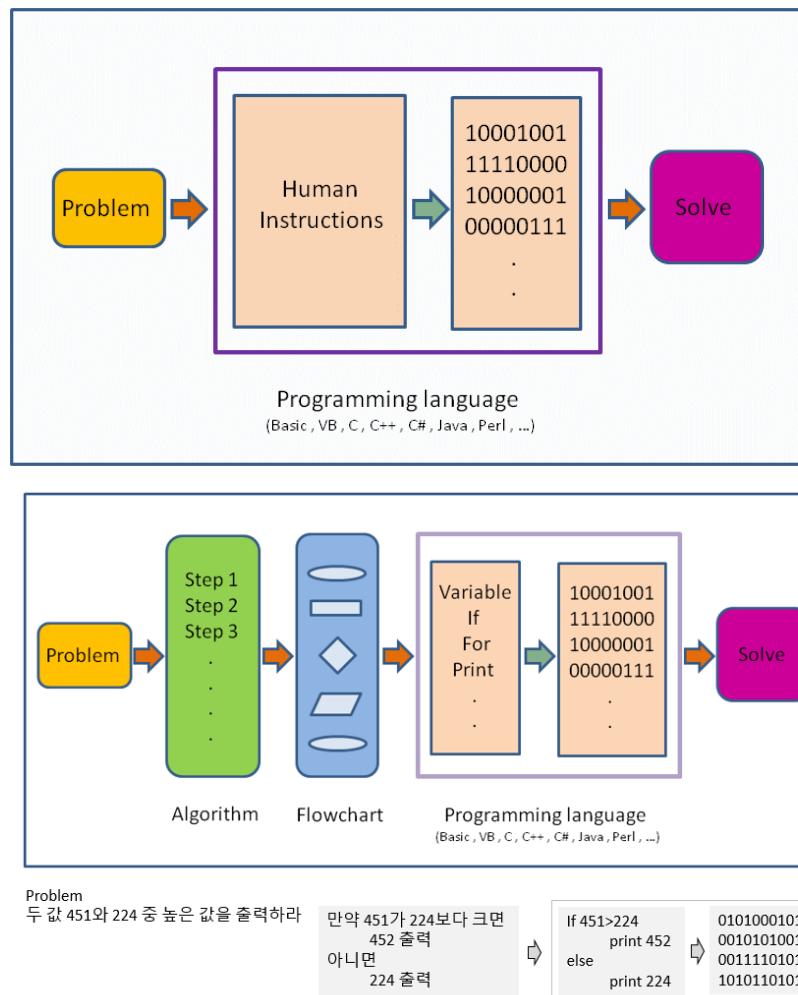
다음 공식들을 계산하는 R 코드를 작성하시오

$$\sqrt{(4 + 3)(2 + 1)}$$

$$2^3 + 3^2$$

$$\frac{0.25 - 0.2}{\sqrt{0.2(1 - 0.2)/100}}$$

## 4.2 What is a programming language



R은 programming language로서 다른 프로그래밍 언어와 같이 몇 가지 공통적 개념을 가집니다 ( , , , , )

### 4.2.1 Terminology

- Session: R 언어 실행 환경
- Console: 명령어 입력하는 창
- Code: R 프로그래밍 변수/제어문 모음
- Object: 변수, 함수 등 프로그래밍에서 사용되는 모든 객체 (Data structure)
  - array: 1D, 2D, 3D, … 형태 값들의 모임
  - vector: 1차원 형태 값들의 모임 combine function c() EX: c(6, 11,

- 13, 31, 90, 92)
- matrix: 2차원 형태 값들의 모임 (같은 타입 값으로 구성)
  - data frame: 2차원 형태 값들의 모임 (다른 타입 값 구성 가능)
  - list: vector, matrix, data.frame 및 list 등 다양한 객체를 원소로 가짐
  - function: 특정 기능 수행, [함수이름, 입력값 (arguments), 출력값 (return)]으로 구성
  - Data (value): 값 - 자료형 (Data type)
    - Integers
    - doubles/numerics
    - logicals
    - characters
    - factor: 범주형
  - Conditionals (조건, 제어):
    - if, ==, & (AND), | (OR) Ex: (2 + 1 == 3) & (2 + 1 == 4)
    - for, while: 반복 수

## 4.3 Data and variables

### 4.3.1 Data

일반적으로 데이터의 의미는 사실을 나타내는 수치입니다.

- 맥도너 정보경제학 (1963)
  - 지혜 (wisdom) : 패턴화된 지식
  - 지식 (knowledge) : 가치있는 정보
  - 정보 (information) : 의미있는 데이터
  - 데이터 (data) : 단순한 사실의 나열

```
library(UsingR)
exec.pay
?exec.pay
```

데이터는 속성에 따라서 다음과 같이 분류할 수 있습니다.

- 범주형 - 질적 데이터, 숫자로 나타낼 수 있으나 의미 없음
  - 명목형 (Nominal) - 사람 이름
  - 순서형 (Ordinal) - 달리기 도착 순서
- 수치형 - 숫자로 나타내며 데이터 속성을 그대로 지님님
  - 구간형 (Interval) - 선수1, 선수2 종점통과 시간
  - 비율형 (Ratio) - 출발시간 기준 종점 통과 시간

이름	등수	도착	걸린시간
둘리	1	13:12	1:12
희동	5	14:30	2:30
길동	2	13:30	1:30
철수	4	14:00	2:00
영희	3	13:50	1:50

- Data type in R
  - Numeric (수치형)
    - \* Discrete (이산형) data - 카운트, 횟수
    - \* Continuous (연속형) data - 키, 몸무게, Cannot be shared
    - \* Date and time
  - Factors (범주형)
    - \* Categories to group the data
    - \* Character data - Identifiers (범주형)

### 4.3.2 Variables

변수는 데이터를 저장하는 공간으로 이해할 수 있습니다.

- Assignment operator ( <- OR = )
  - Valid object name <- value
  - 단축키: Alt + - (the minus sign)
- 내장 변수 Built-in variables

```
x <- 2
y <- x^2 - 2*x + 1
y
x <- "two"
some_data <- 9.8
pi
```

- 변수이름 작성법
  - Characters (letters), numbers, “\_”, “.”
  - A and a are different symbols
  - Names are effectively unlimited in length

```
i_use_snake_case <- 1
otherPeopleUseCamelCase <- 2
some.people.use.periods <- 3
And_aFew.People_RENOUNCEconvention <- 4
```

## 4.4 Object (Data structure)

변수, 함수 등 프로그래밍에서 사용되는 모든 개체를 말합니다.

### 4.4.1 vector

`vector`는 R의 기본 데이터 구조입니다. numeric vector, logical vector, character vector 등 저장되는 값의 타입에 따라 크게 세가지로 나눌 수 있습니다. `class()` 함수를 이용해서 값의 타입을 알아낼 수 있습니다. Combine function인 `c()`를 활용하여 만들며 값을 순차적으로 붙여갈 수 있습니다. 다음과 같은 Univariate (단변량, Single variable)을 표현할 때 사용됩니다.

$$x_1, x_2, \dots, x_n$$

```
x <- c(10.4, 5.6, 3.1, 6.4, 21.7)
class(x)
y <- c("X1", "Y2", "X3", "Y4")
class(y)
z <- c(T, F, F, T)
class(z)
```

#### 4.4.1.1 numeric

numeric 형식의 벡터는 다음과 같은 다양한 편의 함수들을 사용해서 만들 수 있습니다.

```
1:5
seq(1,5, by=1)
seq(0, 100, by=10)
seq(0, 100, length.out=11)
?seq

rep(5, times=10)
rep(1:3, times=4)
rep(1:3, each=3)
```

### Exercises

`odds`라는 이름의 변수에 1부터 100까지의 홀수만을 저장하시오 (`seq()` 함수 사용)

인덱싱은 배열형 (vector, matrix 등) 데이터의 일부 데이터를 참조할 때 사용하는 방법입니다. [와 ]를 사용하며 위치를 나타내는 수로 참조합니다.

```
x[1]
x[1:3]
i <- 1:3
x[i]
```

```
x[c(1,2,4)]
y[3]
```

또한 해당 위치의 이름으로 참조하기도 합니다.

```
head(precip)
precip[1]
precip[2:10]
precip[c(1,3,5)]
precip[-1]
precip["Seattle Tacoma"]
precip[c("Seattle Tacoma", "Portland")]
precip[2] <- 10
```

참고로 vector 들은 다음과 같은 builtin 함수들을 사용해서 해당 변수의 attribute를 알아낼 수 있습니다. attribute에는 원소 이름, 타입, 길이 등 vector형 변수가 가질 수 있는 특성을 말합니다.

```
head(precip)
class(precip)
length(precip)
names(precip)

test_scores <- c(100, 90, 80)
names(test_scores) <- c("Alice", "Bob", "Shirley")
test_scores
```

#### 4.4.1.2 logical

Logical 벡터는 True 또는 False를 원소로 갖는 벡터입니다. 앞글자가 대문자로 시작하는 것을 기억하시고 T 또는 F와 같이 한 문자로 표현할 수도 있습니다. 특정 조건에 대한 판단 결과를 반환할 경우에도 논리값을 사용합니다. 이 경우 조건을 판단 후 인덱싱 방법으로 (which, any, all 등 사용) 해당 값을 뽑아내기도 합니다. 또한 활용이 많은 sample 함수의 사용법을 익혀둡니다.

```
x <- 1:20
x > 13
temp <- x > 13
class(temp)

ages <- c(66, 57, 60, 41, 6, 85, 48, 34, 61, 12)
ages < 30
which(ages < 30)
i <- which(ages < 30)
ages[i]
any(ages < 30)
all(ages < 30)
```

```
random_number <- sample(c(1:10), 2)
```

**Exercises** 1. 1부터 100까지의 수를 evens이라는 이름의 변수에 저장하고 이 중 짝수만을 뽑아내서 출력하시오 (which() 함수 사용)

2. sample 함수를 사용하여 앞서 odds와 evens 변수에서 랜덤하게 1개씩의 샘플을 뽑아서 mynumbers에 저장하시오
3. 어떤 짝수가 뽑혔는지 찾아서 출력하시오 (which와 인덱싱 사용)

#### 4.4.1.3 character

Character(문자형) 벡터의 경우 문자열을 다루는데 자주 쓰이는 paste() 함수의 사용법을 알아두면 편리합니다. paste() 함수는 서로 다른 문자열을 붙이는데 주로 사용됩니다. 참고로 문자열을 나누는 함수는 strsplit()입니다. paste()에서 붙이는 문자 사이에 들어가는 문자를 지정하는 파라미터는 sep이고 strsplit() 함수에서 자르는 기준이 되는 문자는 split 파라미터로 지정해 줍니다 (?split 또는 ?paste 확인).

```
paste("X", "Y", "Z", sep="_")
paste(c("Four", "The"), c("Score", "quick"), c("and", "fox"), sep="_")
paste("X", 1:5, sep="")
paste(c("X", "Y"), 1:10, sep="")

x <- c("X1", "Y2", "X3", "Y4", "X5")
paste(x[1], x[2])
paste(x[1], x[2], sep="")
paste(x, collapse="_")

strsplit("XYZ", split="")
sort(c("B", "C", "A", "D"))
```

#### Exercises

1. m이라는 변수에 “Capital of South Korea is Seoul” 문자열을 저장하고 “Capital of South Korea”를 따로 뽑아내 m2에 저장하시오 (substr() 사용)
2. LETTERS 내장함수에서 랜덤하게 10개의 문자를 뽑아내 myletters 변수에 저장하고 이들을 연결하여 (paste 사용) 하나의 문장(String)을 만드시오
3. myletters 변수의 문자들을 알파벳 순서대로 정렬하고 (sort 사용) 이들을 연결하여 하나의 문장 (String)을 만드시오

#### 4.4.1.4 factor

Factor형은 범주형데이터를 저장하기 위한 object이며 R 언어에서 특별히 만들어져 사용되고 있습니다. factor() 함수를 이용해 생성하며 생성된 객체는 다음과 같이 level이라는 범주를 나타내는 특성값을 가지고 있습니다.

예를 들어 어린이 5명이 각각 빨강, 파랑, 노랑, 빨강, 파랑 색종이를 들고 있을때 색의 종류를 나타내는 값들은 빨강, 파랑, 노랑 입니다. 다섯 명의 아이들이 어떤 색의 색종이를 들고 있는지와는 상관없이 세 가지 범주의 값을 가지는 것 입니다.

```
x <- c("Red", "Blue", "Yellow", "Red", "Blue")
y <- factor(x)
y
```

새로운 범주의 데이터를 추가할 경우 다음과 같이 해당되는 level을 먼저 추가하고 값을 저장해야 합니다.

```
levels(y)
y[1] <- "Gold"
y

levels(y) <- c(levels(y), "Gold")
levels(y)
y
y[1] <- "Gold"
y
```

`factor`는 기본적으로 `level`에 표시된 순서가 위치 (정렬) 순서입니다. 이를 바꾸기 위해서는 다음과 같이 `levels` 함수를 이용해서 순서를 바꿀 수 있습니다.

```
library(MASS)
str(Cars93)
x <- Cars93$Origin
plot(x)
levels(x) <- c("non-USA", "USA")
levels(x)
plot(x)
```

## Exercises

UUU UUC UUA UUG	Phe	UCU UCC UCA UCG	Ser	UAU UAC UAA UAG	Tyr Stop	UGU UGC UGA UGG	Cys Stop Trp
CUU CUC CUA CUG	Leu	CCU CCC CCA CCG	Pro	CAU CAC CAA CAG	His Gln	CGU CGC CGA CGG	Arg
AUU AUC AUA AUG	Ile Met	ACU ACC ACA ACG	Thr	AAU AAC AAA AAG	Asn Lys	AGU AGC AGA AGG	Ser Arg
GUU GUC GUA GUG	Val	GCU GCC GCA GCG	Ala	GAU GAC GAA GAG	Asp Glu	GGU GGC GGA GGG	Gly

1. 아미노산 Phe,

Leu, Ser 를 값으로 갖는 범주형 변수 (factor)를 생성하시오 2. 각 아미노산과 해당 아미노산을 코딩하는 nucleotide triplets (codon)을 어떤 형태의 변수로 저장할 수 있을지 고민해 보시오

#### 4.4.1.5 Missing values

특정 값이 “Not available” 이거나 “Missing value” 일 경우 벡터의 해당 원소 자리에 데이터의 이상을 알리기 위해 NA를 사용합니다. 따라서 일반적인 연산에서 NA가 포함되어 있는 경우 데이터의 불완전성을 알리기 위해 연산의 결과는 NA가 됩니다. `is.na()` 함수는 해당 변수에 NA 값이 있는지를 검사해주는 함수이며 R에는 이 외에도 다음과 같은 특수 값들이 사용되고 있습니다.

- NA: Not available, The value is missing
- NULL: a reserved value
- NaN: Not a number (0/0)
- Inf: (1/0)

```
hip_cost <- c(10500, 45000, 74100, NA, 83500)
sum(hip_cost)
sum(hip_cost, na.rm=TRUE)
?sum
```

#### 4.4.1.6 Useful functions

다음은 벡터형 변수와 같이 쓰이는 유용한 함수들입니다.

```
z <- sample(1:10, 100, T)
head(z)
sort(z)
order(z)
table(z)
```

```
p <- z/sum(z)
round(p, digits=1)
```

`is` 함수를 사용하여 데이터 타입이 사용자가 의도한 타입과 맞는지 검사할 수 있습니다. 콘솔창에서 `is.`를 타이핑한 후 잠시 기다리면 다양한 `is` 함수를 볼 수 있습니다.

```
is.na(1)
is.numeric(1)
is.logical(TRUE)
is.data.frame("A")
is.character("A")
```

`as` 함수는 데이터 타입을 변환해주는 함수입니다.

```
digits <- runif(10)*10
class(digits)
digits_int <- as.integer(digits)
class(digits_int)
digits_char <- as.character(digits_int)
class(digits_char)
digits_num <- as.numeric(digits_char)
class(digits_num)
```

#### 4.4.2 matrix

매트릭스는 2차원 행렬로 같은 형식의 데이터 값 (numeric, character, logical) 으로만 채워진 행렬을 말합니다. 매트릭스를 만드는 방법은 아래와 같으며 `nrow` 와 `ncol` 파라메터에 행과 열의 수를 넣고 각 셀에 들어갈 같은 가장 앞에 위치한 `data` 파라메터에 넣어 줍니다 (`?matrix`로 파라메터 이름 확인). 매트릭스 인덱싱은 매트릭스 안의 값을 저장하거나 참조할때 (빼올때) 사용하는 방법입니다. 매트릭스 변수이름 바로 뒤에 대괄호를 이용해서 제어를 하며 대괄호 안에 콤마로 구분된 앞쪽은 `row`, 뒷쪽은 `column` 인덱스를 나타냅니다.

```
mymat <- matrix(0, nrow=100, ncol=3) # 1
mymat[,1] <- 1:100 # 2
mymat[,2] <- seq(1,200,2) # 3
mymat[,3] <- seq(2,200,2) # 4
```

매트릭스의 `row`나 `column`에 이름이 주어져 있을 경우 이름을 따옴표("")로 묶은 후 참조가 가능합니다. `row`나 `column`의 이름은 `rownames()` 또는 `colnames()`로 생성하거나 변경할 수 있습니다. `row`나 `column`의 개수는 `nrow()` 또는 `ncol()` 함수를 사용합니다.

```
colnames(mymat)
colnames(mymat) <- c("A", "B", "C")
colnames(mymat)
```

```
colnames(mymat)[2] <- "D"
colnames(mymat)
rownames(mymat) <- paste("No", 1:nrow(mymat), sep="")
rownames(mymat)
```

여러 row나 column을 참조할 경우 아래와 같이 combine 함수를 사용하여 묶어줘야 하며 스칼라값을 (임의의 숫자 하나) 더하거나 뺄 경우 vector / matrix 연산을 기본으로 수행합니다.

```
mymat[c(2,3,4,5),2] # 5
mymat[-1] # 6
mysub <- mymat[,2] - mymat[,1] #7
sum(mysub) #8
sum(mysub^2) #8
```

### Exercises

- score라는 변수에 1부터 100까지 중 랜덤하게 선택된 20개의 수로  $10 \times 2$  matrix를 만드시오 (sample() 사용)
- score의 row 이름을 문자형으로 Name1, Name2, …, Name10으로 지정하시오 (paste() 사용)
- score의 column 이름을 문자형으로 math와 eng로 지정하시오
- 이 matrix의 첫번째 컬럼과 두 번째 컬럼의 수를 각각 더한 후 total\_score라는 변수에 저장하시오
- total\_score의 오름차순 순서를 나타내는 인덱스 (order() 함수 사용)를 o라는 변수에 저장하시오
- score를 o순서로 재배치하고 score\_ordered 변수에 저장하시오

### 4.4.3 data.frame

데이터프레임은 형태는 매트릭스와 같으나 컬럼 하나가 하나의 vector형 변수로서 각 변수들이 다른 모드의 값을 저장할 수 있다는 차이가 있습니다. \$ 기호를 이용하여 각 구성 변수를 참조할 수 있습니다. 컬럼 한 줄이 하나의 변수 이므로 새로운 변수도 컬럼 형태로 붙여 넣을 수 있습니다. 즉, 각 row는 샘플을 나타내고 각 column은 변수를 나타내며 각 변수들이 갖는 샘플의 개수 (row의 길이, vector의 길이)는 같아야 합니다. R 기반의 데이터 분석에서는 가장 선호되는 데이터 타입이라고 볼 수 있습니다.

```
data.frame
ids <- 1:10
ids
idnames <- paste("Name", ids, sep="")
idnames
students <- data.frame(ids, idnames)
students
class(students$ids)
class(students$idnames)
```

```

students$idnames
str(students)

students <- data.frame(ids, idnames, stringsAsFactors = F)
class(students$idnames)
students$idnames
students[,1]
str(students)

```

데이터프레임에서는 \$를 사용하여 변수 이름으로 인덱싱이 가능합니다.

```

data frame indexing
students$ids
students[,1]
students[,"ids"]

```

### Exercises

- math라는 변수에 1부터 100까지 중 랜덤하게 선택된 10개의 수를 넣으시오
- eng라는 변수에 1부터 100까지 중 랜덤하게 선택된 10개의 수를 넣으시오
- students라는 변수에 문자형으로 Name1, Name2, …, Name10으로 지정하시오 (paste() 사용)
- math와 eng라는 벡터에 저장된 값들의 이름을 students 변수에 저장된 이름으로 지정하시오
- math와 eng 벡터를 갖는 score라는 data.frame을 만드시오
- math와 eng 변수를 지우시오 (rm() 사용)
- score data frame의 math와 eng를 각각 더한 후 total\_score라는 변수에 저장 하시오

#### 4.4.4 list

리스트는 변수들의 모임이라는 점에서 데이터프레임과 같으나 구성 변수들의 길이가 모두 같아야 하는 데이터프레임과는 달리 다른 길이의 변수를 모아둘 수 있는 점이 다릅니다. 즉, R언어에서 두 변수를 담을 수 있는 데이터 타입은 list와 data frame 두 종류가 있는데 list 변수 타입은 vector 형태의 여러개의 element를 가질 수 있으며 각 vector 길이가 모두 달라도 됩니다. list의 인덱싱에서 [ ]는 리스트를 반환하고 [[]]는 vector element들을 반환합니다.

**Lists**

```
l <- list(x = 1:5, y = c('a', 'b'))
```

A list is a collection of elements which can be of different types.

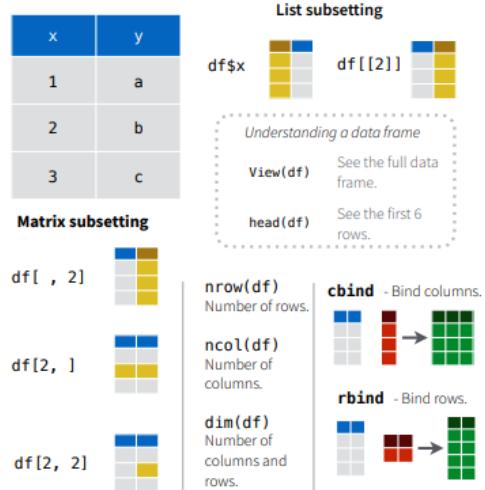
<code>l[2]</code>	<code>l[1]</code>	<code>l\$x</code>	<code>l['y']</code>
Second element of l.	New list with only the first element.	Element named x.	New list with only element named y.

```
list
parent_names <- c("Fred", "Mary")
number_of_children <- 2
child_ages <- c(4, 7, 9)
data.frame(parent_names, number_of_children, child_ages)
lst <- list(parent_names, number_of_children, child_ages)
lst[1]
lst[[1]]
class(lst[1])
class(lst[[1]])
lst[[1]][1]
lst[[1]][c(1,2)]
```

Also see the  
**dplyr** package.

### Data Frames

`df <- data.frame(x = 1:3, y = c('a', 'b', 'c'))`  
A special case of a list where all elements are the same length.



### Exercises

1. 위 아미노산 예제에서 Phe, Leu, Ser 각각의 코돈을 원소로 갖는 세 개의 vector 변수들을 만들고 이를 aalist라는 이름의 하나의 리스트 변수로 만드시오
2. aalist 리스트를 data.frame 형식의 aadf 변수로 만드시오 (데이터 구조를 바꾸어 저장 가능)

## 4.5 Functions

함수(Function)란 사용자가 원하는 기능을 수행하는 코드의 모음으로서 반복적으로 쉽게 사용할 수 있도록 만들어 놓은 코드입니다.

### 4.5.1 A script in R

함수의 개념을 배우기 전에 스크립트를 활용한 명령어 수행을 알아보겠습니다. R 프로그래밍을 통해서 사용자가 원하는 기능을 수행하는 방법은 다음과 같이 스크립트를 만들어서 실행하는 것입니다. 일반적으로 R을 이용한 스크립트 명령을 어떻게 실행하는지 알아보겠습니다. 다음 예제는 입력 값들의 평균을 계산해서 출력해 주는 스크립트 명령입니다. R base 패키지에서 기본으로 제공되는 `mean()`이라는 함수가 있지만 사용하지 않고 `sum()`과 `length()` 함수를 사용했습니다.

```
numbers <- c(0.452, 1.474, 0.22, 0.545, 1.205, 3.55)
cat("Input numbers are", numbers, "\n")
numbers_mean <- sum(numbers)/length(numbers)
out <- paste("The average is ", numbers_mean, ".\n", sep="")
cat(out)
```

상황에 따라 다르긴 하지만 보통 위 스크립트를 실행할 때 R 파일을 하나 만들고 `source()`라는 함수를 사용해서 파일 전체를 한번에 읽어들이고 실행을 시킵니다. 위 코드를 `myscript.R`이라는 새로운 R 파일을 하나 만들고 저장 후 다음과 같이 실행할 수 있습니다. 참고로 위 파일은 현재 Working directory와 같은 위치에 저장해야 합니다.

```
source("myscript.R")
```

그러나 위와 같은 식으로 실행할 경우 다음 몇 가지 문제가 있습니다. 하나는 입력 값이 바뀔 때마다 파일을 열어 바뀐 값을 저장해 줄 필요가 있습니다. 결과 값에 대해서 다른 처리를 하고 싶을 경우 또한 파일을 직접 수정해 주어야 합니다. 또한 모든 변수들이 전역변수로 사용되어 코드가 복잡해질 경우 변수간 간섭이 생길 가능성이 높습니다.

### 4.5.2 Build a function

함수는 특정 데이터를 입력으로 받아 원하는 기능을 수행한 후 결과 데이터를 반환하는 구조를 가집니다. 함수는 일반적으로 다음과 같은 포맷으로 구현할 수 있습니다.

```
my_function_name <- function(parameter1, parameter2, ...){
 ##any statements
 return(object)
}
```

예를 들어 다음과 같은 `my_sine` 함수를 만들 수 있으며 `parameter` (매개변수)는 `x`이고 `y`는 반환값을 저장하는 지역변수입니다.

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

만들어진 함수는 다음과 같이 사용할 수 있습니다. 만들어진 함수는 처음에 한 번 실행해 주어 실행중인 R session에 등록한 후 사용할 수 있습니다. 여기서 함수로 전달되는 값 `pi`는 argument (전달인자) 라고 합니다. 전달인자는 함수에서 정의된 매개변수의 갯수와 같은 수의 전달인자를 입력해 주어야 합니다. 참고로 parameter와 argument는 많은 사람들이 혼동하는 단어입니다. 본 예에서 `my_sine`함수의 괄호 안에 있는 변수 `x`는 parameter이고 `x`에 들어가는 값인 `pi`나 `90`은 argument입니다.

```
my_sine(pi)
my_sine(90)
sin(90)
```

- Terminology

- function name: `my_sine`
- parameter: `x`
- argument: `pi`
- return value: `y`

이제 위 스크립트 (`myscript.R`)에서 사용된 코드를 함수로 바꿔봅니다. `numbers` (전달인자)를 받는 매개변수를 `x`로 하고 함수 이름은 `mymean`이고 평균값 (`numbers_mean`)을 반환하는 함수입니다.

```
numbers <- c(0.452, 1.474, 0.22, 0.545, 1.205, 3.55)

mymean <- function(x){
 cat("Input numbers are", x, "\n")
 numbers_mean <- sum(x)/length(x)
 out <- paste("The average is ", numbers_mean, ".\n", sep="")
 cat(out)
 return(numbers_mean)
}

retval <- mymean(numbers)
cat(retval)
```

`myscript.R`이라는 파일을 열고 작성된 스크립트에 더해서 아래처럼 함수 코드를 만들 경우 `source()` 함수로 함수를 세션으로 읽어오고 바로 사용할 수 있습니다. 위와 같이 함수를 만들 경우 입력 값을 언제든 바꿔서 사용할 수 있고 반환값에 대한 추가적인 연산도 쉽게 수행 할 수 있습니다.

```
new_values <- c(1:10)
retval <- mymean(new_values)
retval
```

### Exercises

1. 변수 `x`에 1, 3, 5, 7, 9를, 변수 `y`에 2, 4, 6, 8, 10을 저장하는 코드를 작성하시오

2. x와 y를 더한 값을 z에 저장하는 코드를 작성하시오
3. mysum 이라는 이름의 함수를 작성하되 두 변수를 입력으로 받아 더한 후 결과를 반환하는 코드를 작성하시오
4. mymean 이라는 이름의 함수를 작성하되 두 변수를 입력으로 받아 평균을 구한 후 결과를 반환하는 코드를 작성하시오

### Exercises

- 1) mysd라는 이름의 (표본)표준편차를 구하는 함수를 myscript.R 파일에 구현하시오 (sd()함수 사용하지 않고, 다음 표준편차 공식 이용)

$$\sigma = \sqrt{\frac{\sum(x - mean(x))^2}{length(x) - 1}}$$

코드는 아래와 같음

```
mysd <- function(x){
 numbers_sd <- sqrt(sum((x - mymean(x))^2)/(length(x)-1))
 return(numbers_sd)
}
```

- 2) 1부터 100까지의 값을 x에 저장하고 mysd 함수를 사용해서 표준편차를 구하시오
- 3) 앞서 작성한 mymean 함수와 mysd 함수를 같이 사용하여 x를 표준화 하고 z로 저장하시오. 표준화 공식은 다음과 같음

$$z = \frac{x - mean(x)}{sd(x)}$$

- 4) x 와 z 변수를 원소로 갖는 y라는 이름의 data.frame을 생성하시오

### 4.5.3 local and global variables

함수를 사용함에 따라서 함수 안에서 사용되는 변수와 함수 밖에서 사용되는 변수들의 경우를 명확히 이해할 필요가 있습니다. 다음 코드를 보면 전역변수 x, y는 지역변수 x, y와 독립적으로 사용되고 있습니다.

```
my_half <- function(x){
 y <- x/z
 cat("local variable x:", x, "\n")
 cat("local variable y:", y, "\n")
 cat("global variable z:", z, "\n")
 return(y)
}
y <- 100
```

```

x <- 20
z <- 30
cat("Global variable x:", x, "\n")
cat("Global variable y:", y, "\n")
cat("Global variable z:", z, "\n")
my_half(5)

my_half <- function(x, z){
 y <- x/z
 cat("local variable x:", x, "\n")
 cat("local variable y:", y, "\n")
 cat("local variable z:", z, "\n")
 return(y)
}

my_half(5, 10)

```

log, sin등의 함수들은 Built-in function으로 같은 이름의 함수를 만들지 않도록 주의합니다.

```

x <- pi
sin(x)
sqrt(x)
log(x)
log(x, 10)
x <- c(10, 20, 30)
x + x
mean(x)
sum(x)/length(x)

```

#### 4.5.4 Vectorized functions

초기에는 R이 다른 프로그래밍 언어에 비해서 경쟁력을 갖는 이유 중 하나가 바로 이 벡터 연산 기능이었습니다. vector 변수에 들어있는 각 원소들에 대해서 특정 함수나 연산을 적용하고 싶을 경우 전통 방식의 C나 Java등의 언어에서는 원소의 개수만큼 반복문을 돌면서 원하는 작업을 수행 했습니다. 그러나 R의 벡터 연산 기능은 별도의 반복문 없이 vector 안에 있는 원소들에 대한 함수 실행 또는 연산을 수행할 수 있습니다.

```

x <- c(10, 20, 30)
x + x
sqrt(x)
sin(x)
log(x)
x-mean(x)

```

```
length(x)
test_scores <- c(Alice = 87, Bob = 72, James= 99)
names(test_scores)
```

### Exercises

다음은 한 다이어트 프로그램의 수행 전 후의 다섯 명의 몸무게이다.

Before	78	72	78	79	105
after	67	65	79	70	93

- 1) 각각을 before 와 after 이름의 변수에 저장 후 몸무게 값의 변화량을 계산하여 diff 라는 변수에 저장하시오
- 2) diff에 저장된 값들의 합, 평균, 표준편차를 구하시오

### Exercises

다음 네 학생이 있으며 “John”, “James”, “Sara”, “Lilly” 각 나이는 21, 55, 23, 53이다. ages라는 변수를 생성하고 각 나이를 저장한 후 who라는 이름의 함수를 만들어서 50살 이상인 사람의 이름을 출력하는 함수를 만드시오.

- ages라는 변수에 나이 저장, c() 함수 이용, vector 형태 저장
- names() 함수 이용해서 각 ages 벡터의 각 요소에 이름 붙이기
- which() 함수 사용해서 나이가 50보다 큰 인덱스 찾고 해당 인덱스 값을 idx에 저장
- ages에서 idx에 해당하는 인덱스를 갖는 값을 sel\_ages에 저장
- names() 함수를 이용해서 sel\_ages의 이름을 sel\_names에 저장
- 위 설명을 참고해서 input이라는 파라미터를 갖고 sel\_names라는 50살 이상인 사람의 이름을 반환하는 who50이라는 이름의 함수 만들기
- who50 함수의 사용법은 who50(ages) 임

## 4.6 Flow control

### 4.6.1 if statements

R에서의 제어문의 사용은 다른 프로그래밍 언어와 거의 유사합니다. 먼저 if 는 다음과 같은 형식으로 사용되며 () 안에 특정 조건 판단을 위한 표현이 들어갑니다.

```
if(condition){
 expr_1
}else{
 expr_2
}
```

특히 condition은 하나의 원소에 대한 조건 판단문으로 T 또는 F 값 하나만을 반환하는 문장이어야 합니다. 위 코드는 만약 condition 조건이 True 이면

`expr_1`를 실행하고 `False`이면 `expr_2`를 실행하라는 명령입니다. `condition` 안에서 사용되는 비교 연산자들은 다음과 같습니다.

<code>! x</code>	logical negation, NOT x
<code>x &amp; y</code>	elementwise logical AND
<code>x &amp;&amp; y</code>	vector logical AND
<code>x   y</code>	elementwise logical OR
<code>x    y</code>	vector logical OR
<code>xor(x, y)</code>	elementwise exclusive OR
<code>&lt;</code>	Less than, binary
<code>&gt;</code>	Greater than, binary
<code>==</code>	Equal to, binary
<code>&gt;=</code>	Greater than or equal to, binary
<code>&lt;=</code>	Less than or equal to, binary

```

x <- 2
if(x%%2 == 1){
 cat("Odd")
}else{
 cat("Even")
}

x <- 5
if(x > 0 & x < 4){
 print("Positive number less than four")
}

if(x > 0) print("Positive number")

x <- -5
if(x > 0){
 print("Non-negative number")
} else if(x <= 0 & x > -5){
 print("Negative number greater than -5")
} else {
 print("Negative number less than -5")
}

if(x > 0)
 print("Non-negative number")
else
 print("Negative number")

```

### 4.6.2 ifelse statements

`if`는 하나의 조건만 비교하는데 사용할 수 있습니다. 그러나 변수에는 여러 값이 벡터형식으로 들어가고 벡터연산을 수행할 경우의 결과도 벡터형식으로 나오지만 `if`문은 이들을 한 번에 처리하기 어렵습니다. `ifelse`는 이러한 단점을 보완하여 여러 값을 한번에 처리할 수 있습니다.

```
ifelse (condition, True , False)
```

`ifelse`의 경우 빠르게 원하는 값을 반환할 수 있으나 조건별로 다른 추가적인 명령의 수행은 불가능하다는 단점이 있습니다.

```
x <- c(1:10)
if(x>10){
 cat("Big")
}else{
 cat("Small")
}

ifelse(x>10, "Big", "Small")
```

### Exercises

다음은 `median` (중간값)을 구하는 공식이며 `x`의 길이가 (`n`이) 홀수일 경우와 짝수일 경우에 따라서 다른 공식이 사용된다. 다음 공식과 코드를 이용하여 `mymedian`이라는 이름의 함수를 만들고 입력 값들의 중간값을 구해서 반환하는 함수를 만드시오. (%% 나머지 연산, `if`문 사용, 아래 중간값 코드 참고)

$$\text{median}(X) = \begin{cases} \frac{1}{2}X\left[\frac{n}{2}\right] + \frac{1}{2}X\left[1 + \frac{n}{2}\right] & \text{if } n \text{ is even} \\ X\left[\frac{n+1}{2}\right] & \text{if } n \text{ is odd} \end{cases}$$

```
sorted_x <- sort(x)
#
retval <- sort_x[n/2]/2 + sort_x[1+(n/2)]/2
#
retval <- sort_x[(n+1)/2]
```

### 4.6.3 for, while, repeat

`for` 문은 반복적으로 특정 코드를 실행하고자 할 때 사용됩니다. 다음과 같은 형식으로 사용할 수 있습니다.

```
for(var in seq){
 expression
}
```

`var`는 반복을 돌 때마다 바뀌는 변수로 {} 안에서 사용되는 지역 변수입니다. `seq`는 vector 형식의 변수로 반복을 돌 때마다 순차적으로 `var`에 저장되는 값을 합니다.

```
x <- 1:10
for(i in x){
 cat(i, "\n")
 flush.console()
}
```

```
sum_of_i <- 0
for(i in 1:10){
 sum_of_i <- sum_of_i + i
 cat(i, " ", sum_of_i, "\n"); flush.console()
}
```

while문도 for문과 같이 반복적으로 특정 코드를 수행하고자 할 때 사용합니다. 사용하는 문법은 다음과 같으며 cond는 True 또는 False로 반복되는 조건문을 넣고 True 일 경우 계속해서 반복하면서 expressions를 수행하며 이 반복은 cond가 False로 될 때 까지 계속됩니다.

```
while(cond){
 expression
}
```

while문을 사용할 경우 다음과 같이 indicator라 불리우는 변수를 하나 정해서 반복 할 때마다 값이 바꿔도록 해 주어야 합니다. 그렇지 않으면 무한 루프를 둘게 되는 문제가 발생합니다.

```
i <- 10
f <- 1
while(i>1){
 f <- i*f
 i <- i-1
 cat(i, f, "\n")
}
f
factorial(10)
```

repeat 명령은 조건 없이 블럭 안에 있는 코드를 무조건 반복하라는 명령입니다. 따라서 블럭 중간에 멈추기 위한 코드가 필요하고 이 명령이 break입니다.

```
repeat{
 expressions
 if(cond) break
}

i <- 10
f <- 1
repeat {
 f <- i*f
 i <- i-1
 cat(i, f, "\n")
 if(i<1) break
}
f
factorial(10)
```

#### 4.6.4 Avoiding Loops

R에서는 가능하면 loop문을 사용하지 않는 것이 좋습니다. 이는 다른 언어들 보다 반복문이 느리게 수행된다는 이유 때문이기도 합니다. 그러나 R에서는 반복문을 수행하는 것 보다 훨씬 더 빠르게 반복문을 수행 한 것과 같은 결과를 얻을 수 있는 다양한 방법들이 제공되고 있습니다. 차차 그런 기법들에 대한 학습을 진행하도록 하겠습니다.

```
x <- 1:1E7
sum(x)
system.time(sum(x))

st <- proc.time()
total <- 0
for(i in 1:length(x)){
 total <- total + x[i]
}
ed <- proc.time()
ed-st
```

#### Exercises

1. 다음 네 사람의 이름과 나이를 데이터로 갖는 `users` 변수를 (data.frame) 만드시오

```
user_score <- c(90, 95, 88, 70)
user_names <- c("John", "James", "Sara", "Lilly")
```

2. 각 사람의 점수가 80보다 작으면 : Fail 크면 : Pass를 출력을 하는 코드를 작성하시오. 예를 들어 John의 점수는 80보다 크므로 John 90: Pass 출력 (`for`, `print` 함수 이용)

## 4.7 Object Oriented Programming (Advanced)

OOP는 객체지향 프로그래밍 이라고 합니다. OOP를 이용해서 프로그래밍으로 풀고자 하는 문제를 좀 더 명확하게 개념을 수립하고 복잡한 코드를 명료하게 만들 수 있습니다. 그런데 R에서 OOP는 다른 언어보다는 좀 더 어려운 개념적인 이해가 필요합니다. S3, S4, 그리고 Reference class 가 있으며 S3, S4는 Generic function을 이용하며 다른 언어에서 사용하는 OOP 개념과는 다릅니다. Reference class는 다른 언어에서 사용하는 OOP 개념과 유사하며 R6 패키지를 이용해서 사용할 수 있습니다.

---

이 저작물은 크리에이티브 커먼즈 저작자표시-비영리-변경금지 4.0 국제 라이선스에 따라 이용할 수 있습니다.

# Chapter 5

## Data transformation

일반적인 데이터 분석은 데이터 전처리(변환), 가시화, 모델링(통계분석)의 반복적인 수행으로 진행될 수 있습니다. R에서는 `data.frame` 형식의 데이터 타입이 주로 사용되며 (최근 `tibble`형식) 따라서 `data.frame` 기반의 데이터를 다루기 위한 다양한 함수를 익힐 필요가 있습니다. 이번 강의에서는 `data.frame` 데이터를 읽거나 쓰는 함수들과 함께 데이터 전처리를 (변환) 위한 함수들을 배워보겠습니다.

앞에서 배웠던 데이터를 저장하는 object의 종류를 먼저 간략히 정리해 봅니다.

- **Vectors** - 같은 타입의 데이터를 (Numeric, character, factor, ...) 저장한 오브젝트로 인덱스는 `[, ]` 사용.
- **Lists** - 여러개의 `vector`를 원소로 가질 수 있으며 각 원소 `vector`들은 문자나 숫자 어떤 데이터 타입도 가능하고 길이가 달라도 됨. `list`의 인덱싱에서 `[]`는 리스트를 반환하고 `[[ ]]`는 `vector`를 반환함.
- **Matrices** - 같은 타입의 데이터로 채워진 2차원 행렬이며 인덱스는 `[i, j]` 형태로 `i`는 row, `j`는 column을 나타냄. 메트릭스의 생성은 `matrix` 명령어를 사용하며 왼쪽부터 column 값을 모두 채우고 다음 컬럼 값을 채워 나가는 것이 기본 설정임. `byrow=T`를 통해 row를 먼저 채울 수도 있음. `row`와 `column` 이름은 `rownames`와 `colnames`로 설정이 가능하며 `rbind`와 `cbind`로 두 행렬 또는 행렬과 백터를 연결할 수 있음 ( `rbind`와 `cbind`의 경우 행렬이 커지면 컴퓨터 리소스 많이 사용함)
- **data.frame** - `list`와 `matrix`의 특성을 모두 갖는 오브젝트 타입으로 `list`와 같이 다른 타입의 `vector`형 변수 여러개가 컬럼에 붙어서 `matrix` 형태로 구성됨. 단, `list`와는 다르게 각 변수의 길이가 (`row`의 길이) 같아야 함. `$` 기호로 각 변수들을 인덱싱(접근) 할 수 있고 `matrix`와 같이 `[i, j]` 형태의 인덱싱도 가능.

## 5.1 Reading and writing

파일에 있는 데이터를 R로 읽어들이거나 쓰는 일은 일반적인 데이터 분석 과정에서 필수적일 수 있습니다. 본 강의에서는 일반적으로 사용하는 텍스트 파일과 엑셀파일을 활용하는 방법을 알아보겠습니다.

### 5.1.1 Text file

면의상 데이터를 쓰는 과정을 먼저 살펴봅니다. UsingR 예제에 있는 데이터 중 batting 데이터는 2002 야구시즌에 선수들의 정보를 모아둔 데이터입니다. str 함수로 데이터 전체적인 구조를 파악한 일부 데이터만을 (홈런 개수와 스트라이크 아웃) 이용해 추가적인 데이터를 생성한 후 별도로 파일에 저장해 보겠습니다.

```
library(UsingR)
data(batting)
str(batting)

mydf <- data.frame(id = batting$playerID,
 team = batting$teamID,
 hr = batting$HR,
 so = batting$SO,
 soperhr = batting$SO/batting$HR)

head(mydf)
```

재미삼아 홈런과 삼진아웃간의 관계를 한 번 알아보겠습니다.

```
plot(mydf$hr, mydf$so)
mycor <- cor(mydf$hr, mydf$so)
fit <- lm(mydf$so ~ mydf$hr)
plot(mydf$hr, mydf$so); abline(fit); text(50, 170, round(mycor,2))
```

위 데이터를 파일로 저장하기 위해서는 write.table 또는 write.csv 함수를 사용할 수 있습니다. 패키지에 따라서 다양한 함수들이 제공되고 있지만 위 두 파일은 utils라는 R의 기본 패키지에 들어있는 함수들로서 가장 많이 사용되는 함수들입니다. ?write.table 등으로 도움말을 보시고 특히 함수의 전달값 (Arguments) 들을 (quote, row.names, col.names, sep) 익혀두시기 바랍니다.

```
write.table(mydf, file="table_write.txt")
write.table(mydf, file="table_write.txt", quote=F)
write.table(mydf, file="table_write.txt", quote=F, row.names=F)
write.table(mydf, file="table_write.txt", quote=F, row.names=F, sep=",")
write.table(mydf, file="table_write.csv", quote=F, row.names=F, sep=",")
```

대부분의 텍스트 파일은 아래와 같이 csv 또는 txt 파일로 저장하여 메모장으로 열어 확인할 수 있으며 읽어올 때 구분자 (sep 파라미터) 나 header를 (header 파라미터) 읽을지 등을 옵션으로 지정할 수 있습니다.

```
dat <- read.csv("Dataset_S1_sub.txt")
head(dat)
```

Dataset\_S1\_sub.txt 파일을 열어보면 다음과 같이 header와 “,”로 구분되어 있는 것을 볼 수 있습니다. `read.csv` 함수의 도움말을 보면 이 함수의 파라미터 `head`와 `sep`이 기본값으로 `T`와 `,`로 되어 있는 것을 볼 수 있습니다. `read.csv` 외에도 `read.table`, `read.delim` 등의 함수를 이용해서 텍스트 파일을 읽어올 수 있습니다.

```
str(dat)
```

### Exercises

1. 위 `mydf`에서 가장 훈련을 많이 순서로 데이터를 정렬하시오
2. 위 정렬된 데이터를 `mydf2`로 저장한 후 csv 형태의 `baseball.csv` 파일로 저장하시오

### 5.1.2 Excel file

텍스트 파일 외에 엑셀파일은 `readxl`이라는 R 패키지를 활용하여 읽거나 쓸 수 있습니다. 패키지는 다음과 같은 방법으로 설치할 수 있으며 `read_excel`이라는 함수를 사용해서 데이터를 읽어들일 수 있습니다. 참고로 이 후 강의에서 배우게 될 `tidyverse` 패키지들 중 `readr` 패키지를 사용하여 엑셀파일 데이터를 다룰 수도 있습니다.

```
install.packages("readxl")
library(readxl)
```

실습 파일은 형광 세포를 배양하여 형광리더기를 이용해 얻어진 실제 데이터이며 `plate_reader.xls`에서 다운로드 받을 수 있습니다. `read_excel` 함수를 이용하여 파일의 내용을 읽어오면 기본 자료형이 `tibble`입니다. `tibble`은 최근 많이 쓰이는 R object로 `data.frame`과 유사하나 입력값의 `type`, `name`, `rowname`을 임으로 바꿀 수 없다는 점이 다릅니다.

```
dat <- read_excel("plate_reader.xls", sheet=1, skip = 0, col_names=T)
```

엑셀파일에는 두 종류의 ( $OD600_{nm}$ , Fluorescence) 데이터가 저장되어 있습니다. 첫 번째 sheet에는 다음처럼 wide형 데이터가 저장되어 있습니다.

