

2강

# 데이터 요약 I

통계·데이터과학과 박서영 교수

# 목차

- 1 변수
- 2 질적 데이터의 요약 - 막대그래프
- 3 양적 데이터의 요약 - 히스토그램
- 4 양적 데이터의 요약 - 점도표, 평균, 분산
- 5 R 패키지 설치

01

# 변수

- ▶ 변수: 각 단위에 대해 관측되는 특성
- ▶ 데이터: 하나 이상의 변수에 대한 관찰값의 모음

ID	성별	나이	학력	몸무게
1	여	42	고졸	65.0kg
2	남	38	대졸	72.3kg
3	남	25	대학원졸	81.1kg
4	여	51	고졸	58.9kg

# 변수의 종류

- ▶ 질적 변수(qualitative variable, 범주형 변수): 유한개의 범주 중 하나의 값을 취하는 변수
- ▶ 양적 변수(quantitative variable): 양적인 수치로 측정되는 변수

ID	성별	나이	학력	몸무게
1	여	42	고졸	65.0kg
2	남	38	대졸	72.3kg
3	남	25	대학원졸	81.1kg
4	여	51	고졸	58.9kg

# 변수의 종류

## ▶ 질적 변수의 종류

- 명목형 변수(nominal variable): 범주들에 의미 있는 순서를 정할 수 없는 질적 변수
- 순서형 변수(ordinal variable): 범주 간의 의미 있는 순서를 정할 수 있는 질적 변수

ID	성별	나이	학력	몸무게
1	여	42	고졸	65.0kg
2	남	38	대졸	72.3kg
3	남	25	대학원졸	81.1kg
4	여	51	고졸	58.9kg

# 변수의 종류

## ▶ 양적 변수의 종류

- 연속형 변수 (continuous variable): 어떤 실수 구간 안의 모든 값을 가질 수 있는 변수
- 이산형 변수 (discrete variable): 취할 수 있는 값을 셀 수 있는 양적 변수

ID	성별	나이	학력	몸무게
1	여	42	고졸	65.0kg
2	남	38	대졸	72.3kg
3	남	25	대학원졸	81.1kg
4	여	51	고졸	58.9kg

# 변수의 분포

- ▶ 변수의 데이터에는 변동(variability)이 있다
- ▶ 변수의 분포: 변수가 취할 수 있는 모든 값에 대해 각 값이 발생하는 빈도를 나열한 것
- ▶ 도수분포표(frequency table): 데이터에서 각 값의 출현빈도나 비슷한 값끼리 묶은 구간별로 관측된 데이터의 개수를 정리한 표



# 도수분포표 예제

## ▶ 한 학급의 학생들의 혈액형 분포

혈액형	학생 수
A형	10
B형	8
AB형	3
O형	9

## ▶ 한 학급의 학생들의 키 분포

키(cm)	학생 수
150 이상 160 미만	4
160 이상 170 미만	11
170 이상 180 미만	13
180 이상 190 미만	2

# 도수분포표 만드는 법

- ▶ 질적변수의 경우: 각 범주에 속하는 단위의 개수를 제시
- ▶ 양적변수의 경우: 계급을 정한 후 각 계급에 속하는 단위의 개수를 제시
  - 계급은 임의로 정할 수 있으나, 각 계급의 폭을 일정하게 하는 것이 좋다
  - 계급의 폭이 너무 좁으면: 계급의 개수가 너무 많아지거나 각 계급의 도수가 너무 작아진다
  - 계급의 폭이 너무 넓으면: 전체적인 분포가 잘 드러나지 않을 수도 있다
  - 각 계급의 경계점에 놓이는 관찰값의 개수가 적어지도록 계급을 정하는 것이 좋다

# 도수분포표의 계급

## 한 학급의 학생들의 키 분포

키(cm)	학생 수
150 이상 160 미만	4
160 이상 170 미만	11
170 이상 180 미만	13
180 이상 190 미만	2

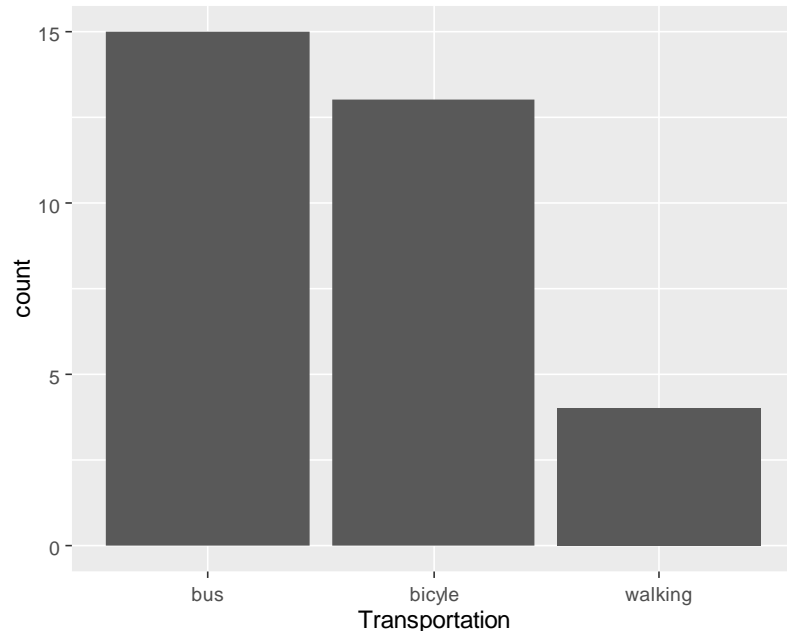
키(cm)	학생 수
150 이상 170 미만	15
170 이상 190 미만	15

키(cm)	학생 수
150 이상 153 미만	1
153 이상 156 미만	0
156 이상 159 미만	3
159 이상 162 미만	0
162 이상 165 미만	4
165 이상 168 미만	3
168 이상 171 미만	4
171 이상 174 미만	5
174 이상 177 미만	5
177 이상 180 미만	3
180 이상 183 미만	1
183 이상 186 미만	1

