통계처리입문

Week02

통계학을 공부하는 이유

- 통계학을 적용한 자료는 신뢰성을 가진다.

- 통계는 실생활에서 광범위하게 활용된다.
- 우리가 접하는 대부분의 정보는 통계학을 통해 가공되어 전달된다.

국내 최고의 미각소유자 vs. 신뢰할 만한 기관

→ 통계는 사회에서 발생할 수 있는 다양한 상황에서
신뢰할 수 있는 자료를 가공해 내고 이를 활용

통계학을 공부하는 이유

- 통계는 의사결정에 필요한 근거 자료를 제시한다.

- 인문/사회과학의 연구(조사)나 실험결과는 다양한 결과를 도출
- 일상생활이나 현상 등을 수치화하기 위한 기준을 토대로 조사와 분석에서의 과학적 접근

획기적 신제품의 홍보방법은 어떻게 해야 할까?

→ 비용 대비 최대의 효과를 얻을 수 있는 의사결정 가능

통계학을 공부하는 이유

- 통계는 현상을 분석하여 실증자료를 제시한다.

- 현상을 분석하여 문제의 해결을 위한 다양한 원인을 찾을 수 있도록 자료제공

(이론) 인간의 소비행동은 자신의 불만족을 만족으로 바꾸려는 행동
Ex. 소비자의 스마트폰 선택 기준은? (디자인, 편의성, 유용성)

→ 3가지 원인에 대해 선택과 집중을 해야 한다면 어디에?

통계학의 정의와 목적

- 통계학의 정의

통계학(statistics)은

수량적인 비교를 기초로 많은 사실을 관찰하고 처리하는 방법을 연구하는 학문

일반적으로 수집되는 데이터가 조사자, 시기, 방법, 목적 등에 따라 다르게 나타나는 불균형적인 데이터이지만, 통계학은 이 안에서 의미를 찾아내고, 실생활에서 적용 가능한 유용성을 찾아내 이를 수치로 표현할 수 있도록 한다.

기술통계(descriptive statistics) :

표본에 대한 분석 결과의 각종 수치들을 활용하여 집단의 특성을 설명

추론통계(inference statistics) : 표본을 활용하여 모집단의 특성을 나타내는 것

통계학의 정의와 목적

- 통계학의 목적

- 의사결정

의사결정은 많은 정보를 지각하고 평가하여 하나를 선택하는 것

∵ 정보와 반응 사이의 다대일 대응으로 나타나므로

→ 여러 가지 대안 가운데 하나를 선택할 때, 기초자료를 제공

- 불확실성의 해소

의사결정을 하게 되면 그 결과가 정확한 것이라 할 수 있는가의 문제

∴ 빅데이터의 개념을 들여와 불확실성을 해소하려는 노력

→ 정보수집의 어려움, 시장의 변화와 대응의 어려움에 대한 극복

통계학의 정의와 목적

- 통계학의 목적

- 요약

다양한 데이터를 신속히 이해할 수 있도록 다양한 형태로 표현

∵ 불확실성의 감소를 위해

→ 반복되어 생산되는 데이터를 정리된 보고서로 표현하여
불확실성이 낮은 상황의 의사결정이 가능하도록 함.