	A	B	C	D	E	F	G	H
1	Plate	Repeat	Well	Type	Time	595nm_kl Time	EGFP_suli	
2	1	1	B01	M	00:00:15.93	0.701	00:01:11.48	67809
3	1	1	B02	M	00:00:16.32	0.752	00:01:11.89	60025
4	1	1	B03	M	00:00:16.69	0.723	00:01:12.30	102745
5	1	1	B04	M	00:00:17.06	0.744	00:01:12.71	99979
3	1	1	B05	M	00:00:17.43	0.706	00:01:13.12	108175
7	1	1	B06	M	00:00:17.80	0.723	00:01:13.53	109575
3	1	1	B07	M	00:00:18.17	0.767	00:01:13.94	76531
9	1	1	B08	M	00:00:18.54	0.777	00:01:14.35	72137
0	1	1	B09	M	00:00:18.91	0.762	00:01:14.76	154549
1	1	1	B10	M	00:00:19.28	0.798	00:01:15.17	128498
2	1	1	B11	M	00:00:19.65	0.793	00:01:15.58	151693
3	1	1	B12	M	00:00:20.02	0.821	00:01:15.99	130526
4	1	1	C01	M	00:00:22.85	0.803	00:01:18.86	42654

프로토콜 상세 내역이 나온 세 번째 시트를 읽을 경우 sheet 옵션을 3로 설정하면 되며 skip=3으로 하고 컬럼 이름을 별도로 사용하지 않으므로 col\_names=T로 하여 읽을 수 있습니다.

```
dat <- read_excel("plate_reader.xls", sheet=3, skip = 3, col_names=F)
```

참고로 엑셀파일로 저장하기 위해서는 csv 파일로 데이터를 writing 한 뒤 Excel로 해당 csv 파일을 열고xlsx 파일로 저장할 수 있습니다.

## 5.2 Subset

R에서 데이터 저장은 data.frame이나 matrix 타입을 일반적으로 사용합니다. 이 데이터의 일부 열 또는 행의 데이터만을 가져와서 별도로 저장하거나 분석이 필요할 경우가 있습니다. 이 때 인덱싱을 사용해서 일부 데이터를 선택하고 사용할 수 있으며 subset 함수도 이러한 선별 기능을 제공합니다. subset은 행과 열 모두를 선별할 수 있는 함수입니다. 다음 airquality 데이터는 1973년 날짜별로 뉴욕의 공기질을 측정한 데이터입니다. NA를 제외한 나머지 데이터만으로 새로운 데이터셋을 만들어 봅시다. is.na함수를 사용하면 해당 데이터가 NA일 경우 TRUE, NA가 아닐 경우 FALSE를 반환해 줍니다.

```
is.na(airquality$Ozone)
ozone_complete1 <- airquality[!is.na(airquality$Ozone),]
ozone_complete1 <- subset(airquality, !is.na(Ozone))
```

위 ozone\_complete1와 ozone\_complete2는 같은 결과를 보입니다. 그러나 ozone\_complete1 보다는 ozone\_complete2 코드가 훨씬 직관적이고 가독성이 높습니다. 특히 airquality\$Ozone로 \$를 사용하여 변수에 접근한 것이 아닌 Ozone이라는 변수 이름을 직접 사용해서 접근함으로써 코드의 간결성과 가독성을 유지할 수 있습니다. 또한 subset의 select 옵션을 이용해서 변수를 선택할 수도 있으며 &(AND)와 |(OR) 연산자를 사용해서 조건을 두 개 이상 설정할 수 있습니다. 아래 select 옵션에서 -는 해당 변수를 제외한다는 의미입니다.

```
ozone_complete3 <- subset(airquality, !is.na(ozone), select=c(ozone, temp, month, day))
ozone_complete4 <- subset(airquality, !is.na(ozone) & !is.na(solar.r), select=c(-month, -day))
```

### Exercises

airquality 데이터에서 Temp와 Ozone 변수로 이루어진 df라는 이름의 `data.frame`을 만드시오 (단 NA가 있는 샘플(열)은 모두 제외하시오)

## 5.3 Merging and Split

`merge` 함수는 두 개 이상의 데이터셋을 통합하는 기능을 수행하는 함수입니다. 특히 `rbind`나 `cbind`와는 다르게, 결합하는 두 데이터에 공통적이거나 한 쪽의 데이터를 기준으로 결합을 수행 합니다. `?merge`를 참고하면 `by`, `by.x`, `by.y`, `all`, `all.x`, `all.y` 등의 옵션으로 이러한 설정을 수행할 수 있습니다. 간단한 예제를 통해서 이해해 보겠습니다.

10명의 사람이 있고 이 사람들의 나이와 성별을 각각 나타낸 두 데이터셋이 있습니다. 그런데 `df1`은 나이만을 `df2`는 성별 정보만을 가지고 있으며 두 정보 모두 제공된 사람은 3명 (인덱스 4,5,6) 뿐입니다. 이제 `merge`를 이용해서 두 데이터셋을 결합해 보겠습니다.

```
merge
df1 <- data.frame(id=c(1,2,3,4,5,6), age=c(30, 41, 33, 56, 20, 17))
df2 <- data.frame(id=c(4,5,6,7,8,9), gender=c("f", "f", "m", "m", "f", "m"))

df_inner <- merge(df1, df2, by="id", all=F)
df_outer <- merge(df1, df2, by="id", all=T)
df_left_outer <- merge(df1, df2, by="id", all.x=T)
df_right_outer <- merge(df1, df2, by="id", all.y=T)
```

만약 두 데이터셋의 `id`가 다를 경우나 각각 다른 기준으로 결합해야 하는 경우는 `by` 대신 `by.x`, `by.y` 옵션을 사용할 수 있습니다.

`split` 함수는 데이터를 특정 기준으로 나누는 역할을 하며 해당 기준은 `factor` 형 벡터 형태로 주어질 수 있습니다. 예를 들어 `airquality` 데이터의 `month` 변수를 기준으로 데이터를 분리해 보겠습니다.

```
str(airquality)
g <- factor(airquality$Month)
airq_split <- split(airquality, g)
class(airq_split)
str(airq_split)
```

위와 같이 `airq_split`은 길이가 5인 (5, 6, 7, 8, 9월) `list` 타입이 되었고 각 요소는 서로 다른 `size`의 `data.frame` 형으로 구성 된 것을 확인할 수 있습니다.

## 5.4 Transformation

R에서 기존 가지고 있는 데이터의 변경은 새로운 변수의 추가, 삭제, 변형과 샘플의 추가, 삭제, 변형을 생각해 볼 수 있습니다. 이러한 기능은 앞에서 배운 `merge`, `split`이나 `rbind`, `cbind`, 그리고 인덱싱을 활용한 값 변경 등의 방법을 이용할 수 있습니다. 또한 가장 직관적으로 필요한 변수들을 기존 데이터셋에서 추출한 후 `data.frame` 명령어를 사용해서 새로운 데이터셋으로 만들어주면 될 것 입니다.

이러한 방법들 외에 `within`을 사용할 경우 특정 변수의 변형과 이를 반영한 새로운 데이터셋을 어렵지 않게 만들수 있습니다. `with` 함수의 사용 예와 함께 `within` 함수를 사용하여 데이터를 변형하는 예를 살펴봅니다. `with`나 `within` 함수는 R을 활용하는데 많이 사용되는 함수들은 아닙니다. 또한 이러한 기능들은 `dplyr` 등의 패키지에서 제공하는 경우가 많아서 필수적으로 익힐 부분은 아닙니다. 그러나 개념적인 이해를 돋기위한 좋은 도구들이며 여전히 고수준의 R 사용자들이 코드에 사용하고 있는 함수들이므로 알아두는 것이 좋습니다.

```
without with
ozone_complete <- airquality[!is.na(airquality$Ozone), "Ozone"]
temp_complete <- airquality[!is.na(airquality$Temp), "Temp"]
print(mean(ozone_complete))
print(mean(temp_complete))

with
with(airquality, {
 print(mean(Ozone[!is.na(Ozone)])))
 print(mean(Temp[!is.na(Temp)])))
})
```

위 `with` 함수에서 보는바와 같이 \$를 이용한 변수 접근 대신 `with`함수 내에서는 `{, }` 안에서 해당 `data.frame`에 있는 변수 이름을 직접 접근할 수 있으며 따라서 코드의 간결함과 가독성이 향상됩니다.

`within` 함수는 `with`함수와 같이 `{, }` 안에서 변수의 이름만으로 해당 변수에 접근이 가능하나 입력된 데이터와 변경된 변수(들)을 반환한다는 점이 다릅니다. 아래 예는 `airquality` 데이터의 화씨 (Fahrenheit) 온도를 섭씨 (Celsius) 온도로 변환해서 새로운 데이터셋을 만드는 코드입니다. `data.frame`을 이용한 코드와 비교해 보시기 바랍니다. 데이터셋 내에서 참조할 변수들이 많아질 경우 `airquality$xxx` 식의 코드를 줄이는 것 만으로도 코드의 가독성과 간결성을 유지할 수 있습니다.

```
newairquality <- within(airquality, {
 celsius = round((5*(Temp-32))/9, 2)
})
head(newairquality)

data.frame
celsius <- round((5*(airquality$Temp-32))/9, 2)
newairquality <- data.frame(airquality, celsius)
head(newairquality)
```

### Exercises

다음 df 의 hour, minute, second로 나누어진 값들을 초 단위로 변환하여 seconds라는 변수에 저장한 후 기존 df에 추가한 df2 데이터셋을 만드시오 (within 함수 이용)

```
df <- data.frame(hour=c(4, 7, 1, 5, 8),
 minute=c(46, 56, 44, 37, 39),
 second=c(19, 45, 57, 41, 27))
```

## 5.5 Babies example

UsingR 패키지의 babies 데이터를 이용해서 산모의 흡연 여부와 신생아 몸무게의 관계를 알아보는 분석을 수행해 보겠습니다. 본 강의를 통해 배우지 않은 내용들이 있지만 코드를 따라 가면서 참고하시기 바랍니다. 우선 UsingR 패키지를 로딩합니다. 산모의 임신 기간이 (gestation) 999로 표기된 데이터는 명백히 에러이며 이들을 NA로 처리합니다.

```
library(UsingR)
head(babies)
a simple way to checkout the data
plot(babies$gestation)
babies$gestation[babies$gestation>900] <- NA
str(babies)
```

아래와 같이 within 함수를 사용해서 babies\$ 를 반복해서 입력해주는 불편함을 줄이고 가독성을 높입니다. 똑같은 방법으로 dwt (아빠의 몸무게) 변수의 에러값들에 대해서도 NA 처리를 할 수 있습니다.

```
new_babies <- within(babies, {
 gestation[gestation==999] <- NA
 dwt[dwt==999] <- NA
})
str(new_babies)
```

smoke 변수는 흡연 여부를 나타내는 범주형 변수로 0, 1, 2, 3 값은 의미가 없습니다. 사람이 읽을 수 있는 label을 붙인 factor 형 변수로 변환하는 코드도 함께 작성해 보겠습니다.

```
str(babies$smoke)
new_babies <- within(babies, {
 gestation[gestation==999] <- NA
 dwt[dwt==999] <- NA
 smoke = factor(smoke)
 levels(smoke) = list(
 "never" = 0,
 "smoke now" = 1,
 "until current pregnancy" = 2,
```

```

 "once did, not now" = 3)
}
str(new_babies$smoke)

```

이제 임신기간과 흡연 여부를 분석해 볼 수 있습니다. 흡연 그룹별로 기간에 차이가 있는지를 알아보는 분석은 t-test나 ANOVA를 사용할 수 있습니다.

```

fit <- lm(gestation~smoke, new_babies)
summary(fit) ## t-test
anova(fit)

```

간단히 결과를 보면 `summary(fit)`은 3가지 t-test의 결과를 보여줍니다. never vs. smoke new 의 경우 t값이 -1.657로 피우지 않은 경우에 비해서 피우는 사람의 임신 기간이 유의하게 줄어들었음을 알 수 있습니다. 그에 비해서 현재 흡연하지 않는 경우 (never vs. until current pregnancy 또는 never vs. once did, not now) 차이가 없는 것으로 나옵니다.

이제 `smoke now` 인 경우 또는 나이가 25세 미만인 경우의 샘플에 대해서 `newdf`를 만들어 봅니다 (`subset` 함수 사용, id, gestation, age, wt, smoke 변수 선택). 이후 `ggplot`을 이용하여 몸무게와 임신기간의 산점도를 그려보면 크게 다르진 않으나 흡연하는 여성 중 몸무게가 적게 나가는 여성에게서 짧은 임신기간을 갖는 경향을 볼 수 있습니다.

```

newdf <- subset(new_babies, (smoke=="smoke now" | smoke == "never") & age < 25, select=-
ggplot(newdf, aes(x=wt, y=gestation, color=smoke)) +
geom_point(size=3, alpha=0.5) +
facet_grid(.~smoke) +
theme_bw()

```

## 5.6 Useful functions

지금까지 배운 여러 R 프로그래밍 기법이나 함수들과 같이 R을 활용한 데이터 분석에서 자주쓰이거나 유용하게 사용되는 함수들을 소개합니다. 먼저 원소들을 비교하여 공통적 또는 유일한 원소들만을 추출해내는 함수들입니다.

```

#match(), %in%, intersect()

x <- 1:10
y <- 5:15
match(x, y)
x %in% y
intersect(x, y)

#unique()
unique(c(x, y))

```

다음은 스트링 관련 함수들로서 서열데이터 분석 등에서 유용하게 활용되는 함수들입니다.

```
#substr()
x <- "Factors, raw vectors, and lists, are converted"
substr(x, 1, 6)

#grep()
grep("raw", x)

#grepl()
grepl("raw", x)
if(grepl("raw", x)){
 cat("I found raw!")
}

x <- paste(LETTERS, 1:100, sep="")
grep("A", x)
x[grep("A", x)]

grepl("A", x)
r <- grepl("A", x)
if(r){
 cat("Yes, I found A")
}else{
 cat("No A")
}

#strsplit()
x <- c("Factors, raw vectors, and lists, are converted", "vectors, or for, strings with")
y <- strsplit(x, split = ,)

#unlist()
unlist(y)

y <- strsplit(x, split = "")
ychar <- unlist(y)
ycount <- table(y2)
ycount_sort <- sort(ycount)
ycount_sort <- sort(ycount, decreasing = T)
ycount_top <- ycount_sort[1:5]
ycount_top_char <- names(ycount_top)

#toupper(), tolower()
toupper(ycount_top_char)
```

### Exercises

built-in 데이터셋 중 state.abb 은 미국의 50개 주에대한 축약어임.

- 1) 이 중 문자 A 가 들어가는 주를 뽑아 x에 저장 하시오 (grep 또는 grep1 사용)
- 2) state.abb 중 위 x에 저장된 이름들을 빼고 y에 저장 하시오 (match() 또는 %in%사용)
- 3) state.abb에 사용된 알파벳의 갯수를 구하고 가장 많이 쓰인 알파벳을 구하시오 (strsplit(), table() 등 사용)

## 5.7 apply

apply는 데이터를 변형하기 위한 함수라기 보다는 데이터를 다룰 때 각 원소별, 그룹별, row, 또는 column 별로 반복적으로 수행되는 작업을 효율적으로 수행할 수 있도록 해주는 함수입니다. apply 계열의 함수를 적절히 사용하면 효율성이나 편리성 뿐만 아니라 코드의 간결성 등 많은 장점이 있습니다. 쉬운 이해를 위해 colMean 함수를 예로 들면 colMean은 column 또는 row 단위로 해당하는 모든 값들에 대해 평균을 계산해주는 함수이고 apply를 사용할 경우 다음과 같이 apply 함수와 mean 함수를 이용해서 같은 기능을 수행할 수 있습니다. 아래는 babies 데이터의 cleanning 된 (위에서 만들었던) new\_babies 데이터에 이어서 수행되는 내용입니다.

```
library(UsingR)
head(babies)
df <- subset(babies, select=c(gestation, wt, dwt))
colMeans(df, na.rm=T)
apply(df, 2, mean, na.rm=T)
```

위와 같이 colMeans와 apply가 똑같은 결과를 보여주고 있습니다. 두 번째 인자인 margin의 값으로 (?apply참고) 여기서는 2가 사용되었으며 margin 값이 1인지 2인지에 따라서 다음과 같이 작동을 합니다.

	열 (2)		
행 (1)	gestation	wt	dwt
	284	120	110
	282	113	148
	279	128	NA
	NA	123	197
	282	108	NA
	286	136	130

`mean`외에도 다양한 함수들이 사용될 수 있으며 아래와 같이 임의의 함수를 만들어서 사용할 수 도 있습니다. 아래 코드에서는 `function(x)...`로 바로 함수의 정의를 넣어서 사용했으나 그 아래 `mysd` 함수와 같이 미리 함수 하나를 만들고 난 후 함수 이름을 이용해서 `apply`를 적용할 수 있습니다.

```
apply(df, 2, sd, na.rm=T)
apply(df, 2, function(x){
 xmean <- mean(x, na.rm=T)
 return(xmean)
})
```

`apply` 함수는 특히 R에서 느리게 작동하는 `loop` (`for`, `while` 등) 문 대신 사용되어 큰 행렬에 대해서도 빠른 계산 속도를 보여줄 수 있습니다.

```
n <- 40
m <- matrix(sample(1:100, n, replace=T), ncol=4)
mysd <- function(x){
 xmean <- sum(x)/length(x)
 tmpdif <- x-xmean
 xvar <- sum(tmpdif^2)/(length(x)-1)
 xsd <- sqrt(xvar)
 return(xsd)
}

for
results <- rep(0, nrow(m))
for(i in 1:nrow(m)){
 results[i] <- mysd(m[i,])
}
print(results)
apply(m, 1, mysd)
```

```
apply(m, 1, sd)
```

apply 함수 외에도 sapply, lapply, mapply 등의 다양한 apply계열 함수가 쓰일 수 있습니다. 먼저 lapply는 matrix 형태 데이터가 아닌 list 데이터에 사용되어 각 list 원소별로 주어진 기능을 반복해서 수행하며 sapply는 lapply와 유사하나 벡터, 리스트, 데이터프레임 등에 함수를 적용할 수 있고 그 결과를 벡터 또는 행렬로 반환합니다.

```
x <- list(a=1:10, b=exp(-3:3), logic=c(T,T,F,T))
mean(x$a)
lapply(x, mean)
sapply(x, mean)

x <- data.frame(a=1:10, b=exp(-4:5))
sapply(x, mean)

x <- c(4, 9, 16)
sapply(x, sqrt)
sqrt(x)

y <- c(1:10)
sapply(y, function(x){2*x})
y*2
```

마지막 예제에서처럼 sapply나 lapply도 임의의 함수를 만들어 적용시킬 수도 있습니다. 자세히 살펴 보면 y는 10개의 값을 갖는 벡터이고 이 벡터의 각 원소(값)에 함수를 반복해서 적용하는 것입니다. 함수에서 x는 각 원소의 값을 차례차례 받는 역할을 하므로 1부터 10까지 값이 함수로 들어가 2를 곱한 수가 반환됩니다. 따라서 벡터연산을 하는 y\*2와 결과가 같으나 원하는 함수를 정의해서 자유롭게 사용할 수 있다는 장점이 있습니다. 리스트의 경우는 다음과 같이 사용합니다.

```
y <- list(a=1:10, b=exp(-3:3), logic=c(T,T,F,T))
myfunc <- function(x){
 return(mean(x, na.rm=T))
}
lapply(y, myfunc)
unlist(lapply(y, myfunc))
```

즉, myfunc의 x가 list y의 각 원소들, y[[1]], y[[2]], y[[3]]를 각각 받아서 mean 연산을 수행해 줍니다. 결과로 각 list 원소들의 평균 값이 반환되며 unlist 함수는 list 형태의 반환 값을 vector 형태로 전환해 줍니다.

### Exercises

다음은 앞에서 수행했던 airquality 데이터를 월별로 나눈 데이터셋임. 이 데이터셋을 이용하여 각 월별로 온도와 오존 농도의 평균값을 저장한 data.frame 형식의 데이터를 만들기 위하여 다음 단계별 과정에 적절한 코드를 작성하시오

```
dataset
g <- factor(airquality$month)
airq_split <- split(airquality, g)
```

- 1) 다음 df의 ozone 평균을 구하는 ozone\_func 함수를 작성하시오 (단 입력은 data.frame 형식의 오브젝트를 받고 출력은 평균값 (정수 값 하나) 출력. mean 함수 사용시 데이터에 NA가 포함되어 있을 경우 na.rm=T 옵션 적용)

```
May data.frame
df <- airq_split[[1]]
#
write your code here for ozone_func function
#
Usage
ozone_func(df)
output
23.61538
```

- 2) lapply와 ozone\_func 함수를 사용하여 airq\_split list 데이터의 월별 ozone 평균 값을 구하고 ozone\_means에 vector 형식으로 저장하시오
- 3) 위 1), 2)와 같은 방법으로 temp\_func 함수를 만들고 월별 temp의 평균값을 temp\_means에 vector 형식으로 저장하시오.
- 4) 위에서 구해진 두 변수값들을 이용하여 air\_means라는 이름의 data.frame으로 저장하시오

### Exercises

1. 다음 코드를 이용해서 파일을 다운로드 하고 myexp에 저장하고 데이터의 구조 및 샘플들의 이름을 확인하시오

```
myexp <- read.csv("https://github.com/greendaygh/kribbr2022/raw/main/examples/gse93819_expression")
```

2. myexp의 1부터 10번째 샘플(컬럼) 데이터를 myexp1으로 11부터 20번째 샘플 데이터를 myexp2로 나누시오
3. myexp1의 row별 평균을 구해서 myexp1mean이 myexp2의 row별 평균을 구해서 myexp2mean에 저장하시오 (apply 이용)
4. myexp1mean과 myexp2mean을 합하여 myexpmean이라는 data.frame을 만드시오 (cbind이용, 주의필요)
5. plot을 이용하여 두 평균들의 산포도를 그리시오
6. myexpmean의 두 변수에 대한 차이를 구하여 mydiff라는 변수에 저장하시오
7. mydiff의 값들에 대한 히스토그램 (막대그래프)을 그리시오

이 저작물은 크리에이티브 커먼즈 저작자표시-비영리-변경금지 4.0 국제 라이선스에 따라 이용할 수 있습니다.

# Chapter 6

## R basic graphics

### 6.1 scatter plot

R에서 `plot` 함수는 가장 기본이 되는 그래프 함수입니다. 아래는 산포도를 그려주는 코드로서 `myxy`가 두 개의 변수(`x1`과 `y1`)를 가지고 있으므로 아래 명령들은 모두 같은 그림을 그려주게 됩니다.

```
x <- c(1:100)
y <- x*2 + rnorm(100)
myxy <- data.frame(x,y)
plot(myxy)
plot(myxy$x, myxy$y)
plot(x=myxy$x, y=myxy$y)
plot(y~x, data=myxy)
```

가장 마지막 명령은 `formula`를 사용한 `plot`으로 첫번째 파라메터 인자로 `formula` 타입이 전달되면 `plot.formula` 함수가 실행되며 `x`, `y` 값이 전달될 경우 `plot.default` 함수가 수행됩니다. R에서는 이렇게 전달되는 파라메터의 타입에 따라서 다른 기능을 하는 함수를 `Generic function`이라고 합니다. 만약 기존 그림에 추가 데이터의 산포를 그리고 싶은 경우 `points`라는 함수를 사용합니다.

```
z <- sample(1:100, 100, replace =T)
points(x, z)
points(x, z, col="red")
```

### 6.2 histogram

`hist` 함수는 데이터들의 분포를 히스토그램으로 그려주는 함수입니다. 히스토그램은 데이터들이 갖는 값을 특정 구간으로 나누고 각 구간에 해당하는 데이터가 몇 개인지

빈도수를 계산하여 막대그래프로 보여줍니다.

```
x <- rnorm(100)
hist(x, br=20, xlim=c(-3,3), main="Main text", xlab="X label")

hist(airquality$Wind, br=50)
hist(airquality$Wind, br=10)
```

### 6.3 boxplot

boxplot (상자 수염 그림)은 데이터의 여러가지 대표값 (중간값 median, 첫번째 사분위수 1st quantile, 세번째 사분위수 3rd quantile, 최소 minimum, 최대값 maximum) 등을 한눈에 볼 수 있도록 만들어놓은 그래프입니다. 수염이 나타내는 값은 최소값이나 최대값이 될 수 있고 또는 하위 1.5 IQR에서 최소 데이터와 상위 1.5 IQR 내에 최고 데이터를 나타낼 수 있으며 이 경우 그 외에 존재하는 값들은 outlier가 됩니다.

```
x <- rnorm(100)
boxplot(x)

r <- boxplot(airquality$Wind)

airquality$Wind[which(airquality$Wind > (1.5*(r$stats[4]-r$stats[2])+r$stats[4]))]

with(airquality, {
 Wind[which(Wind > (1.5*(r$stats[4]-r$stats[2])+r$stats[4]))]
})

with(airquality, {
 val <- (1.5*(r$stats[4]-r$stats[2])+r$stats[4])
 Wind[which(Wind > val)]
})

with(airquality, {
 iqr <- quantile(Wind, 3/4) - quantile(Wind, 1/4)
 val <- 1.5 * iqr + quantile(Wind, 3/4)
 Wind[which(Wind > val)]
})
```

data.frame 타입의 오브젝트에 대해서 boxplot을 그릴 경우 여러 변수의 데이터들의 분포를 한눈에 비교할 수 있습니다.

```
y <- rnorm(100, 1, 1)
#boxplot(y)
xy <- data.frame(x, y)
```

```
boxplot(xy)
class(xy)

mean_vals <- sample(10)
mymat <- sapply(mean_vals, function(x){rnorm(100, x)})
dim(mymat)
boxplot(mymat)
```

## 6.4 barplot

막대그래프는 각 값들을 막대 형태로 나란히 배치하여 서로 비교가 용이하도록 만든 그래프입니다. `table` 함수는 같은 값을 갖는 데이터들이 몇 개나 있는지 테이블을 만들어주는 함수입니다. `rbind`는 두 변수를 `row`를 기준으로 붙여주는 역할의 함수입니다.

```
x <- sample(1:12, 200, replace = T)
tab_x <- table(x)
y <- sample(1:12, 200, replace = T)
tab_y <- table(y)
tab_xy <- rbind(tab_x, tab_y)
barplot(tab_xy)
barplot(tab_xy, beside = T)
barplot(tab_xy, beside = T, col=c("darkblue","red"))
barplot(tab_xy, beside = T, col=c("darkblue","red"), xlab="Month")
barplot(tab_xy, beside = T, col=c("darkblue","red"), xlab="Month", horiz=TRUE)
```

### Exercises

- 1) `iris` 데이터의 꽃받침 (Sepal) 길이와 넓이를 각각 x와 y축으로 하는 산포도를 그리시오
- 2) `iris` 데이터에서 `setosa` 품종의 꽃받침의 (Sepal) 길이와 넓이 데이터를 빨간 점으로 나타내시오
- 3) `iris` 데이터에서 꽃받침과 (Sepal) 꽃잎의 (Petal) 길이의 분포를 그리시오 (`hist` 사용)
- 4) `iris` 데이터에서 꽃받침과 (Sepal) 꽃잎의 (Petal) 넓이의 분포를 그리시오 (`boxplot` 사용)
- 5) `iris` 데이터에서 품종별 꽃받침 (Sepal) 길이의 분포를 그리시오 (`boxplot` 사용)

## 6.5 Draw multiple graphs in the same plot

위 예제들에서 사용한 `high level function`들을 `low level function` (`lines`, `points`, `ablines`, `axis` 등)들과 함께 사용함으로써 원하는 도표 대부분을 그려낼

수 있습니다. 최근 널리 사용되는 `ggplot2` 패키지를 이용한 그래프 사용법 강의에서는 오늘 배우는 그래픽 명령어는 거의 사용하지 않습니다. 그러나 위 함수들은 R의 기본 그래프 함수들로서 단순한 도표에서부터 복잡한 그래픽까지 구현할 수 있는 다양한 유연성을 제공하므로 기본적인 사용법을 정확히 이해하는 것이 좋습니다.

아래 도표는 평균 0, 분산 1인 분포에서 500개의 랜덤한 수를 뽑아 `x`에 저장하고 `x`의 분포를 히스토그램으로 표현한 것입니다. 그리고 `x` 값들과 상관성이 있는 `y`값들을 (`x`에 2를 곱하고 평균 5, 분산 1인 랜덤하게 뽑힌 수를 노이즈로 더함) 생성하고 모든 1000개 값들의 분포를 그린 히스토그램입니다.

```
x <- rnorm(500)
hist(x, 100)
y <- 2*x + rnorm(500, mean=5, sd=1)
z <- c(x,y)
hist(z, br=100)
```

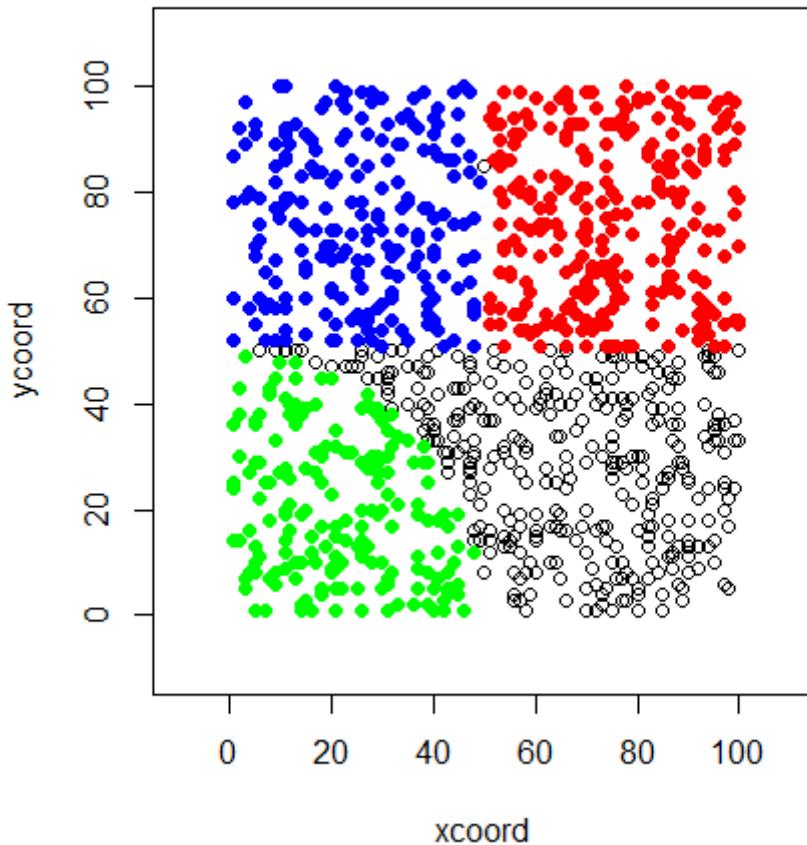
이제 위 `histogram` 그래프에 `density` 함수와 `lines` 함수를 조합하여 확률밀도함수 커브를 그려 넣을 수 있습니다. 이 때 `hist` 함수에 `probability=T` 옵션을 넣어 `y`축의 스케일을 확률밀도함수의 `y` 스케일과 맞춰주어 같은 화면에 그려지도록 했습니다.

```
hist(z, br=100)
hist(z, br=100, probability = T)
zd <- density(z)
lines(zd)
```

또한 아래 그래프는 위에서 생성한 `x`, `y` 값의 산포도를 그리고 `x`축과 `y`축 범위를 `xlim`, `ylim` 파라미터로 조절하는 예제입니다. `?pch` 도움말을 참고하여 다양한 포인트 모양을 선택할 수 있으며 `x` 값이 0 보다 작은 경우의 `index`를 뽑아 해당되는 `x` 값들과 그 값들의 짹이 되는 `y`값들에 대해서만 다시 포인트 그림을 red 색상으로 그려 넣었습니다. `lm`은 linear model의 약자로 회귀 곡선을 구할 때 사용하는 함수이며 이 함수를 `abline`과 조합하여 회귀 직선을 그릴 수 있습니다.

```
plot(x,y, xlim=c(-5, 5), ylim=c(-5, 15), pch=3)
idx <- which(x<0)
points(x[idx], y[idx], col="red")
fit <- lm(y~x)
abline(fit)
```

## Exercises



- 1부터 100까지 수를 랜덤하게 1000개 생성해서 x좌표를 생성하고 xcoord에 저장 하시오 (중복허용)
- 1부터 100까지 수를 랜덤하게 1000개 생성해서 y좌표를 생성하고 ycoord에 저장 하시오 (중복허용)
- x, y 좌표 평면에 xcoord와 ycoord 값을 이용하여 좌표를 (산포도) 그리되 x와 y의 범위가 모두 -10부터 110까지 되도록 지정 하시오 (plot 이용)
- 앞서 문제와 같은 plot에 x가 50보다 크고 y가 50보다 큰 곳에 있는 좌표들에 red closed circle로 표현하시오 (which, points, pch parameter 등 이용, 아래 참고)

```
idx <- which(xcoord>50 & ycoord>50)
points(x=xcoord[idx], y=ycoord[idx], col="red", pch=19)
```

- 앞서 문제와 같은 plot에 x가 50보다 작고 y가 50보다 큰 곳에 있는 좌표들에 blue closed circle로 표현하시오 (which, points, pch parameter 등 이용)
- 앞서 문제와 같은 plot에 원점으로부터 거리가 50 이하인 좌표들을 green closed circle로 표현 하시오

## 6.6 Usuful functions II

```
#match(), %in%, intersect()

x <- 1:10
y <- 5:15
match(x, y)
x %in% y
intersect(x, y)

#unique()
unique(c(x, y))

#substr()
x <- "Factors, raw vectors, and lists, are converted"
substr(x, 1, 6)

#grep()
grep("raw", x)

#grepl()
grepl("raw", x)
if(grepl("raw", x)){
 cat("I found raw!")
}

x <- paste(LETTERS, 1:100, sep="")
grep("A", x)
x[grep("A", x)]

grepl("A", x)
r <- grepl("A", x)
if(r){
 cat("Yes, I found A")
}else{
 cat("No A")
}

#strsplit()
```

```

x <- c("Factors, raw vectors, and lists, are converted", "vectors, or for, strings with")
y <- strsplit(x, split=", ")

#unlist()
unlist(y)

y <- strsplit(x, split="")
ychar <- unlist(y)
ycount <- table(y2)
ycount_sort <- sort(ycount)
ycount_sort <- sort(ycount, decreasing = T)
ycount_top <- ycount_sort[1:5]
ycount_top_char <- names(ycount_top)

#toupper(), tolower()
toupper(ycount_top_char)

```

### Exercises

built-in 데이터셋 중 state.abb 은 미국의 50개 주에 대한 축약어임.

- 1) 이 중 문자 A 가 들어가는 주를 뽑아 x에 저장 하시오 (grep 또는 grep1 사용)
- 2) state.abb 중 위 x에 저장된 이름들을 빼고 y에 저장 하시오 (match() 또는 %in% 사용)
- 3) state.abb에 사용된 알파벳의 갯수를 구하고 가장 많이 쓰인 알파벳을 구하시오 (strsplit(), table() 등 사용)

### Exercises

iris 데이터셋의 각 Species 별로 꽃잎과 꽃받침의 길이와 넓이에 대한 평균값들을 구하고 막대그래프를 그리시오

1. 각 species 별로 데이터를 나누시오 (list 형태)
2. list의 각 원소 (data.frame)의 변수들의 평균을 lapply를 사용하여 구하시오
3. do.call과 rbind 함수로 list의 원소를 통합해 data.frame을 생성하시오
4. barplot을 이용하여 막대그래프를 그리시오

이 저작물은 크리에이티브 커먼즈 저작자표시-비영리-변경금지 4.0 국제 라이선스에 따라 이용할 수 있습니다.



# Chapter 7

## tidyverse

tidyverse (<https://www.tidyverse.org/>)는 데이터 사이언스를 위한 R 기반의 독창적인 패키지들의 모음입니다. Rstudio의 핵심 전문가인 해들리위컴이 (Hadley Wickham) 중심이 되어 만들어 졌으며 기존의 툴보다 쉽고 효율적으로 데이터 분석을 수행할 수 있습니다.



R packages for data science

The tidyverse is an opinionated collection of R packages designed for data science. All packages share an underlying design philosophy, grammar, and data structures.

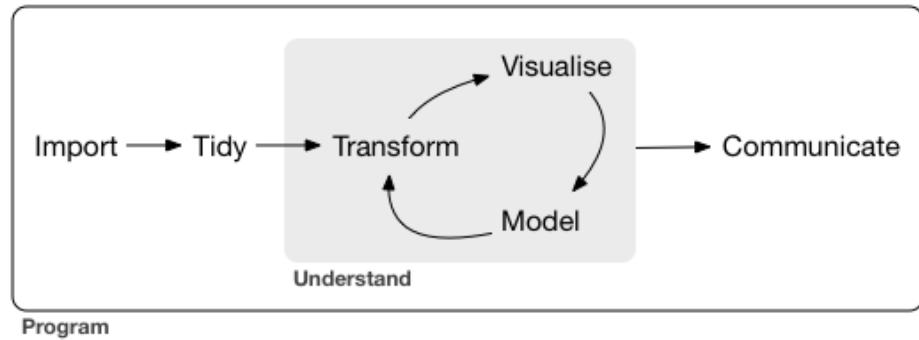
Install the complete tidyverse with:

```
install.packages("tidyverse")
```

데이터사이언스는 넓은 범위의 개념과 방법적인 정도가 있는 것은 아닙니다. 그러나 tidyverse의 목적은 데이터 분석을 위한 핵심이되는 고효율의 툴을 제공하는 것이며 그 철학은 다음과 같은 그림으로 요약할 수 있습니다.

### 7.1 tibble object type

R은 20년 이상된 비교적 오랜 역사를 가진 언어로서 `data.frame` 형태의 데이터 타입이 가장 많이 사용되고 있습니다. 그러나 당시에는 유용했던 기능이 시간이

Figure 7.1: from <https://r4ds.had.co.nz/>

흐르면서 몇몇 단점들이 드러나는 문제로 기존 코드를 그대로 유지한채 package 형태로 단점을 보완한 새로운 형태의 tibble 오브젝트 형식을 만들어 냈습니다. 대부분의 R 코드는 여전히 data.frame 형태의 데이터 탑입을 사용하고 있으나 tidyverse에서는 tibble이 사용되는 것을 참고하시기 바랍니다.

```

library(tidyverse)

tb <- tibble(
 x = 1:5,
 y = 1,
 z = x ^ 2 + y
)

as_tibble(iris)
head(iris)

```

tibble은 data.frame과 다음 몇 가지 점이 다릅니다. data.frame의 경우 탑입을 변환할 때 강제로 값의 탑입을 바꾸거나 내부 변수의 이름을 바꾸는 경우가 있었으나 tibble은 이를 허용하지 않습니다. 샘플들 (row) 이름을 바꿀수도 없습니다. 또한 프린팅할 때 출력물에 나오는 정보가 다르며 마지막으로 data.frame은 subset에 대한 탑입이 바뀔 경우가 있었지만 tibble은 바뀌지 않습니다.

```

x <- 1:3
y <- list(1:5, 1:10, 1:20)

data.frame(x, y)
tibble(x, y)

```

tibble은 컬럼 하나가 벡터형 변수가 아닌 리스트형 변수가 될 수 있다는 것도 data.frame과 다른 점입니다.

```

names(data.frame(`crazy name` = 1))
names(tibble(`crazy name` = 1))

```

또한 다음과 같이 사용되는 변수의 (x) 참조 범위가 다릅니다.

```
data.frame(x = 1:5, y = x ^ 2)
tibble(x = 1:5, y = x ^ 2)

df1 <- data.frame(x = 1:3, y = 3:1)
class(df1)
class(df1[, 1:2])
class(df1[, 1])

df2 <- tibble(x = 1:3, y = 3:1)
class(df2)
class(df2[, 1:2])
class(df2[, 1])
class(df2$x)
```

## 7.2 Tidy data structure

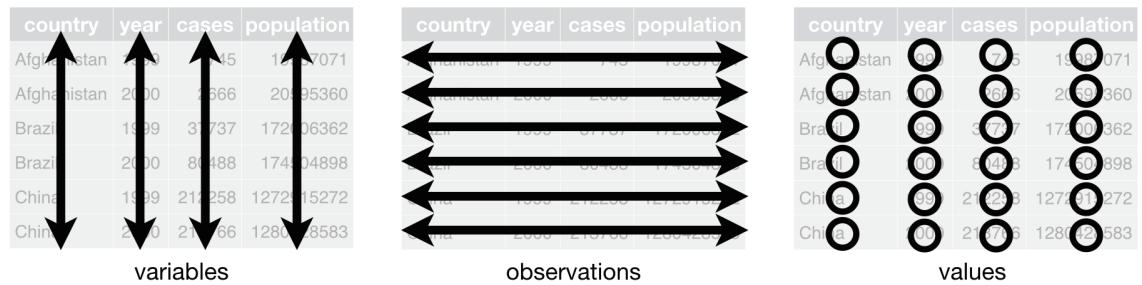
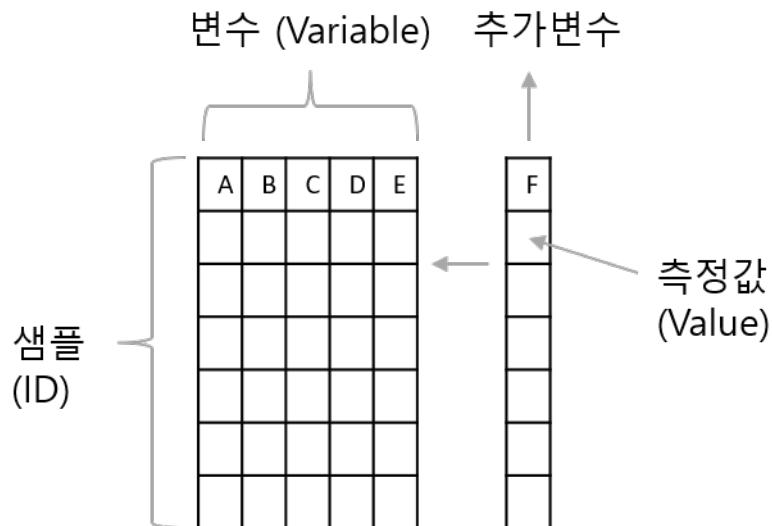
데이터의 변수와 값을 구분하는 일은 적절한 데이터 분석을 위해 필수적인 과정입니다. 특히 복잡하고 사이즈가 큰 데이터일 경우는 더욱 중요할 수 있으나 경험에 의존해서 구분을 하는 것이 대부분입니다. Tidy data는 이러한 변수와 값의 명확한 구분과 활용을 위한 데이터 구조중 하나입니다 (Hadley Wickham. Tidy data. The Journal of Statistical Software, vol. 59, 2014).

country	1999	2000
A	0.7K	2K
B	37K	80K
C	212K	213K

tidy data는 다음과 같은 특징이 있습니다.

- 각 변수는 해당하는 유일한 하나의 column을 가짐
- 각 샘플은 해당하는 유일한 하나의 row를 가짐
- 각 관측값은 해당하는 유일한 하나의 cell을 가짐

Tidy 데이터는 Long형 데이터로 알려져 있기도 합니다. 참고로 Wide형 데이터의 경우 샘플 데이터가 늘어날수록 row에 쌓이고 새로운 변수는 column에 쌓이는 방식으로 데이터가 확장되는 형태입니다. 엑셀에서 볼 수 있는 일반적인 형식으로 다음 그림과 같습니다.

Figure 7.2: from <https://r4ds.had.co.nz/>

Long형 데이터의 경우 ID, variable, value 세가지 변수만 기억하면 되겠습니다. 위 wide형 데이터 경우를 보면 ID, variable, 그리고 value 이 세가지 요인이 주요 구성 요소임을 알 수 있습니다. Long형으로 변환할 경우 샘플을 참조할 수 있는 어떤 변수 (variable)도 ID가 될 수 있으며 2개 이상의 변수가 ID로 지정될 수 있습니다. 참고로 ID를 지정할 경우 해당 ID는 가능하면 중복되지 않는 값들을 갖는 변수를 사용해야 식별자로서 기능을 적절히 수행할 수 있습니다. Long형을 사용할 경우 데이터의 변수가 늘어나도 행의 수만 늘어나므로 코딩의 일관성과 변수들의 그룹을 만들어서 분석하는 등의 장점이 있습니다. 아래는 새로운 변수 F가 추가될 때 long 형 데이터에 데이터가 추가되는 경우를 나타낸 그림입니다.

ID	variable	values
1	B	
1	C	
...	...	
2	B	
2	C	
...	...	

+

1	F	
2	F	
3	F	
4	F	
...	...	

추가변수

### 7.3 Pipe operator

tidyverse 패키지를 활용하기 위해서는 `%>%` 파이프 오퍼레이터의 이해가 필요합니다. 파이프 오퍼레이터의 작동법은 간단히 `%>%`의 왼쪽 코드의 결과를 출력으로 받아 오른쪽 코드의 입력(첫번째 파라미터의 값)으로 받아들이는 작동을 합니다(단축키: Shift+Ctrl+m). 다음 예에서 보면 `sin(pi)` 와 같은 함수의 일반적인 사용법 대신 `pi %>% sin`처럼 사용해도 똑같은 결과를 보여줍니다. `cos(sin(pi))`와 같이 여러 함수를 중첩하여 사용할 경우와 비교해서 코드의 가독성이나 효율 측면에서 크게 향상된 방법을 제공해 줍니다.

```
library(dplyr)

pi %>% sin
sin(pi)
pi %>% sin %>% cos
cos(sin(pi))
```

특히 `%>%`는 이후 설명할 dplyr의 `group_by`, `split`, `filter`, `summary` 등 행렬 편집/연산 함수를 빈번히 다양한 조합으로 쓰게되는 상황에서 더 큰 효과를 발휘할

수 있습니다.

```
x %>% paste("a", sep = "")
```

pipe operator의 왼쪽 구문의 결과가 오른쪽 구문의 첫 번째 파라미터의 입력 값으로 처리된다고 말씀 드렸습니다. 즉, 함수에서 사용되는 파라미터가 여러개일 경우가 있으므로 기본적으로 `%>%` 의 왼쪽 구문의 출력 값은 오른쪽 구문 (함수)의 첫 번째 인자의 입력 값으로 들어가는 것 입니다. 이는 다음 예들을 통해서 명확히 알 수 있습니다. 먼저 `paste`함수는 그 파라미터로 ,로 구분되는 여러개의 입력 값을 가질 수 있습니다. 따라서 다음 코드는 `x`가 `paste`의 첫 번째 파라미터로 들어가게 되어 "1a", "2a", "3a", "4a", "5a"로 `a` 앞에 `x` 값들이 붙어서 출력된 것을 알 수 있습니다.

```
x <- 1:5
x %>% paste("a", sep = "")
```

특정 데이터셋의 컬럼별 평균을 구하고 각 평균의 합을 구할 경우를 생각해 봅시다. R에서는 `colMeans`라는 특별한 함수를 제공하여 컬럼별로 평균을 계산해 줍니다. 그 후 `sum` 함수를 사용하여 최종 원하는 값을 얻을 수 있습니다. 이러한 코드를 `%>%` 오퍼레이터를 사용한 경우의 코드와 비교해 볼 수 있습니다.

```
x <- data.frame(x=c(1:100), y=c(201:300))
sum(colMeans(x))

x <- data.frame(x=c(1:100), y=c(201:300))
x %>% colMeans %>% sum
```

그럼 만약 두 번째 파라미터에 입력으로 왼쪽 구문의 출력을 받아들이고 싶을 경우는 `place holder` 을 사용하면 되겠습니다. `round` 함수는 두 개의 파라미터를 설정할 있 이으며 `digits`라는 두 번째 파라미터에 값을 pipe operator로 넘겨주고 싶을 경우 아래와 같이 표현할 수 있습니다.

```
6 %>% round(pi, digits=.)
round(pi, digits=6)
```

### Exercises

- 1) pipe operator를 사용해서 `airquality`데이터를 `long`형으로 전환하는 코드를 작성하시오 (단 `col` 파라미터에는 `Ozone`, `Solar.R`, `Wind`, `Temp` 변수를 넣음)

- 2) pipe operator를 사용해서 airquality데이터의 Month와 Day 변수(컬럼)을 Date 변수로 병합하는 코드를 작성하시오

## 7.4 Pivoting

일반적으로 얻어지는 데이터의 형태는 wide형이며 이를 Long형으로 변환하기 위해서는 tidyverse 패키지에 속한 tidyr 패키지의 pivot\_longer와 pivot\_wider를 사용합니다. 또한 reshape2 패키지의 melt 함수와 그 반대의 경우 dcast 함수를 사용할 수도 있습니다. 본 강의에서는 tidyr 패키지를 사용합니다. wide형 데이터를 long형으로 변환하거나 long형을 wide형으로 변환하는 작업을 pivoting이라고 합니다.