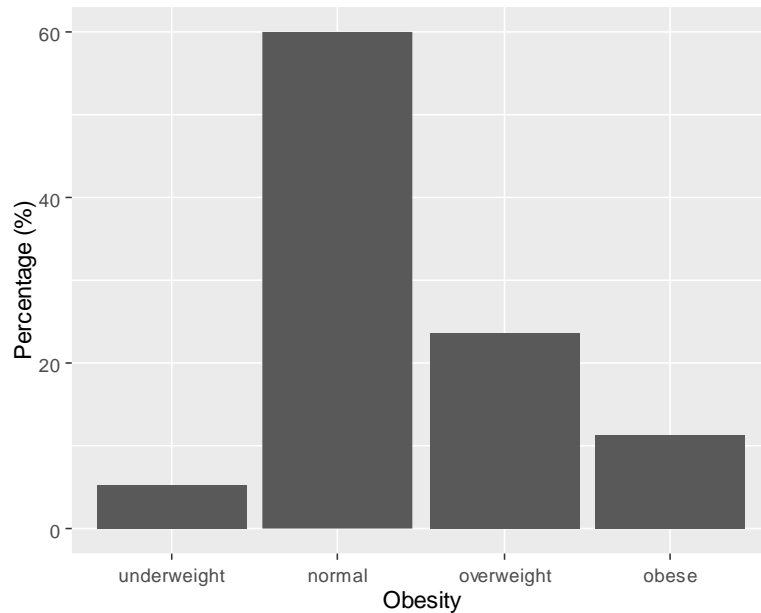
02

# 질적 데이터의 요약

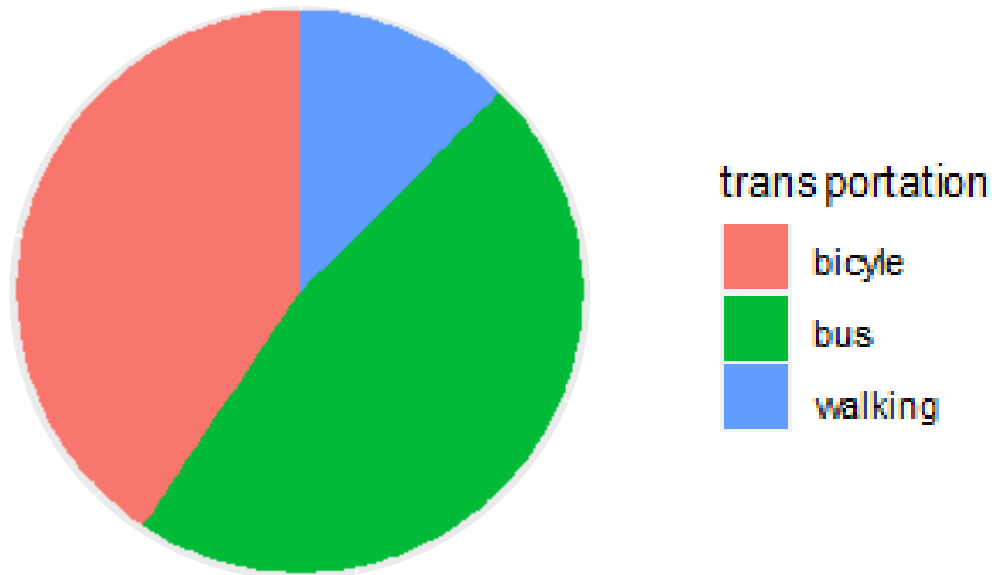
- ▶ 각 범주에 속한 관찰값의 개수 또는 비율을 막대의 길이로 나타낸 그래프
  - 명목형 변수일 때: 큰 빈도부터 작은 빈도, 또는 작은 빈도부터 큰 빈도 순서로 정렬하면 좋다
    - 예제 2-3: 어느 학급 학생들의 등하교 교통수단



- ▶ 각 범주에 속한 관찰값의 개수 또는 비율을 막대의 길이로 나타낸 그래프
  - 순서형 변수일 때: 범주의 순서를 지켜서 그리는 것이 좋다
    - 예제 2-4: 어느 의원 환자들의 비만도 분포



- ▶ 각 범주에 속한 관찰값의 비율의 원의 면적으로 표현한 그래프
- ▶ 막대그래프에 비해서 정보 파악이 어렵기 때문에, 최근에는 선호되지 않는다
  - 예제 2-5: 어느 학급 학생들의 등하교 교통수단



03

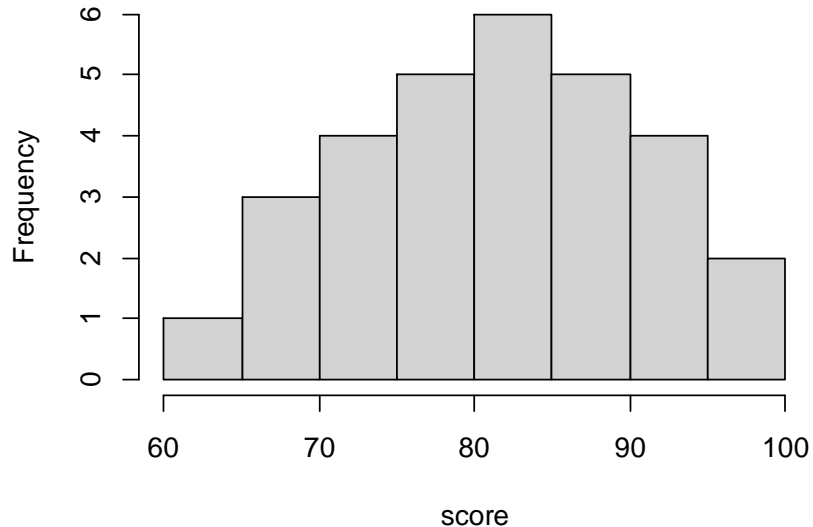
# 양적 데이터의 요약 - 히스토그램



- ▶ 히스토그램, 점도표, 상자그림
- ▶ 평균, 표준편차, 분산
- ▶ 중앙값, 사분위수 범위

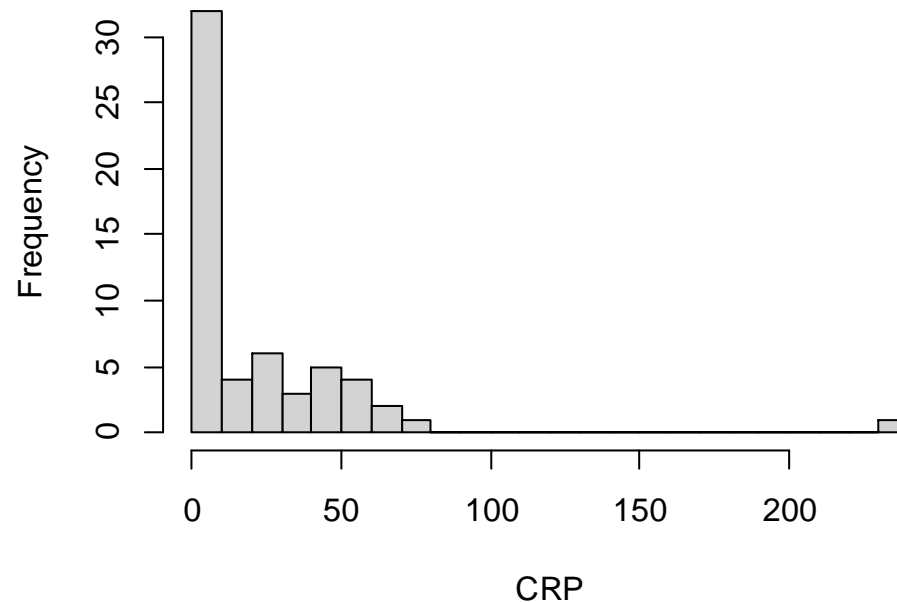
# 히스토그램(histogram)

- ▶ 도수분포표를 그래프로 나타낸 것
- ▶ 계급을 수평축에 표시
- ▶ 각 계급의 도수에 비례하는 넓이의 직사각형
  - 예제 2-6: 어느 학급의 영어점수 분포를 나타낸 히스토그램