The diagram illustrates the pivoting process. On the left, a wide-format table shows 'country' (A, B, C) across years '1999' and '2000'. An arrow points to the right, leading to a long-format table where each country-year combination is a row, with columns 'country', 'year', and 'cases'.

country	1999	2000
A	0.7K	2K
B	37K	80K
C	212K	213K

country	year	cases
A	1999	0.7K
B	1999	37K
C	1999	212K
A	2000	2K
B	2000	80K
C	2000	213K

name value

airquality 데이터는 전형적인 wide형 데이터로 특정 날짜에 네 개의 변수에 해당하는 값을 측정했습니다. 이 데이터를 long형으로 바꿀 경우 ID를 날짜로 하면 데이터들을 식별 할 수 있습니다. 그런데 날짜는 변수가 Month와 Day 두 개로 나누어져 있으므로 다음과 같이 두 변수를 식별 변수로 (ID로) 사용 합니다. 확인을 위해 상위 5개의 데이터만 가지고 형 변환을 진행해 보겠습니다.

```
airquality
```

```
myair <- airquality[1:5,]
myair_long <- pivot_longer(myair, c("Ozone", "Solar.R", "Wind", "Temp"))
myair_long

myair_long <- myair %>%
 pivot_longer(c("Ozone", "Solar.R", "Wind", "Temp"))
myair_long

myair_long2 <- myair %>%
 pivot_longer(c(Ozone, Solar.R, Wind, Temp))
myair_long2
```

```
myair_long3 <- myair %>%
 pivot_longer(!c(Month, Day))
myair_long3
```

생성되는 long형 데이터의 변수 이름인 name과 value는 다음 파라미터를 지정하여 바꿀 수 있습니다.

```
myair_long <- myair %>%
 pivot_longer(c(Ozone, Solar.R, Wind, Temp),
 names_to = "Type",
 values_to = "Observation")
myair_long
```

long형 데이터를 wide형 데이터로 변환 할 수도 있습니다.

```
myair_long %>%
 pivot_wider(
 names_from = Type,
 values_from = Observation)
```

## Exercises

- 1) 다음 데이터가 long형인지 wide형인지 판단하시오
- 2) long형이면 wide형으로 wide형이면 long형으로 변환하시오

```
stocks <- tibble(
 year = c(2015, 2015, 2016, 2016),
 month = c(1, 2, 1, 2),
 profit = c(1.88, 0.59, 0.92, 0.17)
)
```

## Exercises

앞서 gse93819 데이터에서 만든 myexpmean data.frame을 long 형으로 변환하시오

ggplot을 이용한 그래프 작성에는 위와 같은 long형 데이터가 주로 사용됩니다. R을 이용한 데이터 가시화는 dplyr 패키지로 wide형 데이터를 편집하고 pivot\_longer 함수로 long형 데이터로 변환 후 ggplot을 이용하는 방식으로 수행합니다. 두 데이터 포맷에 대한 좀 더 구체적인 내용은 다음 링크를 참고하시기 바랍니다. <https://www.theanalysisfactor.com/wide-and-long-data/>

## 7.5 Separating and uniting

데이터를 분석할 때 하나의 컬럼에 두 개 이상의 변수값이 저장되어 있거나 두 개의 변수를 하나의 컬럼으로 합해야 하는 경우가 종종 있습니다. 전자의 경우 `separate()` 함수를 사용해서 두 변수(컬럼)으로 나누어 줄 수 있으며 후자의 경우 `unite()` 함수를 사용하여 두 변수를 하나의 값으로 병합할 수 있습니다. 다음은 `airquality` 데이터에서 `Month`와 `Day` 변수를 하나의 컬럼으로 병합하여 `Date`라는 변수로 만들어 주는 경우의 예입니다.

```
newairquality <- airquality %>%
 unite(Date, Month, Day, sep=". ")
newairquality
```

`separate()` 함수를 사용하면 다음과 같이 해당 변수의 값을 나누어 다시 두 개의 변수(컬럼)으로 나누어 줄 수 있습니다.

```
newairquality %>%
 separate(col=Date, into = c("Month", "Day"), sep = "\\.")
```

## 7.6 dplyr

`dplyr` (<https://dplyr.tidyverse.org/>) 은 `ggplot2`을 개발한 해들리위컴이 (Hadley Wickham) 중심이 되어 만들어 졌으며 `ggplot2`와 함께 `tidyverse`의 (<https://www.tidyverse.org/>) 핵심 패키지입니다. `dplyr`은 데이터를 다루는 크기나 분석의 속도, 편의성을 향상시켜 새롭게 만들어놓은 패키지입니다. 기존 `apply`와 같은 행렬 연산 기능과 `subset`, `split`, `group` 와 같은 행렬 편집 기능을 더하여 만들어진 도구라고 할 수 있습니다.

`dplyr`의 전신이라 할 수 있는 `plyr` 패키지는 다음과 같이 설명이 되어 있습니다. A set of tools for a common set of problems: you need to split up a big data structure into homogeneous pieces, apply a function to each piece and then combine all the results back together. 즉 `split-apply-combine` 세 가지 동작을 쉽게 할 수 있도록 만들어 놓은 툴입니다. ROI 다른 언어에 비해 데이터 분석에서 주목을 받는 이유로 `split`, `apply` 등의 행렬 연산 함수가 발달한 것을 내세우는데 `dplyr`은 이들을 보다 더 편리하게 사용할 수 있도록 만들어 놓은 것입니다.

이제 `dplyr` 패키지에서 제공하는 함수를 사용해 보겠습니다. `dplyr`을 구성하는 중요한 함수는 다음과 같습니다.

- `select()` - 변수 (columns) 선택
- `filter()` - 샘플 (rows) 선택
- `arrange()` - 샘플들의 정렬 순서 변경
- `mutate()` - 새로운 변수 만들기
- `summarise()` - 대표값 만들기

- `group_by()` - 그룹별로 계산 수행
- `join()` - 두 tibble 또는 `data.frame`을 병합할 때 사용
- 위 함수들과 (특히 `filter`, `select`, `mutate`, `summarise`) 조합하여 (함수 내에서) 사용할 수 있는 helper 함수들이 같이 사용될 수 있습니다 (독립적으로도 사용 가능).
  - `across`
  - `if_any`
  - `if_all`
  - `everything`
  - `starts_with`
  - `end_with`
  - `contains`

이 함수들은 `%>%`와 함께 쓰이면서 강력한 성능을 발휘합니다. `summarise` 함수는 특정 값들의 통계 값을 계산해 주는 함수이며 그 외 함수들은 행렬 편집을 위한 함수들로 보시면 되겠습니다. 간단한 예제를 수행하면서 각각의 기능을 살펴보고 왜 `dplyr`이 널리 사용되고 그 장점이 무엇인지 파악해 보도록 하겠습니다.

`iris` 데이터는 세 종류의 `iris` 품종에 대한 꽃잎과 꽃받침의 `length`와 `width`를 측정해 놓은 데이터입니다. `head`와 `str` 명령어를 `%>%`를 이용해서 데이터를 살펴 봅니다.

```
library(tidyverse)

iris %>% head(10)
iris %>% str
```

### 7.6.1 select

`select()` 는 주어진 데이터셋으로부터 관심있는 변수를 (column) 선택하여 보여줍니다.

```
head(iris)
iris %>% select(Species, everything()) %>% head(5)
iris %>% select(Species, everything())
iris %>% select(-Species)
```

#### Exercises

`babies` 데이터의 변수/구조를 확인해 보고 `id`, `age`, `gestation`, `wt`, `dwt`, `smoke` 변수만을 선택한 새로운 `newbabies` 데이터를 만드시오

다음 helper 함수들은 `select` 함수와 같이 유용하게 쓰일 수 있습니다.

`starts_with("abc")` - “abc”로 시작하는 문자열을 갖는 변수 이름  
`ends_with("xyz")` - “xyz”으로 끝나는 문자열을 갖는 변수 이름  
`contains("ijk")` - “ijk” 문자열을 포함하는 변수 이름  
`matches("(.)\\W1")` - 정규식, 반복되는 문자

```
iris %>% select(starts_with('S'))
iris %>% select(ops = starts_with('S'))
```

아래는 `matches` 함수를 사용한 방법입니다. 좀 더 복잡한 패턴을 적용하여 변수들을 선택할 수 있으며 `grep` 함수를 사용할 경우도 정규식 패턴을 적용할 수 있습니다.

```
iris2 <- rename(iris, aavar = Petal.Length)
select(iris2, matches("(.)\\\\1"))
tmp <- iris[,3:5]
colnames(iris)[grep("^S", colnames(iris))]
iris[,grep("^S", colnames(iris))]
tmp
```

아래 `(.)\\\\1`은 하나의 문자 . 가 (어떤 문자든) 한 번 더 `\\1` 사용된 변수 이름을 말하며 이는 `aavar` 의 `aa`밖에 없으므로 `aavar`가 선택됩니다. `grep`에서 ^ 표시는 맨 처음을 나타내므로 ^S는 S로 시작하는 문자가 되겠습니다. 따라서 `grep("^S", colnames(iris))`의 경우 컬럼 이름 중 S로 시작하는 이름은 `True`로 그렇지 않으면 `False` 값을 리턴합니다.

## 7.6.2 filter

`filter` 함수를 사용해서 원하는 조건의 데이터 (샘플)을 골라낼 수 있습니다.

```
library(dplyr)

head(iris)
iris %>%
 filter(Species=="setosa")

iris %>%
 filter(Species=="setosa" | Species=="versicolor")

iris %>%
 filter(Species=="setosa" & Species=="versicolor")

iris %>%
 filter(Species=="setosa" | Species=="versicolor") %>%
 dim
```

`filter`의 ,로 구분되는 매개변수는 `and` 로직으로 묶인 조건입니다. 지난 강좌에서 보셨듯 R에서 `and`는 `&`, `or`는 `|`, 그리고 `not`은 `!` 으로 사용하면 되며 `filter`에서 ,로 구분된 조건은 `and`와 같다고 보시면 되겠습니다.

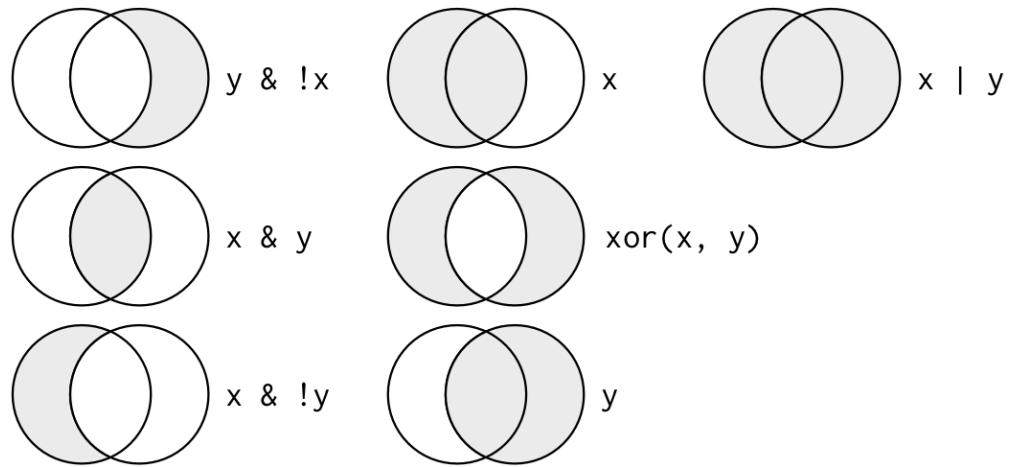


Image from (<https://r4ds.had.co.nz/>)

### Exercises

위 예제에서 만든 `newbabies` 데이터에서 999 값이 들어있는 값을 제외한 새로운 데이터를 만드시오

### 7.6.3 arrange

`arrange()`는 지정된 변수를 기준으로 값의 크기순서로 샘플들의 배열 순서 즉, `row`의 순서를 바꾸는 기능을 수행합니다. 기본으로 크기가 커지는 순서로 정렬이 진행되며 작아지는 순서를 원할 경우 `desc` 함수를 사용할 수 있습니다.

```
iris %>% arrange(Sepal.Length)
iris %>% arrange(desc(Sepal.Length))
iris %>% arrange(Sepal.Length, Sepal.Width)
```

### 7.6.4 mutate

`mutate()` 함수는 새로운 변수를 추가할 수 있는 기능을 제공하며 앞에서 배웠던 `within()`과 비슷하다고 볼 수 있습니다. 아래와 같이 `mutate`함수는 `sepal_ratio`라는 변수를 새로 만들어서 기준 `iris` 데이터들과 함께 반환해 줍니다.

```
iris2 <- iris %>% mutate(sepal_ratio = Sepal.Length/Sepal.Width)
head(iris2)
```

### 7.6.5 summarise

`summarise()`는 `data.frame`내 특정 변수의 값들로 하나의 요약값/대푯값을 만들어 줍니다. `summarise` 함수는 단독으로 쓰이기 보다는 `group_by()` 기능과 병행해서 쓰이는 경우에 유용하게 쓰입니다. `summarise_all()` 함수를 사용하면 모든 변수에

대해서 지정된 함수를 실행합니다. 특히 summarise 함수는 다음과 같이 across, if\_any, if\_all 등의 helper 함수와 조합되어 사용이 가능합니다.

```
iris %>% summarise(mean(Sepal.Length), m=mean(Sepal.Width))
iris %>%
 group_by(Species) %>%
 summarise(mean(Sepal.Width))

iris %>%
 group_by(Species) %>%
 summarise_all(mean)

iris %>%
 group_by(Species) %>%
 summarise(across(everything(), mean))

iris %>%
 group_by(Species) %>%
 summarise_all(sd)

iris %>%
 group_by(Species) %>%
 summarise(across(everything(), sd))
```

### 7.6.6 join

join 함수는 데이터를 병합해주는 기능을 수행하는 함수입니다. 네 가지 종류의 함수가 있으며 (left\_join(), 'right\_join()', 'inner\_join()', 'full\_join()')  
 (key) .by'에서 지정해준 파라미터의 값을 기준으로 기능이 수행 됩니다.

```
df1 <- data.frame(id=c(1,2,3,4,5,6), age=c(30, 41, 33, 56, 20, 17))
df2 <- data.frame(id=c(4,5,6,7,8,9), gender=c("f", "f", "m", "m", "f", "m"))

inner_join(df1, df2, by="id")
left_join(df1, df2, "id")
right_join(df1, df2, "id")
full_join(df1, df2, "id")

vs.
cbind(df1, df2)
```

## 7.7 Code comparison

이제 `split`, `apply`, `combine`을 활용하여 평균을 구하는 코드와 `dplyr` 패키지를 사용하여 만든 코드를 비교해 보도록 하겠습니다. `iris` 데이터를 분석하여 품종별로 꽃받침의 길이 (`Sepal.Length`)의 평균과 표준편차, 그리고 샘플의 수를 구해보는 코드입니다.

`split`은 `factor`형 변수인 `Species`를 기준으로 `iris` 데이터를 나누어 주는 역할을 하며 `lapply`는 `list` 형 데이터인 `iris_split`을 각 리스트의 각각의 원소들에 대해서 임의의 함수 `function(x){...}`를 수행하는 역할을 합니다. 마지막 `data.frame`으로 최종 경로를 `combine` 합니다.

```
iris_split <- split(iris, iris$Species)
iris_means <- lapply(iris_split, function(x){mean(x$Sepal.Length)})
iris_sd <- lapply(iris_split, function(x){sd(x$Sepal.Length)})
iris_cnt <- lapply(iris_split, function(x){length(x$Sepal.Length)})
iris_df <- data.frame(unlist(iris_cnt), unlist(iris_means), unlist(iris_sd))
```

아래는 `dplyr` 패키지를 사용한 코드입니다.

```
iris_df <- iris %>%
 group_by(Species) %>%
 summarise(n=n(), mean=mean(Sepal.Length), sd=sd(Sepal.Length))
```

위에서 보듯 `dplyr` 패키지를 사용할 경우 그 결과는 같으나 코드의 가독성과 효율성면에서 장점을 보여줍니다. `iris` 데이터를 받아서 `Species`에 명시된 그룹으로 나누고 원하는 함수를 타깃 컬럼에 대해서 적용하라는 의미입니다. 다음은 모든 변수에 대한 평균을 구하는 코드입니다.

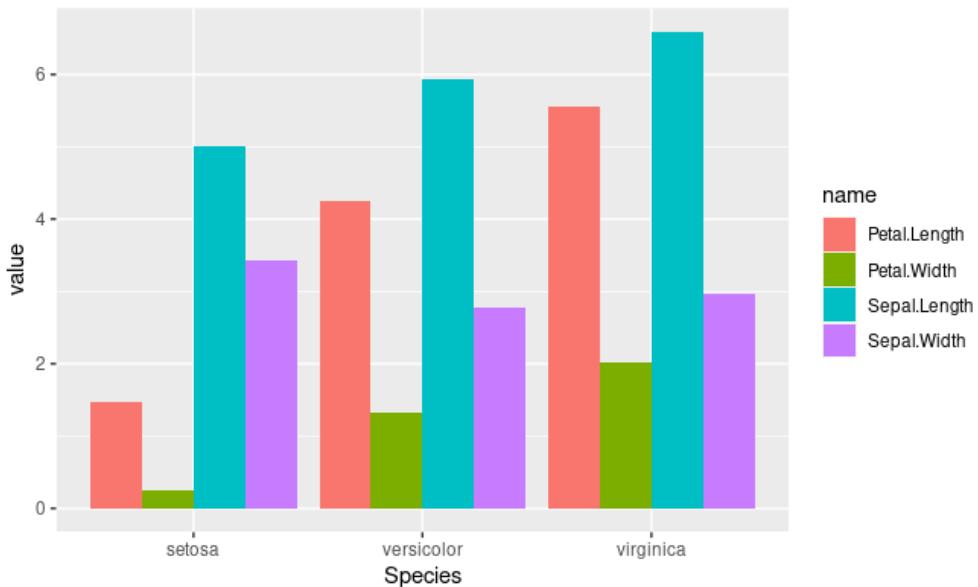
```
iris_mean_df <- iris %>%
 group_by(Species) %>%
 summarise(across(everything(), mean))
```

자세한 `ggplot`의 내용은 다음시간에 학습하겠지만 각 평균에 대한 막대그래프를 그려보겠습니다.

```
library(ggplot2)

iris_mean_df2 <- iris_mean_df %>%
 pivot_longer(-Species)

ggplot(iris_mean_df2, aes(x=Species, y=value, fill=name)) +
 geom_bar(stat="identity", position="dodge")
```



### Exercises

5.5의 babies 데이터를 tidyverse 패키지를 활용하여 다시 분석해 보시오 (모든 경우를 tidyverse 패키지 함수를 사용할 필요는 없음, `select` 를 사용할 경우 `dplyr::select` 로 사용하시오)

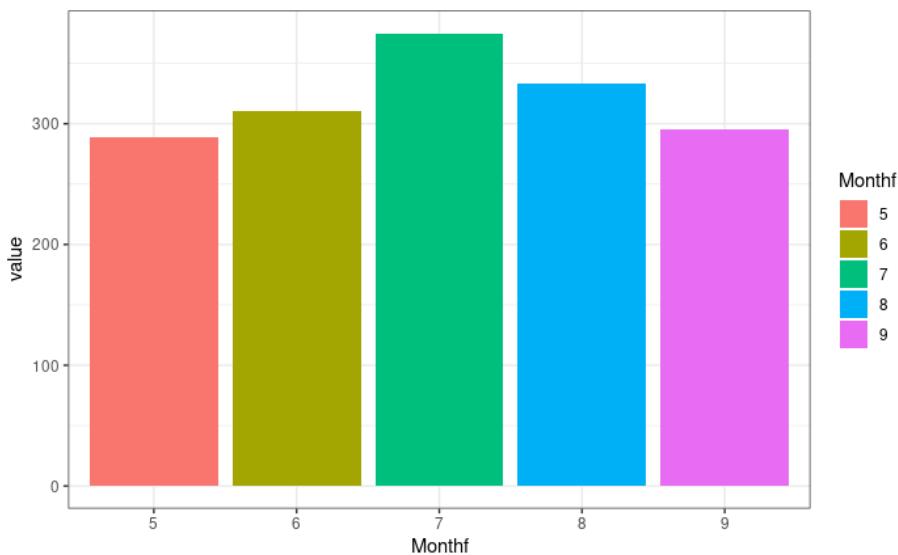
- 1) babies 데이터의 변수를 확인하시오
- 2) babies 데이터의 id, age, gestation, wt, dwt, smoke 변수만을 갖는 newbabies 데이터를 만드시오
- 3) 위 2번에 더해 999로 입력된 데이터를 제외한 newbabies 데이터를 만드시오
- 4) 위 3번에 더해 smoke 데이터를 factor 형으로 변환한 smokef 변수를 추가한 newbabies 데이터를 만드시오
- 5) 위 4번에 더해 25세 미만의 샘플만 갖고 smoke 변수는 제외한 newbabies 데이터를 만드시오

### Exercises

1. airquality 데이터에서 NA가 포함된 샘플 (row)를 제거한 myair라는 데이터셋을 생성하시오
2. 위 1)에 Month 변수를 factor형으로 변환한 Monthf를 추가하고 Month와 Day 변수를 제거한 새로운 데이터셋 myair 데이터를 생성하시오
3. myair 데이터에서 월별로 모든 변수에 (Ozone, Solar.R, Wind, Temp) 대한 평균을 구한 후 myairmean 변수에 저장하시오 (group\_by로 먼저 Monthf를 기준으로 grouping 필요, summarise\_all 사용)

4. 위 3)에 더하여 데이터를 long 형으로 바꾸고 myairmean에 저장하시오
5. ggplot으로 myairmean 데이터의 월별 각 변수들의 평균 값들을 다음과 같은 bar 그래프로 그리시오

```
ggplot(myairmean, aes(x=Monthf, y=value, fill=Monthf)) +
 geom_bar(stat="identity") +
 theme_bw()
```



### Exercises

InsectSprays 데이터는 살충제 6종에 대한 살충력을 (죽은 벌레의 마릿수) 나타내는 데이터이다. 각 살충제별로 평균과 표준편차를 구하시오

### Exercises

dplyr 패키지의 starwars 는 스타워즈 영화에 나오는 등장인물들을 분석한 데이터셋이다. 종족에 따른 키의 평균과 표준편차를 구하시오. (NA 데이터는 제외하고 분석)

### Exercises

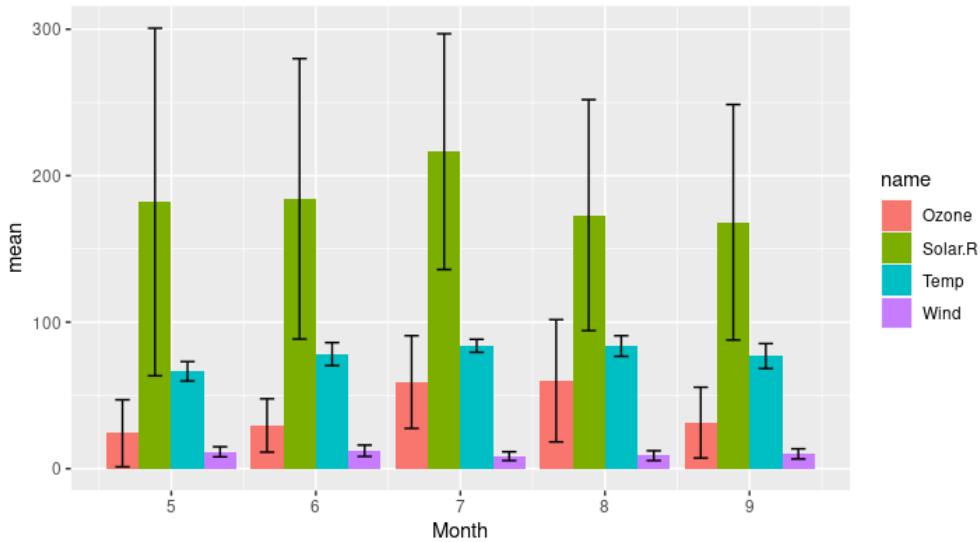
airquality 데이터는 뉴욕주의 몇몇 지점에서의 공기질을 측정한 데이터이다. 데이터에서 NA를 제거하고 각 월별로 평균 오존, 자외선, 풍속, 및 온도에 대한 평균과 표준편차를 구하시오

참고로 errorbar가 있는 막대그래프를 그려보겠습니다. 이를 위해서 먼저 두 테이블을 병합한 후 ggplot2 패키지의 ggplot 함수를 이용해서 그래프를 그립니다. 자세한 ggplot의 내용은 다음시간에 학습하겠습니다.

```
airdata <- left_join(airmean, airsd, by=c("Month", "name"))

ggplot(airdata, aes(x=Month, y=mean, fill=name)) +
```

```
geom_bar(stat="identity", position="dodge") +
 geom_errorbar(aes(ymin=mean-sd, ymax=mean+sd), position=position_dodge(width=0.9), width=0.4)
```



### Exercises

1. 다음 코드를 이용해서 gse93819 실험 관련 파일들을 다운로드하여 저장하고 데이터의 구조 및 샘플들의 이름을 확인하시오

```
myexp <- read.csv("https://github.com/greendaygh/kribbr2022/raw/main/examples/gse93819_expression.csv")
```

2. 샘플들의 정보에 따라서 발현 데이터를 나누시오
3. 각 그룹별 프루브들의 평균과 표준편차를 구하시오

### Exercises

gse103512 데이터도 동일한 방법으로 분석해 보시오

## 7.8 참고 통계

- <https://greendaygh.github.io/Rstat2020/statistical-inference.html#two-sample-significance-tests>
- 정규분포와 t분포 이해
- t 통계량 계산
- t-test 이해

---

이 저작물은 크리에이티브 커먼즈 저작자표시-비영리-변경금지 4.0 국제 라이선스에 따라 이용할 수 있습니다.



# Chapter 8

## Data visualization

이번시간에는 `ggplot2`( <https://ggplot2.tidyverse.org/> )를 이용한 시각화에 대해서 알아봅니다. 데이터를 분석할 때 실제 데이터를 눈으로 확인하는 것은 중요합니다. raw 데이터를 보면서 크기 비교나 분포를 대략적으로 예측할 수 있다면 tool을 사용해서 나오는 결과를 가늠하는 척도가 될 수도 있습니다. `ggplot2` 는 Rstudio 개발팀의 해들리위컴이 (Hadley Wickham) 중심이 되어 만든 데이터 시각화 패키지입니다. 몇 가지 새로운 규칙을 학습해야 하지만 그 활용성이나 성능을 고려한다면 꼭 배워야 할 패키지 중 하나입니다.

### 8.1 Basics

`iris` 데이터를 이용해서 간단하게 barplot을 그려봅니다. `iris` 데이터는 3가지 품종별 꽃잎과 꽃받침의 길이와 넓이를 측정한 데이터입니다. 다음은 꽃잎의 길이와 넓이의 관계를 볼 수 있는 산점도입니다.

```
library(ggplot2)
head(iris)
str(iris)
ggplot(data=iris) +
 geom_point(mapping=aes(x=Petal.Length, y=Petal.Width))
```

위에서는 두 개의 레이어가 있고 `ggplot`에서는 `+`를 이용해서 레이어들을 연결합니다. `ggplot()` 함수 뒤에 다양한 레이어들을 연결할 수 있고 `geom_point()` 함수는 지정한 위치에 산점도 레이어를 추가하는 기능을 합니다. 각 레이어들은 다음과 같은 다양한 기능을 갖는 함수들로 구성될 수 있습니다.

- 데이터 지정 (`ggplot`)
- 색상, 크기, x축의 값, y축의 값 등 심미적 요소 지정 (`aes`)
- 점, 선, 면 등 기하학적 요소 지정 (`geoms`)
- 그릴 통계량 지정 (`stats`)
- 테마, 스케일 지정 (`theme`)

일반적으로 `ggplot`을 이용하여 그래프를 그리는 순서는 다음과 같습니다.

- 어떤 그래프를 그릴지 결정
- `ggplot`의 데이터셋과 aesthetic 설정
- geometric 요소와 적절한 statistics를 설정한 레이어 추가
- 스케일과 테마를 설정한 레이어 추가

`ggplot`만을 실행할 경우 데이터와 x, y 축만 지정한 상태로 어떤 그래프 (히스토그램인지, 산포도인지 등)를 그릴지 명시되어 있지 않아서 아무것도 그리지 않은 상태의 빈 캔버스만 그려지게 되며 `geom_point()` 함수를 즉, 점을 그릴지 선을 그릴지 어떤 통계량을 그릴지 아니면 값 자체를 그릴지 등을 지정해 주고 나서야 비로서 그래프가 그려집니다.

```
ggplot(data=iris, mapping=aes(x=Petal.Length, y=Petal.Width))
?ggplot
ggplot(iris, aes(x=Petal.Length, y=Petal.Width))
ggplot(iris, aes(x=Petal.Length, y=Petal.Width)) + geom_point()
```

`geom_point()`의 도움말을 보면 다음과 같이 `data`, `mapping`, `stat` 등의 파라미터들이 있습니다. 이는 `ggplot`함수에서 설정한 `data`나 `mapping` 정보를 `geom_point`에서 설정하거나 완전히 다른 데이터를 x축과 y축에 그릴 수 있다는 뜻 이기도 합니다.

```
?geom_point
ggplot() +
 geom_point(data=iris, mapping=aes(x=Petal.Length, y=Petal.Width))
```

그런데 위 꽃잎의 길이와 넓이는 세 가지 다른 종류의 봇꽃에 대한 정보입니다. 따라서 각 종에 따라 다른 색이나 기호를 할당하는 것도 `mapping`에서 설정할 수 있습니다.

```
ggplot(iris, aes(x=Petal.Length,
 y=Petal.Width,
 color=Species,
 shape=Species)) +
 geom_point()

ggplot(iris, aes(x=Petal.Length, y=Petal.Width)) +
 geom_point(aes(color=Species, shape=Species))
```

위 산점도들의 `stat`은 `identity`입니다. 즉, 따로 통계량을 계산할 필요 없이 값 그 자체를 사용하겠다는 것 입니다. 히스토그램의 경우 `geom_bar()` 함수로 막대그래프를 그릴 수 있습니다. `geom_bar`의 help페이지를 보면 `stat="count"`로 설정되어 있는 것을 알 수 있습니다. 꽃잎의 넓이에 대한 분포를 예로 구해봅니다. 히스토그램을 그릴경우 변수 한 개의 데이터만 필요하고 y축에는 자동으로 빈도수가 들어가게 되므로 `aes`에서 x만 mapping 해 주면 됩니다.

```
ggplot(iris, aes(x=Petal.Width)) +
 geom_bar()
```

## 8.2 Bar graph

ggplot을 이용한 막대그래프 그리는 방법에 대해서 좀 더 알아보겠습니다. 앞서와 같이 ggplot 함수로 먼저 데이터와 aes로 x축 y축 등을 명시하고 + 오퍼레이터를 사용하여 필요한 레이어를 차례로 추가하면서 그래프를 그릴 수 있습니다. geom\_bar() 함수의 경우 x가 연속형일 경우는 아래와 같이 히스토그램을 그려주기 어렵습니다 (위 iris 예제에서 geom\_bar() 그래프에서는 실제 꽂받침의 width 값은 연속형이 맞으나 관측된 iris 데이터들이 같은 값들이 많은 범주형처럼 되어 있어 히스토그램 그림이 그려졌습니다) 이럴 경우 stat을 bin으로 바꿔주면 해당 범위 안에 있는 값들의 빈도수를 계산하여 히스토그램을 그릴 수 있습니다.

```
dat <- data.frame(x1=rnorm(100))
str(dat)
ggplot(dat, aes(x=x1)) +
 geom_bar()

ggplot(dat, aes(x=x1)) +
 geom_bar(stat="bin", bins=30)
```

x가 이산형인 경우는 stat을 디폴트 값인 count로 설정하여 해당 값들의 빈도수를 그려줄 수 있습니다. 이는 앞서 iris에서 배운 예제와 같습니다.

```
x1 <- sample(1:4, 100, replace = T)
dat <- data.frame(x=x1)
ggplot(dat, aes(x=x)) +
 geom_bar(stat="count")
```

이제 두 개의 변수가 있는 경우를 생각해 봅니다. 두 변수에 대해서 막대그래프를 그릴 경우 다음과 같이 Error: stat\_count() must not be used with a y aesthetic. 에러가 발생할 수 있습니다.

```
x1 <- rnorm(10)
x2 <- rnorm(10)
dat <- data.frame(x1, x2)
ggplot(dat, aes(x=x1, y=x2)) +
 geom_bar()
```

이는 geom\_bar()의 stat이 기본적으로 count로 설정되어 있으므로 생기는 에러입니다. stat을 identity로 설정하면 x1값에 해당하는 x2값을 그려주는 막대그래프를 그릴 수 있습니다. 참고로 이 그래프는 geom\_point와 비슷한 정보를 보여주게 됩니다.

```
x1 <- rnorm(10)
x2 <- rnorm(10)
dat <- data.frame(x1, x2)
ggplot(dat, aes(x=x1, y=x2)) +
 geom_bar(stat="identity")
```

```
ggplot(dat, aes(x=x1, y=x2)) +
 geom_point()
```

다음과 같이 레이어를 추가하여 두 그래프를 같은 화면에 그릴 수도 있습니다. 여기서 `col`과 `size`는 `aes`함수안에서 쓰이지 않았음을 주의하시기 바랍니다. `aes`에서는 데이터와 특정 모양, 색깔을 mapping 해주는 역할을 하고 아래와 같이 지정해 줄 경우 데이터와 상관 없이 해당 레이어의 모든 그래프에 대해서 일괄적으로 적용되게 됩니다.

```
ggplot(dat, aes(x=x1, y=x2)) +
 geom_bar(stat="identity") +
 geom_point(aes(col="red", size=5))
?geom_point
```

또한 다음과 같이 다양한 레이어를 추가하여 필요한 기능을 사용할 수 있습니다. `fill=x1`이라는 코드는 막대그래프의 색을 채울 때 `x1`에 따라서 다른 값들을 채우는 역할을 한다고 보면 되겠습니다.

```
x1 <- as.factor(1:3)
y1 <- tabulate(sample(x1, 100, replace=T))
dat <- data.frame(x1, y1)
ggplot(dat, aes(x=x1, y=y1, fill=x1)) +
 geom_bar(stat="identity")
```

위 그래프에 더해서 `legend`를 없애고 `x`, `y`축 라벨과 스케일, 그리고 타이틀 등을 설정할 수 있습니다.

```
ggplot(dat, aes(x=x1, y=y1, fill=x1)) +
 geom_bar(stat="identity") +
 guides(fill=FALSE) +
 xlab("Discrete cases") +
 ylab("Value") +
 ylim(c(0,50))+
 ggtitle("Bar graph for x:discrete and y:value")
```

### Exercises (revisit)

- 목적: 그룹별로 발현이 다른 유전자 탐색
1. 다음 코드를 이용해서 gse93819 실험 관련 파일들을 다운로드하여 저장하고 데이터의 구조 및 샘플들의 이름을 확인하시오  

```
myexp <- read.csv("https://github.com/greendaygh/kribbr2022/raw/main/examples/gse93819")
myexp
```
  2. 샘플들의 정보에 따라서 (그룹별로) 발현 데이터를 나누시오 (필요하지 않을 수 있음)
  3. 각 그룹별 프루브들의 평균과 표준편차를 구하시오

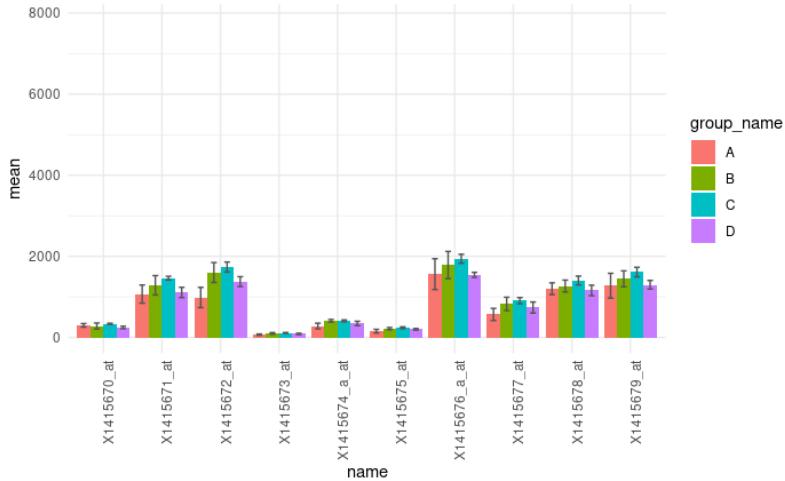
- 분석 목적에 따른 샘플, 변수, 값 구분하기
- 평균을 구할 때 샘플의 특정 변수에 대한 평균을 구함
- 발현 데이터와 메타데이터 두 개의 테이블로 만들어 작업

```
require(knitr)
require(kableExtra)

mytable <- myexp[1:4, 1:4]
mytable

mytablet <- data.frame(t(myexp[1:4, 1:4]))
mytablet
```

4. 각 프루브의 평균값을 bargraph로 그리고 표준편차로 에러바를 그리시오



## 8.3 Line graph

다음으로 ggplot을 이용한 line graph를 그리는 방법을 알아봅니다. Line graph는 geom\_line이라는 함수를 사용해서 그릴 수 있으며 stat의 사용법은 앞서 bar graph와 같습니다.

```
x1 <- c(12, 21, 40)
x2 <- c(33, 10, 82)
dat <- data.frame(x1, x2)
ggplot(dat, aes(x=x1, y=x2)) +
 geom_line()
```

아래와 같이 그려지는 선의 두께를 조절하거나 레이어를 추가하는 방법으로 점을 추가로 그려볼 수 있습니다. fill의 경우 특정 도형에 채워지는 색을 의미합니다. 도형에 대한 자세한 종류는 ?pch라는 도움말로 살펴보실 수 있습니다.

```
ggplot(dat, aes(x=x1, y=x2)) +
 geom_line(size=2) +
 geom_point(size=4, pch=21, fill="white") +
 guides(fill=FALSE) +
 ylim(c(0, 100)) +
 xlab("Continuous cases") + ylab("Value") +
 ggtitle("Line graph for x:continuous and y:continuous")
```

위 경우는 x와 y가 모두 연속형 데이터일 경우입니다. x는 이산형, y가 연속형일 경우 일반적으로 bar graph를 이용하여 그래프를 그리게 됩니다. 그런데 이런 bar의 높이에 해당하는 값들을 서로 선으로 연결하고 싶은 경우가 있습니다.

```
x1 <- as.factor(c(1:3))
y1 <- c(33, 10, 82)
dat <- data.frame(x1, y1)
str(dat)
ggplot(dat, aes(x=x1, y=y1)) +
 geom_bar(stat="identity") +
 xlab("Discrete cases") + ylab("Value") +
 ylim(c(0,100))+
 ggtitle("Line plot for x:discrete and y:continuous")
```

이 때는 다음과 같이 aes의 group이라는 파라미터를 설정하여 두 점 이상을 연결할 수 있습니다. 만약 group으로 나타낼 수 있는 변수가 없을 경우 group=1이라고 명시해 주고 선을 그릴 수 있으며 이 경우 모든 값들이 같은 1 그룹에 있는 것으로 간주됩니다. 1이라는 것은 하나의 예이며 어떤 숫자나 문자가 와도 괜찮습니다.

```
ggplot(dat, aes(x=x1, y=y1, group=1)) +
 geom_bar(stat="identity") +
 geom_line() +
 xlab("Discrete cases") + ylab("Value") +
 ylim(c(0,100))+
 ggtitle("Line plot for x:discrete and y:continuous")
```

위에서와 같은 방법으로 point와 bar 등을 같이 그려줄 수 있습니다.

```
ggplot(dat, aes(x=x1, y=y1, group=1)) +
 geom_bar(aes(fill=x1), stat="identity") +
 geom_line(size=2, color="gray") +
 geom_point(size=4, pch=21, fill="white") +
 xlab("Discrete cases") + ylab("Value") +
 ylim(c(0,100))+
 ggtitle("Line for x:discrete and y:value")
```

여기서는 fill 옵션이 geom\_bar에 하나 geom\_point에 하나씩 쓰였는데 geom\_bar에서 사용된 fill은 bar에 채워지는 색을 x1의 값에 따라 바꾸겠다는 것을 의미하고 (aes함수 안에 사용) geom\_point의 fill은 데이터에 상관 없이 모두

white로 채우라는 명령입니다. 각 geometry에 따라서 필요한 옵션이 다르므로 각각의 geom\_xxx를 사용할 때 상황에 맞게 사용하시면 되겠습니다.

## 8.4 Smoothing

산포도는 앞서와 같이 데이터를 점으로 표현한 그래프입니다. Smoothing은 관측된 데이터를 이용하여 모형을 추정하는데 사용되는 통계적 방법이며 이를 그래프로 표현하여 추세선을 그릴 수 있습니다. 예를 들어 몸무게와 키라는 두 변수의 관계를 알아보고자 할 때 산포도를 그리고 Smoothing을 통해 점들의 평균값을 이어주는 방법으로 모형을 추정하고 추세선을 그릴 수 있습니다.

mtcars 데이터는 1974년 미국 자동차 잡지에서 추출한 데이터로서 당시 다양한 모델의 자동차에 대한 성능을 저장하고 있습니다 (?mtcars로 자세한 정보를 볼 수 있음). 이 데이터를 이용해서 연비와 마력 (horsepower) 두 변수의 관계를 그래프로 그려보겠습니다. 직관적으로 생각하면 두 변수는 반비례 할 것으로 기대됩니다. ggplot을 활용해서 두 변수의 산포도를 그리고 smoothing을 수행해 보도록 하겠습니다.

```
ggplot(mtcars, aes(x=mpg, y=hp)) +
 geom_point()
```

위와 같이 mtcars는 data.frame이므로 ggplot으로 바로 받아서 x축과 y축 mapping에 필요한 변수들 이름을 직접 할당하고 geom\_point함수를 이용해서 간단히 산포도를 그릴 수 있습니다. 이 산포도만으로도 mpg와 hp 두 변수간의 관계가 역함수 관계임을 알 수 있고 또한 선형이 아닌 것도 알 수 있습니다. 이제 위 그림에 geom\_smooth() 함수를 이용해서 (모형) 적합 곡선 (또는 추세선)을 그려봅니다.

```
ggplot(mtcars, aes(x=mpg, y=hp)) +
 geom_point() +
 geom_smooth()
```

간단히 geom\_smooth() 한 줄을 추가하여 추세선을 그렸으며 경고 메세지에서 볼 수 있듯이 알고리즘은 loess 모형을 사용했고 공식은 (formula는)  $y \sim x$ 로, 즉, y축 변수를 반응변수로 x축 변수를 설명변수로 설정하여 그려졌습니다. 직선의 공식  $y = ax + b$ 를 생각해 보시면 무슨 의미인지 이해가 더 쉬울듯 합니다. ?geom\_smooth로 보면 알 수 있듯이 모형을 적합하는 알고리즘 옵션을 lm, glm, loess 등 다양하게 설정할 수 있으며 auto로 하게 되면 데이터의 크기나 형식에 맞춰서 방법을 자동으로 선택해서 그려주게 됩니다. se 옵션은 기본적으로 TRUE 값을 가지며 위 그림에서 볼 수 있는 선분 주위의 회색 구간으로 신뢰구간을 그려주는 옵션입니다. span 옵션은 loess 모형의 smoothing 정도를 조절할 수 있는데 이는 직접 바꿔가면서 실습을 해보면 이해에 도움이 되겠습니다.

```
ggplot(mtcars, aes(x=mpg, y=hp)) +
 geom_point() +
 geom_smooth(se=FALSE, span=0.2)
```

위와 같이 span 옵션을 작게 설정할 수록 관측된 데이터(점)에 선분(모형)이 가까이

불게 됩니다. 이를 과대적합 (overfitting)이라고 하며 간단히 설명하면 관측된 데이터에만 너무 잘맞는 모형을 만드는 경우를 말합니다. 이럴 경우 새롭게 관측된 데이터는 모형의 예측값과 잘 맞지 않게 됩니다.

이번에는 모의 데이터를 생성해서 그래프를 그려보겠습니다. 네 개 학급에 있는 학생들의 키와 몸무게를 저장한 데이터를 만들어 봅니다. 이 경우 몇 개의 변수가 필요할지 생각해 보시기 바랍니다. 키와 몸무게 그리고 학급을 나타내는 변수 3개가 필요하며 키와 몸무게는 정수형, 그룹을 나타내는 변수는 문자형이나 factor형으로 나타내면 되겠습니다. 각 학급의 학생수는 50명으로 총 200명의 학생이 있는 것으로 하며 각 그룹별로 키나 몸무게의 차이는 없고 키가 큰 사람은 몸무게가 많이 나가는 것으로 합니다. 키와 몸무게 사이에는 다음과 같은 연관성을 만들어 줍니다.  
 $height = weight + N(100, 10)$

```
weights <- rnorm(200, 75, 5)
heights <- weights + rnorm(200, 100, 5)
classes <- sample(c("A", "B", "C", "D"), size=length(heights), replace = T)
mydata <- data.frame(heights, weights, classes)
str(mydata)
```

이제 위 데이터를 이용해서 몸무게와 키의 산포도와 추세선을 그려보고 추가로 그룹별로 다른 색의 점으로 표현해 보겠습니다.

```
ggplot(mydata, aes(x=weights, y=heights, color=classes)) +
 geom_point() +
 geom_smooth()
```

그런데 위와 같은 코드를 실행하면 그룹마다 다른 점과 smooth 선분이 그려집니다. 우리가 원하는 그림은 단지 점만 그룹별로 다른 색으로 표현하고 추세선은 전체 학생들에 대해서 하나의 선분만 그려지길 원합니다. 이제 우리가 알아야 할 부분은 각 레이어마다 mapping을 지정할 수 있다는 것이고 이 원리를 이해한다면 다음과 같이 geom\_point에서는 color를 mapping해 주고 geom\_smooth에서는 지정해주지 않으면 됩니다.

```
ggplot(mydata) +
 geom_point(aes(x=weights, y=heights, color=classes)) +
 geom_smooth(aes(x=weights, y=heights))
```

그리고 중복되는 부분을 줄여줄 수도 있습니다. 즉, ggplot에서 지정하는 mapping은 하위 layer에 모두 적용이 되며 각 layer마다 다른 mapping 특성을 부여하고 싶을 경우 해당 layer의 mapping 함수 (aes)를 이용하여 설정할 수 있다는 점을 기억하시기 바랍니다.

```
ggplot(mydata, aes(x=weights, y=heights)) +
 geom_point(aes(color=classes)) +
 geom_smooth()
```

## 8.5 Statistics and positions

앞서 smoothing 곡선은 실제 데이터에서 관측된 값이 아닌 계산된 값을 그래프에 표현한 것 입니다. 막대그래프에서도 y축 count 값은 관측된 값이 아닌 빈도수를 계산한 값이고 boxplot의 경우도 중간값, 1,3사분위수 등 통계량을 표현해 주는 그래프입니다. 이는 대부분 통계 분석용 소프트웨어에서 제공되는 기능으로 통계량을 가시화 해주는 역할을 합니다. ggplot2에서도 각 geom 레이어에 stat이라는 옵션을 통해 이러한 통계량을 그래프로 표현할 수 있습니다. 예를 들어 앞서 생성한 키, 몸무게 데이터에서 키의 분포를 보기 위한 히스토그램을 그리면 geom\_histogram을 사용할 수 있고 이 레이어의 stat 옵션의 기본값은 "bin"입니다 (?geom\_histogram 참고).

```
library(tidyverse)

weights <- rnorm(200, 75, 5)
heights <- weights + rnorm(200, 100, 5)
classes <- sample(c("A", "B", "C", "D"), size=length(heights), replace = T)
mydata <- data.frame(heights, weights, classes)
str(mydata)

ggplot(mydata, aes(x=heights)) +
 geom_histogram()
```

경고 문구의 bins=30은 기본 stat옵션이 bin인데 bins옵션은 null로 되어 있기 때문에 경고가 발생한 것이고 30으로 강제 할당해서 그린다는 메세지입니다. bins 옵션을 다르게 해서 테스트 해보시기 바랍니다. 또한 stat="identity"로 그래프를 그린 경우는 데이터 값을 그대로 그린다는 것도 다시 기억해 보시기 바랍니다.

```
ggplot(mydata, aes(x=heights)) +
 geom_histogram(stat="identity")

ggplot(mydata, aes(x=heights, y=weights)) +
 geom_histogram(stat="identity")
```

또 다른 예를 위해 앞서 키 몸무게 데이터에 혈액형 변수를 추가해 보겠습니다. 혈액형은 위 4개 학급에 관계 없이 A, B, O, AB 네 그룹으로 나눌 수 있으며 200명의 학생들에게 랜덤하게 할당하도록 합니다.

```
bloodtype <- sample(c("A", "B", "O", "AB"), nrow(mydata), replace=T)
mynewdata <- data.frame(mydata, bloodtype)
str(mynewdata)
```

위와 같이 새로운 변수 bloodtype이 factor형으로 추가되어 새로운 data.frame을 생성하도록 했습니다. 이제 각 학급별로 몇 명의 혈액형 타입을 갖는 학생들이 있는지를 막대그래프로 표현해 보도록 하겠습니다. 혈액형의 타입별로 다른 색으로 막대를 칠하도록 해봅니다. 막대그래프의 색은 fill옵션으로 채울 수 있고 막대그래프는 geom\_bar 그리고 이 레이어의 stat은 기본값이 count이므로 따로 명시하지 않은 채로 다음과 같이 코드를 작성할 수 있습니다.

```
ggplot(mynewdata, aes(x=classes, fill=bloodtype)) +
 geom_bar()
```

그런데 위와같이 그래프가 위로 쌓여서 보입니다. 이는 `geom_bar`의 `position` 기본값이 `stack`으로 되어있어서 보이는 현상입니다 (`?geom_bar`참고). 옆으로 나란히 막대를 위치시킨 후 크기를 비교하기 위해서 `position="dodge"`를 사용합니다. 또한 막대그래프에 칠해지는 색의 투명도를 `alpha` 옵션을 사용해 변경할 수 있습니다.

```
ggplot(mynewdata, aes(x=classes, fill=bloodtype)) +
 geom_bar(alpha=0.5, position="dodge")
```

다음과 같이 간단히 한 줄만 추가하여 위 막대그래프의 위치를 가로로 전환하거나 Coxcomb chart로 그릴수도 있습니다.

```
ggplot(mynewdata, aes(x=classes, fill=bloodtype)) +
 geom_bar(position="dodge") +
 coord_flip()
```

```
ggplot(mynewdata, aes(x=classes, fill=bloodtype)) +
 geom_bar(position="dodge") +
 coord_polar()
```

참고로 위 Coxcomb 그래프의 경우는 해석이 어렵거나 x, y축의 라벨링에 혼돈이 올수 있으니 정보 전달이 명확하도록 그래프의 옵션들을 추가하거나 용도에 맞게 사용할 필요가 있습니다.

### Exercises

- gse93819 데이터에서 두 그룹의 프루브들에 대해서 산포도를 그리시오

```
myexp <- read.csv("https://github.com/greendaygh/kribbr2022/raw/main/examples/gse93819.csv")
myexp

myexpt <- data.frame(t(myexp[1:100,]))

groupid <- rep(c("A", "B", "C", "D"), each=5)
mysample <- data.frame(sample_name=names(myexp), group_name=groupid)
mysample

myexp2 <- myexpt %>%
 rownames_to_column(var = "sample_name") %>%
 left_join(mysample, c("sample_name")) %>%
 dplyr::select(sample_name, group_name, everything()) %>%
 group_by(group_name)

expmean <- myexp2 %>%
```

```

summarise(across(where(is.numeric), mean)) %>%
pivot_longer(-group_name, values_to = "mean")

tmpd <- expmean %>%
pivot_wider(values_from=mean)

expmeant <- as_tibble(t(tmpd[, -1]))
names(expmeant) <- tmpd$group_name

ggplot(expmeant, aes(x=A, y=B)) +
 geom_point() +
 geom_smooth(method='lm')+
 scale_x_continuous(trans='log10')

```

## 8.6 Facets

산점도의 예에서 위와 같이 다른 색이나 모양으로 그리기 보다는 종 별로 다른 켄버스에 별도의 산점도를 그려야 할 경우가 있습니다. 이럴때 사용하는 함수가 `facet_wrap()`이나 `facet_grid()`입니다. 보통 범주형 자료에 대해서 적용할 수 있으며 `facet_wrap()`은 하나의 변수에 대해서 그림을 나눠그릴때 사용하고 `facet_grid()`는 두 개 변수의 조합에 의한 그래프들을 그릴 때 사용합니다. 위 봇꽃 예에서는 3가지 종을 나타내는 변수 `Species`를 이용하면 되겠습니다. `facet_wrap()`함수에는 `~`를 이용한 formula를 사용합니다.

```

ggplot(iris, aes(x=Petal.Length, y=Petal.Width)) +
 geom_point(aes(color=Species, shape=Species)) +
 facet_wrap(~Species, nrow=2)

```

만약 두 개의 범주형 변수에 대해서 x, y축 각각으로 나누고 싶을 때는 `facet_grid()`를 사용할 수 있습니다. `iris` 데이터는 하나의 범주형 변수와 네 개의 숫자형 변수로 구성되어 있습니다 (`str(iris)` 확인). 여기에 랜덤하게 0과 1을 갖는 범주형 변수 하나를 추가해 보겠습니다.

```

str(iris)
mycate <- sample(c(0,1), nrow(iris), replace=T)
myiris <- data.frame(iris, mycate)
str(myiris)

```

이제 `mycate`와 `Species` 두 범주형 변수에 대해서 facet 그래프를 그려보면 다음과 같습니다. `facet_grid()`함수를 사용하면 되며 x와 y축의 변수는 `~`를 활용한 formula를 사용합니다. 즉 `~` 왼편의 변수는 y축 오른편의 변수는 x축으로 구성되어집니다. 새로운 `myiris`라는 데이터를 만들었으므로 `iris` 대신 `myiris`를 사용합니다.

```
ggplot(myiris, aes(x=Petal.Length, y=Petal.Width)) +
 geom_point(aes(color=Species, shape=Species)) +
 facet_grid(Species~mycate)
```

만약 하나의 변수에 대해서 x축이나 y축 하나에만 나열하고 싶은 경우 다음처럼 . 을 사용하면 됩니다.

```
ggplot(myiris, aes(x=Petal.Length, y=Petal.Width)) +
 geom_point(aes(color=Species, shape=Species)) +
 facet_grid(.~mycate)

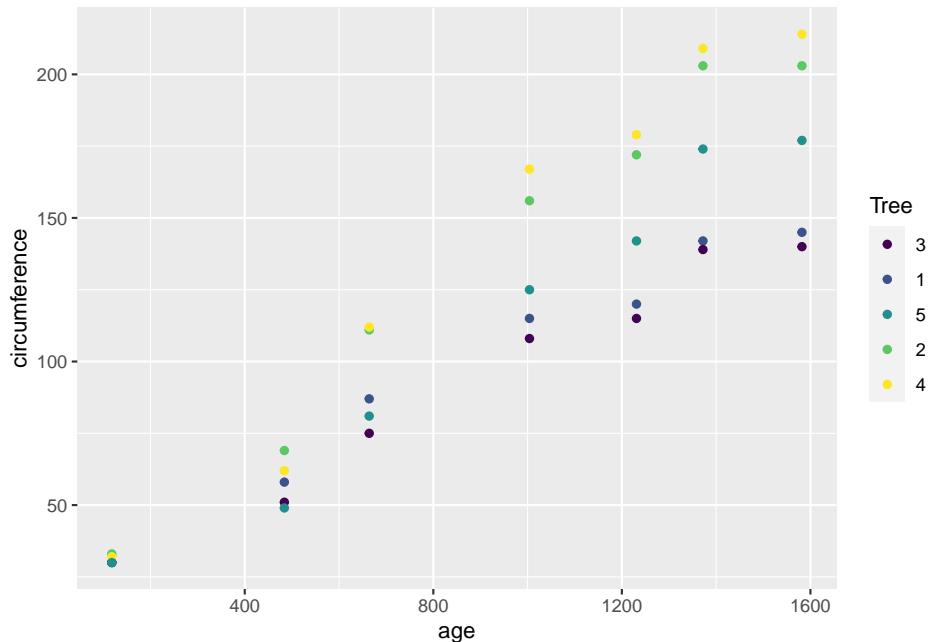
ggplot(myiris, aes(x=Petal.Length, y=Petal.Width)) +
 geom_point(aes(color=Species, shape=Species)) +
 facet_grid(Species~.)

ggplot(myiris, aes(x=Petal.Length, y=Petal.Width)) +
 geom_point(aes(color=Species, shape=Species)) +
 facet_grid(.~Species)
```

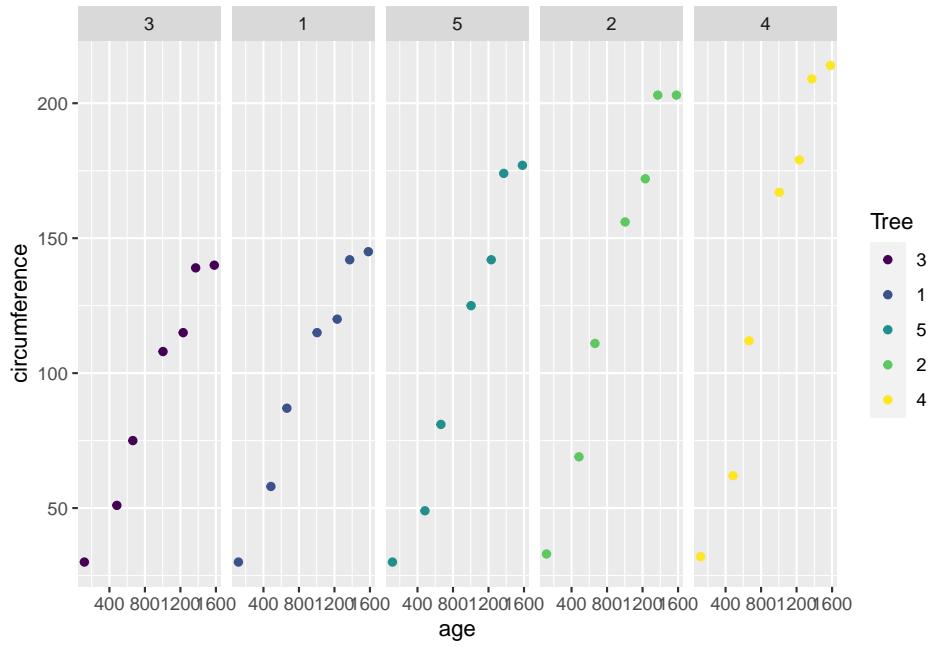
### Exercises

Orange 데이터셋은 다섯 그루의 오랜지 나무에 대한 시간에(age-days) 따른 성장을(circumference) 기록한 데이터임.

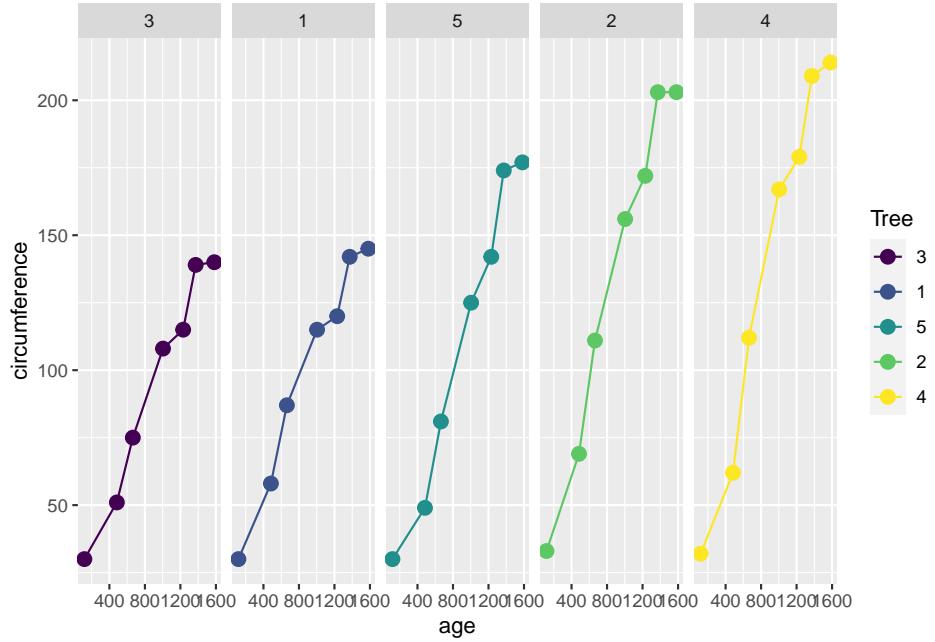
- age와 circumference 를 각각 x와 y축으로 하는 산점도를 그리는 코드를 작성하시오 (ggplot 이용, 나무별로 다른 색 사용)



- 2) 나무별로 다른 캔버스에 age와 circumference를 x와 y축으로 하는 산점도를 그리는 코드를 작성하시오 (ggplot, facet\_grid이용)

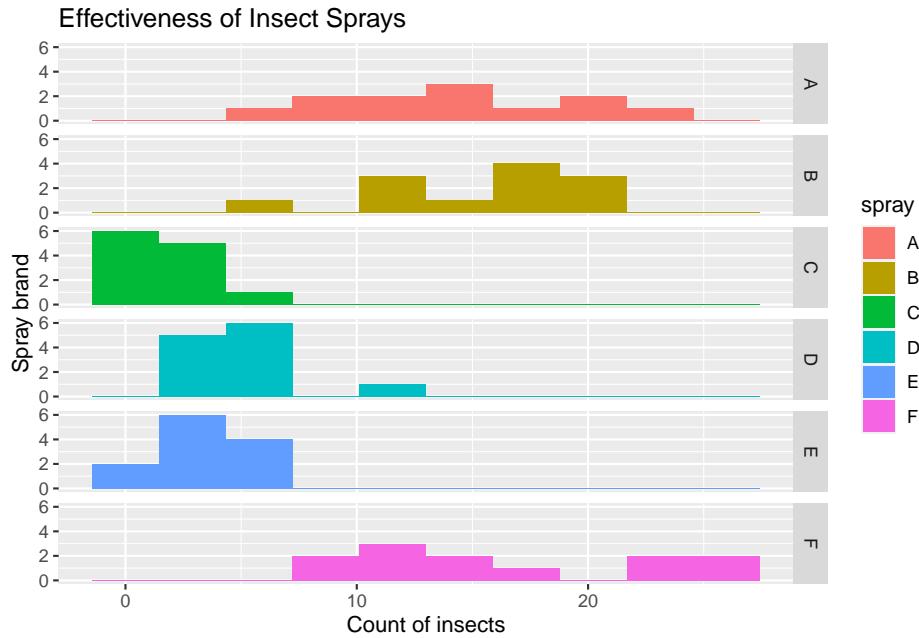


- 3) 2)에서 그려진 나무별 산점도에 다음과 같이 선분을 추가한 그래프를 그리는 코드를 작성 하시오



### Exercises

InsectSprays는 제초제의 효능에 관한 데이터이다. 다음과 같은 plot을 그리는 코드를 작성 하시오



## 8.7 Themes, Labels, and Scales

Theme은 data관련 요소들 외의 것들에 대한 설정을 위해서 사용됩니다. 즉, 제목이나 라벨, 배경, 범례 등의 색, 위치, 크기, 모양 등을 설정하는데 사용합니다. 주의할 부분은 해당 텍스트 등 데이터를 변경하는 것이 아니고 보여지는 모습만을 바꿀 수 있다는 것 입니다. 텍스트 설정은 labs를 사용합니다. 예제를 가지고 몇 가지 실습을 해 보겠습니다. 먼저 labs라는 명령어로 x축, y축, Title 등을 설정할 수 있습니다. 참고로 xlab(), ylab() 등의 함수도 x축, y축 라벨을 설정하는데 사용될 수 있지만 여기서는 labs만을 사용하도록 합니다.

```
ggplot(mynewdata, aes(x=classes, fill=bloodtype)) +
 geom_bar(position="dodge") +
 labs(x='Four classes',
 y='Number of students',
 title='Blood type distribution',
 subtitle = 'Blood type distribution from the 200 students',
 fill='Blood Types')
```

위 코드에서 labs에서 설정할 수 있는 옵션은 title, subtitle과 x축, y축 라벨 그리고 범례의 title까지 가능합니다. 특히 ggplot 명령에서 aes(fill=bloodtype)이

사용되었으므로 범례의 title은 fill="Blood types"로 설정해야 하며 만약 aes(color=bloodtype)으로 사용되었을 경우에는 color="Blood types"으로 설정합니다. 참고로 범례의 label을 설정하는 방법은 다음과 같이 scale\_fill\_discrete 함수의 labels 옵션을 사용하면 됩니다. element\_blank()는 텍스트를 공백으로 설정할 때 사용합니다. 아래 나올 scale 관련 내용과 함께 이해하시면 좋습니다.

```
ggplot(mynewdata, aes(x=classes, fill=bloodtype)) +
 geom_bar(position="dodge") +
 scale_fill_discrete(name=element_blank(), labels=c("A type", "AB type", "B type", "O type"))
```

이제 본격적으로 Theme으로 그래프를 장식해 보도록 합니다. Theme 관련된 옵션들은 <https://ggplot2.tidyverse.org/reference/theme.html> 이곳을 참고하시기 바랍니다. 여기서 mapping은 그대로인채로 모양 등의 설정을 바꿔가면서 그래프의 형태를 확인하는 작업이 반복되므로 다음과 같이 myplot이라는 변수에 기본이 되는 ggplot 코드를 저장하고 이후 + 연산자를 사용해서 옵션을 바꿔가며 편리하게 코드를 재사용 할 수도 있습니다.

```
myplot <- ggplot(mynewdata, aes(x=classes, fill=bloodtype)) +
 geom_bar(position="dodge") +
 labs(x='Four classes',
 y='Number of students',
 title='Blood type distribution',
 subtitle = 'Blood type distribution from the 200 students',
 fill='Blood Types')
myplot + theme_bw()
```

위 theme\_bw() 함수는 theme의 세부 사항 몇 가지를 미리 설정해 놓아서 (배경을 white 색, 눈금을 회색으로 바꾸는 등) theme 설정을 위한 일련의 과정을 한번에 수행하도록 만든 함수입니다. theme을 이용한 설정은 plot, axis, legend, panel, facet 등에 적용할 수 있으며 따라서 다음 코드와 같이 해당하는 요소를 참고할 때 . 기호로 구분된 옵션 이름을 사용합니다. 값을 지정할 때에는 element\_xxx의 패턴으로 이루어진 함수를 사용합니다. 다음은 각각 plot과 panel 배경색을 바꾸는 코드입니다.

```
myplot + theme(plot.background = element_rect(fill="gray"))
myplot + theme(panel.background = element_rect(fill="gray"))

myplot +
 theme(
 panel.background = element_rect(fill="gray"),
 plot.background = element_rect(fill="gray")
)
```

또한 축이나 라벨 텍스트의 모양도 바꿀 수 있습니다.

```
myplot +
 theme(
 axis.line = element_line(arrows = arrow(angle = 15, length = unit(.15,"inches"))),
```

```

 axis.text = element_text(face = "bold", size = 12, angle = 30),
 axis.text.x = element_text(color="blue", size=18)
)

myplot +
 theme(
 plot.title=element_text(size=18, face = "bold", color="red", hjust=0.5),
 plot.subtitle = element_text(size=18, face = "bold", color="gray")
)

```

위 예제 외에도 다양한 그래프를 그릴 수 있으며 모든 사용법을 외워서 사용하기보다는 사용할 때마다 필요한 함수와 옵션을 찾아서 사용하다 보면 점차 익숙해질 것입니다. 가장 정확한 참고 자료는 공식 reference 페이지를 참고하면 좋으며 <https://ggplot2.tidyverse.org/reference/index.html> 이 외에도 다른 사람들이 만들어 놓은 그래프를 <https://exts.ggplot2.tidyverse.org/> 참고해서 원하는 목적으로 맞는 코드를 가져다 사용할 수 있습니다.

본 장에서 마지막으로 소개할 내용은 Scale입니다. 앞서 어떤 데이터를 x축, y축 또는 group이나 color로 맵핑할지를 결정하는 함수가 aes였다면 scale은 어떻게 (위치, 색상, 크기, 모양 등) 맵핑할 것인가를 설정하는 방법입니다. 함수 형태는 `scale_<aesthetic>_<type>`이며 <aesthetic>과 <type>에 해당하는 (미리 지정된) 단어를 넣어주면 되겠습니다. 예를 들어 앞서 예제에서 `fill=bloodtype`로 혈액형 데이터를 막대그래프의 색을 칠하는데 사용했다면 `scale_fill_manual` 함수로 어떤 색을 칠할지를 정해주는 방식입니다. 다음 몇 가지 예를 실습해 보고 이해해 봅니다.

```

myplot +
 scale_fill_manual(values = c("orange", "skyblue", "royalblue", "blue"))

myplot +
 scale_fill_brewer(palette="BrBG")

```

두 번째 `scale_fill_brewer`의 경우는 brewer라는 (<https://colorbrewer2.org/>) 미리 지정된 색의 조합을 가져와 사용하는 방식입니다. `?scale_fill_brewer`의 Palettes 섹션을 보시면 사용 가능한 팔레트의 이름이 나와 있으며 위 예제에서는 BrBG라는 이름의 팔레트를 사용했습니다. 아래는 viridis라는 이름의 팔레트이며 (<https://bids.github.io/colormap/>) 이러한 팔레트는 R 뿐만 아니라 python, Matlab 등의 다른 프로그래밍 언어에서도 사용할 수 있도록 라이브러리를 제공하고 있습니다.

```

myplot +
 scale_fill_viridis_d()

```

참고로 앞서 설명한 바와 같이 `aes(fill=bloodtype)`이 사용되었으므로 `scale_fill_viridis_d`을 사용했으며 만약 `aes(color=bloodtype)`으로 사용되었을 경우에는 이에 맞는 `scale_fill_viridis_d`으로 설정해야 합니다. 맵핑된 데이터가 연속형일 경우에는 (위 학급 예제의 혈액형은 4개의 혈액형으로 나뉘는 범주형 데이터임) `scale_fill_gradient`, `scale_fill_distiller` 등의

연속형 데이터에 맞는 scale 함수를 사용해야 합니다. 또한 데이터의 스케일이 log나 지수 단위일 경우에도 일 때에도 scale\_x\_log10() 등의 함수를 이용해서 x축 또는 y축의 스케일을 변경해줄 수 있습니다. 다음은 간단한 형태의 로그 분포 데이터를 생성하고 히스토그램을 그리는 코드입니다.

```
mydf <- data.frame(x=rlnorm(1000, log(10), log(2.5)))
p <- ggplot(mydf, aes(x=x)) +
 geom_histogram()
p
```

위 히스토그램의 x축을 로그 스케일로 전환하고자 할 때 다음과 같이 scale\_x\_log10() 함수를 추가하면 됩니다.

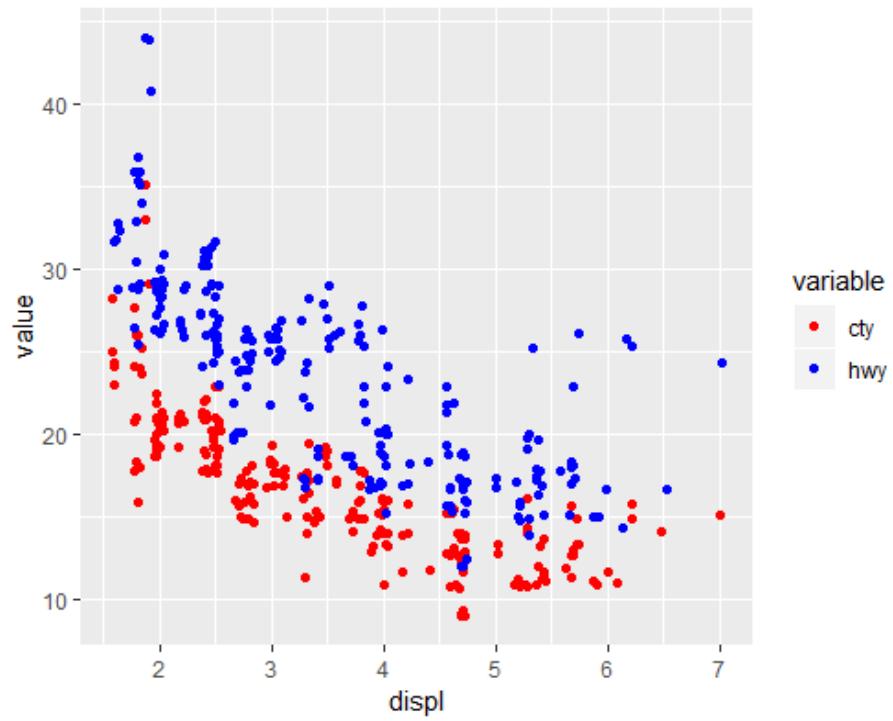
```
p + scale_x_log10()
```

## Exercises

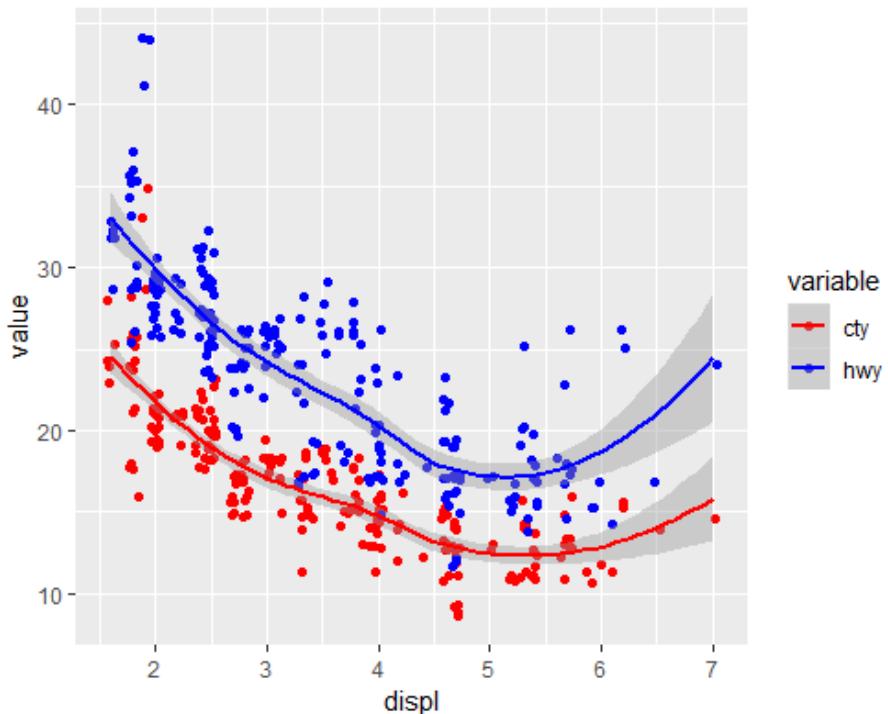
mpg 데이터셋은 38종 자동차의 연비 데이터임. 이 데이터셋을 이용하여 다음 그래프를 그리시오

- 1) 엔진 배기량과 (displ) 도심연비 (cty)를 비교하는 산포도를 그리고 어떤 연관성이 있는지 설명하시오
- 2) 위 산포도의 점들은 실제로는 한 개 이상의 데이터가 겹쳐서 표현된 경우가 많음. ggplot2에서는 이러한 문제를 극복하기 위해서 position="jitter"라는 옵션을 사용할 수 있음. 이 옵션을 적용한 코드를 작성하시오.
- 3) 위 그래프에 배기량과 (displ) 고속도로연비 (hwy) 산포도를 추가하여 다음과 같이 scale\_color\_manual() 함수를 사용해서 “red”와 “blue”로 점들을 표현한 그래프를 그리시오.

```
mydf <- data.frame(displ=mpg$displ, cty=mpg$cty, hwy=mpg$hwy) %>%
 pivot_longer(cols=c("cty", "hwy"), names_to="type")
str(mydf)
```



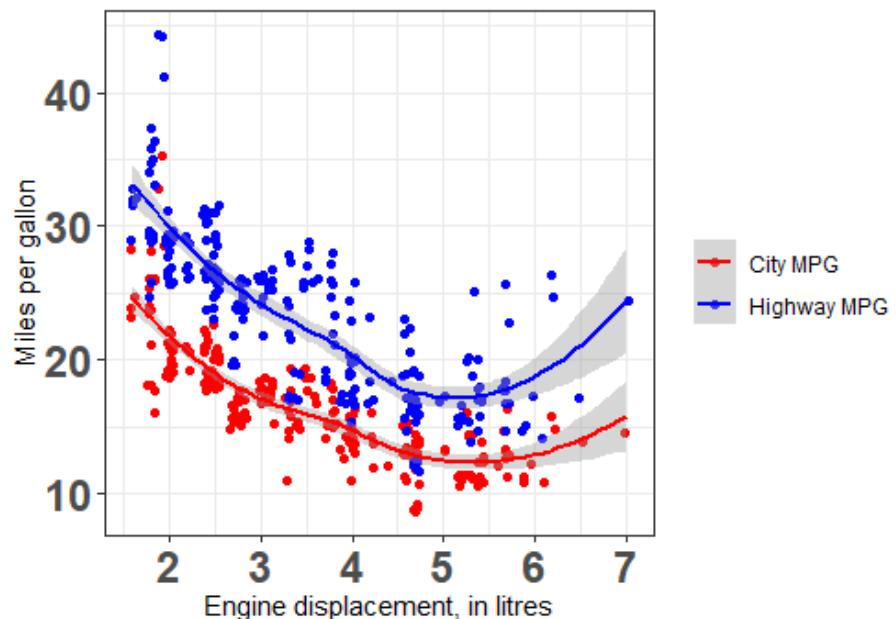
- 4) 다음과 같이 배기량과 고속도로/도심 연비의 관계를 나타내는 추세선을 추가하시오 (geom\_smooth 이용)



- 5) 아래 그림과 같이 Theme을 `theme_bw()`를 사용하고 추가로 Title, subtitle, x축, y축 라벨, 그리고 범례의 Title을 변경하시오. (범례의 라벨 설정은 `scale_color_manual`에서 `labels=c("City MPG", "Highway MPG")`으로 설정, 범례의 title을 지울때는 `name=element_blank()`, Title의 텍스트 크기는 20, x축, y축의 라벨 텍스트 크기는 18로 설정)

## Engine displacement vs. MPG

City MPG vs. Highway MPG



# Chapter 9

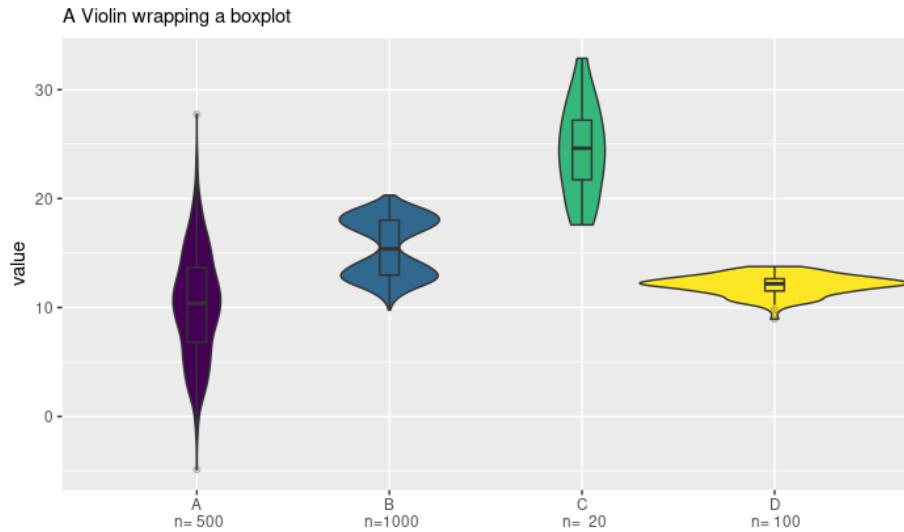
## ggplot2 examples

인터넷에서 찾은 다음 사이트의 예제를 보면서 다양한 그래프 예제를 실행해 보겠습니다. 코드는 조금씩 변형된 부분이 있으니 참고 부탁 드립니다.

- <https://www.r-graph-gallery.com/ggplot2-package.html>
- <http://r-statistics.co/Top50-Ggplot2-Visualizations-MasterList-R-Code.html>
- <https://www.datanovia.com/en/blog/ggplot-examples-best-reference/>

### 9.1 Violin plot

- [https://www.r-graph-gallery.com/violin\\_and\\_boxplot\\_ggplot2.html](https://www.r-graph-gallery.com/violin_and_boxplot_ggplot2.html)



```

library(tidyverse)
library(viridis)

create a dataset
data <- data.frame(
 name=c(rep("A",500), rep("B",500), rep("B",500), rep("C",20), rep('D', 100)),
 value=c(rnorm(500, 10, 5), rnorm(500, 13, 1), rnorm(500, 18, 1), rnorm(20, 25, 4),
))

data %>% str

ggplot(data, aes(x=name, y=value, fill=name)) +
 geom_violin(width=1.4) +
 geom_boxplot(width=0.1, alpha=0.2)

sample summary
sample_size = data %>%
 group_by(name) %>%
 summarize(num=n())

xlab <- sample_size %>%
 apply(1, function(x) paste0(x, collapse="\n n="))

apply(sample_size, 1, function(x) paste0(x, collapse="\n n="))

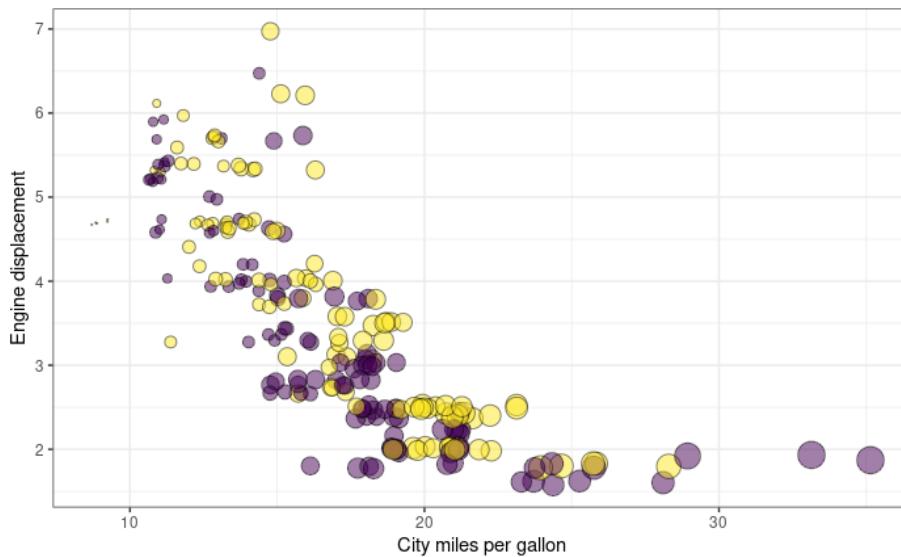
ggplot(data, aes(x=name, y=value, fill=name)) +
 geom_violin(width=1.4) +

```

```
geom_boxplot(width=0.1, alpha=0.2) +
scale_fill_viridis(discrete = TRUE) +
scale_x_discrete(labels=xlab) +
theme(
 legend.position="none",
 plot.title = element_text(size=11)
) +
ggtitle("A Violin wrapping a boxplot") +
xlab("")
```

## 9.2 Bubble plot

- <https://www.r-graph-gallery.com/320-the-basis-of-bubble-plot.html>



```
mpg %>% str

Most basic bubble plot
ggplot(mpg, aes(x=cty, y=displ, size = hwy)) +
 geom_point(alpha=0.7, position="jitter")

ggplot(mpg, aes(x=cty, y=displ, size = hwy)) +
 geom_point(alpha=0.3, position="jitter") +
 scale_size(range = c(.1, 7), name="")

ggplot(mpg, aes(x=cty, y=displ, size = hwy, color=year)) +
 geom_point(alpha=0.3, position="jitter") +
```

```

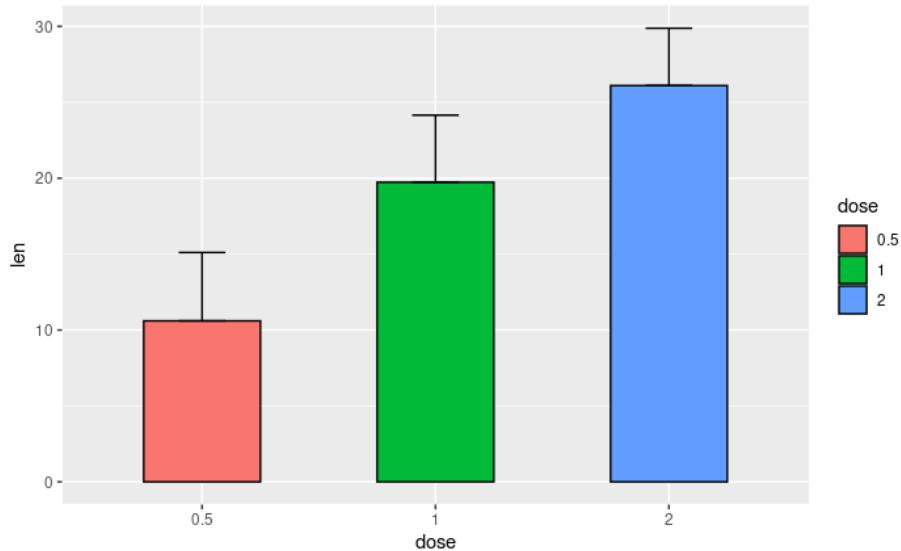
scale_size(range = c(.1, 7), name="")

mpg %>%
 mutate(yearf = factor(year)) %>%
 ggplot(aes(x=cty, y=displ, size=hwy, color=yearf)) +
 geom_point(alpha=0.3, position="jitter") +
 scale_size(range = c(.1, 7), name="")

mpg %>%
 mutate(yearf = factor(year)) %>%
 ggplot(aes(x=cty, y=displ, size=hwy, fill=yearf)) +
 geom_point(alpha=0.5, position="jitter", shape=21) +
 scale_size(range = c(.1, 7), name="") +
 scale_fill_viridis(discrete=TRUE, guide=FALSE, option="D") +
 theme_bw() +
 ylab("Engine displacement") +
 xlab("City miles per gallon") +
 theme(legend.position = "none")

```

### 9.3 Barplot with errorbars



```

ToothGrowth %>% str

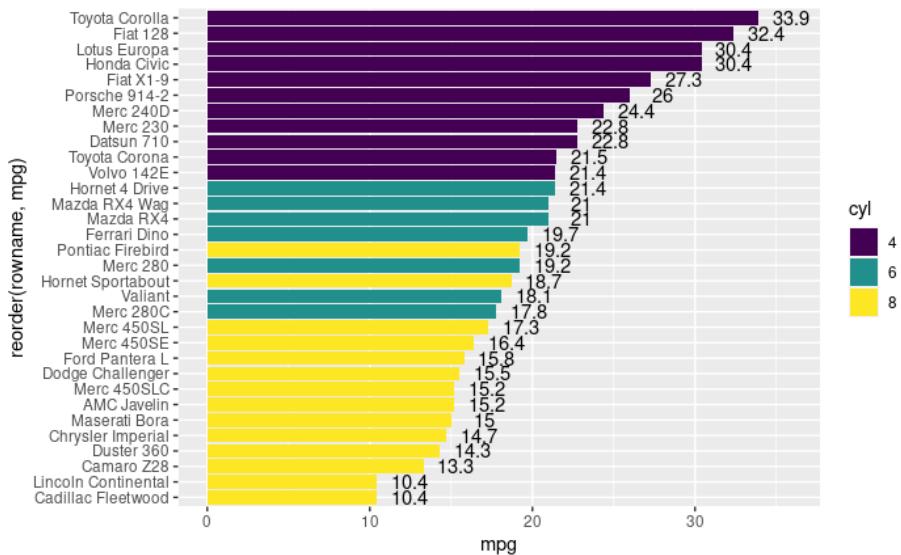
df <- ToothGrowth %>%
 mutate(dose = as.factor(dose))
df %>% str

```

```
summary
df_summary <- df %>%
 group_by(dose) %>%
 summarise(sd = sd(len, na.rm = TRUE), len = mean(len))
df_summary

ggplot(df_summary, aes(x=dose, y=len, fill=dose)) +
 geom_bar(stat = "identity", color = "black", width = 0.5) +
 geom_errorbar(aes(ymin = len, ymax = len+sd), width = 0.2)
```

## 9.4 horizontal barplot



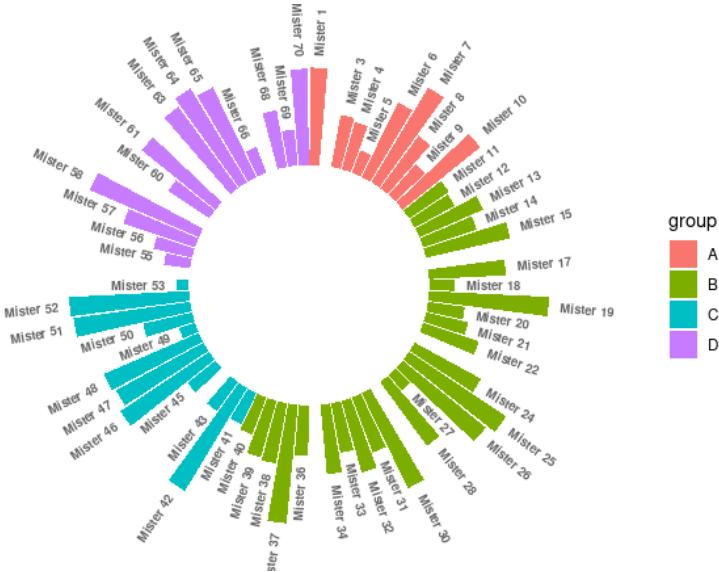
```
df <- mtcars %>%
 rownames_to_column() %>%
 as_data_frame() %>%
 mutate(cyl = as.factor(cyl)) %>%
 select(rowname, wt, mpg, cyl)
df

change fill color by groups and add text labels
ggplot(df, aes(x = reorder(rowname, mpg), y = mpg)) +
 geom_col(aes(fill = cyl)) +
 geom_text(aes(label = mpg), nudge_y = 2) +
```

```
coord_flip() +
scale_fill_viridis_d()
```

## 9.5 Circular barplot

- <https://www.r-graph-gallery.com/297-circular-barplot-with-groups.html>



```
Create dataset
n <- 70
data <- data.frame(
 id = seq(1, n),
 individual=paste("Mister ", seq(1,n), sep=""),
 group=c(rep('A', 10), rep('B', 30), rep('C', 14), rep('D', n-10-30-14)),
 value=sample(seq(10,100), n, replace=T)
)
data %>% str

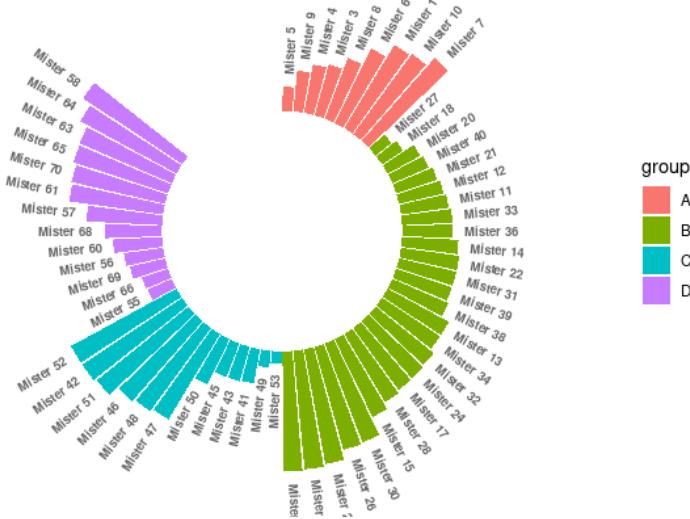
introduce NA
empty_bar_idx <- sample(1:n, 10)
data[empty_bar_idx,c(2:4)] <- c(NA, NA, NA)

label_data <- data
number_of_bar <- nrow(label_data)
angle <- 90 - 360 * (label_data$id-0.5) /number_of_bar # I subtract 0.5 because the
label_data$hjust <- ifelse(angle < -90, 1, 0)
```

```
label_data$angle <- ifelse(angle < -90, angle+180, angle)
```

```
data %>%
 ggplot(aes(x=as.factor(id), y=value, fill=group)) +
 geom_bar(stat="identity") +
 ylim(-100,120) +
 theme_minimal() +
 theme(
 axis.text = element_blank(),
 axis.title = element_blank(),
 panel.grid = element_blank(),
 plot.margin = unit(rep(-1,4), "cm")
) +
 coord_polar(start = 0) +
 geom_text(data=label_data, aes(x=id, y=value+10, label=individual, hjust=hjust), color="black",
```