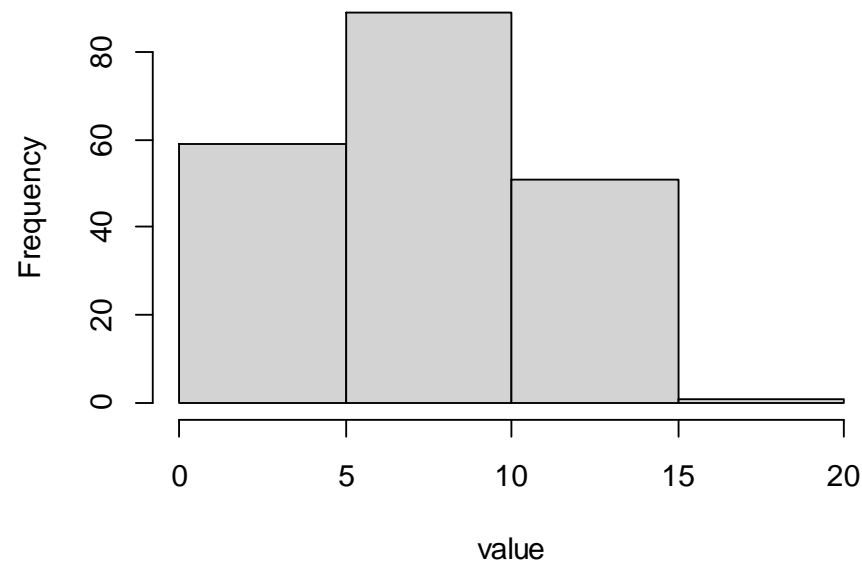
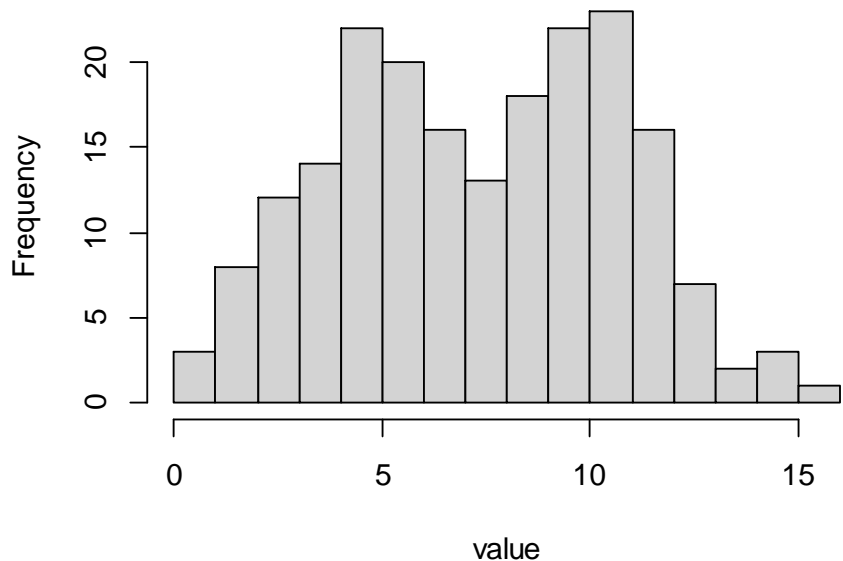


# 히스토그램과 특이점

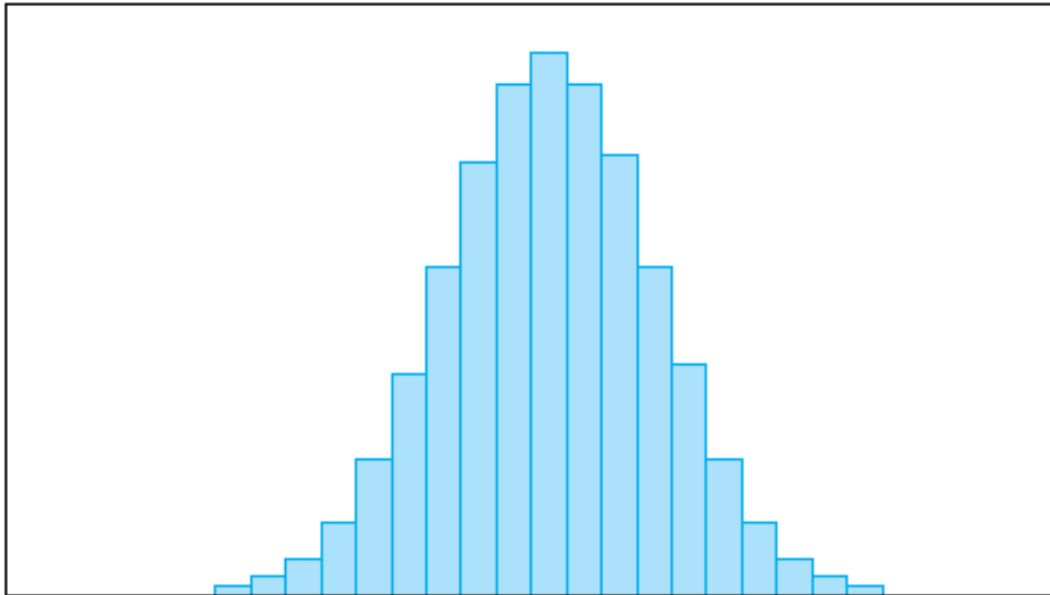
- ▶ 히스토그램을 이용하면 특이점을 쉽게 찾을 수 있다
- ▶ 특이점(outlier): 대부분의 데이터로부터 멀리 떨어져 있는 관찰값
  - 예제 2-7: 어느 의원 환자의 C-반응 단백질의 분포를 나타낸 히스토그램



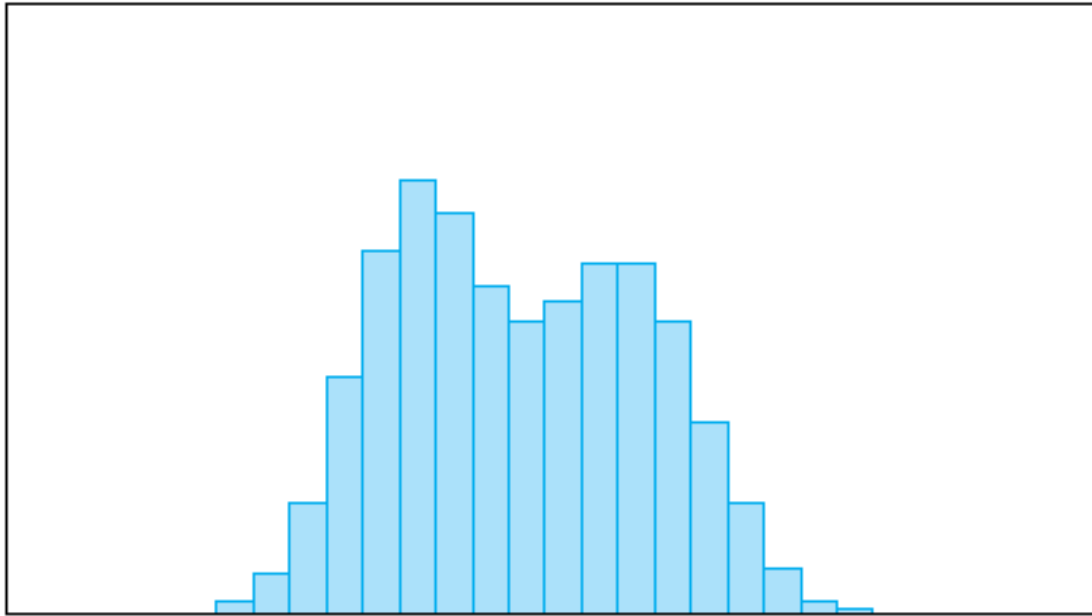
- ▶ 히스토그램을 이용하면 전체적인 분포를 한눈에 파악할 수 있다
- ▶ 주의점: 같은 데이터라도 계급의 폭에 따라 분포의 특성이 달라보일 수 있다
  - 예제 2-8: 같은 데이터, 다른 계급 폭



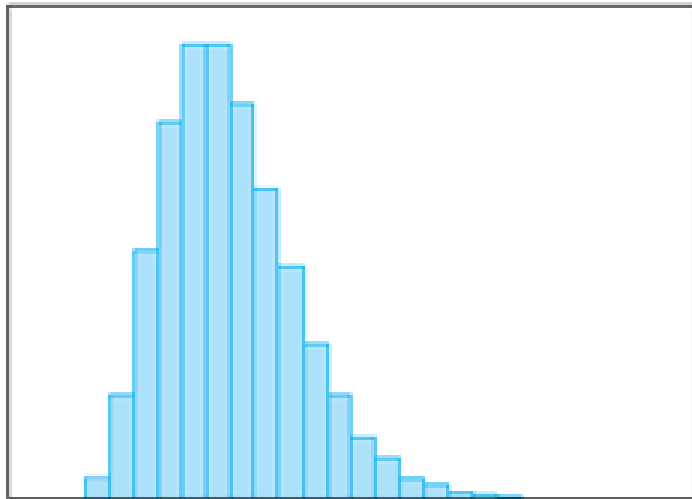
- ▶ 종 모양 분포(bell-shaped distribution): 좌우 대칭이고 데이터가 가운데에 모여있다



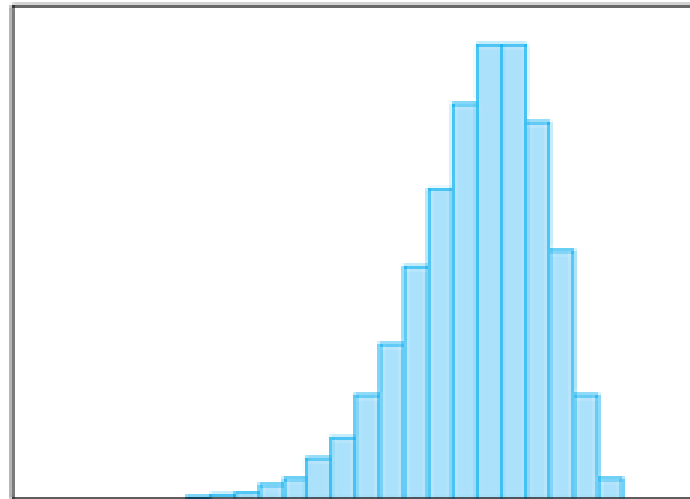
- ▶ 쌍봉우리형 분포(bimodal distribution): 2개의 봉우리 주변으로 데이터가 모여있는 분포



- ▶ 치우친 분포(skewed distribution): 비대칭으로 한쪽 꼬리가 다른 쪽 꼬리보다 긴 분포.
  - 왼쪽으로 치우친 (right-skewed) 분포: 오른쪽 꼬리가 더 길다
  - 오른쪽으로 치우친 (left-skewed) 분포: 왼쪽 꼬리가 더 길다

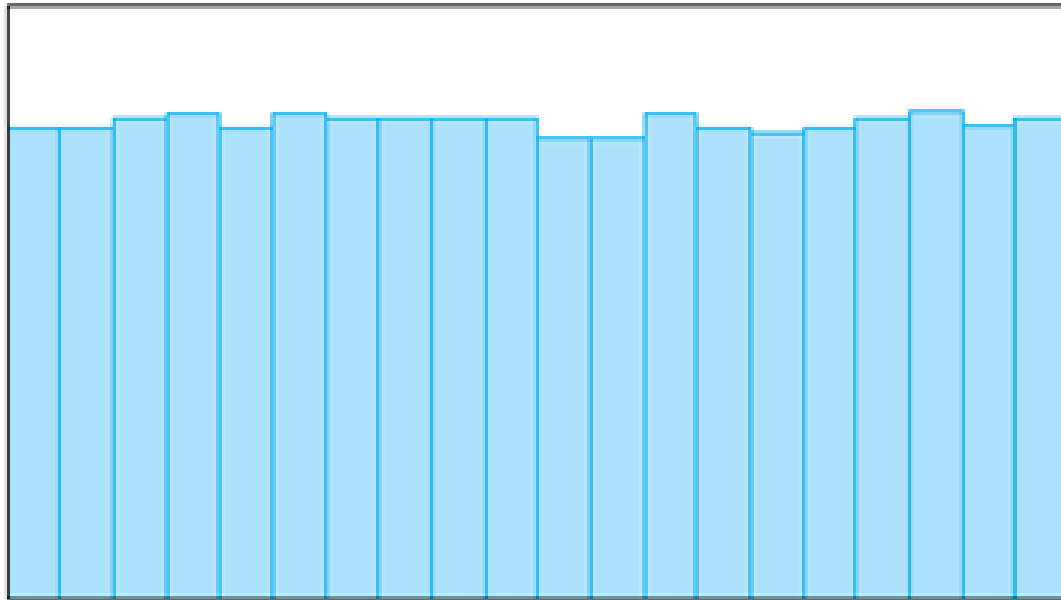


왼쪽으로 치우친 분포



오른쪽으로 치우친 분포

- ▶ 균등분포(uniform distribution): 어떤 범위 내의 값이 고르게 나타나는 분포



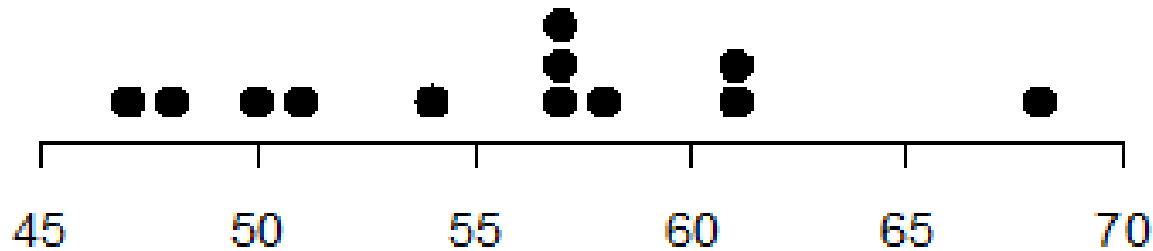


04

# 양적 데이터의 요약 - 점도표, 평균, 분산

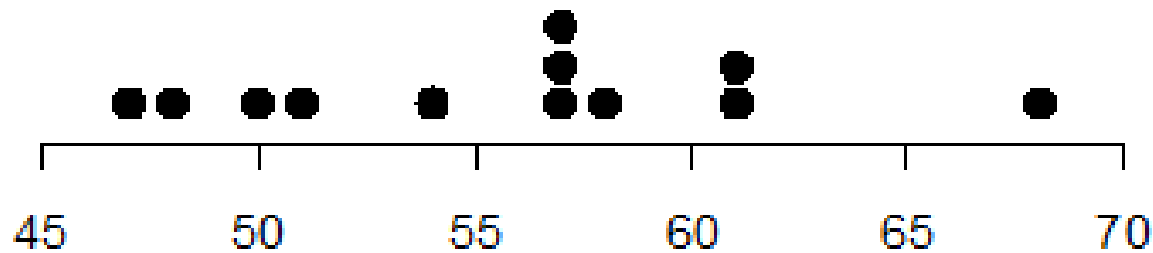
- ▶ 수평선 위에 데이터 값에 해당하는 위치에 점을 찍는 그래프
- ▶ 데이터가 작을 때 유용하다
- ▶ 관찰값의 개수가 20~30개를 넘어가면 너무 복잡해진다
  - 예제: 어느 봉사단체 회원들의 연령을 나타낸 점도표

데이터: 57, 61, 45, 57, 48, 58, 57, 61, 54, 50, 68, 51



- ▶ 양적 데이터의 관찰값들을 대표하는 수치는 무엇일까?
- ▶ 데이터의 퍼진 정도를 나타내는 수치는 무엇일까?