데이터 정렬 후 plot



```
data2 <- data %>%
 arrange(group, value) %>%
 mutate(id2=1:n())
```

```
label_data2 <- data2
number_of_bar <- nrow(label_data2)
angle <- 90 - 360 * (label_data2$id2-0.5) / number_of_bar # I subtract 0.5 because the letter
label_data2$hjust <- ifelse(angle < -90, 1, 0)
label_data2$angle <- ifelse(angle < -90, angle+180, angle)
```

```

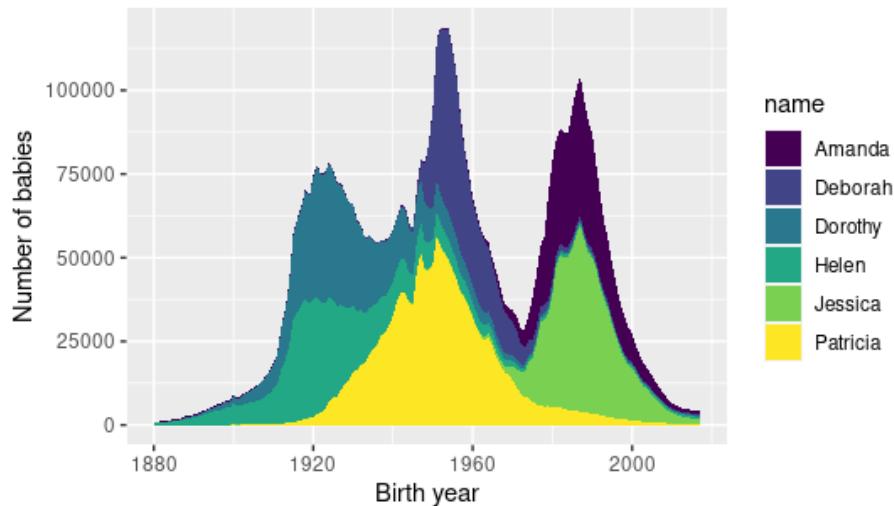
data2 %>%
 ggplot(aes(x=as.factor(id2), y=value, fill=group)) +
 geom_bar(stat="identity") +
 ylim(-100,120) +
 theme_minimal() +
 theme(
 axis.text = element_blank(),
 axis.title = element_blank(),
 panel.grid = element_blank(),
 plot.margin = unit(rep(-1,4), "cm")
) +
 coord_polar(start = 0) +
 geom_text(data=label_data2, aes(x=id2, y=value+10, label=individual, hjust=hjust), c

```

## 9.6 Stacked area chart

- <https://www.data-to-viz.com/caveat/stacking.html>

Popularity of American names in the previous 30 years



```

library(babynames)

babynames %>% str

Load dataset from github

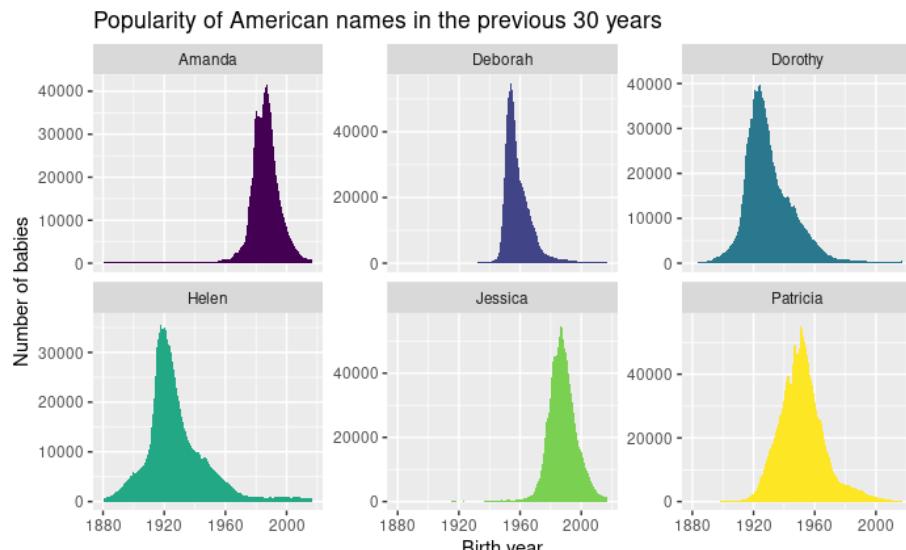
```

```

data <- babynames %>%
 filter(name %in% c("Amanda", "Jessica", "Patricia", "Deborah", "Dorothy", "Helen")) %>%
 filter(sex=="F")

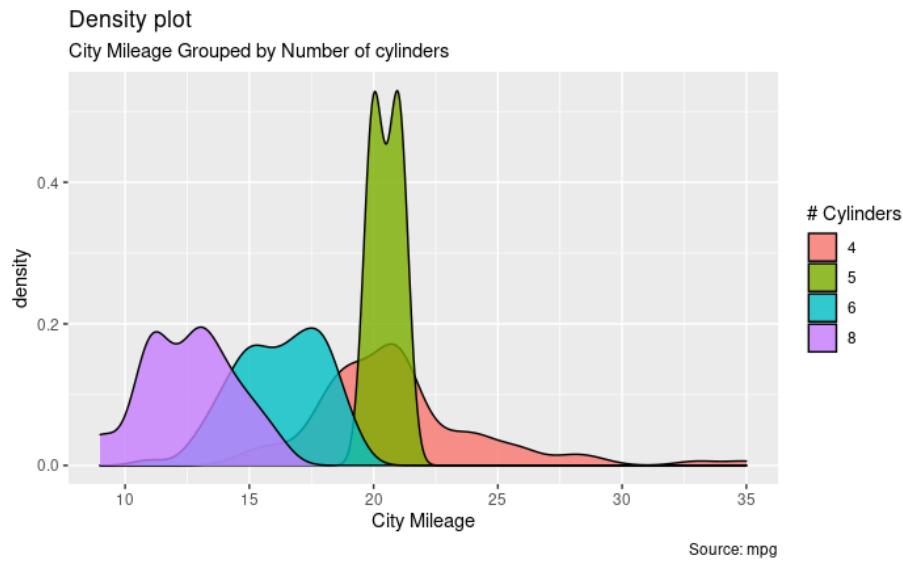
Plot
p <- data %>%
 ggplot(aes(x=year, y=n, fill=name, text=name)) +
 geom_area() +
 scale_fill_viridis(discrete = TRUE) +
 ggtitle("Popularity of American names in the previous 30 years") +
 theme() +
 xlab("Birth year") +
 ylab("Number of babies")
p

```



```
p + facet_wrap(~name, scale="free_y")
```

## 9.7 Density plot



```
Plot
g <- ggplot(mpg, aes(cty))
g + geom_density(aes(fill=factor(cyl)), alpha=0.8) +
 labs(title="Density plot",
 subtitle="City Mileage Grouped by Number of cylinders",
 caption="Source: mpg",
 x="City Mileage",
 fill="# Cylinders")
```

## 9.8 Waffle chart

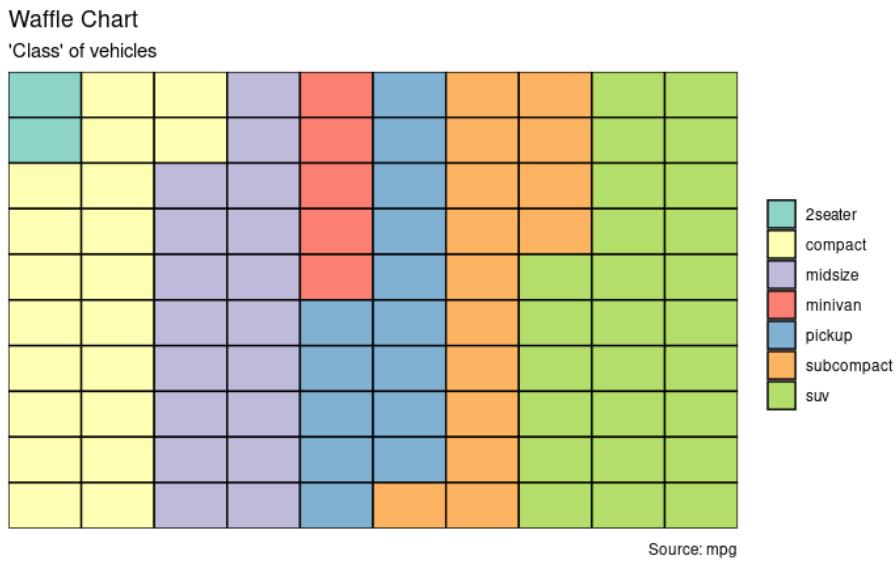
- <http://r-statistics.co/Top50-Ggplot2-Visualizations-MasterList-R-Code.html#Waffle%20Chart>

```
var <- mpg$class # the categorical data

Prep data (nothing to change here)
nrows <- 10
df <- expand.grid(y = 1:nrows, x = 1:nrows)
categ_table <- round(table(var) * ((nrows*nrows)/(length(var))))
categ_table

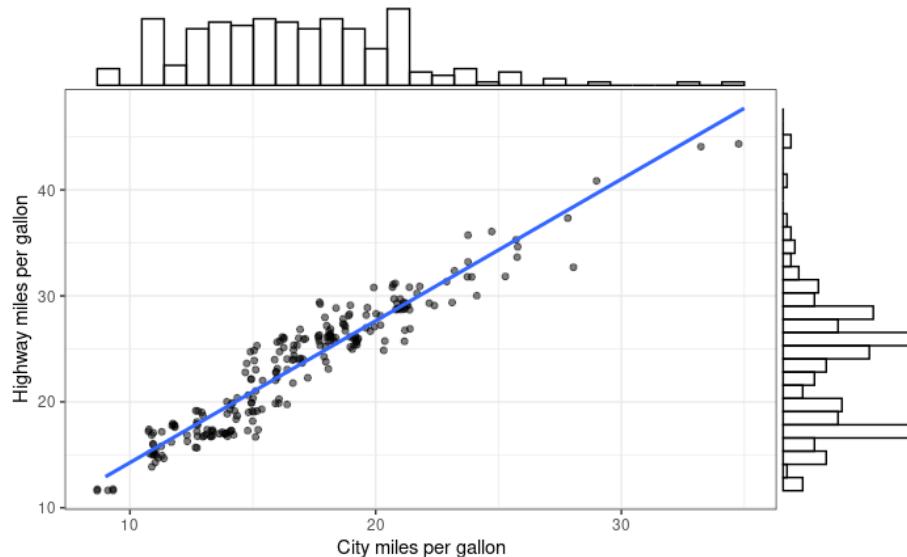
df$category <- factor(rep(names(categ_table), categ_table))
```

```
NOTE: if sum(categ_table) is not 100 (i.e. nrow > 2), it will need adjustment to make the sum to 100
Plot
df %>% str
ggplot(df, aes(x = x, y = y, fill = category)) +
 geom_tile(color = "black", size = 0.5)
```



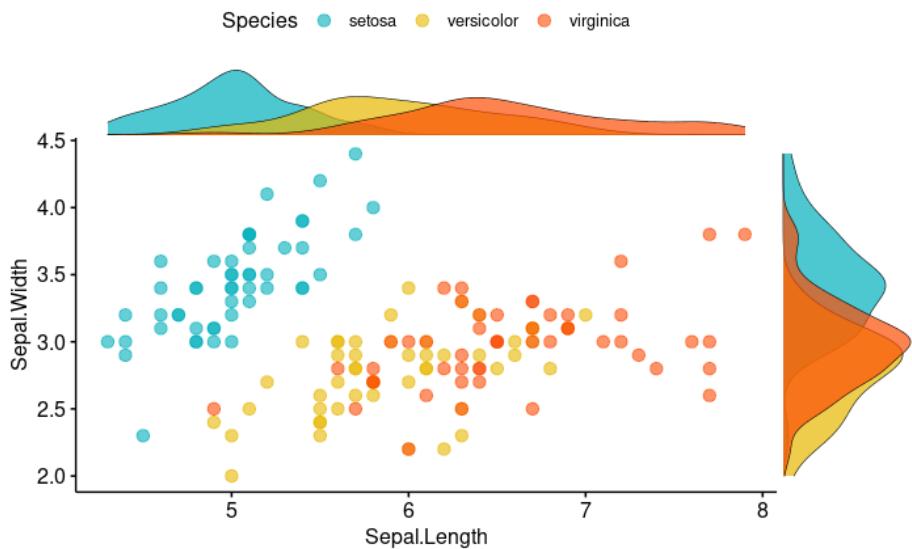
```
ggplot(df, aes(x = x, y = y, fill = category)) +
 geom_tile(color = "black", size = 0.5) +
 scale_x_continuous(expand = c(0, 0)) +
 scale_y_continuous(expand = c(0, 0), trans = 'reverse') +
 scale_fill_brewer(palette = "Set3") +
 labs(title = "Waffle Chart", subtitle = "'Class' of vehicles",
 caption = "Source: mpg") +
 theme(plot.title = element_text(size = rel(1.2)),
 axis.text = element_blank(),
 axis.title = element_blank(),
 axis.ticks = element_blank(),
 legend.title = element_blank(),
 legend.position = "right")
```

## 9.9 Marginal histogram



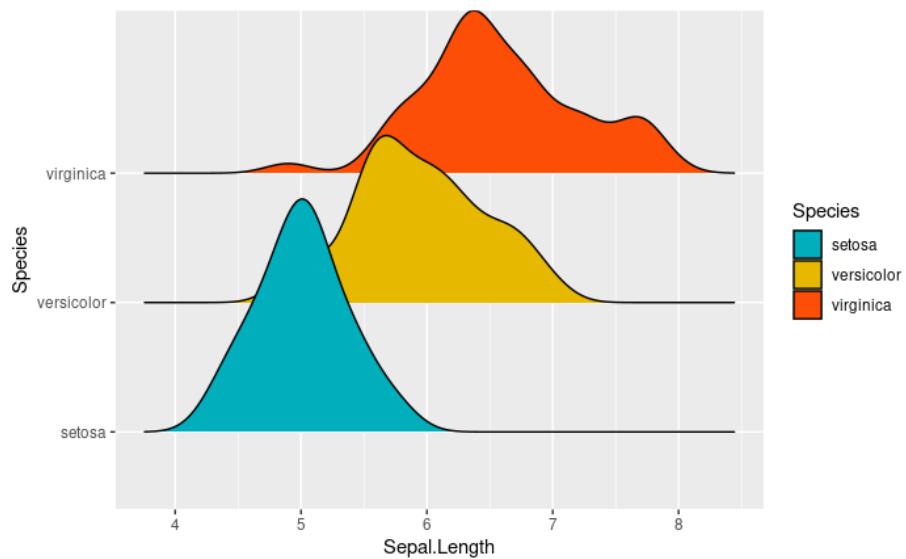
```
library(ggExtra)

Scatterplot
p <- ggplot(mpg, aes(x=cty, y=hwy)) +
 geom_point(position="jitter", alpha=0.5) +
 geom_smooth(method="lm", se=F) +
 theme_bw() +
 theme(
 legend.position = "none"
) +
 xlab("City miles per gallon") +
 ylab("Highway miles per gallon")
p
ggMarginal(p, type = "histogram", fill="transparent")
ggMarginal(p, type = "density", fill="transparent")
```



```
library(ggpubr)
Grouped Scatter plot with marginal density plots
ggscatterhist(
 iris,
 x = "Sepal.Length",
 y = "Sepal.Width",
 color = "Species",
 size = 3,
 alpha = 0.6,
 palette = c("#00AFBB", "#E7B800", "#FC4E07"),
 margin.params = list(fill = "Species", color = "black", size = 0.2)
)
```

## 9.10 Density ridgeline plots



```
library(ggridges)
ggplot(iris, aes(x = Sepal.Length, y = Species)) +
 geom_density_ridges(aes(fill = Species)) +
 scale_fill_manual(values = c("#00AFBB", "#E7B800", "#FC4E07"))
```

---

이 저작물은 크리에이티브 커먼즈 저작자표시-비영리-변경금지 4.0 국제 라이선스에 따라 이용할 수 있습니다.

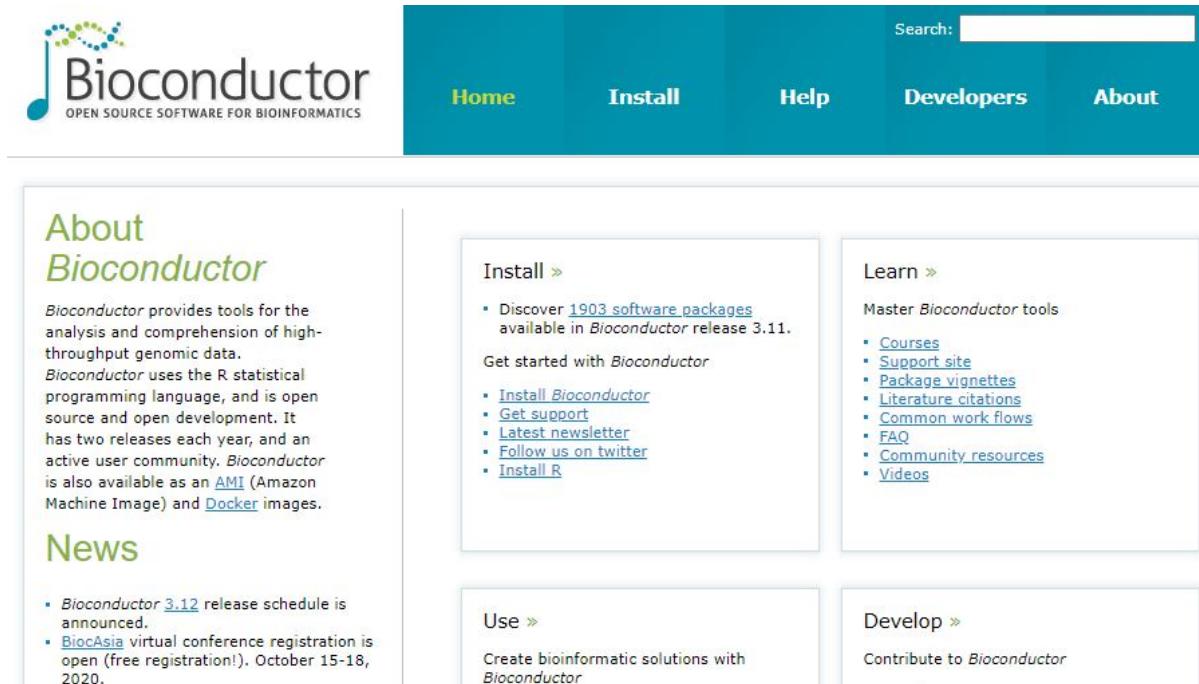
# Chapter 10

## Bioconductor

- <https://www.bioconductor.org>

Bioconductor는 바이오인포메틱스를 위한 R기반의 데이터, 메소드, 그리고 패키지들의 모음입니다. 2002년 microarray 데이터 분석을 위한 플랫폼으로 시작되었으며 현재 2000개 이상의 패키지로 구성되어 있습니다. R은 분산형 오픈소스이나 Bioconductor는 Full-time developer들에 의해서 유지되고 있습니다. CRAN에 배포되지 않고 CRAN에 비해 더 많은 필수 자료들 (vignettes 등)이 필요하며 높은 수준으로 quality control이 되고 있습니다. Bioconductor는 6개월마다 예정된 릴리스를 통해 모든 bioconductor 패키지가 충돌없이 조화롭게 작동하도록 유지되고 있습니다.

사용 가능한 패키지들은 이곳을 참고하시면 되겠습니다.



The screenshot shows the Bioconductor website's 'About' page. At the top, there is a navigation bar with links for Home, Install, Help, Developers, and About. A search bar is also present. The main content area has a sidebar on the left with sections for 'About Bioconductor' (describing it as open-source software for bioinformatics) and 'News' (mentioning the 3.12 release schedule and BioAsia registration). The main content area contains sections for 'Install' (with a link to 1903 software packages), 'Learn' (with links to Courses, Support site, Package vignettes, Literature citations, Common work flows, FAQ, Community resources, and Videos), 'Use' (with a link to Create bioinformatic solutions with Bioconductor), and 'Develop' (with a link to Contribute to Bioconductor).

Bioconductor 코어 개발 그룹은 사용자들이 지놈스케일 데이터를 더 편리하게 다루를 수 있도록 데이터의 구조를 개발하고 있습니다. Bioconductor의 주요 기능은 다음과 같습니다.

- 지놈스케일의 서열이나 발현등 대용량 유전자형 데이터 관리 및 통계적 분석을 위한 툴 제공
- 분자수준의 현상과 생장이나 질병 등 표현형수준의 관계를 규명하기 위한 정량 데이터 통합 및 관리

## 10.1 Packages

메인화면 > Use > Software, Annotation, Experiment

- Software: 데이터 분석을 위한 알고리즘/툴 모음
- Annotation: 유전자 symbol/ID mapping, gene ontology 기반 유전자 분류, 유전체상에서 exon, transcript, gene 등의 위치, 단백질 기능 등. Annotation > Packagetype 참고
- Experiment data: 검증된 실험 데이터
- Workflow: 특정 데이터 분석을 위한 프로세스 모음 RNA-seq, ChIP seq, copy number analysis, microarray methylation, classic expression analysis, flow cytometry 등



**Bioconductor**  
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Home      Install      Help      Developers      About

Search:

[Home](#) » [BiocViews](#)

## All Packages

**Bioconductor version 3.13 (Release)**

Autocomplete biocViews search:

Packages found under Software:

Rank based on number of downloads: lower numbers are more frequently downloaded.

Show All entries	Search table: <input type="text"/>		
Package	Maintainer	Title	Rank
<a href="#">BiocGenerics</a>	Bioconductor Package Maintainer	S4 generic functions used in Bioconductor	1
<a href="#">BiocVersion</a>	Bioconductor Package Maintainer	Set the appropriate version of Bioconductor packages	2
<a href="#">S4Vectors</a>	Bioconductor Package Maintainer	Foundation of vector-like and list-like containers in Bioconductor	3
<a href="#">IRanges</a>	Bioconductor Package Maintainer	Foundation of integer range manipulation in Bioconductor	4
<a href="#">Biobase</a>	Bioconductor Package Maintainer	Biobase: Base functions for Bioconductor	5
<a href="#">zlibbioc</a>	Bioconductor Package Maintainer	An R packaged zlib-1.2.5	6
<a href="#">GenomeInfoDb</a>	Bioconductor Package Maintainer	Utilities for manipulating chromosome names, including modifying them to follow a particular naming style	7
<a href="#">XVector</a>	Hervé Pagès	Foundation of external vector representation and manipulation in Bioconductor	8
<a href="#">DelayedArray</a>	Hervé Pagès	A unified framework for working transparently with on-disk and in-memory array-like datasets	9
<a href="#">AnnotationDbi</a>	Bioconductor Package Maintainer	Manipulation of SQLite-based annotations in Bioconductor	10
<a href="#">GenomicRanges</a>	Bioconductor Package Maintainer	Representation and manipulation of genomic intervals	11

Annotation 리소스는 다음과 같이 몇 단계의 레벨로 구분할 수 있습니다.

- ChipDb: 가장 낮은 단계, Affymatrix Chip 정보
- OrgDb: 특정 생물(Organism)의 기능적 annotations
- TxDb/EnsDb: 전사체 정보, 위치 정보
- OrganismDb: meta-packages for OrgDb, TxDb
- BSgenome 특정 생물의 실제 염기 정보
- Others GO.db; KEGG.db
- AnnotationHub:
- biomaRt:

Bioconductor에서 제공하는 패키지를 설치하기 위해서는 BiocManager를

먼저 설치하고 해당 패키지를 설치하시기 바랍니다. BiocManager에는 available()이라는 함수로 (특정 문자가 포함된) 사용 가능한 패키지를 검색할 수도 있습니다. 예를 들어 IRanges라는 패키지를 설치할 경우 bioconductor 상단 오른쪽의 Search 나 software package list의 검색창에서 IRanges를 입력하여 해당 패키지를 찾고 다음과 같이 설치를 수행합니다.

```
if (!requireNamespace("BiocManager", quietly = TRUE))
 install.packages("BiocManager")

BiocManager::install("IRanges")
.libPaths()
```

### Exercises

OrganismDb는 meta-package의 형태로 OrgDb, TxDb, 그리고 GO.db 패키지들을 포함하는 정보를 가지고 있음. OrganismDB 중 인간의 정보를 가진 Homo.sapiens를 찾아 설치하시오

## 10.2 Learning and support

각 패키지는 제목, 저자, 유지관리자, 설명, 참조, 설치법 등의 정보가 포함된 landing page가 있으며 패키지 내 함수들은 상세한 설명과 예제가 제공됩니다. 예를 들어 IRanges의 landing page를 참고하세요. vignettes는 bioconductor의 중요한 특징 중 하나로 R 코드와 함께 패키지를 사용하는 방법에 대한 상세한 설명을 제공하는 문서입니다.

```
library(IRanges)

vignette(package="IRanges")
browseVignettes("IRanges")
vignette("IRangesOverview", package="IRanges")

ir1 <- IRanges(start=1:10, width=10:1)
ir1
class(ir1)
methods(class="IRanges")

example(IRanges)
?IRanges
??IRanges
```

메인페이지 ➤ Learn ➤ Support site 게시판에는 관련된 여러 QnA 들이 있어서 유사 문제에 대한 도움을 받을 수 있습니다.

### 10.3 OOP - Class, Object and Method

객체지향프로그래밍 (OOP)은 복잡한 문제를 프로그래밍할 때 발생되는 코드의 복잡성을 해결할 수 있는 하나의 방안으로 1990년대부터 많이 사용되었습니다.

R도 객체지향 프로그래밍 언어입니다. 그런데 R은 다른 언어들에 비해서 좀 어려운 (다른) 개념으로 사용됩니다. R에서 사용하는 Class에는 크게 base type, S3, S4, RC, 그리고 R6 등 다양한 타입이 있고 이 중 S3를 많이 사용해 왔으며 S3의 단점을 보완한 S4 형식의 class와 R6를 주로 사용합니다 (?). 본 강의에서는 S3 형식의 class만 다루도록 하겠습니다.

클래스를 사용하는 이유는 여러가지가 있겠지만 복잡한 개념의 데이터를 구조화하고 쉽게 관리하기 위해서 사용한다고 보면 될 것 같습니다. 여러분이 알아야 할 개념은 Class와 Object 그리고 Method입니다. 사실 R의 모든것이 Object이고 이러한 Object들의 정의가 Class입니다.

```
df <- data.frame(x=c(1:5), y=LETTERS[1:5])
df
class(df)
```

위에서 df는 변수라고 부르지만 object이기도 합니다. df의 class는 data.frame입니다. 클래스는 누구든 원하는 만큼 얼마든지 만들 수 있습니다.

```
class(df) <- "myclass"
df
class(df)

class(df) <- c("data.frame", "myclass")
df
class(df)
```

그런데 모든 object들이 OOP 유래는 아닙니다 base object들이 그 예입니다.

```
x <- 1:10
class(x)
attr(x, "class")

mtcars
attr(mtcars, "class")
```

method는 위와 같은 클래스들에 특화된 어떤 기능을 하는 함수라고 생각하시면 됩니다.

```
mt <- matrix(1:9, 3,3)
df <- data.frame(1:3, 4:6, 7:9)

class(mt)
class(df)
str(mt)
```

```
str(df)

diamonds <- ggplot2::diamonds

summary(diamonds$carat)
summary(diamonds$cut)

methods(class="data.frame")
```

위 summary, str 등이 generic function이라 불리는 method들입니다. class마다 사용 가능한 method가 어떤 정보가 있는지 알기 위해서 methods()라는 함수를 사용합니다. R의 객체지향프로그래밍에 대한 상세한 내용은 Advanced R를 참고하세요.

### Exercises

다음 두 종류의 객체에 대해서 class 가 integer 일 경우 평균을 계산하고 character일 경우 비율을 계산하는 (table 함수 사용) mysummary 함수를 만드시오

```
x <- c(1:10)
y <- c("A", "G", "G", "T", "A")
```

## 10.4 Bioconductor의 OOP

bioconductor에서 다루는 genome 스케일의 experiment나 annotation은 대표적인 복잡한 데이터 중 하나입니다. Bioconductor에서 OOP 개념은 다음과 같습니다.

- class - 복잡한 생물학적 데이터 구조의 틀 정의
- object - 특정 클래스가 특정 구현된 실체
- method - 특정 클래스에 대한 기능 수행

예를 들어 앞에서 설치한 Homo.sapiens의 class인 OrganismDb 살펴보면 다음과 같습니다.

```
library(Homo.sapiens)
class(Homo.sapiens)
?OrganismDb
```

The OrganismDb class is a container for storing knowledge about existing Annotation packages and the relationships between these resources. The purpose of this object and its associated methods is to provide a means by which users can conveniently query for data from several different annotation resources at the same time using a familiar interface.

```
homo_seq <- seqinfo(Homo.sapiens)
class(homo_seq)
?Seqinfo
```

A Seqinfo object is a table-like object that contains basic information about a set of genomic sequences. ...

```
length(homo_seq)
seqnames(homo_seq)
```

bioconductor에는 대용량 정보가 object 형태로 구조화되어 저장되어 있으며 library()함수로 읽어올 수 있고 다양한 함수로 해당 object의 정보를 읽어올 수 있습니다.

### Exercises

Homo.sapiens 정보에서 상위 10개 유전자와 상위 10개 exon을 구하시오

---

이 저작물은 크리에이티브 커먼즈 저작자표시-비영리-변경금지 4.0 국제 라이선스에 따라 이용할 수 있습니다.



# Chapter 11

## Biostrings

High-throughput sequencing 데이터를 포함한 DNA나 Amino acid와 같은 생물학적 서열은 Bioconductor의 다양한 패키지들에 의해서 분석될 수 있으며 특히 Biostrings 패키지는 생물학적 서열을 효과적으로 활용하기 위한 핵심 도구로 활용됩니다.

### 11.1 Working with sequences

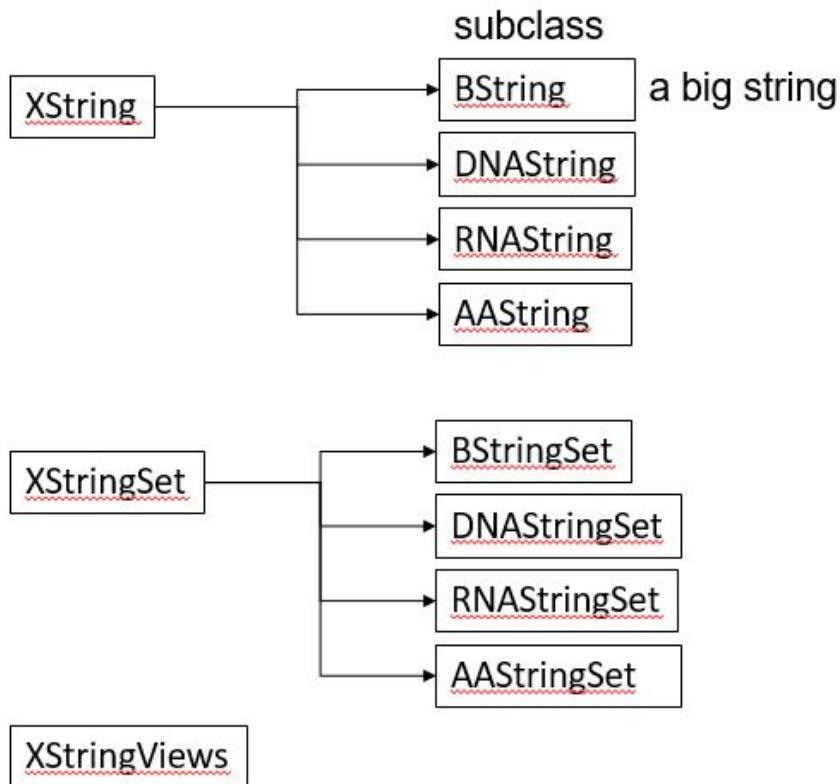
Biostrings는 DNA, RNA, amino acids와 같은 생물학적 string을 다루기 위한 다양한 함수를 제공하는 패키지입니다. 특히 서열에서의 패턴 탐색이나 Smith-Waterman local alignments, Needleman-Wunsch global alignments 등의 서열 비교함수를 제공하여 간단한 서열 분석에 자주 활용되는 패키지입니다 (?). Biostrings 패키지의 설치 방법은 아래와 같습니다.

```
if (!requireNamespace("BiocManager", quietly = TRUE))
 install.packages("BiocManager")

BiocManager::install("Biostrings")

library(Biostrings)
```

Biostrings 패키지는 기본적으로 XString, XStringSet, XStringViews 3가지의 class를 정의하고 있습니다. XString은 DNA나 RNA, AA 등 생물학적 서열 한 가닥을 다루기위한 클래스이며 XStringSet은 여러 가닥을 다루기위한 클래스입니다.



DNAString 함수를 이용해서 객체를 만들어낼 수 있으며 ‘A’, ‘C’, ‘G’, ‘T’ 외에 ‘-’ (insertion), ‘N’ 을 허용합니다.

```

dna1 <- DNAString("ACGT?")
dna1 <- DNAString("ACGT-N")
dna1[1]
dna1[2:3]

dna2 <- DNAStringSet(c("ACGT", "GTCA", "GCTA"))
dna2[1]
dna2[[1]]
dna2[[1]][1]

```

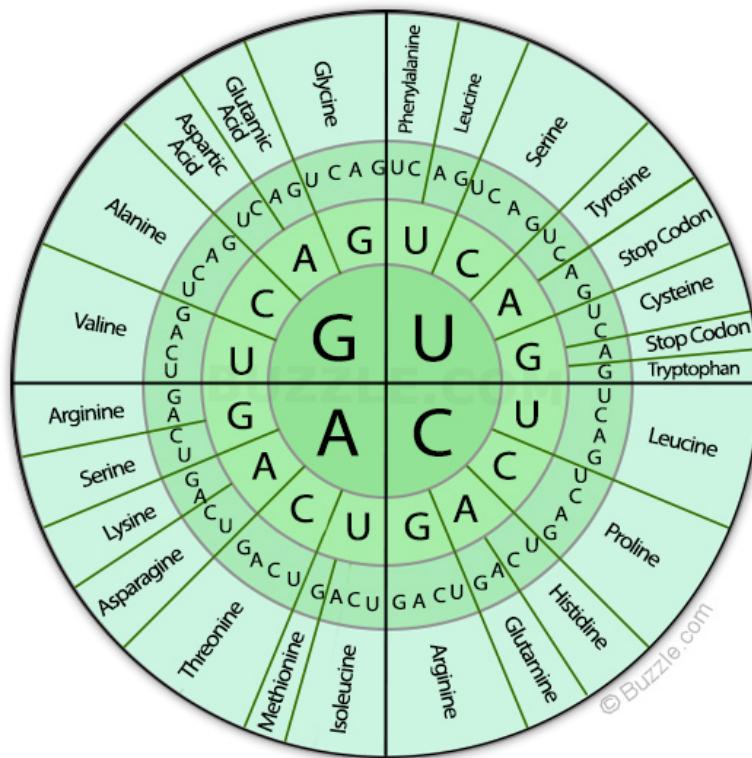
다음 내장변수 들은 Biostrings 패키지를 로드하면 자동으로 저장되는 변수들로 생물학적 서열을 미리 정의해 놓았습니다. IUPAC (International Union of Pure and Applied Chemistry, 국제 순수·응용 화학 연합)

```

DNA_BASES
DNA_ALPHABET

```

IUPAC\_CODE\_MAP  
GENETIC\_CODE



To decode the codon, move from the center circle towards the periphery.

위 변수들을 이용하면 다음처럼 `sample()` 함수를 이용해서 랜덤하게 DNA 서열을 얻을 수 있습니다. `DNA_BASES`가 4개 길이를 갖는 벡터인데 이 중 10개를 뽑으려면 `replace=T`로 해야 합니다.

```
x0 <- sample(DNA_BASES, 10, replace = T)
x0
s1 <- "ATG"
s2 <- "CCC"
s3 <- paste(s1, s2, sep="")
s3
x1 <- paste(x0, collapse="")
x1
```

관련 함수는 Cheat sheet 참고

### 11.1.1 XString

XString 클래스는 DNAString과 RNAString, AAString의 subclass로 나눌 수 있습니다. DNAString class에서 length 함수는 핵산의 갯수를 (DNAStringSet 타입의 변수에서 length는 DNA 가닥의 갯수) 계산하며 핵산의 갯수는 nchar 함수로 얻어낼 수 있습니다. toString은 DNAString 타입을 단순 문자열로 변환해주는 함수이며 상보서열, 역상보서열 등의 정보도 complement, reverseComplement 등을 사용하여 찾아낼 수 있습니다.

```
x0 <- paste(sample(DNA_BASES, 10, replace = T), collapse="")
x1 <- DNAString(x0)
class(x0)
class(x1)
length(x1)
toString(x1)
complement(x1)
Biostrings::complement(x1)
reverseComplement(x1)
```

DNAString의 인덱싱은 vector (string)과 같으며 DNAStringSet은 list의 인덱싱과 같습니다.

```
indexing
x1[1]
x1[1:3]
subseq(x1, start=3, end=5)
subseq(x1, 3, 5)

letter frequency
alphabetFrequency(x1, baseOnly=TRUE, as.prob=TRUE)
letterFrequency(x1, c("G", "C"), as.prob=TRUE)
```

#### Exercises

1. 개시코돈과 스탑코돈을 포함한 30개 길이를 갖는 랜덤 유전자서열을 하나 만드시오
2. AA\_ALPHABET은 IUPAC에서 정의된 아미노산 서열 알파벳이 저장된 내장변수임. “M”과 “\*”를 포함하는 10개 길이를 갖는 랜덤 유전자서열을 하나 만드시오

### 11.1.2 XStringSet

XStringSet역시 DNAStringSet, RNAStringSet, 그리고 AAStringSet으로 나눌 수 있으며 DNAStringSet class는 여러개의 DNAString 을 모아 놓은 집합이라고 보면 됩니다. length 함수는 DNA string의 갯수이며 width 또는 nchar 함수로 각 string의 길이를 구할 수 있으며 이 외 대부분의 DNAString 에서 사용되는 함수가 동일하게 사용될 수 있습니다.

```

x0 <- c("CTC-NACCAGTAT", "TTGA", "TACCTAGAG")
x1 <- DNAStringSet(x0)
class(x0)
class(x1)
names(x1)
names(x1) <- c("A", "B", "C")
length(x1)
width(x1)
subseq(x1, 2, 4)
x1[[1]]
x1[1]

x3 <- DNAString("ATGAGTAGTTAG")
x4 <- c(x1, DNAStringSet(x3))
x4[-1]
x4
alphabetFrequency(x1, baseOnly=TRUE, as.prob=TRUE)
letterFrequency(x1, c("G", "C"), as.prob=TRUE)
rowSums(letterFrequency(x1, c("G", "C"), as.prob=TRUE))
subseq(x4, 2, 4)

```

RNA나 아미노산 역시 동일한 방식으로 적용 가능하며 `c` 함수를 이용해서 `XStringSet`으로 변환 가능합니다.

```

x1 <- paste(sample(AA_ALPHABET, 10, replace = T), collapse="")
x2 <- paste(sample(AA_ALPHABET, 10, replace=T), collapse="")

x3 <- AAString(x1)
x4 <- AAString(x2)

AAStringSet(c(x1, x2))
AAStringSet(c(x3, x4))

```

### Exercises

1. 시작코돈과 종결코돈이 있는 길이 36bp 짜리 DNA (랜덤) 서열을 하나 만드시오
2. 위와 같은 랜덤서열 10개 만들어서 `DNAStringSet`으로 변환하시오

아래는 가장 직관적으로 생각할 수 있는 `for`를 이용한 방법입니다. 즉, 10개 저장소를 갖는 `x0` 변수를 미리 생성해 두고 `for` 문을 돌면서 서열을 하나씩 만들어 저장하는 방법입니다.

```

x0 <- rep("", 10)
for(i in 1:length(x0)){

```

```

tmp <- paste(sample(DNA_BASES, 30, replace = T), collapse="")
x0[i] <- paste("ATG", tmp, "TAG", sep="")
}
x0

```

위 코드를 함수로 만들어 보겠습니다. random dna를 만들 때 길이만 다를 뿐 같은 코드를 반복해서 사용하고 있습니다. 이럴 경우 DNA 길이를 사용자가 정해주도록 input parameter로 하고 해당 파라미터를 받아 DNA를 만들어 주는 함수를 만들어 사용하면 편리합니다.

```

data(DNA_BASES)
random_dna <- function(len){
 tmp <- paste(sample(DNA_BASES, len, replace = T), collapse="")
 x0 <- paste("ATG", tmp, "TAG", sep="")
 return(x0)
}
random_dna(len=30)
random_dna(len=40)

```

파라미터로 넘겨진 len 값이 sample 함수의 len에 사용된 것을 참고하세요.

이제 길이 30bp짜리 10개의 서열을 반복해서 만들 때 위 함수를 앞서와 같이 for문을 이용하여 10번 반복해서 실행해 주면 같은 결과를 얻습니다. 위와 같이 함수를 만들어 두면 언제든 DNA 서열을 만들 때 재사용 할 수 있습니다.

```

x0 <- rep("", 10)
for(i in 1:length(x0)){
 x0[i] <- random_dna(30)
}
x0

```

그런데 R에는 apply 와 같은 행렬연산 함수가 있어서 for문을 사용하지 않고 편리하게 반복문을 실행할 수 있습니다. replicate 함수는 apply와 같은 기능으로 list나 vector 변수에 대해서 사용할 수 있습니다. 즉, 다음과 같이 사용자가 원하는 함수를 반복해서 실행하고 반복 수 만큼의 길이를 갖는 결과를 반환합니다.

```

x0 <- replicate(10, random_dna(30))
x0
x1 <- DNAStringSet(x0)
x1

```

### 3. 위 생성한 10개 서열의 GC 비율을 계산하고 bar그래프를 그리시오

위 x0 스트링들을 XStringSet으로 바꾸고 GC 비율을 구한 후 bargraph를 그리겠습니다. gc\_ratio가 G와 C의 비율값을 저장한 10x2 테이블이므로 x축에 10개의 서열과 각 서열의 GC비율을 나타내고 y축에 비율 값을 그리는 것으로 생각한 후 ggplot의 aes와 파라미터를 적절히 지정해 줍니다.

bar plot using ggplot2

```

x1 <- DNAStringSet(x0)
gc_ratio1 <- letterFrequency(x1, c("G", "C"), as.prob=TRUE)
gc_ratio2 <- rowSums(gc_ratio1)
barplot(gc_ratio2, beside=T)

names(gc_ratio2) <- paste("seq", 1:length(gc_ratio2), sep="")
barplot(gc_ratio2, beside=T)

data.frame(gc_ratio2) %>%
 rownames_to_column() %>%
 ggplot(aes(x=rownname, y=gc_ratio2, fill=rownname)) +
 geom_bar(stat="identity") +
 scale_y_continuous(limits = c(0, 1)) +
 scale_fill_brewer(palette = "green") +
 theme_bw()

```

### 11.1.3 XStringView

Biostrings의 또 다른 class인 XStringView는 XString class의 DNA, RNA, AA서열을 사용자가 원하는대로 볼 수 있는 인터페이스를 제공합니다. 사용법은 다음과 같습니다.

```

x2 <- x1[[1]]
Views(x2, start=1, width=20)
Views(x2, start=1, end=4)
Views(x2, start=c(1,3), end=4)
Views(x2, start=c(1,3,4), width=20)
Views(x2, start=c(1,3,4), width=20)
i <- Views(x2, start=c(1,3,4), width=20)

```

다음과 같이 한 서열에 대한 여러 부분의 서열 조각도 볼 수 있으며 gaps 함수는 매개변수로 주어진 서열 view의 구간을 제외한 나머지 구간의 서열을 보여주는 함수입니다. successiveviews 함수는 처음 서열부터 매개변수 width에 주어진 갯수 만큼의 서열을 보여주며 rep() 함수를 이용해서 서열의 처음부터 끝까지 보여주는 기능을 합니다.

```

v <- Views(x2, start=c(1,10), end=c(3,15))
gaps(v)

successiveViews(x2, width=20)
successiveViews(x2, width=rep(20, 2))
successiveViews(x2, width=rep(20, 3))

```

### Exercises

1. 1000bp 길이의 랜덤 DNA 서열을 만들고 40bp 단위의 길이로 보는 코드를 작성하시오.

앞서 만들어둔 `random_dna()` 함수를 사용하면 되며 `successiveViews` 함수를 사용해야 하므로 `DNAString`으로 변환이 필요하며 서열의 길이에 따라서 `rep()`을 이용하여 반복 횟수를 자동 계산합니다.

## 11.2 Sequence read and write

`Biostrings` 패키지의 `readDNAStringSet()`이나 `writeXStringSet`을 사용하면 기본 DNA/RNA/AA 서열의 읽고 쓰기가 가능하며 `fasta`와 `fastq` 등의 파일타입으로 적용이 가능합니다.

```
x1 <- DNAStringSet(x0)
writeXStringSet(x1, "myfastaseq.fasta", format="fasta")

names(x1) <- "myfastaseq"
writeXStringSet(x1, "myfastaseq.fasta", format="fasta")

myseq <- readDNAStringSet("myfastaseq.fasta", format="fasta")
myseq
```

`successiveViews`로 나눈 여러개의 DNA 조각을 `myfastaseqs.fasta`에 저장하고 다시 읽을 수 있습니다.

```
myseqs <- DNAStringSet(sv)
names(myseqs) <- paste("myseqs", 1:length(myseqs), sep="")
writeXStringSet(myseqs, "myfastaseqs.fasta", format="fasta")
```

## 11.3 Sequence statistics

`oligonucleotideFrequency` 는 `width`와 `step`이라는 옵션에 따라서 해당 서열의 모든 핵산의 수를 세어주는 함수입니다. 다음에 사용되는 `yeastSEQCHR1`는 `Biostrings` 패키지에 포함된 내장 데이터로서 `yeast`의 첫 번째 염색체 정보를 담고 있습니다.