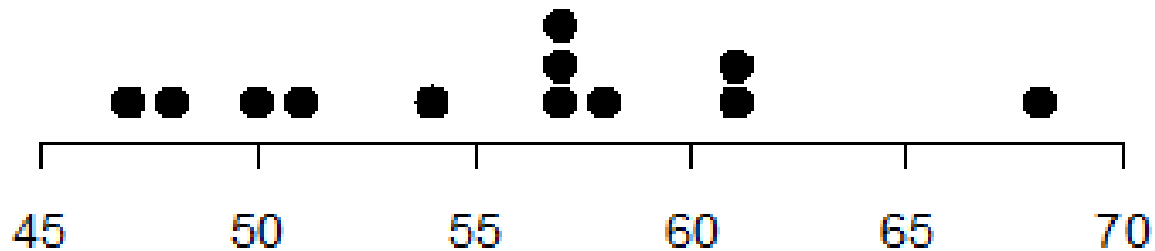
데이터: 57, 61, 45, 57, 48, 58, 57, 61, 54, 50, 68, 51

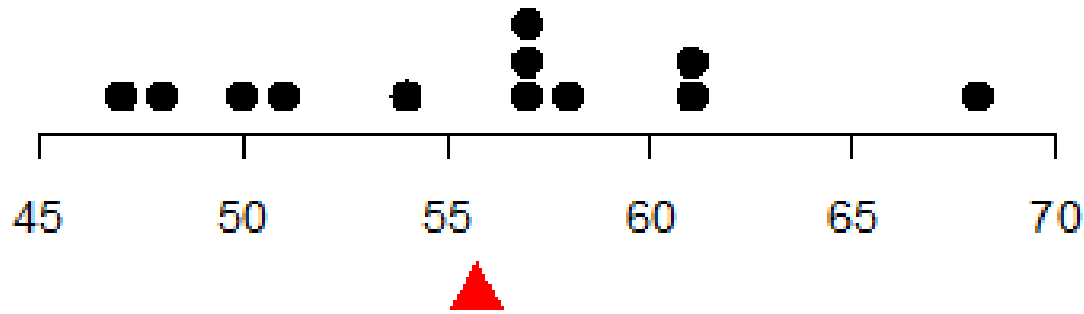


# 최빈값(mode)

- ▶ 관찰값 중에서 발생빈도가 가장 높은 값
- ▶ 여러개일 수도 있고, 하나도 없을 수도 있다

데이터: 57, 61, 45, 57, 48, 58, 57, 61, 54, 50, 68, 51





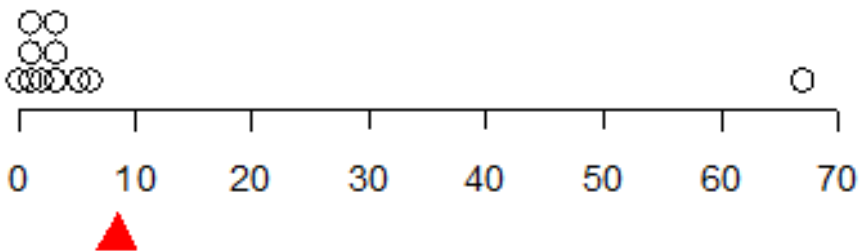
- ▶ 점도표를 시소 위에 물체가 놓여있는 것으로 생각하면, 시소가 평형을 이루는 무게 중심의 위치가 데이터를 대표한다고 생각할 수 있다
- ▶ 평균(mean): 양적 변수의 분포의 균형을 이루는 무게중심의 위치에 해당하는 값

- ▶ 양적 변수의 분포의 균형을 이루는 무게중심의 위치에 해당하는 값
- ▶ 어떤 변수의 관찰값의 총합을 관찰값의 개수로 나눈 값
- ▶ 표본 크기가  $n$ 인 표본 데이터의 관찰값을  $x_1, x_2, \dots, x_n$ 이라고 할 때, 표본 평균은  $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$  이다

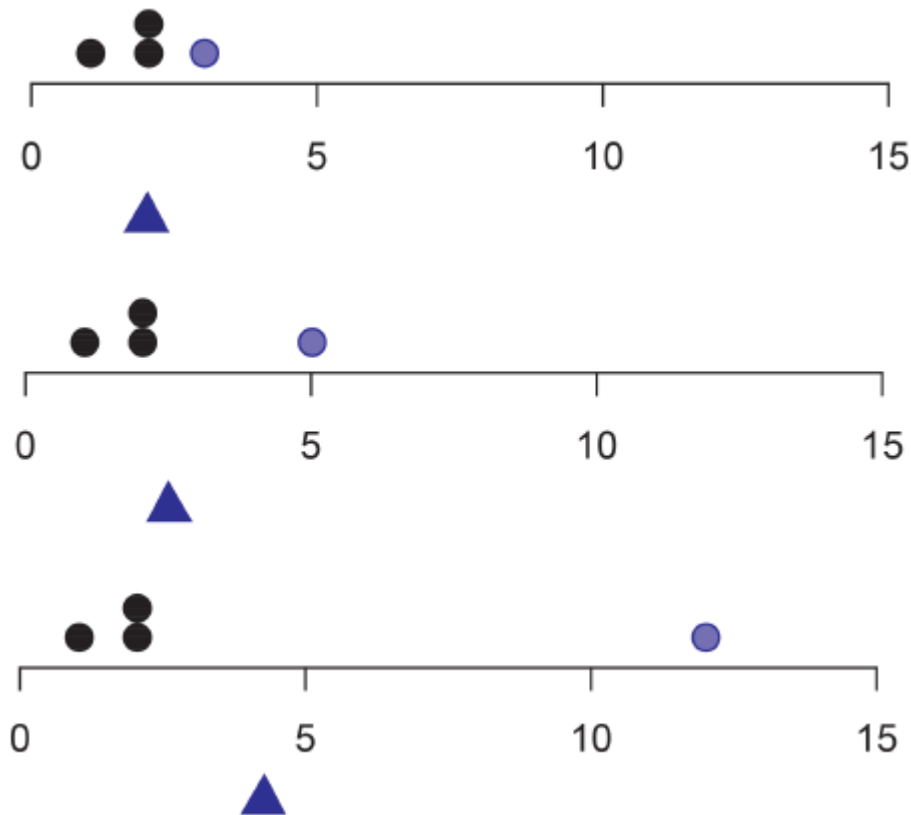
- ▶ 표본데이터가 기울어진 분포를 가졌거나 특이점이 있는 경우, 평균이 데이터 전체를 잘 대표하지 못한다
- ▶ 특이점의 영향을 크게 받는다
  - 예제: 어떤 학급의 각 학생이 한달 동안 읽은 책 수

데이터: 6, 0, 1, 3, 1, 5, 2, 3, 1, 3, 67

$$\text{평균} = \frac{6+0+1+3+1+5+2+3+1+3+67}{11} = 8.36$$



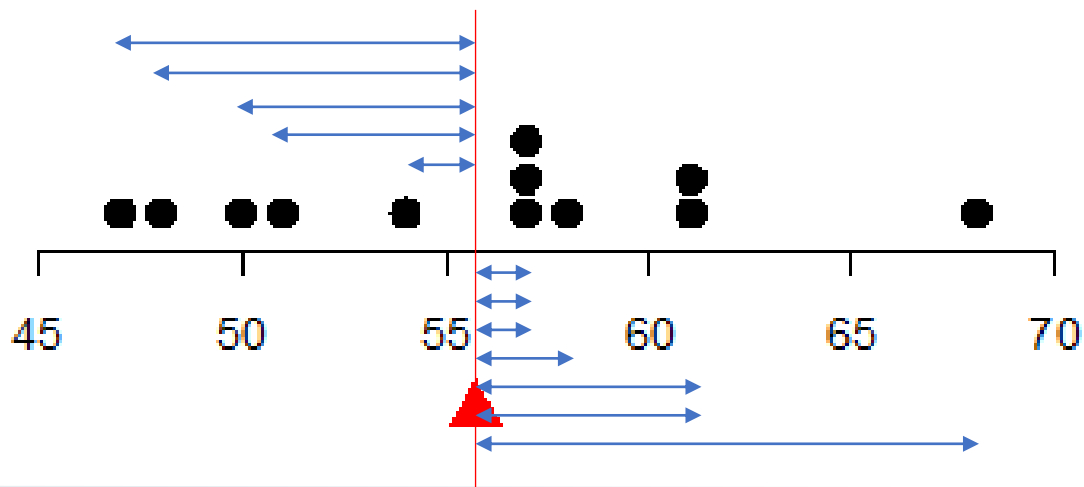
- ▶ 데이터의 분포가 좌우 대칭인 경우 평균은 분포의 가운데에 위치한다
- ▶ 데이터 중 하나라도 한쪽으로 치우치면 평균은 특이점 쪽으로 움직이게 된다





# 분산과 표준편차

- ▶ 편차: 관찰값 - 평균
- ▶ 분산(variance): 편차의 제곱의 평균
  - 표본분산  $s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$
- ▶ 표준편차(standard deviation): 분산의 제곱근
  - 표본표준편차  $s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$



# 분산과 표준편차 예제 2-11

학생 10명이 1분당 할 수 있는 윗몸일으키기 개수가 다음과 같다. 표본분산과 표본표준편차를 구하시오.

25, 41, 35, 8, 52, 23, 32, 37, 42, 28

▶ 평균  $\bar{x} = \frac{25+41+35+8+52+23+32+37+42+28}{10} = 32.3$

▶ 편차와 편차의 제곱

관찰값 ( $x_i$ )	편차 ( $x_i - \bar{x}$ )	편차제곱 ( $(x_i - \bar{x})^2$ )
25	-7.3	53.29
41	8.7	75.69
...	...	...
28	-4.3	18.49
계	0.0	1336.1

▶ 표본분산  $s^2 = \frac{1336.1}{10-1} = 148.5$ , 표본표준편차  $s = \sqrt{148.5} = 12.2$

## 분산과 표준편차

- ▶ 분산, 표준편차가 크면 데이터가 평균을 중심으로 광범위하게 분포되어 있다는 뜻
- ▶ 분산, 표준편차가 작으면 데이터가 평균을 중심으로 조밀하게 모여 있다는 뜻
- ▶ 분산, 표준편차는 특이점의 영향을 많이 받는다
- ▶ 분산의 단위 = 데이터 측정단위의 제곱
- ▶ 표준편차의 단위 = 데이터 측정단위

## 변이계수(coefficient of variation)

- ▶ 변수 2개 이상의 변동을 비교할 때 분산이나 표준편차를 비교하면 공평한 비교일까?
- ▶ 예) 두부 가격의 표준편차 ≪ 아파트 가격의 표준편차
- ▶ 변동을 비교할 때는 측정 단위나 데이터 중심위치의 차이를 고려해야한다
- ▶ 변이계수: 표준편차를 평균으로 나눈 값

- ▶ 다음은 만 21세 남자 그룹과 만 9세 남아 그룹의 체중의 평균과 표준편차이다. 어느 그룹의 체중의 변동이 더 크다고 할 수 있는가?

	평균	표준편차
만 21세 남자	72kg	11kg
만 9세 남자	32kg	7kg

- 만 21세 남자 변이계수 =  $11/72 = 0.153$
- 만 9세 남자 변이계수 =  $7/32 = 0.219$

05

# R 패키지 설치

- ▶ R 자체에 내장되지 않은, 사용자들이 개별적으로 만들어낸 함수들의 모음
- ▶ 누구나 새로운 패키지를 만들어서 공유할 수 있다
- ▶ CRAN(<https://cran.r-project.org>)에서 Packages>Table of available packages, sorted by name 선택하면 공개된 모든 패키지를 볼 수 있다

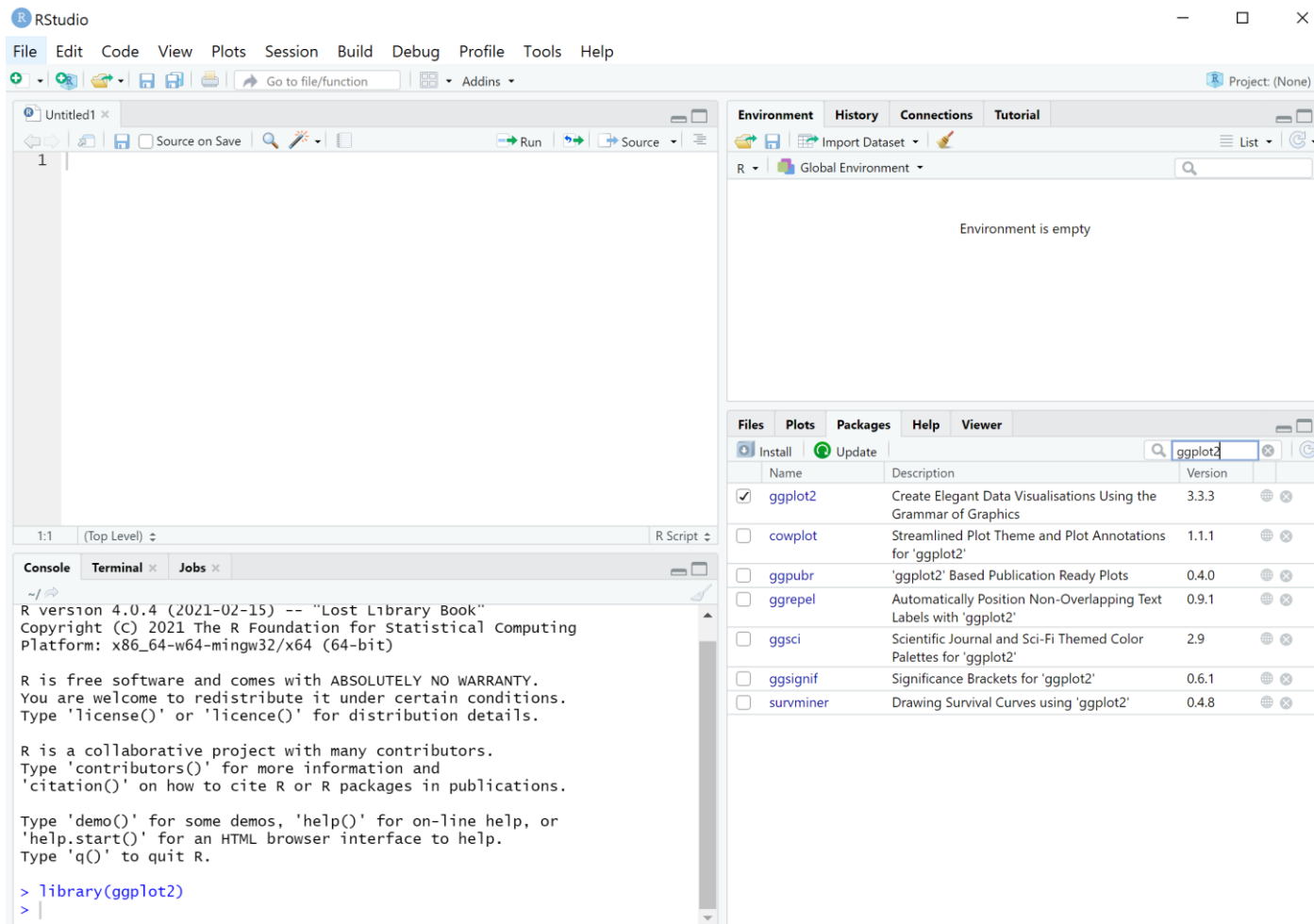
- ▶ Wilkinson의 The grammar of graphics의 원칙에 따라 그래프를 만들 수 있는 함수들의 모음
- ▶ 기본 구조에 레이어를 추가하는 방식으로 원하는 그래프의 형태를 지정한다
- ▶ 디테일을 상세하게 지정하지 않아도 자동으로 예쁜 그래프를 그려준다