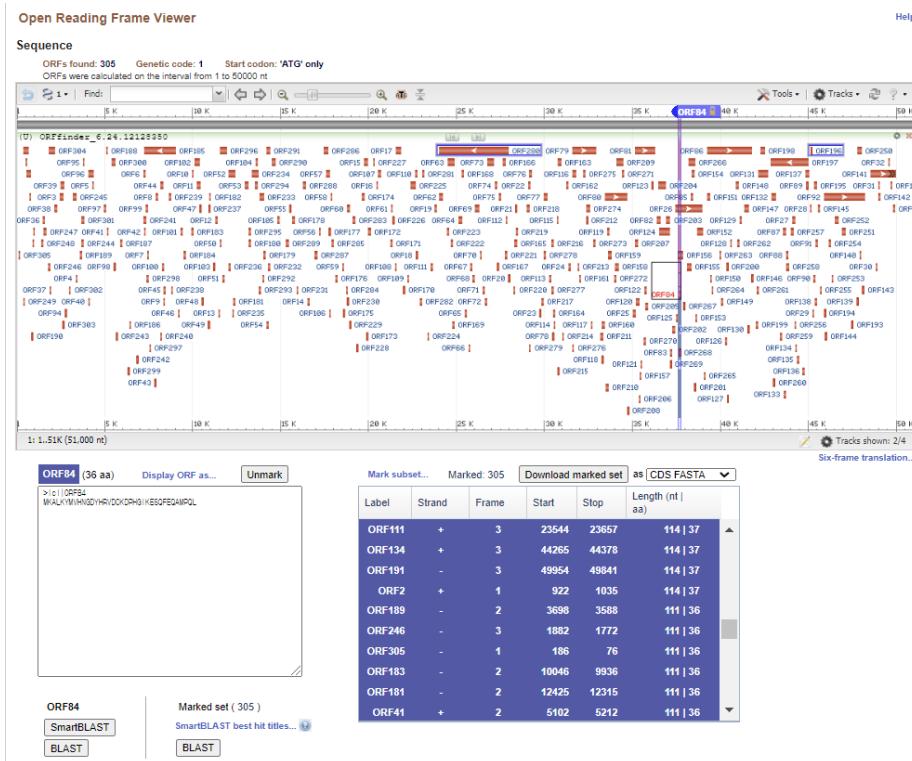
```
data(yeastSEQCHR1) #Biostrings
yeast1 <- DNAString(yeastSEQCHR1)

oligonucleotideFrequency(yeast1, 3)
dinucleotideFrequency(yeast1)
trinucleotideFrequency(yeast1)

tri <- trinucleotideFrequency(yeast1, as.array=TRUE)
tri
```

아미노산 정보를 얻기 위해서 ORF를 찾아보겠습니다. `yeast`의 첫 번째 염색체에 대한 정보는 `annotation`이 되어 있지만 학습을 위해 툴을 사용하겠습니다. 이미

많은 종류의 ORF 탐색 툴이 나와있지만 본 강의에서는 NCBI에서 제공하는 orffinder를 사용하도록 하겠습니다.



```
my_ORFs <- readDNAStringSet("yeast1orf.cds")
hist(nchar(my_ORFs), br=100)
codon_usage <- trinucleotideFrequency(my_ORFs, step=3)
global_codon_usage <- trinucleotideFrequency(my_ORFs, step=3, simplify.as="collapsed")

colSums(codon_usage) == global_codon_usage
names(global_codon_usage) <- GENETIC_CODE[names(global_codon_usage)]
codonusage2 <- split(global_codon_usage, names(global_codon_usage))
global_codon_usage2 <- sapply(codonusage2, sum)
```

yeast 첫 번째 염색체에 대한 정보는 bioconductor annotationData OrgDb 또는 bioconductor annotationData TxDb 에서 찾아볼 수 있습니다.

```
#BiocManager::install("org.Sc.sgd.db")
library(org.Sc.sgd.db)
class(org.Sc.sgd.db)
?org.Sc.sgd.db
ls("package:org.Sc.sgd.db")
columns(org.Sc.sgd.db)
```

```
mykeys <- keys(org.Sc.sgd.db, keytype = "ENTREZID")[1:10]
AnnotationDbi::select(org.Sc.sgd.db,
 keys=mykeys,
 columns = c("ORF", "DESCRIPTION"),
 keytype="ENTREZID")
```

### TxDb

```
BiocManager::install("TxDb.Scerevisiae.UCSC.sacCer3.sgdGene")
library(TxDb.Scerevisiae.UCSC.sacCer3.sgdGene)
class(TxDb.Scerevisiae.UCSC.sacCer3.sgdGene)
columns(TxDb.Scerevisiae.UCSC.sacCer3.sgdGene)
methods(class=class(TxDb.Scerevisiae.UCSC.sacCer3.sgdGene))
ygenes <- genes(TxDb.Scerevisiae.UCSC.sacCer3.sgdGene)

library(tidyverse)

mydat <- global_codon_usage2 %>%
 data.frame %>%
 rownames_to_column %>%
 rename(codon = "rowname", freq = ".")

ggplot(mydat, aes(x=codon, y=freq)) +
 geom_bar(stat="identity")

mydat
AMINO_ACID_CODE[mydat$codon]
```

### Exercises

AMINO\_ACID\_CODE를 이용해서 위 그래프의 1약자를 3약자로 변환, 라벨을 세로로 90도 회전, y축 라벨 “Frequency”, x축 라벨 “Amino acid code”, theme 옵션 “theme\_bw” 등을 적용하여 다시 그림을 그리시오 (revisit ggplot2)

## 11.4 Pattern matching

Biostrings 패키지에는 하나의 subject 서열에 특정 pattern이 존재하는지 탐색하는 matchPattern함수를 제공합니다. 만약 여러개의 subject 서열에서 하나의 pattern을 찾을 경우에는 vmatchPattern함수를 사용하고 하나의 subject 서열에 여러개의 pattern을 찾는 경우에는 matchPDict 함수를 사용합니다.

```
length(coi)
hits <- matchPattern("ATG", yeast1, min.mismatch=0, max.mismatch=0)
hits
class(hits)
```

```
methods(class="XStringViews")
ranges(hits)

hits <- vmatchPattern("ATG", my_ORFs, min.mismatch=0, max.mismatch=0)
stack(hits)
```

---

이 저작물은 크리에이티브 커먼즈 저작자표시-비영리-변경금지 4.0 국제 라이선스에 따라 이용할 수 있습니다.



# Chapter 12

## Tools for sequences

### 12.1 Sequences from NCBI

전세계 연구자들이 서열 데이터를 분석하는데 가장 많이 이용하는 사이트 중 하나가 NCBI이며 따라서 NCBI에서는 연구자들이 데이터베이스에 접근하기 위한 편리한 방법을 제공하고 있고 그 중 하나가 Entrez입니다.

R에서도 Entrez 기능을 도입한 package들이 제공되고 있으며 그 중 하나가 `rentrez`입니다. <https://www.ncbi.nlm.nih.gov/books/NBK25500/> 이 곳의 Downloading Full Records 를 참고하시면 좋습니다. Entrez는 대략적으로 다음 9개의 유ти리티를 제공합니다.

- EInfo (database statistics)
- ESearch (text searches)
- EPost (UID uploads)
- ESummary (document summary downloads)
- EFetch (data record downloads)
- ELink (Entrez links)
- EGQuery (global query)
- ESpell (spelling suggestions)
- ECitMatch (batch citation searching in PubMed)

이 중 ESearch, EPost, ESummary, EFetch 등이 많이 사용하는 유ти이며 정보를 다운로드 받을 경우는 EFetch 를 주로 사용하게 됩니다. `rentrez` 는 위와 같은 NCBI Eutils API를 활용하여 R 환경에서 탐색이나 다운로드 등 NCBI 데이터베이스와 상호작용이 용이하도록 만들어 놓은 tool입니다. `rentrez` landing page `entrez_dbs` 명령은 NCBI에서 제공하는 데이터베이스의 리스트를 볼 수 있으며 특정 DB에 대한 설명은 `entrez_db_summary`를 사용하면 되겠습니다. `entrez_search`는 각종 키워드를 사용한 검색 기능을 제공합니다.

```
library(rentrez)
require(Biostrings)
```

```

entrez_dbs()
entrez_db_summary("nuccore")

covid_paper <- entrez_search(db="pubmed", term="covid19")
covid_paper$ids

names(covid_paper)
covid_paper$ids

covid_link <- entrez_link(db="all", id=covid_paper$ids, dbfrom="pubmed")
names(covid_link)
names(covid_link$links)
head(covid_link$links$pubmed_pubmed)

```

`entrez_search`에서 검색어를 입력하는 방식은 이곳을 참고하세요. 검색으로 찾아진 특정 오브젝트(객체)에 대한 내용은 `entrez_summary` 함수를 사용하여 조회할 수 있으며 `extract_from_esummary`로 조회된 아이템들에 대한 정보를 추출할 수 있습니다. 특정 id에 대한 서열 등 다양한 타입의 데이터를 실제로 다운로드 받는 기능은 `entrez_fetch` 함수가 제공하고 있습니다. `entrez_fetch` 함수의 `rettype` 옵션에서 지원하는 데이터 타입을 다운로드 받을 수 있으며 `rettype` (return type)의 자세한 정보는 Eutils table 또는 NCBI Eutils 페이지를 참고하시기 바랍니다.

```

popset database is a collection of related DNA sequences derived from population
katipo_search <- entrez_search(db="popset", term="Latrodectus katipo[Organism]")
katipo_search$ids

katipo_summs <- entrez_summary(db="popset", id=katipo_search$ids)
names(katipo_summs)
katipo_summs$`41350664`
class(katipo_summs)
methods(class="esummary_list")

titles <- extract_from_esummary(katipo_summs, "title")
unname(titles)

print(katipo_summs)
katipo_summs$`1790798044`$gi

COI_ids <- katipo_search$ids[c(2,6)]
trnL_ids <- katipo_search$ids[4]
COI <- entrez_fetch(db="popset", id=COI_ids, rettype="fasta")
trnL <- entrez_fetch(db="popset", id=trnL_ids, rettype="fasta")

```

```

write(COI, "COI.fasta")
write(trnL, "trnL.fasta")

#library(Biostrings)
coi <- readDNAStringSet("COI.fasta")
trnL <- readDNAStringSet("trnL.fasta")

```

### Exercises

뎅기바이러스 서열 4종에 대한 NCBI의 accession 번호가 다음과 같음 NC\_001477, NC\_001474, NC\_001475, NC\_002640 해당 DNA 서열을 fasta 형식으로 nuccore 데이터베이스에서 다운로드 하시오. (참고로 strwrap 함수 사용법을 익혀두면 좋습니다)

### Exercises

1. popset 데이터베이스에서 “Covid-19” 단어가 들어간 유전자 40개를 찾고 (entrez\_search에서 retmax=40 옵션 사용) 이들의 요약 정보 중 title 속성을 출력하시오 (entrez\_summary와 extract\_from\_esummary 함수 사용).
2. 위 결과에서 찾아진 유전자들 각각이 몇 개의 서열 샘플에 (population) 대해서 연구된 것인지 각각의 서열을 fasta 형태로 다운로드 받고 샘플의 개수에 대한 barplot을 그리시오
  - summary\_record 결과를 받아서 extract\_from\_esummary로 title을 추출 후 data.frame으로 변환
  - tidyverse의 rownames\_to\_column() 함수로 uid 정보 변수로 변환, my-data 이름으로 저장
  - entrez\_fetch 함수로 모든 uid에 대한 샘플 서열 fasta 파일 다운로드 후 파일 저장 (write 함수 사용)
  - readDNAStringSet 함수로 읽은 후 앞서 title 정보 비교를 통해서 앞서 my-data 와 병합
  - 각 uid 별로 몇 개의 서열 샘플이 있는지 정보를 추출 후 barplot 그리기

### Exercises

Comparative sequence analysis of SARS-CoV-2 suggests its high transmissibility and pathogenicity 논문을 참고하여 COVID-19 서열의 NCBI accession 번호를 찾고 nuccore 데이터베이스에서 fasta 포맷과 genbank 포맷의 정보를 다운로드 하시오. (데이터는 “covid\_table.csv” 파일에 저장되어 있음)

```

covid <- data.frame(
 species = c(rep("Human", 7), c("Civet", "Civet"), rep("Bat", 3), "Pangolin"),
 coronavirus = c("SARS-CoV-2", "SARS-CoV-2", "SARS-CoV-1", "SARS-CoV-1", "SARS-CoV-1", "H-CoV-OC43",
 isolate = c("Wuhan Hu-1", "USA-WA-1", "Urbani", "Tor2", "GD03T10013", "UK/London", "EMC-2012", "EMC-2012",
 year = c("2020", "2020", "2002", "2002", "2003", "2011", "2011", "2003", "2004", "2015", "2013",
 gbacc = c("NC_045512.2", "MN985325.1", "AY278741.1", "AY274119.3", "AY525636.1", "KU131570.1", "M
 write.csv(covid, file = "covid_table.csv", quote = F, row.names=F)

```

species	coronavirus	isolate	year	gbacc
Human	SARS-CoV-2	Wuhan Hu-1	2020	NC_045512.2
Human	SARS-CoV-2	USA-WA-1	2020	MN985325.1
Human	SARS-CoV-1	Urbani	2002	AY278741.1
Human	SARS-CoV-1	Tor2	2002	AY274119.3
Human	SARS-CoV-1	GD03T10013	2003	AY525636.1
Human	H-CoV-OC43	UK/London	2011	KU131570.1
Human	MERS-CoV	EMC-2012	2011	NC_019843.3
Civet	SARS-CoV	SZ3	2003	AY304486.1
Civet	SARS-CoV	Civet007	2004	AY572034.1
Bat	SL-CoV	ZXC21	2015	MG772934.1
Bat	SL-CoV	WIV16	2013	KT444582.1
Bat	SL-CoV	RaTG13	2013	MN996532.1
Pangolin	SL-CoV	MP789	2020	MT084071.1

```
require(kableExtra)
#> Loading required package: kableExtra

download.file(url = "https://raw.githubusercontent.com/greendaygh/kribbr2022/main/covid19.csv")
covid19 <- read.csv("covid_table2.csv")
kable_classic(kable(covid19))
```

## 12.2 Align two sequences

서열 정렬은 match, mismatch, penalty 등의 scoring rule을 기반으로 최적의 score를 갖는 서열 정렬을 찾는 알고리즘입니다. Biostrings 패키지에는 두 개의 서열에 대해 local, global alignment를 수행할 수 있는 pairwiseAlignment 함수를 제공하고 있습니다. 첫 번째 파라메터는 pattern이며 두 번째는 subject로서 pattern은 query로서 해당 서열이 subject (target)에 있는지를 보는 것과 같습니다.

```
covid19seq
?pairwiseAlignment
aln <- pairwiseAlignment(covid19seq[1], covid19seq[2])
class(aln)
methods(class="PairwiseAlignmentsSingleSubject")
?PairwiseAlignmentsSingleSubject
```

위에서 서열 정렬이 된 결과를 DNAString class의 변수에 저장한 후 해당 class에서 제공하는 다양한 함수를 동일하게 적용할 수 있습니다. 또한 writePairwiseAlignments 함수는 두 서열의 비교 결과를 보기 좋게 출력해주는

기능을 수행합니다. `summary` 함수를 사용하면 염기가 다른 곳의 위치를 출력해주며 `consensusString`은 50% 초과하는 서열에 대한 문자열을 출력해 줍니다. 이 외에도 `score`, `consensusMatrix` 등 다양한 `help` 함수들이 있습니다.

```
alnseqs <- c(alignedPattern(aln), alignedSubject(aln))
class(alnseqs)

aln

writePairwiseAlignments(aln, block.width=50)
writePairwiseAlignments(aln, file="covidalign.txt", block.width=30)

summary(aln)

consensusMatrix(aln)[1:4,1:10]
consensusString(aln)

score(aln)
```

참고로 아래와 같이 별도의 정보 테이블을 만들고 서열의 이름은 간단히 `id`만 사용해서 분석하는 것이 더 효율적입니다. 문자열을 다루는 코드를 익혀두시기 바랍니다.

```
names(covid19seq)

ids <- strsplit(names(covid19seq), split=" ") %>%
 lapply(function(x){x[1]}) %>%
 unlist

titles <- strsplit(names(covid19seq), split=" ") %>%
 lapply(function(x){
 paste(x[-1], collapse=" ")
 }) %>%
 unlist

covid19info <- data.frame(ids, titles)
names(covid19seq) <- covid19info$ids

aln <- pairwiseAlignment(covid19seq[1], covid19seq[2])
writePairwiseAlignments(aln, block.width=50)
```

## 12.3 Multiple sequence alignment

Multiple sequence alignment(MSA) tool은 서열 데이터의 양과 계산량의 문제로 linux 기반 commandline 프로그램들이 많습니다. 대표적으로 CLUSTAL-Omega, MUSCLE. window 기반 환경에서는 docker 등을 활용해서 관련 분석을 수행할 수

있습니다.

### 12.3.1 msa

msa 패키지는 위 Clustal나 MUSCLE 등의 프로그램에 대한 R 인터페이스를 제공하며 Linux 뿐만 아니라 모든 운영체제에서 수행될 수 있도록 만들어 둔 패키지입니다.

```
library(msa)

alnmsa <- msa(subcovid19seq)
class(alnmsa)
alnmsa

print(alnmsa, show="complete")
?msaPrettyPrint
msaPrettyPrint(alnmsa, output="pdf", showNames="none", showLogo="top", askForOverwrite=
myconseq <- msaConsensusSequence(alnmsa)
```

참고로 정렬된 서열 출력물에 표시되는 ?는 판단하기 모호한 위치를 나타냅니다. 예를 들어 A와 T 가 5:5로 나타난 위치는 ?로 표시됩니다.

MsaDNAMultipleAlignment class는 Biosting 패키지의 DNAString class를 상속받은 클래스로서 다음과 같이 DNAStringSet class로 변환해서 분석도 가능합니다.

```
alnseq <- DNAStringSet(aln)
class(alnseq)
alnseq
myconseq2 <- ConsensusSequence(alnseq)
```

위 alignment 결과에서 관심있는 특정 위치만을 선택해서 임의의 분석을 수행하고 싶은 경우 마스킹 함수를 사용할 수 있습니다. 이 기능 역시 MsaDNAMultipleAlignment class는 Biostring 패키지의 DNAString class 모두에 적용이 가능합니다. IRanges 함수는 뒤에서 더 상세히 설명하도록 하겠습니다.

```
colM <- IRanges(start=1, end=300)
colmask(alnmsa) <- colM
alnmsa
msaConsensusSequence(alnmsa)
alphabetFrequency(alnmsa)

alphabetFrequency(unmasked(alnmsa))
```

### 12.3.2 DECIIPHER

DECIIPHER 패키지는 서열 alignment나 primer design 등을 수행할 수 있는 패키지로 다음과 같이 별도 메모리에 서열을 저장하고 빠르게 alignment를 수행할

수 있어서 중소 규모의 서열에 대한 분석으로 유용하게 사용될 수 있습니다. 다음은 관련 서열을 SQLite 데이터베이스에 저장하고 그 내용을 쉽게 볼 수 있는 기능들입니다. dbDisconnect 함수를 실행하면 모든 저장된 데이터가 사라지며 매모리는 다시 사용할 수 있게 됩니다.

```
library(DECIPHER)

dbConn <- dbConnect(SQLite(), ":memory:")
Seqs2DB(covid19seq, "XStringSet", dbConn, "covid19")
BrowseDB(dbConn)

l <- IdLengths(dbConn)
Add2DB(l, dbConn)
BrowseDB(dbConn)

Seqs2DB(trnl, "XStringSet", dbConn, identifier = "trnl")
BrowseDB(dbConn)

Seqs2DB(coi, "XStringSet", dbConn, "coi")
BrowseDB(dbConn)

l <- IdLengths(dbConn, identifier = "trnl")
Add2DB(l, dbConn)
BrowseDB(dbConn)

dbDisconnect(dbConn)
```

다중서열을 정렬하는 AlignSeqs 함수와 BrowseSeqs 함수를 활용해서 html 형태로 정렬된 서열을 볼 수 있습니다. 특히 patterns 옵션을 사용해서 원하는 서열이 존재하는지도 확인할 수 있습니다.

```
extract sequences
covid19seq2 <- SearchDB(dbConn, identifier = "covid19")
subcovid19seq <- subseq(covid19seq2, 1, 2000)
aln <- AlignSeqs(subcovid19seq)
aln
class(aln)

BrowseSeqs(aln, colWidth = 100)
BrowseSeqs(aln, colWidth = 100, patterns=DNAStringSet(c("ACTG", "CSC")))
BrowseSeqs(aln, colWidth = 100, patterns="-", colors="black")
?BrowseSeqs

aln2 <- AlignTranslation(subcovid19seq, type="AAStringSet")
BrowseSeqs(aln2, colWidth = 100)
BrowseSeqs(aln2, colWidth = 100, highlight=1)
```

AlignSeqs 함수의 결과가 DNAStringSet 클래스이므로 앞서 수행한 마스킹 등의 기능을 동일하게 적용 가능합니다.

DECIPHER 패키지의 DigestDNA 함수를 이용하면 enzyme digestion을 시뮬레이션할 수 있는 기능을 활용할 수 있습니다. 단 숫자 등 필요없는 문자를 제거하기 위해서 stringr 패키지를 사용합니다.

```
data(RESTRICTION_ENZYMES)
RESTRICTION_ENZYMES
rsite <- RESTRICTION_ENZYMES["BsmBI"]
rsite <- RESTRICTION_ENZYMES["BsaI"]

d <- DigestDNA(rsite, covid19seq2[1])
unlist(d)
#writeXStringSet(unlist(d), file="covid19bsmbi.fasta")
pos <- DigestDNA(rsite, covid19seq2[1], type="positions")
unlist(pos)

library(stringr)
library(stringi)

BrowseSeqs(covid19seq2[1], colWidth = 100, patterns=rsite)
sub("(^[[alpha:]]*).*", "", rsite)

rsite2 <- paste(str_extract_all(rsite, "[[A-Z]]", simplify = T), collapse="")
rsite3 <- as.character(reverseComplement(DNAString(rsite2)))
BrowseSeqs(covid19seq2[1], colWidth = 100, patterns=c(rsite2, rsite3))
```

### Exercises

1. 앞서 다운로드 받은 *Latrodectus katipo* 서열 데이터를 읽어들이고 100bp 단위로 출력하시오
2. MSA 를 수행하고 정렬된 결과를 100bp 단위로 출력하시오
3. ConsensusSequence 함수를 이용하여 정렬된 결과로부터 consensus 서열을 추출하시오

참고로 염기와 아미노산의 Standard Ambiguity Codes는 각각 다음과 같습니다.

<b>Code</b>	<b>Represents</b>	<b>Complement</b>
A	Adenine	T
G	Guanine	C
C	Cytosine	G
T	Thymine	A
Y	Pyrimidine (C or T)	R
R	Purine (A or G)	Y
W	weak (A or T)	W
S	strong (G or C)	S
K	keto (T or G)	M
M	amino (C or A)	K
D	A, G, T (not C)	H
V	A, C, G (not T)	B
H	A, C, T (not G)	D
B	C, G, T (not A)	V
X/N	any base	X/N
-	Gap	-

- DNA ambiguity code

Nucleotide	Symbol	3-Let	Amino Acid	IUB
(Adenosine) A	A	Ala	Alanine	GCX
C or G or T/U	B	Asx	Aspartate or Asparagine	RAY
(Cytidine) C	C	Cys	Cysteine	UGY
A or G or T/U	D	Asp	Aspartate	GAY
-	E	Glu	Glutamate	GAR
-	F	Phe	Phenylalanine	UIY
(Guanosine) G	G	Gly	Glycine	GGX
A or C or T/U	H	His	Histidine	CAY
(Inosine) I	I	Ile	Isoleucine	AUH
-	J	-	-	-
G or T/U	K	Lys	Lysine	AAR
-	L	Leu	Leucine	UUR,CUX,YUR
A or C	M	Met	Methionine	AUG
unknown base	N	Asn	Asparagine	AAY
-	O	-	-	-
-	P	Pro	Proline	CCX
-	Q	Gln	Glutamine	CAR
(Purine) A or G	R	Arg	Arginine	CGX,AGR,MGR
C or G	S	Ser	Serine	UCX,AGY
(Thymidine) T	T	Thr	Threonine	ACX
(Uridine) U	U	-	-	-
A or C or G	V	Val	Valine	GUX
A or T/U	W	Trp	Tryptophan	UGG
unknown base	X	unknown amino acid		XXX
(Pyrimidine) C or T/U	Y	Tyr	Tyrosine	UAY
-	Z	Glx	Glutamate or Glutamine	SAR
no base (deletion/gap)	.	no amino acid (deletion/gap)	-	-
-	*	End	terminator	UAR,URA

- Amino acid ambiguity code

## 12.4 Phylogenetic trees with clustering

다중서열비교 결과는 계통학에서 널리 쓰이며 msa나 DECIPHER 패키지에서 얻어진 결과를 계통학의 tree 형태로 가시화할 수 있습니다. tree 형태의 가시화를 위해 다양한 포맷의 파일이 개발되었고 treeio 패키지는 이를 다양한 포맷의 파일을 쉽게 변환하기 위해 만들어진 패키지입니다. 다음은 Newick tree 포맷의 예입니다.

```
((t2:0.04,t1:0.34):0.89,(t5:0.37,(t4:0.03,t3:0.67):0.9):0.59);
```

가장 널리 사용되는 포맷은 phylo 형태로서 이는 ape라는 phylogenetic 분석의 대표적인 패키지에서 처음 제안되어 사용되고 있습니다. 최근 ggplot 형태의

ggtree, reference이 개발되어 계통도를 좀더 세밀하게 그릴 수 있으며 ggtree는 phylo 형태의 포맷을 주로 사용합니다.

```
if (!requireNamespace("BiocManager", quietly = TRUE))
 install.packages("BiocManager")

BiocManager::install("ggtree")
#BiocManager::install("treeio")
```

Phylogenetic tree는 서열간의 유사도(거리)를 기반으로 분석할 수 있습니다. 앞서 사용한 mas나 DECIPHER 패키지로 얻어진 MSA 결과는 모두 Biostrings 패키지의 XStringSet (DNAStringSet, AAStringSet 등) 클래스입니다. 따라서 XStringSet의 거리를 계산해주는 Biostrings::stringDist 함수나 DECIPHER::DistanceMatrix가 사용될 수 있습니다. 참고로 phylo class는 as.tibble 함수를 이용해서 테이블 형태로 변환, 활용할 수 있습니다.

```
library(ape)
library(ggtree)

alnmsa <- msa(subcovid19seq)
mydist <- stringDist(DNAStringSet(alnmsa))
clust <- hclust(mydist)
class(clust)

mytree <- as.phylo(clust)
ggtree(mytree, layout="circular") +
 geom_tiplab()

as.tibble(mytree)
```

DECIPHER 패키지에는 XStringSet 서열의 거리를 계산해주는 DistanceMatrix 함수가 있습니다. 이 함수를 이용하면 역시 같은 패키지에서 제공하는 IdClusters 함수를 이용해서 유사한 서열끼리 묶어주는 tree 를 만들 수 있습니다. dendrogram는 str 함수 활용이 가능합니다.

```
dm <- DistanceMatrix(subcovid19seq)
class(dm)

clust <- IdClusters(dm, cutoff=10, method="NJ", showPlot=F, type="dendrogram")
class(clust)
methods(class="dendrogram")
plot(clust)
str(clust)
```

dendrogram class는 hclust를 거쳐 phylo class 형태로 변환 후 ggtree 패키지를 활용할 수 있습니다.

```
convert to dendrogram -> hclust -> phylo
cl <- as.hclust(clust)
py <- as.phylo(cl)
class(py)
ggtree(py)
as.tibble(py)
```

ggtree를 활용하면 다양한 레이아웃을 활용할 수 있고 레이아웃에 대한 정보는 Layouts of a phylogenetic tree 이 곳을 참고하시면 되겠습니다.

```
tree <- rtree(n = 20)
ggtree(tree)

ggplot(tree) +
 geom_tree() +
 theme_tree()

ggtree(tree, branch.length="none")

ggtree(tree, layout="circular") +
 geom_tiplab(color="firebrick")

ggtree(tree, layout="circular") +
 geom_tiplab(size=3, aes(angle=angle))

ggtree(tree, layout="circular", branch.length="none") +
 geom_tiplab(size=3, aes(angle=angle))

ggtree(tree) +
 theme_tree2()

ggtree(tree, layout="circular") +
 geom_tiplab() +
 theme(plot.margin = unit(c(100,30,100,30), "mm"))

ggsave("myphylo.pdf", width = 50, height = 50, units = "cm", limitsize = FALSE)
```

covid19 example

```
covid19seq <- readDNAStringSet("covid19.fasta", format="fasta")[1:6]
subcovid19seq <- subseq(covid19seq, 1, 2000)
names(subcovid19seq) <- sapply(strsplit(names(subcovid19seq), " "), function(x){return

alnmsa <- msa(subcovid19seq)
mydist <- stringDist(DNAStringSet(alnmsa))
clust <- hclust(mydist)
```

```
mytree <- as.phylo(clust)

ggtree(mytree, layout="circular") +
 geom_tiplab(color="firebrick", size=3)

ggtree(mytree) +
 geom_tiplab(color="firebrick", size=3) +
 theme_tree2(scale=0.1)
```

특정 그룹을 highlight 하기 위해서 `geom_hilight` 함수를 사용합니다.

```
ggtree(tree) +
 theme_tree2() +
 geom_tiplab() +
 geom_hilight(node=20, fill="steelblue", alpha=.4)
```

노드를 알아보기 위해서 `tidytree`를 사용할 수 있습니다.

```
as.tibble(tree)

d <- data.frame(node=c(20, 20, 22), type=c("T1", "T1", "T2"))

ggtree(tree) +
 theme_tree2() +
 geom_tiplab() +
 geom_hilight(d, aes(node=node, fill=type), alpha=.4) +
 scale_fill_manual(values=c("steelblue", "darkgreen"))
```

### 12.4.1 Facet Utilities

`geom_facet` 와 `facet_widths` 를 사용하면 추가 판넬에 tree를 그릴 수 있습니다.

```
tree <- rtree(30)

p <- ggtree(tree, branch.length = "none") +
 geom_tiplab() +
 theme(legend.position='none')

a <- runif(30, 0,1)
b <- 1 - a
df <- data.frame(tree$tip.label, a, b)
df2 <- pivot_longer(df, -tree.tip.label)

p2 <- p + geom_facet(panel = 'bar', data = df2, geom = geom_bar,
 mapping = aes(x = value, fill = as.factor(name)),
 orientation = 'y', width = 0.8, stat='identity') +
```

```
xlim_tree(9)

facet_widths(p2, widths = c(1, 2))
```

## 12.5 BLAST result analysis

BLAST를 로컬컴퓨터에 설치하거나 docker를 이용해서 활용할 수 있으나 본 강의에서는 직접 BLAST를 수행하는 대신 NCBI에서 실행한 BLAST 출력물을 분석하는 경우에 대해서 설명을 진행하겠습니다. 예시로는 PET를 분해하는 단백질로 알려진 IsPETase의 서열과 유사한 서열을 찾아서 분석해 보겠습니다. IsPETase 정보는 다음과 같습니다 Genes encoding *I. sakaiensis* 201-F6 IsPETase (WT PETase) (accession number: A0A0K8P6T7).

### Exercises

- 1) NCBI BLAST 사이트에서 A0A0K8P6T7 단백질에 대한 BLASTp를 수행하시오 (db: nr)
- 2) 결과물을 (100개) 다운로드 하고 (fasta와 hit table 각 1개 파일씩) 작업디렉토리로 복사하시오
- 3) fasta 와 hit 데이터를 각각 읽어들이시오
- 4) 100개의 서열을 DECIPHER 패키지를 활용해서 정렬하고 100bp 단위로 출력해보시오
- 5) align된 결과에서 consensus 서열을 추출하고 각 위치별로 어떤 아미노산이 많은지 bar 그래프를 그려보시오
- 6) align된 결과를 ggtrree 패키지를 사용해서 phylogenetic를 그리시오

---

이 저작물은 크리에이티브 커먼즈 저작자표시-비영리-변경금지 4.0 국제 라이선스에 따라 이용할 수 있습니다.

# Chapter 13

## Tools for genome analysis

### 13.1 genbank file

#### Exercises

1. NC\_045512.2는 우한에서 발생한 코로나바이러스의 accession number임. entrez\_fetch 함수를 사용하여 nuccore 데이터베이스에서 genbank 정보를 다운로드 받으시오
2. 받은 텍스트를 covid19wuhan.gb라는 파일로 저장하시오

genbank 파일은 DNA 및 단백질 서열을 저장하는데 사용되는 서열 파일 포맷으로서 하나 이상의 시퀀스에 대한 정보와, 주석, 특정 서열 구간의 feature 정보와 메타 데이터도 포함합니다. NCBI에서 개발했으며 표준 DNA 및 단백질 서열 파일 형식으로 공공 데이터베이스 등 널리 사용되고 있습니다. R의 genbankr 패키지는 genbank 타입의 데이터를 읽는 readGenBank 함수를 제공합니다.

```
require(genbankr)

covid19 <- readGenBank("covid19wuhan.gb")
covid19
methods(class="GenBankRecord")
cds(covid19)
exons(covid19)

covid19seq <- getSeq(covid19)
```

### 13.2 IRanges

유전체 데이터의 대부분을 차지하는 정보는 전체 지놈 서열 중 어디서 어디까지가 유전자 또는 coding sequence이고 그 번역된 정보가 무엇인지 설명하는 정보

입니다. 즉, 일련의 feature에 대한 위치와 특성 정보를 분석하는 것이 효율적인 지놈을 분석하기 위해 필수입니다. `bioconductor` 에서는 이러한 유전체 정보를 효율적으로 분석하고 가시화하기 위한 방법들이 다양하게 개발되어 왔으며 `IRanges` 와 `GenomicRanges` 라는 패키지가 대표적으로 사용될 수 있습니다.

`IRanges`는 간격을 나타내는 임의의 숫자 세트이며 지놈상에 위치한 특정 feature들의 간격이나 통계적 수치들을 효율적으로 나타내기 위해서 만들어진 패키지입니다 (?). 임의의 feature에 대한 시작, 끝, 넓이를 나타내는 숫자들이 리스트로 이루어져 있습니다.

```
library(IRanges)

ir <- IRanges(start = c(1,3,5), end = c(3,5,7))
ir

ir <- IRanges(start = 1:10, width = 10:1)
ir
class(ir)
methods(class="IRanges")
?IRanges
```

`IRange` 는 `Rle` (run-length encoding format, 런 렌스 부호화) class 를 사용하며 일종의 압축 방법입니다. 예를 들어 GATTGCCCCCTAG 라는 서열이 있다고 하면 이를 그대로 text 파일에 저장하지 않고 GAT2GC6TAG 라고 표현함으로써 용량을 줄이는 압축의 기능을 합니다. `GenomicRange`는 이러한 `Rle` 개념을 사용하기 위해서 `Rle`라는 기본 함수를 사용합니다.

```
library(IRanges)

x <- "GATTGCCCCCTAG"
y <- unlist(strsplit(x, split=""))
yrle <- Rle(y)
yrle

runLength(yrle)
runValue(yrle)
nrun(yrle)

x <- Rle(values = c(1:3), lengths = c(1:3))
x
class(x)
#methods(class="Rle")

convert Rle to IRanges
xrange <- IRanges(start(x), end(x))
xrange
```

IRange 생성과는 반대로 IRange 객체로부터 몇몇 함수를 사용하여 정보를 추출할 수 있습니다.

```
ir <- IRanges(start = c(1,3), end = c(4,5))
ir

start(ir)
end(ir)
width(ir)
disjointBins(ir)

ir <- IRanges(start = c(1,3,6), end = c(4,5,7))
ir
bins <- disjointBins(ir)
bins
```

이러한 정보를 가시화하는 가장 간단한 방법은 ggbio라는 패키지를 사용하는 것입니다.

```
library(ggbio)

autoplot(ir)

autoplot(ir) +
 theme_bw()

autoplot(ir, aes(fill=width)) +
 theme_bw()
```

특히 disjoin과 reduce 함수는 overlap 되어 있는 구간의 분석을 수행하는데 유용하게 활용 됩니다.

```
ir2 <- disjoin(ir)
ir2
autoplot(ir)
autoplot(ir2)

ir3 <- IRanges::reduce(ir)
ir3
autoplot(ir3)
```

## Exercises

- 1) 구간의 길이가 각각 100, 15, 30, 45 인 IRange 객체를 만드시오
- 2) 1부터 100까지의 전체 구간에서 시작 위치가 각각 1, 15, 30, 60 이면서 길이가 20 인 IRange 객체를 만드시오

### 13.3 GenomicRanges

GenomicRanges는 지놈상의 위치정보와 Bioconductor에서 제공하는 다양한 high-throughput 정보들을 같이 표현하기 위해서 만들어진 패키지입니다.

GRanges 함수를 이용해서 생성할 수 있으며 `browseVignettes("GenomicRanges")`나 `methods()` 함수를 이용해서 관련된 기능을 찾아서 사용할 수 있습니다.

```
library(GenomicRanges)

gr <- GRanges(
 seqnames = "chr1",
 ranges = IRanges(1:10, 101:110),
 strand = rep("+", 10)
)
gr
class(gr)
```

다양한 함수 사용법을 보여줍니다.

```
gr <- GRanges(
 seqnames = Rle(c("chr1", "chr2", "chr1", "chr3"), c(1, 3, 2, 4)),
 ranges = IRanges(101:110, end = 111:120, names = head(letters, 10)),
 strand = Rle(strand(c("-", "+", "*", "+", "-")), c(1, 2, 2, 3, 2)),
 score = 1:10,
 GC = seq(1, 0, length=10))
gr

seqnames(gr)
ranges(gr)
strand(gr)

granges(gr)
mcols(gr) #meta data
seqlengths(gr)

seqlengths(gr) <- c(249250621, 243199373, 198022430)
names(gr)
```

ggbio의 `autoplot`을 사용하여 IRange와 같이 가시화 할 수 있으며 `split` 함수를 사용하면 지정된 규칙에 따라 Grange를 나눌 수 있습니다.

```
autoplot(gr)
sp <- split(gr, rep(1:2, each=5))
```

#### Exercises

위 결과에서 chromosome 별로 항목을 나눈 Grange list를 만드시오

### Exercises

위에서 받은 NC\_045512.2는 우한에서 발생한 코로나바이러스 서열에 대한 CDS 추출, 서열 비교, 가시화 등을 수행하시오

## 13.4 plyranges

위 GenomicRanges 데이터를 dplyr 형태로 좀 더 쉽게 다루기 위한 패키지가 plyranges입니다.

```
library(plyranges)

covid19cds
gcr <- rowSums(letterFrequency(cdsseqs, c("G", "C"), as.prob=T))

covid19cds %>%
 dplyr::select(gene, product) %>%
 mutate(gc = gcr) # %>%
 #filter(grepl(pattern = "ORF", gene))
```

### Exercises

위에서 계산된 GC 비율로 bar 그래프를 그리되 product를 라벨로 지정하여 그리시오

## 13.5 Visualization

```
require(ggplot2)
require(gggenes)
require(plyranges)

targetr <- covid19cds

mcols(targetr)$gene

plyranges::summarise(targetr)

plotdf1 <- data.frame(molecule=seqnames(targetr),
 gene=mcols(targetr)$gene,
 start=start(targetr),
 end=end(targetr),
 strand=case_when(
 as.vector(strand(targetr))=="+" ~ TRUE,
 as.vector(strand(targetr))== "-" ~ FALSE
))
```

```
)
```

```
ggsave("plotdf1.pdf", width = 10, height = 10)
ggplot(plotdf1, aes(xmin = start, xmax = end, y = molecule, label=gene, fill = gene, fontface = gene))
 geom_gene_arrow(
 arrowhead_height = unit(3, "mm"),
 arrowhead_width = unit(1, "mm")
 arrowhead_height = grid::unit(12, "mm"),
 arrowhead_width = grid::unit(6, "mm"),
 arrow_body_height = grid::unit(12, "mm")
) +
 geom_gene_label(align = "left", height = grid::unit(19, "mm"), grow=TRUE) +
 theme_genes() +
 theme(legend.position="none")
```

다음은 대장균 지놈 NC\_000913.3 gb 파일을 다운로드 받고 지놈 전체를 가시화하는 코드입니다.

```
library(rentrez)
library(genbankr)

tmps <- entrez_fetch("nuccore", id="NC_000913.3", rettype="gbwithparts")
write(tmps, "ecoli-mg1655.gb")
ecoligb <- readGenBank("ecoli-mg1655.gb")

ecoli_cds <- cds(ecoligb)
ecoli_cds

p.txdb <- autoplot(ecoli_cds)
p.txdb

#library(igvR)
ecoli_cds
ggbio() +
 circle(ecoli_cds, geom = "ideo", fill = "gray70") +
 circle(ecoli_cds, geom = "scale", size = 5) +
 circle(ecoli_cds, geom = "text", aes(label = locus_tag), vjust = 0, size = 3) +
 theme(
 axis.text.x = element_text(angle=90)
)

gr1 <- granges(ecoli_cds)
gr2 <- granges(ecoli_cds)
mcols(gr2)$test <- rnorm(length(ecoli_cds))
ggplot() +
 layout_circle(ecoli_cds, geom = "ideo", fill = "gray70", radius = 9, trackWidth = 1)
```

```
layout_circle(ecoli_cds, geom = "scale", size = 3, trackWidth = 1, scale.n=20) +
 layout_circle(gr1, geom = "rect", color = "steelblue", radius = 5) +
 layout_circle(gr2, geom = "bar", aes(y=test), size = 3, trackWidth = 1, scale.n=20, radius = 4)
```

---

이 저작물은 크리에이티브 커먼즈 저작자표시-비영리-변경금지 4.0 국제 라이선스에 따라 이용할 수 있습니다.

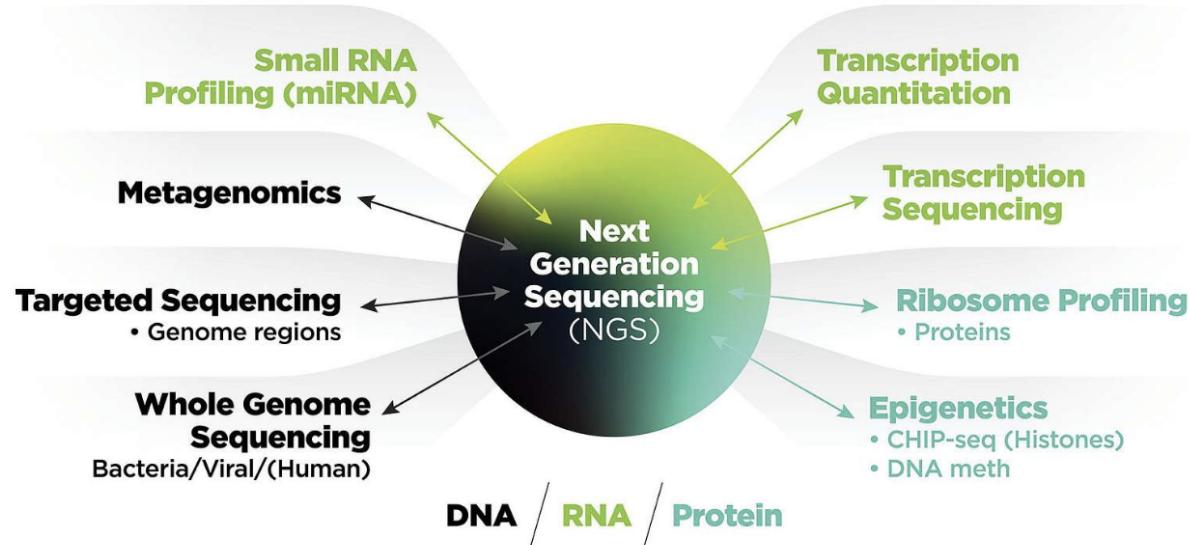


# Chapter 14

## High-throughput data

### 14.1 Next-generation sequencing

Next-generation sequencing (NGS) 는 DNA나 RNA 서열을 해독하는 기술로 2005년 개발될 초기에는 기존의 Sanger sequencing과는 다르게 여러 DNA 가닥을 동시에 해독하는 특징으로 “massively-parallel sequencing” 으로 불리우기도 했습니다.



Source: Mehta NAL, Dow DJ, Battram AM. 2011. DNA sequencing technologies and emerging applications in drug discovery. European Pharmaceutical Review website. <https://www.europeanpharmaceuticalreview.com/article/10409/dna-sequencing-technologies-and-emerging-applications-in-drug-discovery/>. Accessed May 4, 2020.

Note: Colors used in the diagram are an adaptation of the original.

NGS에 대한 자세한 설명은 illumina 사에서 제공하는 튜토리얼의 다음 사이트들을 참고하시기 바랍니다.

- Sequencing Fundamentals
- Sequencing Illumina Technology
- Illumina Sequencing by Synthesis

(Short read) NGS 워크플로는 다음과 같은 네 단계를 순차적으로 수행합니다.

각 단계별로 보면 다음과 같습니다.

위 단계 중 Secondary analysis에 해당하는 분석이 일반적으로 우리가 수행하는 RNA-Seq 등의 분석입니다. 시퀀싱 장비에서 읽힌 이미지 정보는 Binary Base Call (BCL) 파일로 변환됩니다. 우리가 일반적으로 다루는 FASTQ 파일은 서열 정보와 quality 정보를 text 형태로 저장한 파일로서 BCL 파일로부터 만들어집니다.

## 14.2 FASTQ preprocessing

FASTQ 파일에는 타깃 서열정보뿐만 아니라 바코드나 인덱스 등의 서열이 포함되어 있습니다.

The Illumina sequencing workflow can be divided into four separate processes:

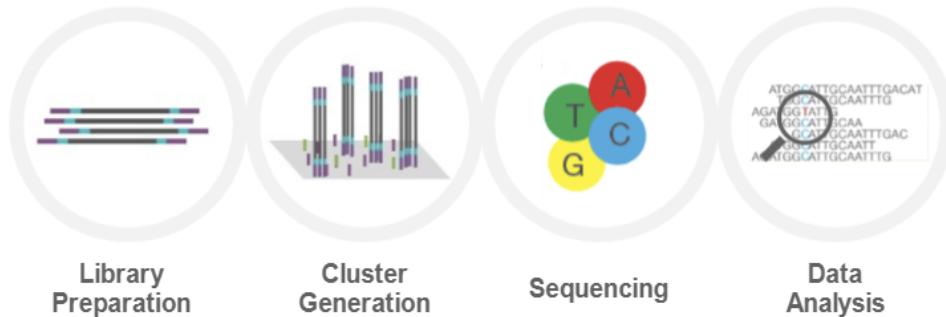
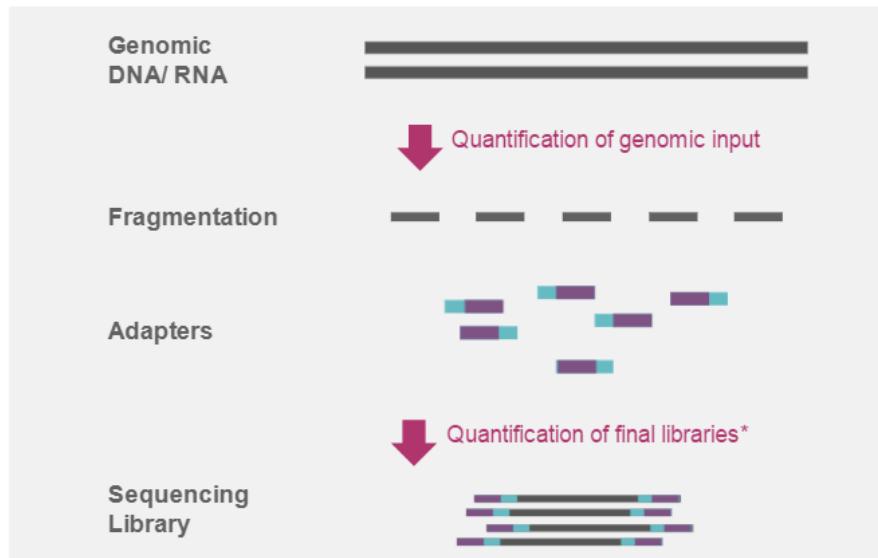


Figure 14.1: NGS workflow, figures from Illumina

### Library Preparation Overview

The library preparation process generally involves the following steps:

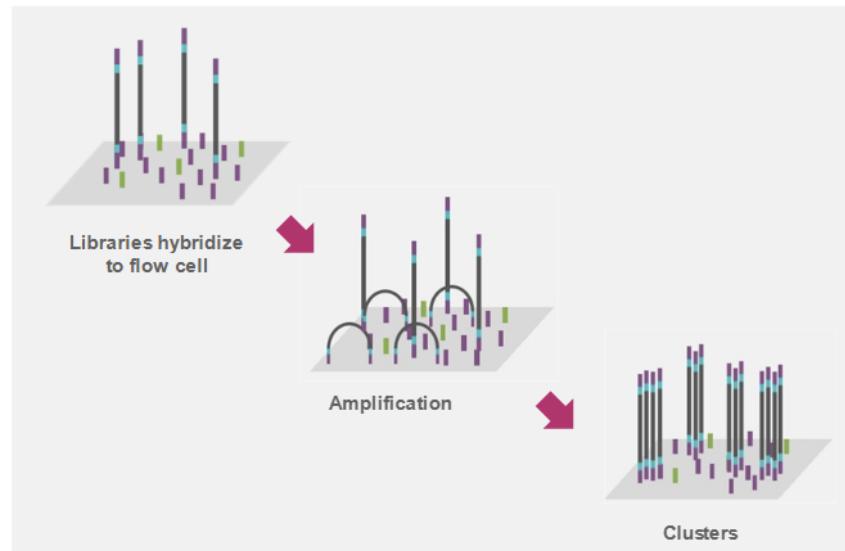


\* Library quantification methods may vary depending on your selected library preparation kit.