# R 패키지 설치하는 법 1

## R 패키지 설치

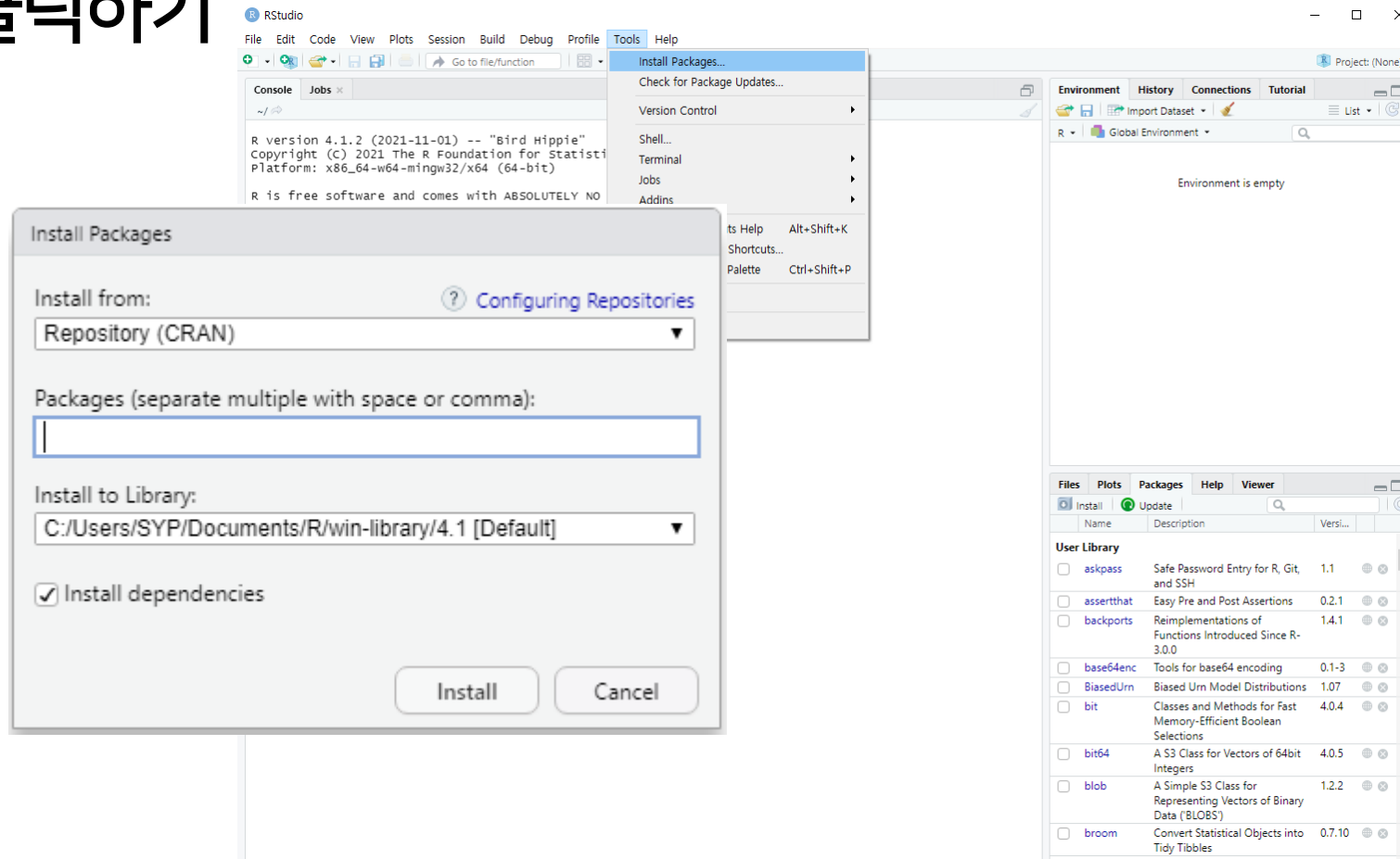
- RStudio의 오른쪽 아래 Packages 창에서 원하는 패키지 이름 검색 후, 체크박스 선택하고 Install 클릭하기