Figure 14.2: Library Prep, figures from Illumina

### Cluster Generation Overview

Cluster generation is the process by which each fragment in a library is cloned into thousands of identical copies.\*



\* Details of the cluster generation process may vary depending on flow cell architecture.

Figure 14.3: cluster generation, figures from Illumina

### Four-Channel Sequencing

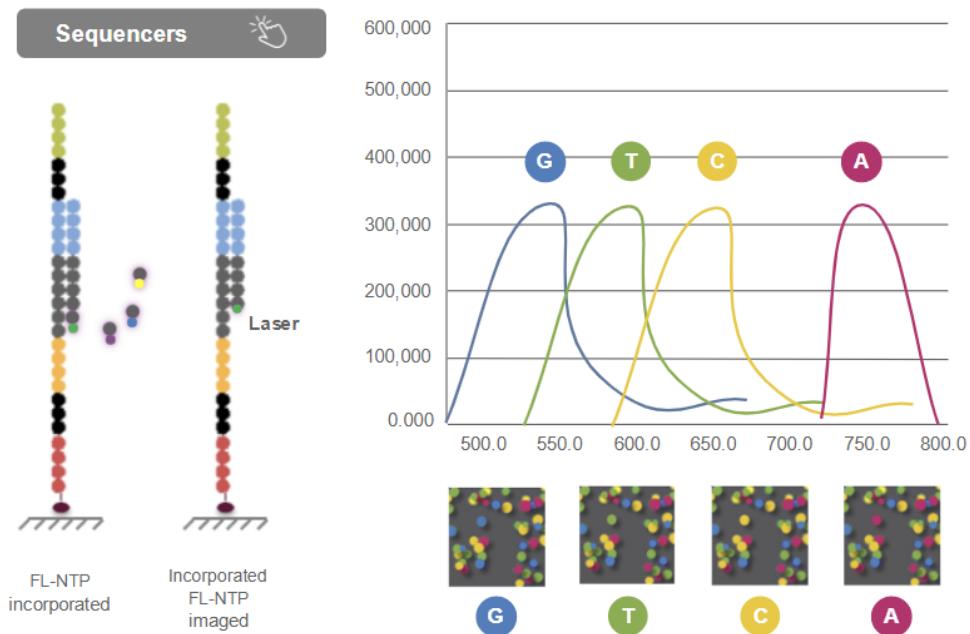


Figure 14.4: Sequencing, figures from Illumina

### Data Analysis

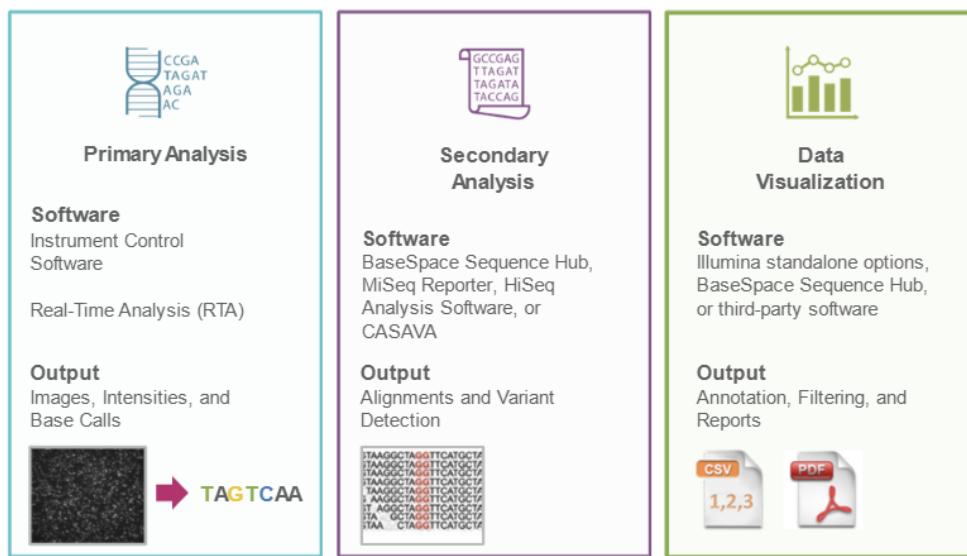


Figure 14.5: Data analysis, figures from Illumina

### Pipeline for data analysis

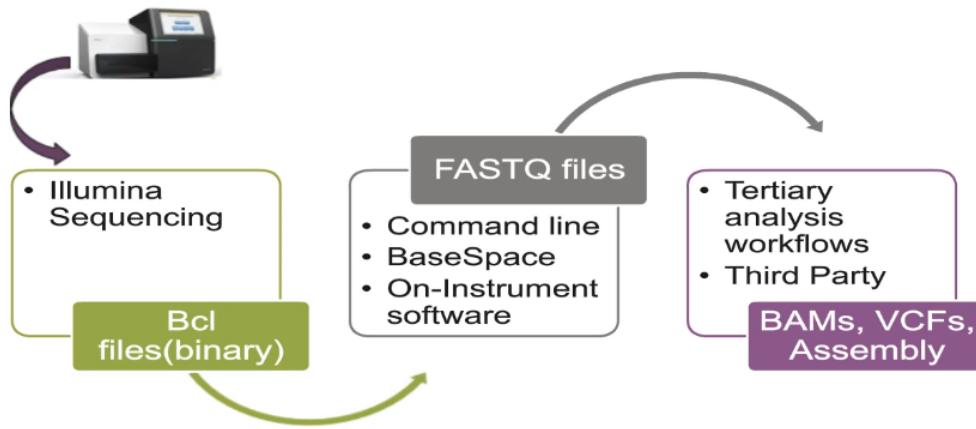


Figure 14.6: Data analysis, figures from Illumina

따라서 분석을 위해서는 위 서열들을 제거하고 quality에 따라서 read 들을 필터링 하는 작업이 필요합니다. 기존에는 linux 스크립트 기반의 소프트웨어들이 사용되었으나 본 강의에서는 Rstudio에서 바로 설치해서 활용할 수 있는 Rfastq 패키지를 사용하겠습니다. Rfastq는 quality control과 polyX trimming, adapter trimming, paired-end reads merging 등의 기능을 제공하고 있습니다.

```
if (!require("BiocManager", quietly = TRUE))
 install.packages("BiocManager")

BiocManager::install("Rfasttp")
```

examples 디렉토리 생성 후 예시 fastq 파일 다운로드, rfastq 실행으로 로딩과 필터링을 수행합니다. 참고로 Q10은 약 90%의 정확도, Q20은 약 99%의 정확도, Q30은 약 99.9% 정확도를 갖는 read의 개수입니다.

```
library(Rfastp)

download.file(url = "https://github.com/greendaygh/kribbr2022/raw/main/fastq/SRR11549087_1.fastq"

fqfiles <- dir(path = "examples", pattern = "*.fastq")

#?rfastp
fastq_report <- rfastp(read1 = file.path("examples", fqfiles[1]),
 outputFastq = file.path("examples", paste0("filtered_", fqfiles[1])))

round(qcSummary(fastq_report), 2)
```

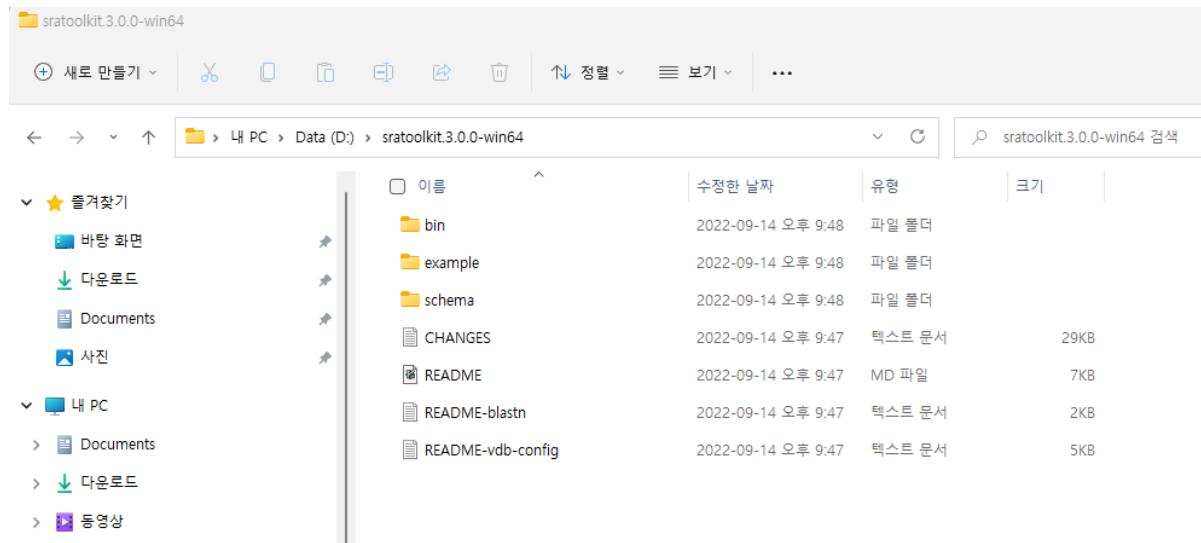
## 14.3 NGS database

Sequence Read Archive

SRA (Sequence Read Archive)는 High-throughput 시퀀싱 데이터의 공개 데이터베이스 중 가장 큰 규모의 미국 국립 보건원(NIH)의 1차 데이터베이스로서 서열데이터 뿐만 아니라 메타데이터, 유전체, 및 환경 데이터를 포함합니다. NCBI와 EBI(European Bioinformatics Institute), DDBJ(DNA Database of Japan) 같은 국제적 제휴를 통해 세 기관에서 제출 받은 데이터는 서로 공유되고 있습니다.

간략한 사용법은 NBK569238 또는 SRA download 문서 이곳을 참고하시기 바랍니다.

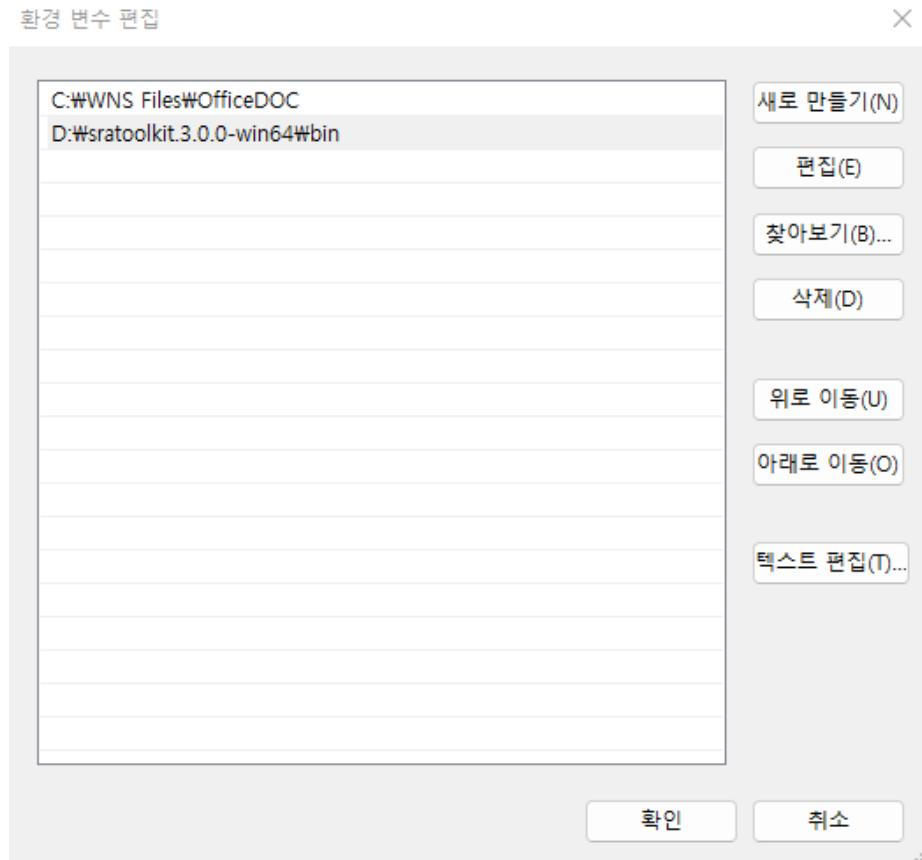
데이터를 다운로드 할 수 있는 NCBI SRA Toolkit을 제공하며 이 중 MS Windows 64 bit architecture 를 다운로드 받아 압축을 풀고 사용할 적절한 디렉토리로 옮겨 줍니다. 여기서는 D:\sratoolkit.3.0.0-win64\ 곳에 이동해 두었고 전체 디렉토리 구성을 다음과 같습니다.



명령을 어느 디렉토리에서나 사용하고 싶다면 위 경로의 bin 디렉토리를 path로 잡아주는 과정이 필요합니다. 다음 위치로 이동 후 “내PC > 속성 > 고급 시스템 설정 > 환경변수” 를 클릭하면 다음 창이 생성됩니다.



Path를 선택후 편집을 클릭하면 다음 화면이 생성되고 새로만들기를 누른 후 D:\sratoolkit.3.0.0-win64\bin라고 입력해주고 모든 창에서 엔터를 눌러주면 되겠습니다.



이제 파일 탐색기로 파일을 다운로드 받을 작업 디렉토리로 이동한 후 주소창에 cmd이라고 입력해서 프롬프트가 있는 명령창을 실행합니다.

fastq-dump.exe를 사용해서 다운로드 받을 수 있으며 최근에는 fasterq-dump를 사용해서 더욱 빠르게 다운로드를 받을 수 있습니다.

```
C:\Windows\System32\cmd.exe
Microsoft Windows [Version 10.0.22000.8561]
(c) Microsoft Corporation. All rights reserved.

D:\lecture>fastq-dump

Usage:
 /D//sratoolkit.3.0.0-win64//bin//fastq-dump.exe [options] <path> [<path>...]
 /D//sratoolkit.3.0.0-win64//bin//fastq-dump.exe [options] <accession>

Use option --help for more information

/D//sratoolkit.3.0.0-win64//bin//fastq-dump.exe : 3.0.0

D:\lecture>fasterq-dump

Usage:
 fasterq-dump.exe <path> [options]

Options:
 -o|--outfile output-file
 -O|--outdir output-dir
 -b|--bufsize size of file-buffer dflt=1MB
 -c|--curcache size of cursor-cache dflt=10MB
 -m|--mem memory limit for sorting dflt=100MB
 -t|--temp where to put temp. files dflt=curr dir
 -e|--threads how many thread dflt=6
```

뒤에서 설명할 GEO 데이터베이스에서 GSE148719 데이터를 다운로드 해보겠습니다. 위 링크를 클릭해서 들어가면 화면 하단의 SRA Run Selector라는 링크가 있고 이를 클릭하면 다음과 같은 화면이 보입니다.

The screenshot shows the NCBI SRA Run Selector interface. At the top, a banner informs users that SRA data is now in the cloud, with a link to revert to the old Run Selector. The main header includes the NCBI logo and the title "SRA Run Selector". Below the header, there's a search bar with the accession number "PRJNA625493" and a "Search" button.

**Common Fields**

BioProject	PRJNA625493
Consent	PUBLIC
Assay Type	RNA-Seq
AvgSpotLen	100
Center Name	GEO
DATASTORE filetype	BAM, SRA
DATASTORE provider	GS, NCBI, S3
DATASTORE region	gs.US, ncbi.public, s3.us-east-1
Genotype	contain pZS*plasmid, with mCherry (RFP) under pLtetO-1 promoter

**Select**

	Runs	Bytes	Bases	Download	Cloud Data Delivery	Compu
Total	4	1.60 Gb	5.08 G	Metadata or Accession List		
Selected	0	0	0	Metadata or Accession List or JWT Cart	Deliver Data	Gal

**Found 4 Items**

	Run	1	BioSample	2	Bases	3	Bytes	4	Experiment	5	GEO_Accession	6	Sample Name	7	source_name
<input type="checkbox"/>	<input checked="" type="checkbox"/> SRR11549076	SAMN11401217	1.27 G	108.11 Mb	SRR11549076	GSM1177185	GSM1177185	E.coli strain chromosome							

Metadata (SraRunTable.txt) 와 Accession list (SRR\_Acc\_List.txt)를 파일 형태로 다운로드 받은 후 적절한 전처리 후 사용하면 되겠습니다.

```
prefetch --option-file SRR_Acc_List.txt
```

만약 하나의 fastq 데이터만 다운로드 받을 경우 다음과 같습니다.

```
prefetch SRR11549076
```

이후 fasta 파일로 변환해 줍니다

```
fasterq-dump --split-files SRR11549076
```

100000개 read만 별도로 저장

```
fastq-dump -X 10000 --split-files SRR11549076
```

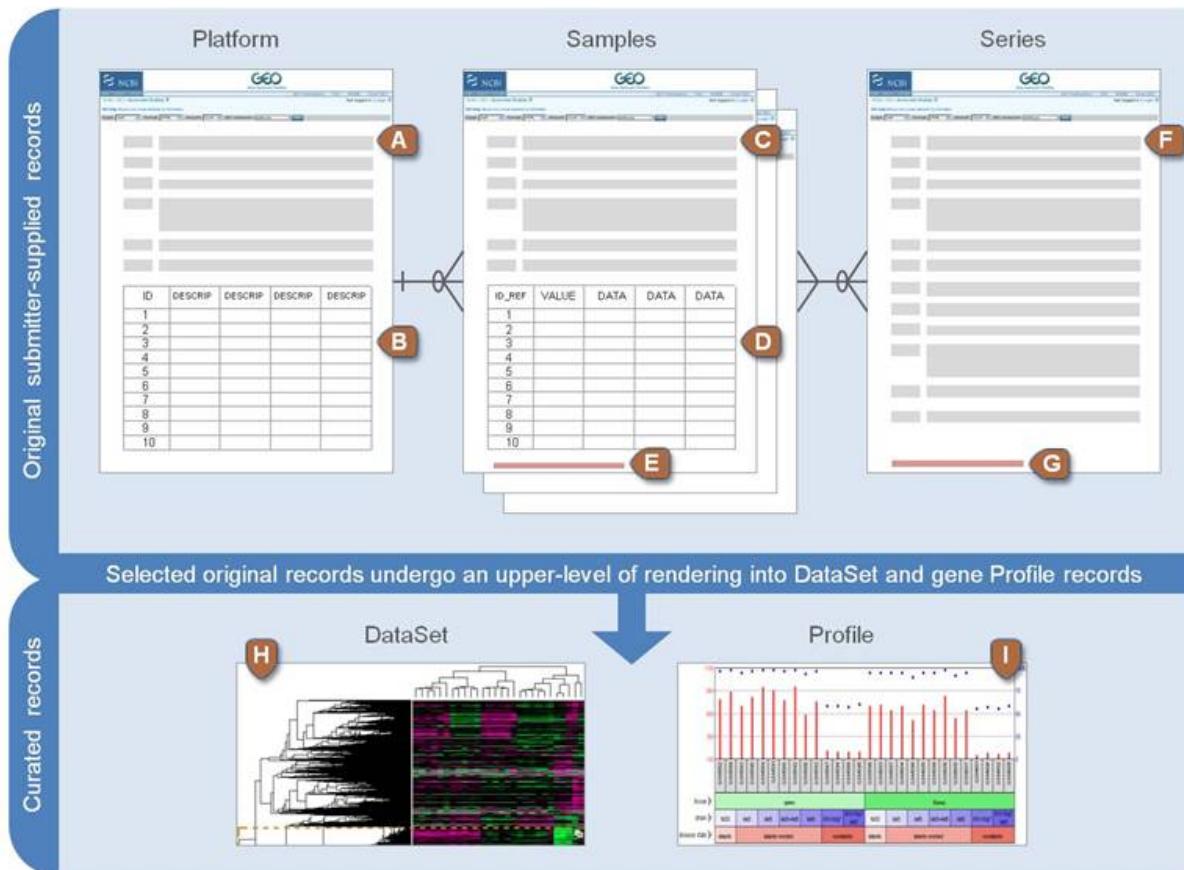
Gene expression omnibus (GEO)

GEO는 microarray, next-generation sequencing 등의 high-throughput 유전체 데이터를 보유한 공공 저장소입니다.

- 대규모 기능유전체 데이터베이스
- 데이터 기탁 쉽게 만들고 고수준 QC 유지
- 사용하기 쉬운 인터페이스 유지

### GEO

Platform, Sample, Series로 구성되어 있으며 Platform은 사용된 어레이 플랫폼에 대한 설명과 데이터 테이블로 구성되어 있습니다. GPLXXX 형태의 GEO 액세스 번호가 할당되면 하나의 플랫폼은 많은 샘플들에 사용될 수 있습니다. Sample은 개별 샘플이 처리된 조건 등의 설명이 있는 테이블로 구성되며 GSMxxx 형태의 GEO 등록 번호가 할당됩니다. Sample은 하나의 Platform만 참조 가능하며 여러 Series에 포함될 수 있습니다. Series는 관련된 샘플을 그룹화하고 전체 연구의 주요 설명을 제공합니다. GEO 등록 번호 GSExxx가 할당됩니다.



위 세 가지 타입 외에 Datasets 이 있으며 Datasets은 GDSxxx 아이디를 가집니다. 앞서 Series (GSExxx) 데이터가 연구자들이 업로드한 raw 데이터라고 한다면

Datasets (GDSxxx)는 관리자들에 의해 큐레이션된 데이터로 볼 수 있습니다. 브라우저를 통해 쉽게 검색할 수 있습니다. Bioconductor에서는 GEOquery라는 패키지로 관련 파일들을 다운로드 받을 수 있습니다.

```
if (!requireNamespace("BiocManager", quietly = TRUE))
 install.packages("BiocManager")

BiocManager::install("GEOquery")

library(GEOquery)
#browseVignettes("GEOquery")
```

#### The GDS class

```
gds <- getGEO(filename=system.file("extdata/GDS507.soft.gz", package="GEOquery"))
class(gds)
methods(class=class(gds))
Table(gds)
Columns(gds)
```

The GSM class - 샘플의 실제 측정값과 실험 조건 등 샘플별 정보 포함. 참고로 MAS 5.0 알고리즘은 서열의 Perfect-Match (PM)과 Mismatch (MM)를 이용해서 유전자의 발현을 정량화 하는 방법으로 (logged) PM-MM의 평균으로 계산함.

```
gsm <- getGEO(filename=system.file("extdata/GSM11805.txt.gz", package="GEOquery"))
methods(class=class(gsm))
head(Meta(gsm))
Table(gsm)
Columns(gsm)
```

#### The GPL class - 사용된 칩의 기본 Annotation 정보

```
gpl <- getGEO(filename=system.file("extdata/GPL97.annot.gz", package="GEOquery"))
gpl
```

#### The GSE class - 관련된 샘플, annotation 들의 집합 (실험)

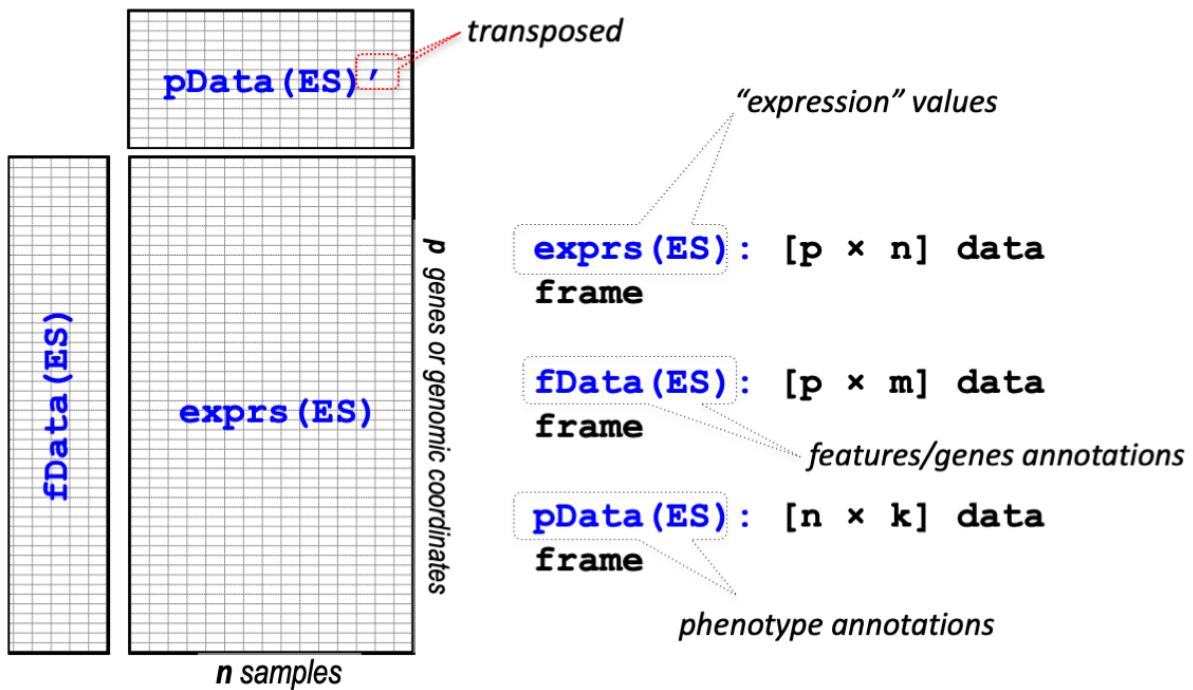
```
gse <- getGEO(filename=system.file("extdata/GSE781_family.soft.gz", package="GEOquery"))
methods(class=class(gse))
Meta(gse)
head(GSMList(gse))
gsm <- GSMList(gse)[[1]]
Meta(gsm)
Table(gsm)
Columns(gsm)

GPLList(gse)
gpl <- GPLList(gse)[[1]]
```

```
class(gpl)
```

## 14.4 ExpressionSet

Biobase 패키지는 지놈 데이터를 관리하기 위한 표준화된 데이터 구조 class인 ExpressionSet를 제공합니다. ExpressionSet은 HT assay 데이터와 실험 meta를 포함하고 있습니다.



출처BS831 lecture note

GES 데이터 받기 GSE2553

```
gse2553 <- getGEO('GSE2553', GSEMatrix=TRUE)
gse2553
class(gse2553)
class(gse2553[[1]])
mygse <- gse2553[[1]]
?ExpressionSet
methods(class=class(mygse))
mypdata <- pData(mygse)
myfdata <- fData(mygse)
myexdata <- exprs(mygse)
```

```

rownames , data.frame
as.data.frame(myexdata)

filtering
table(mypdata$description)

mypdata

class(myexdata)

```

GDS 데이터를 ExpressionSet class로 변환하기

```

gds <- getGEO(filename=system.file("extdata/GDS507.soft.gz", package="GEOquery"))
class(gds)
eset <- GDS2eSet(gds, do.log2=TRUE)
eset
pData(eset)

```

\*\* Example \*\*

다음 예제는 GEOquery 패키지에 있는 데이터셋을 활용하여 간단한 DEG 분석을 수행하는 코드로서 DEG 분석의 원리 이해와 해석을 위해 학습하는 예제입니다. 다음 장에 소개되는 통계(추정과 검정)을 먼저 참고하고 실습해 보아도 좋겠습니다.

GEOquery 패키지에 포함된 GDS507 dataset을 읽어서 ExpressionSet으로 변환합니다. ExpressionSet class를 활용하는 edgeR, DESeq2 등의 패키지를 사용하지는 않지만 변환 후 필요한 데이터를 추출해서 사용하겠습니다.

```

gds <- getGEO(filename = system.file("extdata/GDS507.soft.gz", package="GEOquery"))
gds
eset <- GDS2eSet(gds, do.log2=TRUE)
eset

myexp <- data.frame(exprs(eset))
myfeature <- fData(eset)
mypheno <- pData(eset)

```

위 세 종류의 데이터 테이블을 추출한 후 샘플들 두 그룹별로 평균을 계산하기 위해서 matrix transpose가 필요합니다. tidyverse는 (대부분의 통계 데이터는) row에 샘플이 위치하고 column에 feature (변수)가 있는 반면 위 myexp는 (ExpressionSet) 특성상 샘플이 컬럼에 위치하므로 transpose 수행 후 평균을 계산할 필요가 있습니다.

```

table(mypheno$disease.state)

transpose
mydat1 <- myexp %>%

```

```

rownames_to_column() %>%
pivot_longer(cols = -rowname) %>%
pivot_wider(names_from = rowname, values_from = value)

mydat2 <- mypheno %>%
dplyr::select(sample, disease.state) %>%
left_join(mydat1, by = c("sample" = "name"))

```

이 후 그룹별로 평균을 계산후 다시 transpose 시켜줍니다.

```

mymean <- mydat2 %>%
group_by(disease.state) %>%
summarise(across(where(is.numeric), mean))

mymean2 <- mymean %>%
pivot_longer(-disease.state) %>%
pivot_wider(names_from=disease.state, values_from = value)

```

두 그룹의 평균 값에 대한 각 유전자(feature)들의 산포도를 그릴 수 있습니다.

```

ggplot(mymean2, aes(x=normal, y=RCC)) +
geom_point()

```

위 데이터는 feature 별로 normal과 RCC 값들의 평균을 가지고 있습니다. 유사한 방법으로 표준편차 값을 구한 후 t값을 계산할 수 있으나 코드가 길어지게 됩니다. 기존 lecture에서는 apply 함수를 활용해서 t.test를 수행하여 p-value를 구할 수 있었으나 더 간단히 아래와 같이 tidyverse 타입의 함수를 사용해서 t.test를 사용할 수도 있습니다.

```

ttestval <- mydat2 %>%
#dplyr::select(1:10) %>%
summarise(across(where(is.numeric), function(x){
z <- t.test(x[disease.state=="normal"], x[disease.state=="RCC"])
c(z$p.value, z$statistic)
}))

ttestval

```

평균값과 pvalue, tstatistic 등의 값을 하나의 테이블로 만들기 위해서 위 ttestval 데이터를 transpose 시킨 후 mymean2 데이터와 병합합니다.

```

ttestvalt <- ttestval %>%
mutate(rnames = c("pvalue", "tstat")) %>%
#column_to_rownames(var="rnames") %>%
pivot_longer(-rnames) %>%
pivot_wider(names_from = rnames)

finaldat <- mymean2 %>% left_join(ttestvalt, by="name")

```

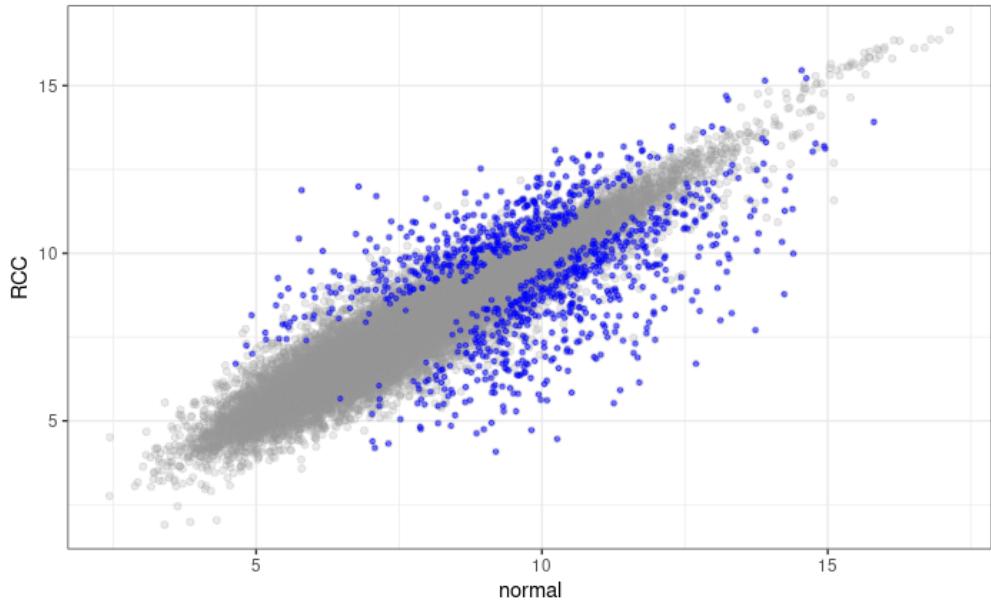
```
finaldat
```

유의한 데이터를 선별하고 가시화 합니다.

```
sigdat <- finaldat %>%
 filter(pvalue < 0.001)

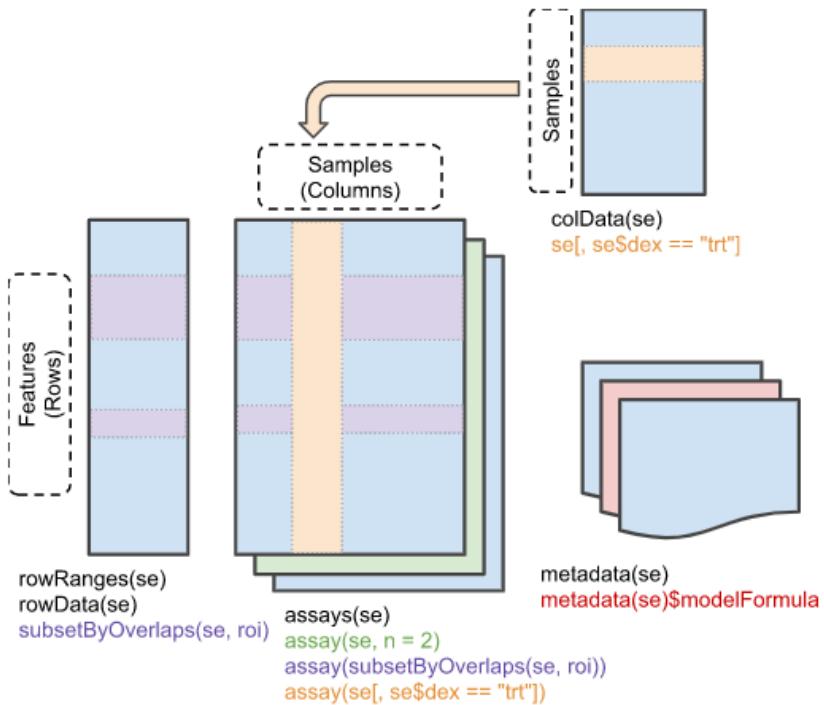
ggplot(finaldat, aes(x=normal, y=RCC)) +
 geom_point(alpha=0.2, color="#999999") +
 geom_point(data=sigdat, aes(x=normal, y=RCC), color="blue", alpha=0.5, shape=20) +
 theme_bw()
```

참고로 아래 결과는 p-value 가 0.001 이하인 probe들을 표현한 결과로서 정확한 결과 도출을 위해서는 multiple testing correction을 수행 후 수정된 유의확률을 이용할 필요가 있습니다.



## 14.5 SummarizedExperiment

ExpressionSet은 일반적으로 행이 feature (유전자) 인 마이크로어레이 기반 실험 및 유전자 발현 데이터에 사용되었습니다. 그러나 유전체 분석을 위해서는 유전자 정보 외에도 유전체상의 위치 정보 등이 필요하며 이는 앞서 배운 GenomicRanges 형태의 데이터가 필요합니다. 따라서 최근에는 새로운 버전인 SummarizedExperiment class가 SummarizedExperiment 개발되어 사용되고 있습니다.



```

library(SummarizedExperiment)

#if (!requireNamespace("BiocManager", quietly = TRUE))
install.packages("BiocManager")

#BiocManager::install("airway")

library(airway)
data(airway, package="airway")
se <- airway
se
?RangedSummarizedExperiment

assay data
assay(se)

Row (features)
rowRanges(se)

Column (sample)
colData(se)

```

```
Experiment-wide metadata
metadata(se)

SummarizedExperiment 생성
nrows <- 200
ncols <- 6
counts <- matrix(runif(nrows * ncols, 1, 1e4), nrows)
rowRanges <- GRanges(rep(c("chr1", "chr2"), c(50, 150)),
 IRanges(floor(runif(200, 1e5, 1e6)), width=100),
 strand=sample(c("+", "-"), 200, TRUE),
 feature_id=sprintf("ID%03d", 1:200))
colData <- DataFrame(Treatment=rep(c("ChIP", "Input"), 3),
 row.names=LETTERS[1:6])

se <- SummarizedExperiment(assays=list(counts=counts),
 rowRanges=rowRanges, colData=colData)

assay(se)

Row (regions-of-interest) data
rowRanges(se)

Column (sample) data
colData(se)

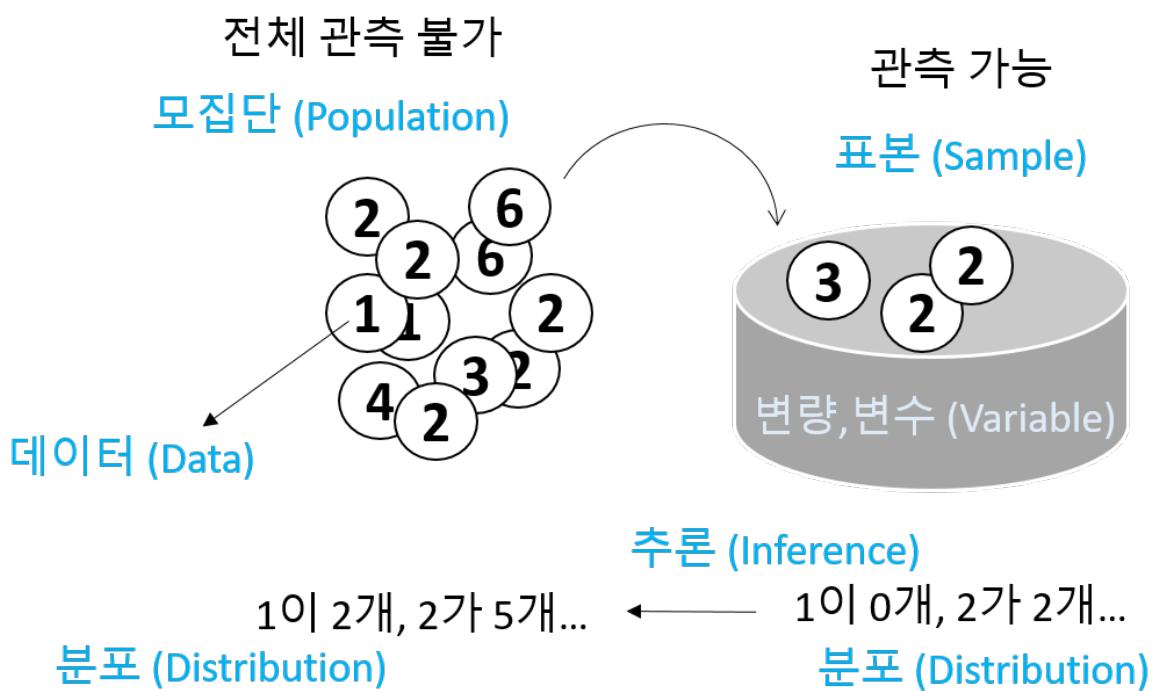
Experiment-wide metadata
metadata(se)
```

---

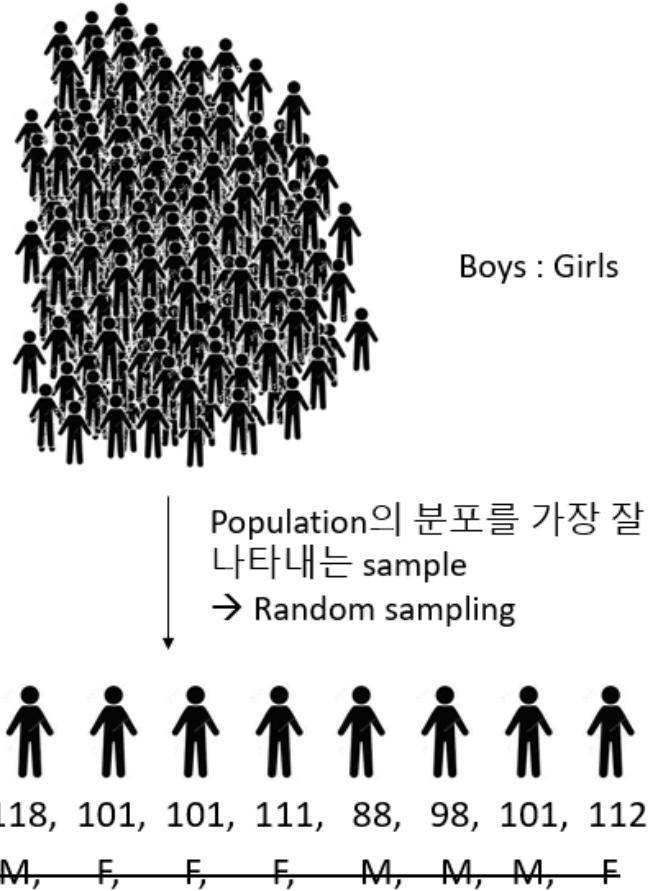
이 저작물은 크리에이티브 커먼즈 저작자표시-비영리-변경금지 4.0 국제 라이선스에 따라 이용할 수 있습니다.

# Chapter 15

## Inference and test



# Sampling



통계적 추정이란 모집단으로부터 임의 추출된 표본을 이용하여 모집단을 추정하는 과정을 의미합니다.

- 모집단 (population) - 전체 대상
- 모수 (Parameter) - 모집단의 분포를 설명하는 대푯값
- 표본 (sample) - 모집단으로부터 임의 추출된 관측값의 모음
- 통계량 (statistics) - 표본의 평균, 분산과 같은 대푯값으로 표본의 특징을 수치화한 값
- 확률변수 (random variable) - 확률적으로 따라 값이 결정되는 변수
- 확률분포 및 확률 질량/밀도 함수
- 표본분포 (sampling distribution) - 통계량의 분포

다음은 표준정규분포 모집단에서 (모수:  $\mu = 0, \sigma = 1$ ) 16개 표본을 임의 추출하여

평균을 (통계량)  $\bar{x}$  구하고 이 과정을 10번 반복한 상황을 표현한 그림으로 통계적 추론의 과정을 보여 줍니다. 즉, 표본을 뽑고 그 평균을 (표본평균) 구하는 과정을 반복할 경우 표본평균의 평균이 모평균에 수렴하고 표본평균의 분산이 표본들의 분산보다 더 작다는 것을 보여줍니다.

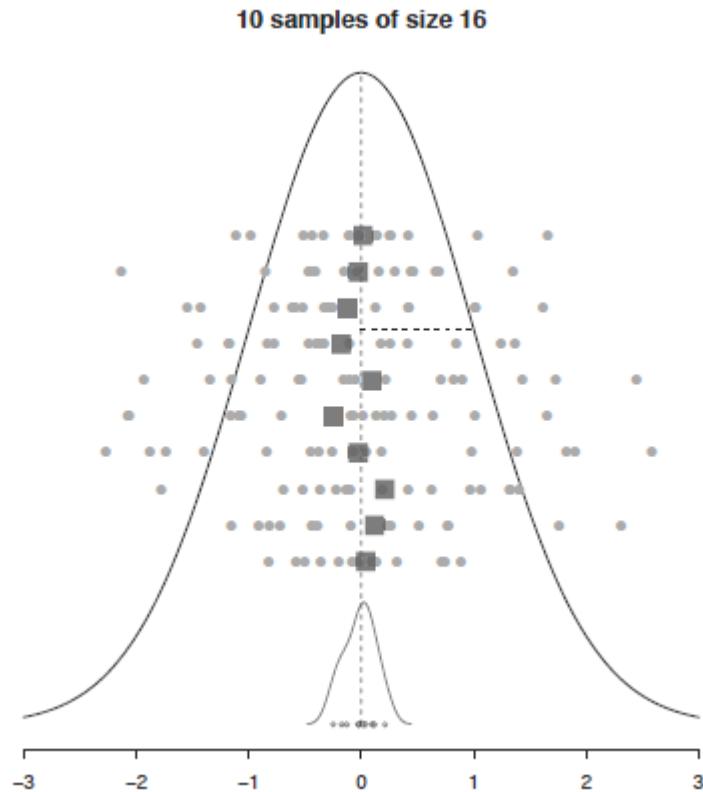


Figure 15.1: UsingR for introductory statistics, 243 페이지, 동그라미: 표본  $x$ , 사각형: 표본평균  $\bar{x}$ , 점선: 모평균 ( $\mu$ ), 하단 밀도 그래프: 표본평균의 분포

- 어떤 임의 표본에 대해서  $\bar{x}$ 의 표본분포는  $\mu$  근처에 위치
- 어떤 임의 표본에 대해서  $\bar{x}$ 의 표본분포의 표준편차는  $\sigma/\sqrt{n}$ 로 표현 ( $\sigma$ 는 모분산, 표본들의 분산보다 작음)
- 모분포가 정규분포이면  $\bar{x}$ 도 정규분포임

## 15.1 Simulation

이번 장에서는 시뮬레이션을 통해 추정의 개념을 이해하는 것을 목적으로 합니다. 확률과 공식의 유도를 통한 추정과정의 이해도 중요하지만 컴퓨팅 기반의

시뮬레이션도 통계적 추정의 개념을 이해하는데 큰 도움이 될 수 있습니다. R에서 분포관련한 시뮬레이션은 앞서 소개한 d, r, p, q 함수를 이용할 수 있습니다.