# R 패키지 설치하는 법 2

## R 패키지 설치

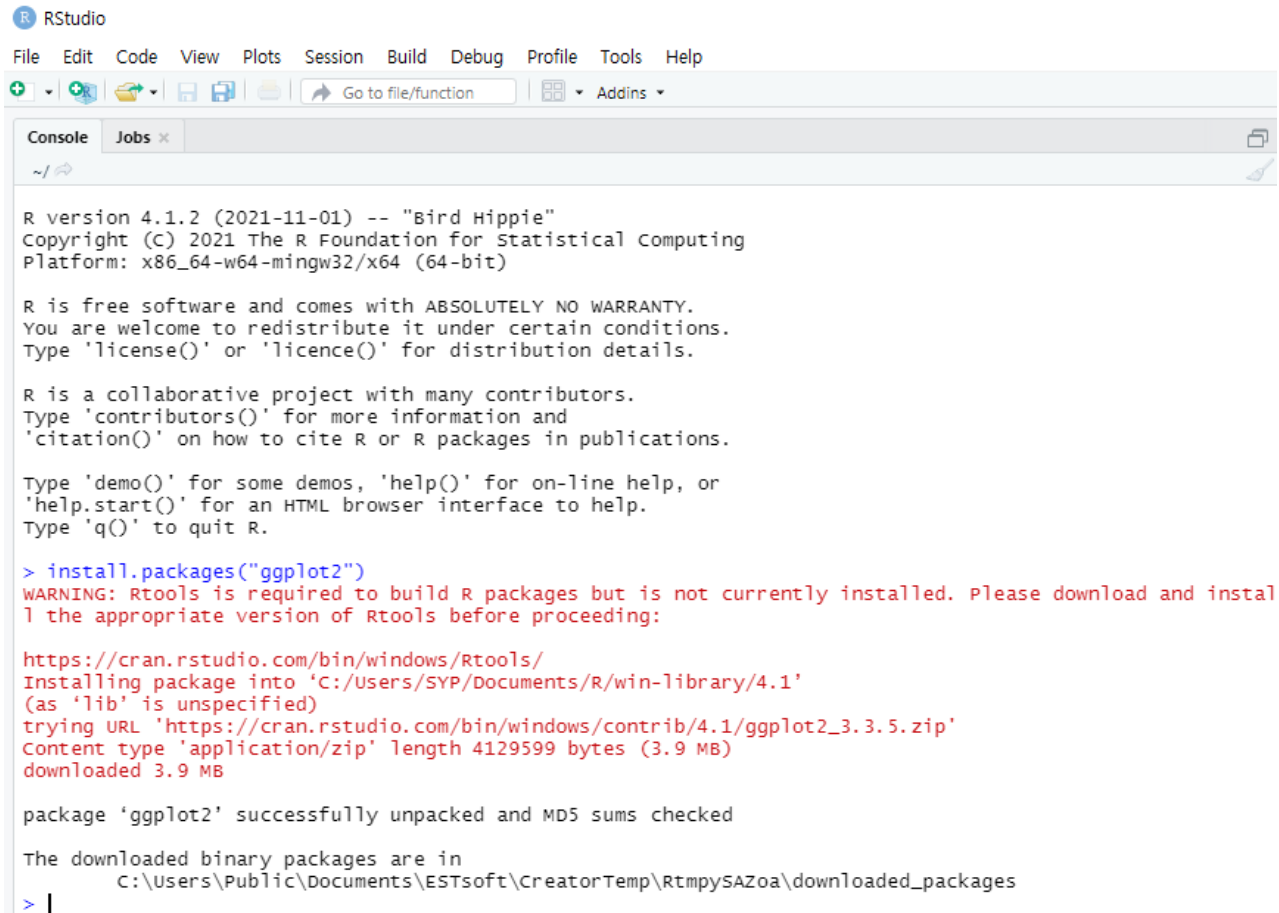
- ▶ RStudio의 위쪽 메뉴에서 Tools>Install Packages 메뉴를 선택하고 대화창에 원하는 패키지 이름 입력, Install 클릭하기



# R 패키지 설치하는 법 3

## R 패키지 설치

### ▶ 콘솔에 `install.packages("원하는 패키지 이름")` 입력



RStudio interface showing the console output for installing the `ggplot2` package. The console displays the R version (4.1.2), copyright information, and a warning that `Rtools` is required for building R packages. The installation process is shown, including the URL, package path, and successful unpacking.

```
R version 4.1.2 (2021-11-01) -- "Bird Hippie"
Copyright (c) 2021 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> install.packages("ggplot2")
WARNING: Rtools is required to build R packages but is not currently installed. Please download and install
the appropriate version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/
Installing package into 'C:/Users/SYP/Documents/R/win-library/4.1'
(as 'lib' is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.1/ggplot2_3.3.5.zip'
Content type 'application/zip' length 4129599 bytes (3.9 MB)
downloaded 3.9 MB

package 'ggplot2' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
C:\Users\Public\Documents\ESTsoft\CreatorTemp\RtmpySAZoa\downloaded_packages
> |
```

- ▶ 패키지를 설치한 후, 반드시 '로드(load)'해야 사용할 수 있다
- ▶ 로드하는 명령어: `library(ggplot2)`
- ▶ 한번 설치한 패키지는 (일부러 지우거나 R을 업그레이드 하지 않는 한) 없어지지 않으므로 재설치가 필요없다
- ▶ 한번 로드한 패키지는 RStudio를 닫으면 주기억장치에서 사라진다. 따라서 RStudio를 닫았다가 다시 열 경우, 필요한 패키지를 다시 로드해야한다
  - 따라서 패키지를 로드하는 명령어를 스크립트에 저장하는 것이 좋다

## ▶ 기본 형태

```
ggplot(data = <DATA>) +  
  <GEOM_FUNCTION>(mapping = aes(<MAPPINGS>))
```

- ▶ `ggplot()` 은 먼저 자료의 좌표축을 만든다
- ▶ geom function은 `mapping = aes()` 구문을 통해 x축과 y축 변수를 지정한다
- ▶ 그래프의 종류에 따라 다른 geom function을 사용한다
- ▶ 주의: "+"는 항상 라인의 마지막에 위치해야 한다

## 교재 예제 2-3의 막대그래프

```
library(ggplot2)
library(forcats)

transp<-c("bicycle", "bus", "bus", "walking", "bus", "bicycle", "bicycle",
          "bus", "bus", "bus", "bicycle", "bus", "bicycle", "bicycle", "walking",
          "bus", "bus", "bicycle", "bicycle", "walking", "walking",
          "bicycle", "bus", "bus", "bus", "bus", "bus", "bicycle",
          "bus", "bus", "bicycle", "bicycle", "bicycle")

dat1<-data.frame(transp)

ggplot(data=dat1) + geom_bar(mapping=aes(x=transp)) + xlab("Transportation")

ggplot(data=dat1) + geom_bar(mapping=aes(x=fct_infreq(transp))) +
xlab("Transportation")
```

## 교재 예제 2-4의 막대그래프

```
obesity<-factor(c("underweight", "normal", "overweight", "obese"),  
               levels=c("underweight", "normal", "overweight", "obese"))  
count<-c(6, 69, 27, 13)  
perc<-count/sum(count)*100  
dat2<-data.frame(obesity, count, perc)  
  
ggplot(data=dat2) + geom_bar(mapping=aes(x=obesity, y=perc),  
stat="identity") + xlab("Obesity") + ylab("Percentage (%)")
```

## 교재 예제 2-5의 원그래프

```
table(transp)
dat3<-data.frame(transportation=c("bus", "bicycle", "walking"), count=c(15, 13, 4))

ggplot(data=dat3) + geom_bar(mapping=aes(x="", y=count, fill=transportation),
stat="identity") +
    coord_polar("y", start=0) + xlab("") + ylab("")

ggplot(data=dat3) + geom_bar(mapping=aes(x="", y=count, fill=transportation),
stat="identity") +
    coord_polar("y", start=0) + xlab("") + ylab("") +
    theme(axis.text = element_blank(),
          axis.ticks = element_blank(),
          panel.grid = element_blank())
```



# 정리하기

- 변수는 질적 변수와 양적 변수로 나뉜다. 질적 변수에는 명목형 변수, 순서형 변수가 있고, 양적 변수에는 연속형 변수와 이산형 변수가 있다.
- 변수의 분포를 나타내기 위하여 각 값의 출현빈도나 비슷한 값끼리 묶은 구간별로 관측된 데이터의 개수를 정리한 표를 도수분포표라고 한다.
- 막대그래프, 히스토그램, 점도표를 이용하여 데이터를 요약할 수 있다.
- 특이점은 대부분의 데이터로부터 멀리 떨어져있는 관찰값이다.
- 평균은 분포의 무게 중심으로서 관찰값의 총합을 표본크기로 나눈 값이다. 분산은 편차의 제곱의 평균이고, 표준편차는 분산의 제곱근이다.

## 3강

# 다음시간안내

# 데이터 요약 II