[EXERCISE]  $N(0, 1)$ 의 분포를 dnorm()을 이용해 그리시오 ( $-4 \leq x \leq 4$ )

```
library(tidyverse)

x <- seq(-4, 4, by=0.01)
y <- dnorm(x, 0, 1)
dat <- data.frame(x, y)
ggplot(dat, aes(x, y)) +
 geom_line()
```

지수분포의 경우, 파라미터 ( $\lambda$ ) 값에 따른 그래프 변화

```
x <- seq(0, 4, by=0.01)
y1 <- dexp(x, 1)
y2 <- dexp(x, 2)
y3 <- dexp(x, 3)
dat <- data.frame(x, y1, y2, y3)
datlong <- dat %>% pivot_longer(cols=c(y1, y2, y3))
ggplot(datlong, aes(x=x, y=value, col=name)) +
 geom_line(size=2)
```

[EXERCISE] 표준정규분포로부터 16개의 표본을 뽑아 평균을 구하고 각 표본과 평균값들을  $y = 1$  위치에 점으로 표현하시오 (rnorm() 사용)

```
nsample <- 16
x <- rnorm(nsample, 0, 1)
y <- rep(1, nsample)
xbar <- mean(x)
dat <- data.frame(x, y)
ggplot(dat, aes(x, y)) +
 geom_point() +
 geom_point(aes(x=mean(x), y=1), colour="blue", size=5, shape=15)
```

[EXERCISE] 위 예제와 같이 표준정규분포로부터 16개의 표본을 뽑아 평균을 구하는 과정을 두 번 반복하되 두 번째 데이터는  $y = 0.9$  위치에 표현하시오

```
nsample <- 16
x <- rnorm(nsample*2, 0, 1)
y <- c(rep(1, nsample), rep(0.9, nsample))
g <- factor(c(rep(1, nsample), rep(2, nsample)))
dat <- data.frame(x, y, g)

ggplot(dat, aes(x, y)) +
 geom_point() +
 geom_point(aes(x=mean(x[1:nsample]), y=1), colour="blue", size=5, shape=15) +
 geom_point(aes(x=mean(x[(nsample+1):length(x)]), y=0.9), colour="blue", size=5, shape=15)
```

```
scale_y_continuous(limits=c(0, 1.2))
```

[EXERCISE] 위 예제를 10번 반복하되 각 반복 데이터는 각각  $y = 1, 0.9, 0.8, \dots, 0.1$  위치에 그리시오

```
nsample <- 16
nrep <- 10

x <- rnorm(nsample*nrep, 0, 1)
tmpy <- seq(0.1, 1, length.out=nrep)
y <- rep(tmpy, each=nsample)
?rep
g <- factor(y)

dat <- data.frame(x, y, g)
head(dat)
sample means
dat_mean <- dat %>%
 group_by(g) %>%
 summarise(mean=mean(x))
head(dat_mean)

ggplot() +
 geom_point(data=dat, aes(x, y)) +
 geom_point(data=dat_mean,
 aes(x=mean, y=as.numeric(as.character(g))),
 colour="blue",
 size=5,
 shape=15) +
 theme_bw()
```

[EXERCISE] 위 예제에서 사용된 샘플들의 정규분포 곡선과  $\bar{x}$ 의 분포를 같이 그리시오 ( $-4 \leq x \leq 4$ , 앞서 예제의 dat와 dat\_mean 사용)

```
head(dat)
head(dat_mean)

x <- seq(-4, 4, by=0.01)
distribution of the samples
y <- dnorm(x, mean(dat$x), sd(dat$x))
distribution of the sample means
y2 <- dnorm(x, mean(dat_mean$mean), sd(dat_mean$mean))
dat2 <- data.frame(x, y, y2)
dat2_long <- dat2 %>%
 pivot_longer(cols=c(y, y2))
```

```
head(dat2_long)

ggplot(dat2_long, aes(x=x, y=value, color=name)) +
 geom_line(size=1.2) +
 labs(y="Density") +
 theme_bw() +
 scale_color_manual(name="Type",
 labels=c("Samples", "Sample Means"),
 values=c("red", "blue"))
```

위와 같이 표본들의 분포보다 표본평균들의 분포가 분포가 더 중심에 가깝다는 것을 볼 수 있습니다.

## 15.2 z-statistics with simulation

정규분포에서 추출된 표본들의 평균(표본평균)은  $n$ 의 수가 많지 않더라도 정규분포를 따릅니다 ( $n$ 이 충분히 많은 경우, 중심극한정리에 의해서 모집단의 분포와 상관없이 표본평균의 분포는 정규분포입니다). 통계적 유의성 판단의 기본 룰은 특정 확률변수  $X$ 의 분포를 가정한 후에 특정 사건의 관측한 값이  $X$ 의 확률분포 어디에 위치하는지 찾고 확률을 계산하여 해당 사건이 일어날만한 일이였으면 가정이 맞는 것으로 사건이 일어날 확률이 적게 나오면 가정이 틀린 것으로 판단하는 것입니다.

유사한 방법으로 모집단이 정규분포로 알려진 표본들을 가지고 표본평균을 구했을 때 이 표본평균이 정규분포에서 어디에 위치하는지와 그 확률을 계산하여 관측한 표본평균의 유의성을 판단할 수 있습니다. 이 과정에 사용하는 통계량이 z-score (z값)입니다. 관측값을 z-score로 변환해줄 경우 표준정규분포 ( $N(0, 1)$ )로부터 확률을 쉽게 계산할 수 있습니다.

$$z = \frac{\bar{x} - \mu}{(\sigma / \sqrt{n})}$$

z-score는 표본평균에 대한 z-score로서 모평균을 빼주고 모분산/ $\sqrt{n}$  값을 사용합니다.

[EXERCISE] A 제과업체에는 그들이 생산하는 사탕의 평균 무게가 평균 100g이고 표준편차가 16인 정규분포를 따른다고 주장한다. 그런데 소비자들은 이 사탕의 평균 무게가 100g보다 낮다고 의심을 하고 표본 10개를 추출하고 평균을 구했더니 90g이 관측되었다. z-score를 계산하고 표준정규분포에서 위치를 표시하시오.

```
zstat <- function(x, mu, sigma){
 z <- (mean(x)-mu)/(sigma/sqrt(length(x)))
 return(z)
}
```

```

xobs <- c(90, 75, 89, 103, 95, 110, 73, 93, 92, 80)
z <- zstat(xobs, 100, 16)

x <- seq(-5, 5, length.out=100)
y <- dnorm(x, 0, 1)
distribution of the sample means
dat <- data.frame(x, y)
p <- ggplot(dat, aes(x=x, y=y)) +
 geom_area(data=filter(dat, x < z), fill="red") +
 geom_line(size=1.2) +
 labs(y="Density") +
 theme_bw() +
 # geom_segment(aes(x=-1, xend=z, y=0.05, yend=0),
 # arrow = arrow(length = unit(0.1, "inches")),
 # size=1) +
 annotate("text", label=round(z,2), x=-0.7, y=0.07)

p

```

정규분포에서 관측값보다 더 작은 값이 관측될 확률 ( $p(\bar{X} < 90) = p(Z < -1.98)$ )은 `pnorm` 함수를 사용해서 구할 수 있으며 시뮬레이션을 이용할 수도 있습니다.

```

p + geom_segment(aes(x=-2.2, xend=-2.2, y=0.1, yend=0.01),
 arrow = arrow(length = unit(0.1, "inches")),
 size=1) +
 annotate("text", label=round(pnorm(z, 0, 1),3), x=-2.2, y=0.12)

simulation
n <- 1000
x <- rnorm(n, 0, 1)
sum(x < z)/n

```

이제  $p(X < 90) = 0.024$  값의 의미를 생각해 봅니다. 이 확률은 회사측이 주장하는  $\mu = 100$ 이라는 가설을 전제로 합니다. 즉,  $p(X < 90 | \mu = 100) = 0.024$ 입니다. 이는  $X < 90$ 라는 사건이 굉장히 낮은 확률로 일어났다고도 볼 수 있으나 가설이 틀렸다고 보는 것이 합리적입니다. 따라서 회사측이 주장하는 사탕 평균 무게 100g의 주장을 기각하며 소비자측의 주장  $\mu < 100$  즉 사탕이 100g 보다 작다는 주장을 강하게 지지하는 결과입니다.

### 15.3 t-statistics

위와 같은 z-score를 계산하기 위해서는 모표준편차가 필요하지만 모분산은 일반적으로 알려져있지 않기 때문에 z-score를 사용한 검정의 활용은 한정적입니다. 모표준편차 대신 표본의 표준편차를 사용하는 통계량이 t-statistic입니다. t 통계량은 t분포를 가지며 t분포는  $n$  이 무한에 가까워지면 표준정규분포와 같아집니다. 표본의 표준편차가 모표준편차보다 작은 경우 t 통계량 값이 z 값보다 커지게 되어 분포 양측 tail쪽 값이 많아지고 더 두꺼운 tail 분포 모양을 가지게 됩니다.

$$t = \frac{\bar{x} - \mu}{(s/\sqrt{n})}$$

시뮬레이션을 통해 분포를 그려보겠습니다.  $N(0, 1)$  분포에서 랜덤하게  $n=\{4, 10, 20, 50, 100, 1000\}$  개의 표본을 뽑는 과정을 1000회 반복한 후 boxplot을 그려보겠습니다.

```
tstat <- function(x, mu){
 (mean(x)-mu)/(sd(x)/sqrt(length(x)))
}

mu <- 0
sigma <- 1
M <- 1000
n <- c(4, 10, 20, 50, 100, 1000)

tstat_array <- replicate(M,
 sapply(n, function(x){
 tstat(rnorm(x, mu, sigma), mu)
 }))
dim(tstat_array)

transposition
tstat_array <- t(tstat_array)
dim(tstat_array)
colnames(tstat_array) <- as.character(n)
boxplot(tstat_array)

tstat_df_long <- as.data.frame(tstat_array) %>%
 pivot_longer(cols=everything())
ggplot(tstat_df_long, aes(x=name, y=value)) +
 geom_boxplot()

tstat_df_long <- as.data.frame(tstat_array) %>%
 pivot_longer(cols=everything()) %>%
 mutate(name=fct_relevel(name, "4", "10", "20", "50", "100", "1000"))
```

```
ggplot(tstat_df_long, aes(x=name, y=value)) +
 geom_boxplot()
```

[EXERCISE] 두 그룹 데이터에서 임으로 10명을 두 번 뽑아 그 평균의 차이를 계산하시오

```
caf <- c(245, 246, 246, 248, 248, 248, 250, 250, 250, 252)
no_caf <- c(242, 242, 242, 244, 244, 245, 246, 247, 248, 248)
dat <- c(caf, no_caf)
obs <- mean(caf) - mean(no_caf)

x <- sample(dat, 10, replace=T)
y <- sample(dat, 10, replace=T)

mean(x) - mean(y)
```

[EXERCISE] 위 예제의 과정을 1000번 반복하고 계산된 차이값들로 분포를 그리시오(for 문 이용)

```
diff_vals <- rep(0, 1000)
for(i in 1:1000){
 x <- sample(dat, 10, replace=T)
 y <- sample(dat, 10, replace=T)
 diff_vals[i] <- mean(x) - mean(y)
}

ggplot(data.frame(diff_vals), aes(x=diff_vals)) +
 geom_histogram()
```

[EXERCISE] 분포에서 실제 관측한 3.5 값의 위치를 표시하고 관측값보다 더 극단적인 경우가 나올 경우의 비율을 계산하시오 (위 예제 코드의 연속)

```
emp_pval <- sum(diff_vals > obs)/length(diff_vals)

textstring <- paste("p(X > ", obs, ") = ", emp_pval, sep="")
ggplot(data.frame(diff_vals), aes(x=diff_vals)) +
 geom_histogram() +
 # geom_segment(aes(x = obs,
 # y = 30,
 # xend = obs,
 # yend = 5),
 # arrow = arrow(),
 # size=2) +
 annotate("text",
 label = obs,
 x = obs,
 y = 35,
```

```

 size = 5) +
annotate("text",
 label = textstring,
 x = 2.5,
 y = 100,
 size = 5) +
labs(x="X", y="Count")

```

관측된 차이 3.5는 가능한 차이값들을 모두 그려본 분포에서 가장자리에 위치합니다. 관측값이 중심에 가까울수록 흔하게 관측되는 것으로 두 그룹간 차이가 랜덤하게 나누어도 높은 확률로 관측 가능한 값이라는 의미입니다. 반면 가장자리에 위치할수록 그룹간의 차이가 랜덤이 아닌 특정 요인이 작용해서 발생한 사건으로 해석할 수 있습니다.

$p(X > 3.5)$ 는 위 사건이 발생한 경우(3.5)보다 극단적으로 큰 값의 사건이 발생할 확률을 말하며 이는  $1 - p(X \leq 3.5)$ 이며  $p(X \leq 3.5)$ 는 누적분포함수, qnorm으로 구할 수 있습니다. 예제에서는 이 값이 0.003으로 관측값 3.5는 랜덤으로는 거의 일어나기 힘든 확률의 사건임을 알 수 있습니다.

통계적 유의성 검정의 측면에서 생각해 보면 위 두 그룹의 데이터를 랜덤하게 섞은 후 그룹을 다시 나누는 것은 그룹간 차이가 없다는 것을 가정하는 것 입니다. 즉,  $\mu_1 = \mu_2$ 이며 이 상태에서  $X = \mu_1 - \mu_2$ 인 확률변수라 할 때  $p(X > 3.5)$ , 즉,  $p(\mu_1 - \mu_2 > 3.5 | \mu_1 = \mu_2)$ 를 계산 한 값입니다. 이 값이 0.001이라는 것은 희박한 확률로 3.5가 관측되었다고 볼 수 있으나 가정이 틀렸다고 보는 것이 더욱 합리적입니다. 따라서  $\mu_1 = \mu_2$ 를 받아들이지 않고 (기각하고)  $\mu_1 \neq \mu_2$ 를 지지하는 확률이 높아지게 되는 것 입니다.

## 15.4 Significance test

유의성 검정은 분포를 가정한 상태에서 모수에 대한 특정 값을 추정한 후 (점추정) 해당 값이 가정된 분포로부터 관측될 확률을 계산하여 가설에 대한 판단을 수행합니다.

유의성검정을 재판 과정의 검사와 배심원 입장으로 생각하면 이해가 쉬울 수 있습니다. 검사는 피의자가 유죄임을 주장하며 배심원들을 설득합니다. 배심원들은 피의자가 유죄라는 확정적 증거가 없는 한 무고하다는 가정을 하고 있으며 증거가 많아질수록 자신들이 가정한 무죄가 아닐 확률은 점점 적어집니다. 즉, 확률이 충분히 작으면 무죄라는 가정을 버리고 검사의 주장을 받아들이게 됩니다.

## 15.5 Errors in significance tests

다른 예를 들어봅시다. 어떤 영업사원이 A 회사에서 판매하는 기계의 영점이 평균 0으로 맞춰져 있다고 주장을 하며 해당 기계를 팔고 있습니다. 실제로 판매하는 기계의 영점이 0으로 맞춰져 있을 때 (0이 참) 우리 입장에서는 영업사원의 말은 당장 증명할 수 없는 가설일 뿐입니다. 그래도 그 가설을 믿고 (채택하고) 기계를

구입할 경우 오류 없는 정상적인 거래가 됩니다. 그런데 우리가 영점이 2라고 의심을 하며 영업사원의 말을 믿지 않고 (가설을 기각하고) 기계를 구입하지 않는다면 오류가 발생한 것입니다. 이 상황을 그래프로 알아봅니다 (편의상 기계 영점의 분산은 1이라고 가정).

```
x1 <- seq(-6, 6, by=0.01)
y1 <- dnorm(x1, 0, 1)
z <- data.frame(x1, y1)
pval <- round(1-pnorm(2,0,1), 4)
ggplot(z) +
 geom_line(aes(x1, y1), color="purple") +
 geom_vline(xintercept = 2) +
 geom_hline(yintercept = 0) +
 geom_area(data=filter(z, x1 > 2),
 aes(x1, y1),
 fill="#80008055") +
 annotate("text", x=3, y=0.1, label=pval)
```

위 그래프에서 실제 기계들의 평균 영점이 0임인 분포만을 생각하면 면적의 넓이 0.0228은 2보다 큰 영점을 가지는 기계가 생산될 확률입니다. 이 실제 사실에 더하여 “영점이 0이다”라는 가설의 분포를 생각하면 (가설과 사실의 분포가 겹쳐있음) 면적의 넓이 0.0228 부분은 2를 기준으로 가설을 받아들이지 않는 (기각하는) 경우로 볼 수 있으며 결국 실제 사실도 받아들이지 않는 오류를 범할 확률을 나타냅니다. 이 오류를 우리는 “제1종오류 ( $\alpha$ )”라고 합니다.

이제 영업사원의 가설이 거짓일 경우를 생각해 봅니다. 즉, 어떤 이유로 A회사 기계의 영점이 평균 3일 경우 영업사원의 “영점이 0이다”라는 주장을 거짓이 됩니다. 이 상황에서도 우리는 두 가지 경우를 생각할 수 있습니다. 영업사원의 가설을 믿고 (채택하고) 기계를 구입할 경우는 오류가 발생하는 상황과 영점이 2라는 의심으로 영업사원의 가설을 믿지않고 (기각하고) 기계를 구입하지 않는 올바른 판단을 한 상황입니다.

```
x1 <- seq(-6, 6, by=0.01)
y1 <- dnorm(x1, 0, 1)
x2 <- seq(-1, 11, by=0.01)
y2 <- dnorm(x2, 3, 1)
z <- data.frame(x1, y1, x2, y2)
#pval <- round(1-pnorm(2,0,1), 4)
ggplot(z) +
 geom_line(aes(x1, y1), color="blue") +
 geom_line(aes(x2, y2), color="red") +
 geom_vline(xintercept = 2) +
 geom_hline(yintercept = 0) +
 geom_area(data=filter(z, x1 > 2),
 aes(x1, y1),
 fill="#0000ff55") +
```

```
geom_area(data=filter(z, x2 < 2),
 aes(x2, y2),
 fill="#ff000055") +
 annotate("text", x=3, y=0.1, label=pval)
```

이 때는 실제 사실의 분포와(red) 가설의 분포가(blue) 다릅니다. 실제 사실의 분포 입장에서 2를 기준으로 가설을 기각하는 상황은 올바른 판단을 하는 상황입니다. 그러나 2를 기준으로 가설을 받아들이는 경우, 실제 사실은 받아들이지 않게 되는 오류가 발생합니다. 이 오류를 “제2종오류 ( $\beta$ )”라고 합니다.

일반적으로 실제 사실은 모집단의 모수와 같이 알 수 없는 값입니다. 따라서 우리는 가설의 분포를 가지고 판단을 하게되며 이 때  $\alpha$ 와  $\beta$  오류는 위 그림이 보여주는 것처럼 서로 trade off 관계에 있게 됩니다. 즉, 임의의 가설을 기반으로 특정 관측값의 유의성을 판단할 때 제1종오류를 최소화 하려하면 제2종오류는 최대화되고 그 반대로 제2종오류를 최소화 하면 제1종오류는 오히려 커지게 되는 것입니다.

유의성검정에서  $H_0$ 는 귀무가설(Null hypothesis)이고  $H_1$ 을 대립가설(alternative hypothesis)이라 합니다. 일반적으로  $H_1$ 이 사람들이 관심있는 주장이고 유의성검정을 위해서 사람들의 주장의 반대인  $H_0$ 를 가정합니다. 만약  $H_0$  가정 하에서 만들어진 통계량의 관측될 확률이 작으면 가정이 틀린 것으로  $H_0$ 를 기각하고 사람들이 주장하는  $H_1$ 을 채택합니다. 여기서 계산된 통계량이 관측될 확률을 유의확률(p-value) 이라 하며 유의성검정은 p-value를 계산하는 것과 같습니다. 위 그림에서 0.0228 값이 p-value입니다.

$$\text{p-value} = P(\text{test statistic is the observed value or is more extreme} | H_0)$$

p-value의 크고 작음을 판단하는 대략적인 범위는 다음과 같습니다.

<i>p</i> -value range	significance stars	common description
[0,0.001]	***	extremely significant
(0.001,0.01]	**	highly significant
(0.01,0.05]	*	statistically significant
(0.05,0.1]	.	could be significant
(0.1,1]		not significant

Table 9.1: Level of significance for range of *p*-values.

유의성검정에서는  $H_0$ 가 참인지 거짓인지 판별하기 보다는 유의수준(significance level,  $\alpha$ )이라는 기준에 따라서  $H_0$ 를 기각할지 안할지를 판단하게 됩니다. 위 영점 예제에서 제1종오류  $\alpha$ 가 유의수준과 같은 의미입니다. 일반적인 유의수준은 0.01, 0.05, 0.1 정도로 p-value가 이들 값보다 작게 나오면  $H_0$ 를 기각합니다. 앞서 기계 영점에 대한 예제에서와 같이  $\alpha$ 를 기준으로  $H_0$ 를 기각 할 경우 두 가지 오류, 제1종오류 (type-I error)와 제2종오류 (type-II error)가 발생할 수 있습니다.

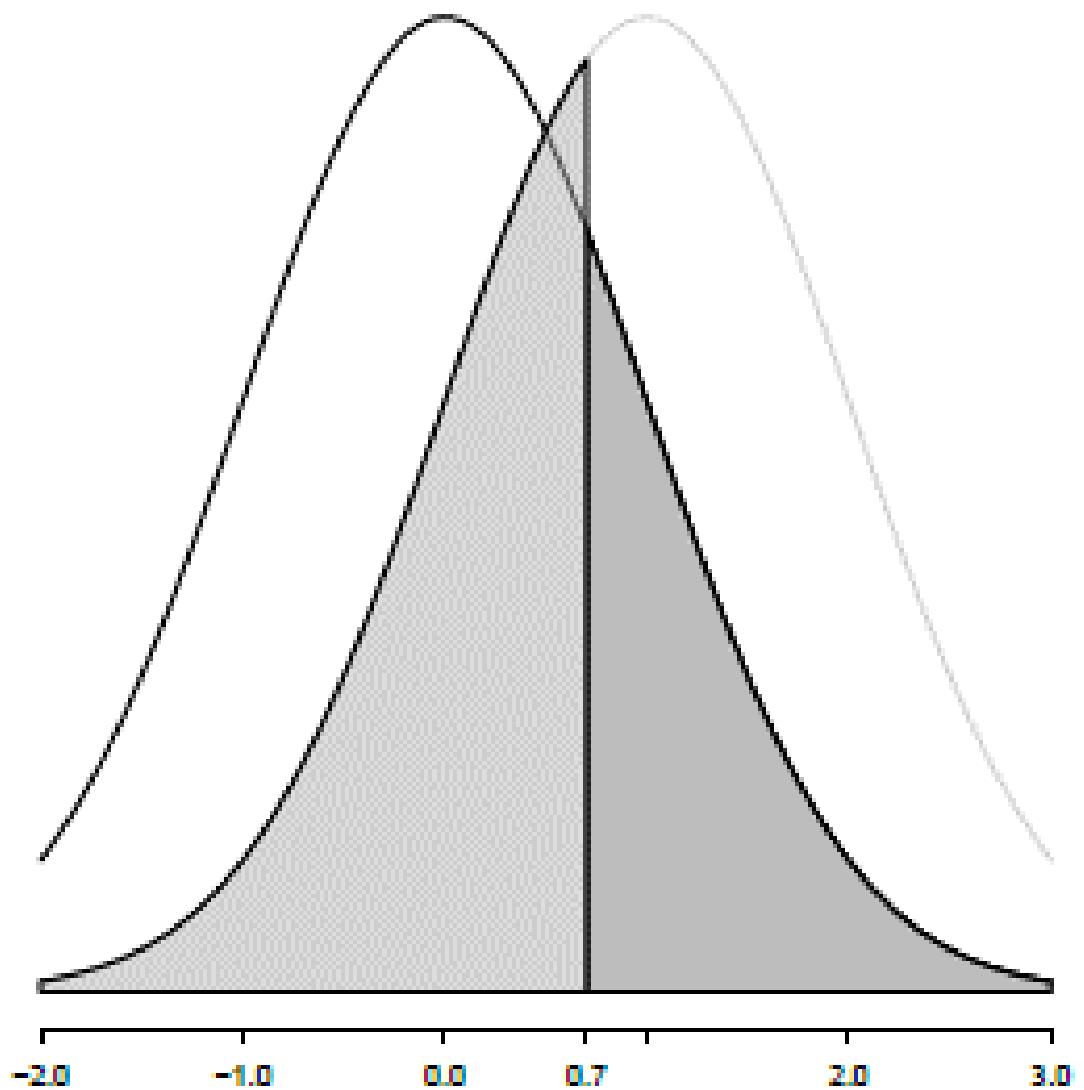
Table of error types		Null hypothesis ( $H_0$ ) is	
		True	False
Decision About Null Hypothesis ( $H_0$ )	Fail to reject	Correct inference (True Negative) (Probability = $1 - \alpha$ )	Type II error (False Negative) (Probability = $\beta$ )
	Reject	Type I error (False Positive) (Probability = $\alpha$ )	Correct inference (True Positive) (Probability = $1 - \beta$ )

- 귀무가설 참 -> 채택 (o)
- 귀무가설 참 -> 기각 (오류)
- 귀무가설 거짓 -> 채택 (오류)
- 귀무가설 거짓 -> 기각 (o)

앞서 재판의 경우를 예로 들면 제1종오류는 죄가 없는 사람( $H_0$ 가 참)을 죄가 있다고 판단 ( $H_0$ 기각) 하는 경우로 가능하면 일어나서는 안되는 상황입니다. 따라서  $\alpha$ 는 보수적인 기준으로 정하게 되나  $\alpha$ 가 작아지면 자동적으로  $\beta$ 가 큰 값으로 결정되어 두 오류를 동시에 작게 만족시키는 유의수준은 정하기 어렵습니다. 따라서 가능한 제1종 오류를 작게 유지하면서 power ( $1 - \beta$ )를 가능한 높게 되도록 검정을 디자인할 필요가 있습니다.

[Example] 어떤 기계의 영점이  $N(0,1)$ 의 분포를 가지고, 영점이 맞지 않을 경우  $N(1,1)$ 의 분포를 가진다고 한다. 기계로 부터 측정한 값이 0.7일 경우 기계의 영점이 맞춰져 있는지 아닌지를 판단하시오

$$H_0 : \mu = 0 \text{ vs } H_1 : \mu = 1$$



만약 영점이 맞춰진 상태에서 관측된 값이라면  $Z = 0.7 - \mu/sd = 0.7$  이므로 p-value는  $1 - \text{pnorm}(0.7, 0, 1) = 0.2419$  이므로 가설을 기각할만한 증거가 충분치 않습니다. 즉,  $H_0 : \mu = 0$ 를 받아들이는 상황인 것인데 그렇지만 0.7은 분명히 0보다는 1에 가까운 ( $H_1$ ) 값입니다. 이러한 경우에 1 standard deviation 대신  $1/\sqrt{10}$  값을 사용하면 훨씬 더 명확한 판단을 내릴 수 있습니다.

```
1 - pnorm(0.7, 0, 1/sqrt(10))
```

즉, p-value가 충분히 작으므로 귀무가설  $H_0$ 를 기각하고 대립가설을 지지하게 됩니다. 1 standard deviation unit 대신  $1/\sqrt{10}$  unit을 사용함으로써 더욱 명확한 판단을 내릴 수 있게 된 것입니다.

위와 같은 p-value는 가설을 검정하는데 사용되는 핵심 기준이 되며 가설을 검정하기 위한 p-value 계산법은 일반적으로 다음과 같습니다.

- 데이터에 맞는 분포를 정함 (모수 정의)
- $H_0$ 와  $H_1$ 를 정함
- 검정 통계량 정의
- 데이터 수집
- 검정 통계량 계산
- p-value 계산

유의성 검정의 목적은 추정한 모수가 얼마나 통계적으로 유의한지를 판단하기 위한 것입니다. 모형에 (분포) 따라서 모수가 달라지므로 다음과 같이 몇 가지 경우에 대한 유의성 검정 방법들이 있습니다.

## 15.6 Significance test for the mean (t-test)

이번에는 미지의 모평균에 대한 검정을 수행하는 방법을 알아봅니다. 검정 방법은 앞서 배운 검정 과정과 유사하며 통계량은 신뢰구간을 학습할 때 배웠던 t통계량과 같습니다.

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu > \mu_0, H_1 : \mu < \mu_0, H_1 : \mu \neq \mu_0$$

$$T = \frac{\bar{x} - E(\bar{x}|H_0)}{SE(\bar{x}|H_0)} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{\text{observed} - \text{expected}}{SE}$$

이제 데이터  $x_1, x_2, \dots, x_n$ 을 얻고 이로부터  $t = (\bar{x} - \mu_0)/(s/\sqrt{n})$ 을 구할 경우 p-value는 다음과 같습니다.

[Example] 새 SUV 자동차의 연비가 17miles / gallon 으로 알려져 있다. 소비자 그룹에서는 그러나 이보다 낮은 것으로 의심하고 있다. 다음 데이터들이 관측되었을 때 해당 신차의 연비가 17mile 보다 작은지 검정하시오.

$$p\text{-value} = \begin{cases} P(T \leq t | H_0) & H_A : \mu < \mu_0, \\ P(T \geq t | H_0) & H_A : \mu > \mu_0, \\ P(|T - \mu_0| \geq |t - \mu_0| | H_0) & H_A : \mu \neq \mu_0. \end{cases}$$

Figure 15.2: page304

```
mpg <- c(11.4, 13.1, 14.7, 14.7, 15, 15.5, 15.6, 15.9, 16, 16.8)
xbar <- mean(mpg)
s <- sd(mpg)
n <- length(mpg)
tstat <- (xbar-17)/(s/sqrt(n))
```

$$H_0 : \mu = 17$$

$$H_1 : \mu < 17$$

$$T = \frac{14.87 - 17}{1.582/3.162} = -4.284$$

```
x <- seq(-5, 5, length=100)
y <- dt(x, df=n-1)
dat <- data.frame(x, y)
ggplot(dat, aes(x, y)) +
 geom_line() +
 geom_vline(xintercept = tstat)

pt(tstat, df=9, lower.tail=T)
```

[EXERCISE] 위 예제를 `t.test` 를 사용해서 구현하시오

## 15.7 Two sample significance tests

두 그룹의 데이터 (표본)을 가지고 있을 때 두 그룹이 통계적으로 차이가 있는지를 검증하는 방법으로 (코호트 데이터, Case-control 데이터) 확률 분포를 이용한 통계적 검증을 알아보겠습니다.

카페인(커피)이 초초한 상태를 유발하는가? 라는 질문에 답하기 위해서 다음 데이터를 얻었습니다. 다음 값들은 커피를 제공한 그룹과 그렇지 않은 그룹의

손가락 탭핑 횟수를 비디오로 분석한 데이터입니다. 이럴 경우 일반적으로 두 그룹의 평균의 차이를 비교합니다.

```
coff <- c(245, 246, 246, 248, 248, 248, 250, 250, 250, 252)
nocoff <- c(242, 242, 242, 244, 244, 245, 246, 247, 248, 248)
obsdiff <- mean(coff) - mean(nocoff)
obsdiff
```

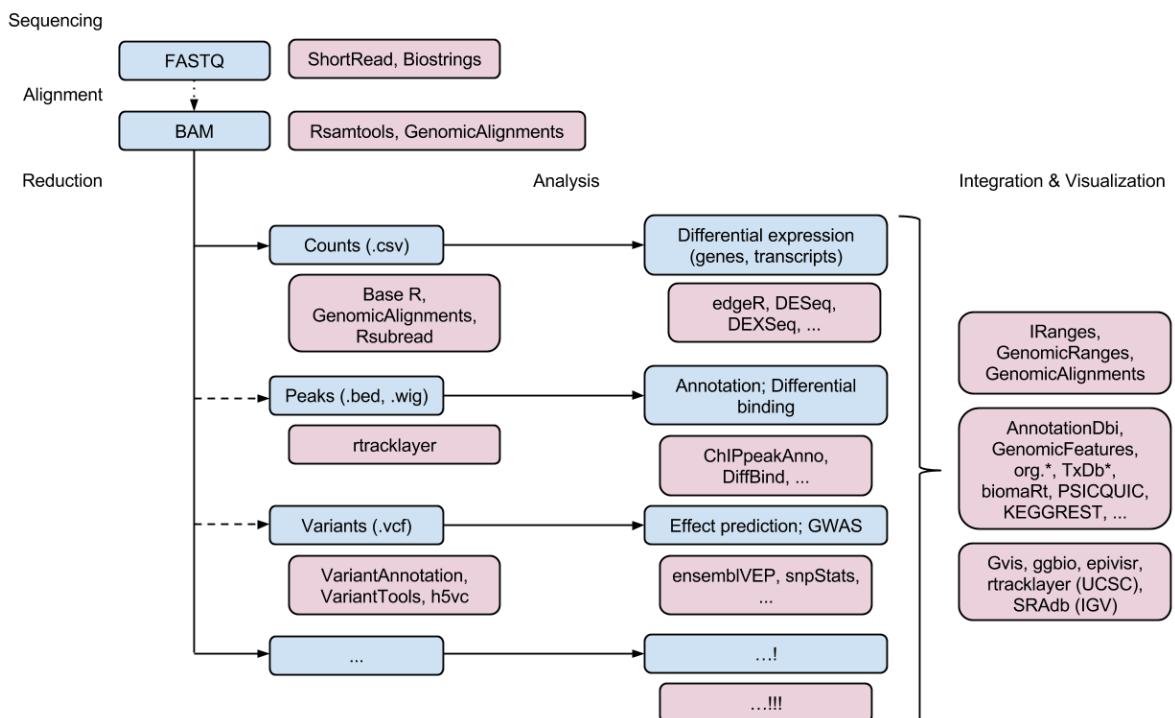
차이는 3.5가 나왔지만 이 차이가 얼마나 통계적으로 유의한지를 알아야 합니다.  
(보충 필요)



# Chapter 16

## DEG analysis

### 16.1 High-throughput genomic data analysis



## 16.2 DEG analysis with bioconductor

Differentially Expressed Gene 분석은 전통적 two-channel microarray나 RNA-Seq 데이터를 활용한 분석입니다. Genome reference에 fastq 파일의 read들을 맵핑하고 맵핑된 read들을 카운팅하여 해당 유전자의 발현을 정량화하고 이를 기준이 되는 발현값과 비교하여 질병이나 조건에 따라 다른 발현값을 갖는 유전자를 찾는 방법입니다.

Reference 서열 정보와 발현된 mRNA 서열을 분석한 fastq 파일이 필요하며 BWA, Bowtie, tophat 등의 linux 스크립트 기반 software들이 있으며 Bioconductor에서도 다양한 분석 툴이 소개되고 있습니다. 참고로 진핵세포의 경우 splicing 등을 고려한 mapping 기술이 필요합니다.

## 16.3 Creating a reference genome

```
library(gentools)
library(Biostrings)

download.file(url="https://github.com/greendaygh/kribbr2022/raw/main/ecoli-mg1655.gb",

ecoli <- readGenBank("examples/ecoli-mg1655.gb")
ecoliseq <- getSeq(ecoli)

maxlen <- 1000000
ecoliseqsub <- subseq(ecoliseq, 1, maxlen)
names(ecoliseqsub) <- "K-12"
writeXStringSet(ecoliseqsub, "examples/ecolisub.fasta")
```

## 16.4 Creation of indexing

Rsubread 패키지를 활용해서 align을 수행합니다. align 전에 reference genome의 index를 생성해야 하며 지늘이 클 경우 오랜 시간이 소요될 수 있으니 주의가 필요합니다.

```
if (!require("BiocManager", quietly = TRUE))
 install.packages("BiocManager")

BiocManager::install("Rsubread")
```

index 생성은 Rsubread 패키지의 buildindex 함수를 사용하며 memory 옵션을 이용해서 데이터 사이즈에 따라 할당되는 메모리를 조절할 수 있습니다. 기본은 8GB로 memory = 8000 입니다. indexSplit 옵션으로 크로모좀별로 인덱스를 분리해서 메모리 효율을 증가시킬 수 있습니다.

```
library(Rsubread)

buildindex(basename = file.path("examples", "ecoliexample"),
 reference = file.path("examples", "ecolisub.fasta"))
```

## 16.5 RNA-Seq alignment (Mapping)

Rsubread 패키지를 활용해서 mapping을 수행합니다. align 함수를 사용하며 splicing 여부에 따라 옵션이 조금씩 다를 수 있습니다. If the annotation is in GTF format, it can only be provided as a file. If it is in SAF format, it can be provided as a file or a data frame.

```
library(Rsubread)

alignstat <- align(file.path("examples", "ecoliexample")
 , readfile1 = file.path("examples", "filtered_SRR11549076_1.fastq_R1.fastq.gz")
 , output_file = file.path("examples", "ecoliexample.BAM")
 , nthreads = 6)

?alignstat

#alignstat
```

## 16.6 sorting

SAM 파일은 Sequence alignment data를 담고 있는 텍스트 파일(.txt)로 각 내용은 탭(tab)으로 분리되어 alignment, mapping 정보를 담고 있습니다. BAM 파일은 SAM 파일의 binary 버전으로 동일한 정보를 담고 있으며 이들 파일을 다루기 위해서는 SAMtools 소프트웨어가 필요합니다. R에서는 SAMtools의 R 버전인 Rsamtools 패키지를 활용할 수 있습니다.

```
library(Rsamtools)

sortBam(file = file.path("examples", "ecoliexample.BAM")
 , destination = file.path("examples", "sorted_ecoliexample.BAM"))

indexBam(files = file.path("examples", "sorted_ecoliexample.BAM.bam"))
```

## 16.7 visualization

IGV를 활용하여 mapping 파일 가시화가 가능합니다.

## 16.8 Counting in gene models

```
library(GenomicAlignments)
library(plyranges)

ecolicds <- cds(ecoli)
ecolicds_sub <- ecolicds %>%
 filter(end < maxlen)
seqlengths(ecolicds_sub) <- maxlen

mybam <- BamFile("examples/sorted_ecoliexample.BAM.bam", yieldSize = 100000)
myresult <- summarizeOverlaps(ecolicds_sub, mybam, ignore.strand = T)

class(myresult)

tmp <- assay(myresult)

rowRanges(myresult)
colData(myresult)
metadata(myresult)

?summarizeOverlaps
```

---

이 저작물은 크리에이티브 커먼즈 저작자표시-비영리-변경금지 4.0 국제 라이선스에 따라 이용할 수 있습니다.

# Chapter 17

## DEG Excercise

### 17.1 Counting with multiple RNAseq datasets

- <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE136101>
- [https://www.ncbi.nlm.nih.gov/Traces/study/?acc=PRJNA561298&o=acc\\_s%3Aa](https://www.ncbi.nlm.nih.gov/Traces/study/?acc=PRJNA561298&o=acc_s%3Aa) 사이트에서 metadata, accession list 파일 다운로드
- RNAseq 파일 다운로드

```
prefetch --option-file SRR_Acc_List.txt
• fastq 파일 가져오기
fastq-dump -X 100000 --split-files SRR10009019/SRR10009019.sralite
fastq-dump -X 100000 --split-files SRR10009020/SRR10009020.sralite
fastq-dump -X 100000 --split-files SRR10009021/SRR10009021.sralite
fastq-dump -X 100000 --split-files SRR10009022/SRR10009022.sralite
fastq-dump -X 100000 --split-files SRR10009023/SRR10009023.sralite
fastq-dump -X 100000 --split-files SRR10009024/SRR10009024.sralite
dirnames <- dir(file.path("examples", "ecoli"), pattern = "^\$SRR10")
```

- QC

```
library(Rfastp)

fastq_report <- rfastp(
 read1 = file.path("examples", "ecoli", "SRR10009019_1.fastq"),
 read2 = file.path("examples", "ecoli", "SRR10009019_2.fastq"),
 outputFastq = file.path("examples", "ecoli", "filtered_SRR10009019.fastq"))

fastq_report <- rfastp(
 read1 = file.path("examples", "ecoli", "SRR10009020_1.fastq"),
```

```

read2 = file.path("examples", "ecoli", "SRR10009020_2.fastq"),
outputFastq = file.path("examples", "ecoli", "filtered_SRR10009020.fastq"))

fastq_report <- rfastp(
 read1 = file.path("examples", "ecoli", "SRR10009021_1.fastq"),
 read2 = file.path("examples", "ecoli", "SRR10009021_2.fastq"),
 outputFastq = file.path("examples", "ecoli", "filtered_SRR10009021.fastq"))

fastq_report <- rfastp(
 read1 = file.path("examples", "ecoli", "SRR10009022_1.fastq"),
 read2 = file.path("examples", "ecoli", "SRR10009022_2.fastq"),
 outputFastq = file.path("examples", "ecoli", "filtered_SRR10009022.fastq"))

fastq_report <- rfastp(
 read1 = file.path("examples", "ecoli", "SRR10009023_1.fastq"),
 read2 = file.path("examples", "ecoli", "SRR10009023_2.fastq"),
 outputFastq = file.path("examples", "ecoli", "filtered_SRR10009023.fastq"))

fastq_report <- rfastp(
 read1 = file.path("examples", "ecoli", "SRR10009024_1.fastq"),
 read2 = file.path("examples", "ecoli", "SRR10009024_2.fastq"),
 outputFastq = file.path("examples", "ecoli", "filtered_SRR10009024.fastq"))

```

- Reference 서열 준비

```

library(genbankr)
library(Biostrings)

download.file(url="https://github.com/greendaygh/kribbr2022/raw/main/ecoli-mg1655.gb",

ecoli <- readGenBank("examples/ecoli-mg1655.gb")
ecoliseq <- getSeq(ecoli)

maxlen <- 1000000
ecoliseqsub <- subseq(ecoliseq, 1, maxlen)
writeXStringSet(ecoliseqsub, "examples/ecoli/ecolisub.fasta")

```

- index 생성

```

library(Rsubread)

buildindex(basename = file.path("examples", "ecoli", "ecolimedia"),
 reference = file.path("examples", "ecoli", "ecolisub.fasta"))

```

- mapping

```
alignstat <- align(file.path("examples", "ecoli", "ecolimedia")
 , readfile1 = file.path("examples", "ecoli", "filtered_SRR10009019.fastq_R1.f
 , readfile2 = file.path("examples", "ecoli", "filtered_SRR10009019.fastq_R2.f
 , output_file = file.path("examples", "ecoli", "ecolimedia_19.BAM")
 , nthreads = 6)

alignstat <- align(file.path("examples", "ecoli", "ecolimedia")
 , readfile1 = file.path("examples", "ecoli", "filtered_SRR10009020.fastq_R1.f
 , readfile2 = file.path("examples", "ecoli", "filtered_SRR10009020.fastq_R2.f
 , output_file = file.path("examples", "ecoli", "ecolimedia_20.BAM")
 , nthreads = 6)

alignstat <- align(file.path("examples", "ecoli", "ecolimedia")
 , readfile1 = file.path("examples", "ecoli", "filtered_SRR10009021.fastq_R1.f
 , readfile2 = file.path("examples", "ecoli", "filtered_SRR10009021.fastq_R2.f
 , output_file = file.path("examples", "ecoli", "ecolimedia_21.BAM")
 , nthreads = 6)

alignstat <- align(file.path("examples", "ecoli", "ecolimedia")
 , readfile1 = file.path("examples", "ecoli", "filtered_SRR10009022.fastq_R1.f
 , readfile2 = file.path("examples", "ecoli", "filtered_SRR10009022.fastq_R2.f
 , output_file = file.path("examples", "ecoli", "ecolimedia_22.BAM")
 , nthreads = 6)

alignstat <- align(file.path("examples", "ecoli", "ecolimedia")
 , readfile1 = file.path("examples", "ecoli", "filtered_SRR10009023.fastq_R1.f
 , readfile2 = file.path("examples", "ecoli", "filtered_SRR10009023.fastq_R2.f
 , output_file = file.path("examples", "ecoli", "ecolimedia_23.BAM")
 , nthreads = 6)

alignstat <- align(file.path("examples", "ecoli", "ecolimedia")
 , readfile1 = file.path("examples", "ecoli", "filtered_SRR10009024.fastq_R1.f
 , readfile2 = file.path("examples", "ecoli", "filtered_SRR10009024.fastq_R2.f
 , output_file = file.path("examples", "ecoli", "ecolimedia_24.BAM")
 , nthreads = 6)

#alignstat
```

- Sorting

```

library(Rsamtools)

sortBam(file = file.path("examples", "ecoli", "ecolimedia_19.BAM")
 , destination = file.path("examples", "ecoli", "sorted_ecolimedia_19.BAM"))

sortBam(file = file.path("examples", "ecoli", "ecolimedia_20.BAM")
 , destination = file.path("examples", "ecoli", "sorted_ecolimedia_20.BAM"))

sortBam(file = file.path("examples", "ecoli", "ecolimedia_21.BAM")
 , destination = file.path("examples", "ecoli", "sorted_ecolimedia_21.BAM"))

sortBam(file = file.path("examples", "ecoli", "ecolimedia_22.BAM")
 , destination = file.path("examples", "ecoli", "sorted_ecolimedia_22.BAM"))

sortBam(file = file.path("examples", "ecoli", "ecolimedia_23.BAM")
 , destination = file.path("examples", "ecoli", "sorted_ecolimedia_23.BAM"))

sortBam(file = file.path("examples", "ecoli", "ecolimedia_24.BAM")
 , destination = file.path("examples", "ecoli", "sorted_ecolimedia_24.BAM"))

indexBam(files = file.path("examples", "ecoli", "sorted_ecolimedia_19.BAM.bam"))
indexBam(files = file.path("examples", "ecoli", "sorted_ecolimedia_20.BAM.bam"))
indexBam(files = file.path("examples", "ecoli", "sorted_ecolimedia_21.BAM.bam"))
indexBam(files = file.path("examples", "ecoli", "sorted_ecolimedia_22.BAM.bam"))
indexBam(files = file.path("examples", "ecoli", "sorted_ecolimedia_23.BAM.bam"))
indexBam(files = file.path("examples", "ecoli", "sorted_ecolimedia_24.BAM.bam"))

```

- Counting

```

library(GenomicAlignments)
library(plyr)

ecolicds <- cds(ecoli)
ecolicds_sub <- ecolicds %>%
 filter(end < maxlen)
seqlengths(ecolicds_sub) <- maxlen

filelist <- c(file.path("examples", "ecoli", "sorted_ecolimedia_19.BAM.bam"),
 file.path("examples", "ecoli", "sorted_ecolimedia_20.BAM.bam"),
 file.path("examples", "ecoli", "sorted_ecolimedia_21.BAM.bam"),
 file.path("examples", "ecoli", "sorted_ecolimedia_22.BAM.bam"),
 file.path("examples", "ecoli", "sorted_ecolimedia_23.BAM.bam"),
 file.path("examples", "ecoli", "sorted_ecolimedia_24.BAM.bam")
)
mybam <- BamFileList(filelist)

```

```

myresult <- summarizeOverlaps(ecolicds_sub, mybam, ignore.strand = T)

class(myresult)
assay(myresult)
rowRanges(myresult)
colData(myresult)
metadata(myresult)

• Log transform and normalization
• 가정: 샘플마다 RNA 발현 정도는 유사하며 총 량은 같음.

library(tidyverse)

mydata <- assay(myresult)
mygenes <- rowRanges(myresult)
boxplot(mydata)

mydatat <- mydata %>%
 data.frame() %>%
 rownames_to_column() %>%
 pivot_longer(-rowname) %>%
 pivot_wider(names_from = rowname, values_from = value)

mymeanval <- mydatat %>%
 summarise(across(where(is.numeric), mean))
mysdval <- mydatat %>%
 summarise(across(where(is.numeric), sd))

mydf <- mymeanval %>%
 bind_rows(mysdval) %>%
 as.data.frame
rownames(mydf) <- c("mean", "sd")

mydft <- mydf %>%
 rownames_to_column() %>%
 pivot_longer(-rowname) %>%
 pivot_wider(names_from = rowname, values_from = value)

ggplot(mydft, aes(x=mean, y=sd)) +
 geom_point() +
 scale_x_log10() +
 scale_y_log10()

```

- installation DESeq2

```
if (!require("BiocManager", quietly = TRUE))
 install.packages("BiocManager")
```

```
BiocManager::install("DESeq2")
```

- create DESeq2 object

```
library(DESeq2)
```

```
metaData <- data.frame(Group = c("control", "control", "control", "case", "case", "case"))
metaData
```

```
dds <- DESeqDataSetFromMatrix(countData = mydata,
 colData = metaData,
 design = ~Group,
 rowRanges = mygenes)
```

```
dds2 <- DESeq(dds)
```

```
counts(dds2)
```

```
counts(dds2, normalized = T)
```

- Mean variance

```
plotDispEsts(dds2)
```

```
?plotDispEsts
```

- DESeq results

```
myres <- results(dds2, contrast = c("Group", "control", "case"))
myres
```

```
summary(myres)
```

```
plotMA(myres)
```

- Multiple testing

- Bonferroni = pvalue \* total genes tested

- Benjamini-Hockberg = (Pvalue\*total genes)/rank of pvalue