- 연관성 파악

요약된 보고서에서 주요한 항목들 간의 연관성을 파악한 경쟁우위의 확보

∴ 의사결정권자에게 항목 간 연관성을 제시하여 미래의 계획을 지원

→ 다양한 자료는 의사결정에 있어 세부적 판단에도 기여

통계학의 정의와 목적

- 통계학의 목적

- 예측

인과관계 파악을 통해 패턴을 찾아내고 이러한 패턴을 통해 추세를 판단

∴ 다양한 변수의 대입과 삭제를 통해 예측 가능

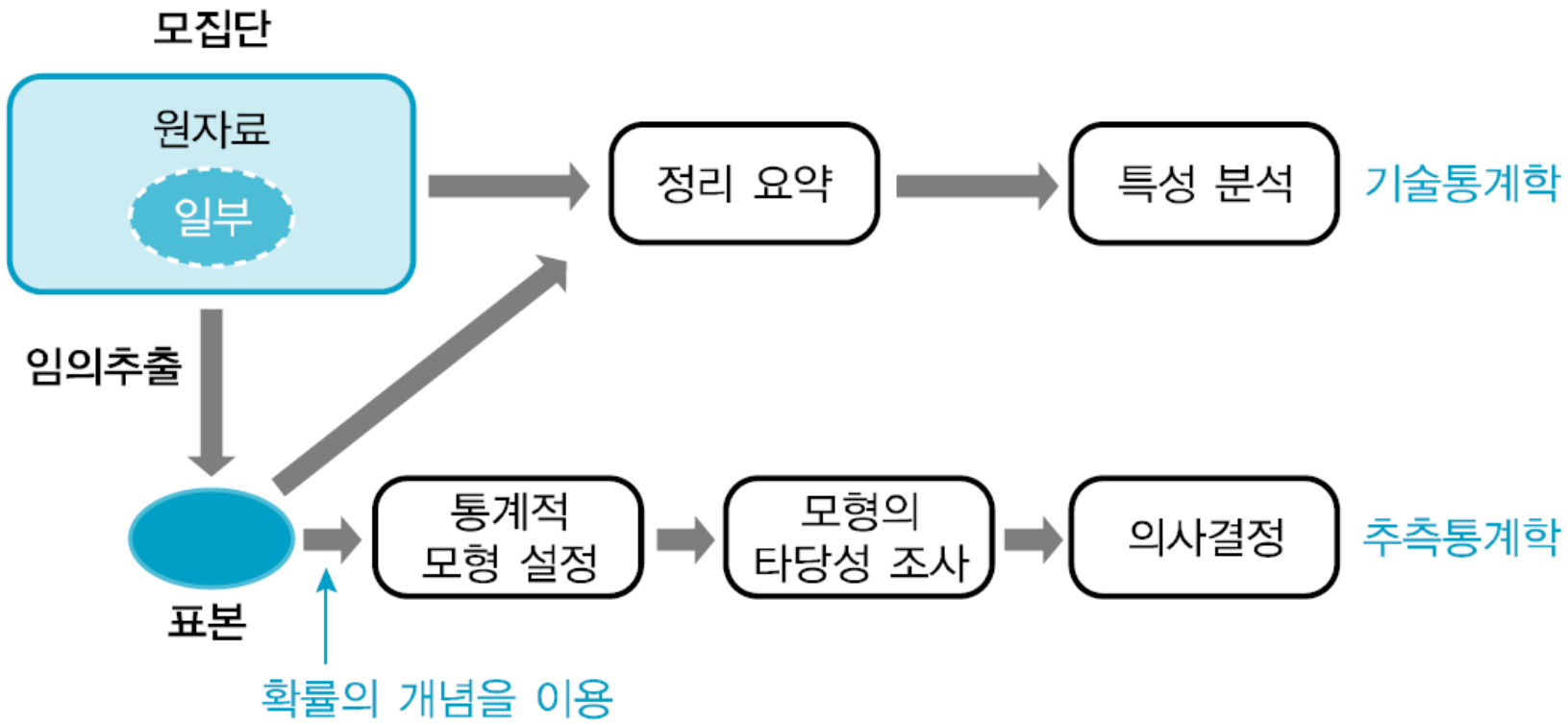
→ 다양한 계량 기법과 여러 변수들을 활용하여
최소의 비용으로 최대의 수익을 얻을 수 있는 조합 확인

기술통계학과 추측통계학

- **기술통계학**descriptive statistics
  : 자료를 수집하고 정리하여, 표 또는 그래프나 그림 등으로 나타내거나 자료가 갖는 수치적인 특성을 분석하고 설명하는 방법을 다루는 통계학의 한 분야

- **추측통계학**inferential statistics
  : 표본을 대상으로 얻은 정보로부터 모집단에 대한 불확실한 특성을 과학적으로 추론하는 방법을 다루는 통계학의 한 분야

기술통계학과 추측통계학



모집단과 표본

□ 모집단

모집단(population) : 통계분석 방법을 적용할 관심 대상의 전체 집합

예 모든 대한민국 여성

2016년에 수입된 모든 쇠고기

A 쇼핑몰 회원 전체

B 통신회사 전체 가입자

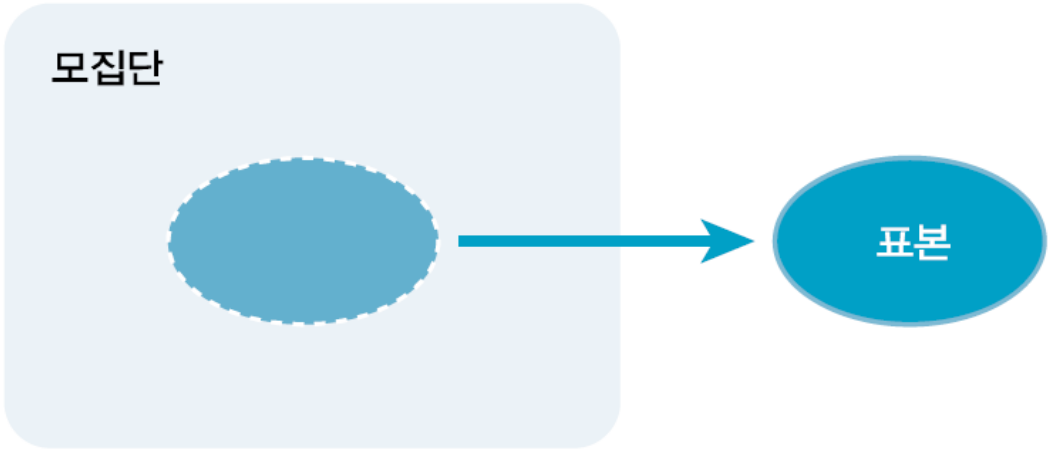
→ 물리적인 한계로 인해 모집단 전체를 전수조사하기는 쉽지 않다.

□ 표본

표본(sample) : 과학적인 절차를 적용하여 모집단을 대표할 수 있는 일부를 추출하여 직접적인 조사 대상이 된 모집단의 일부

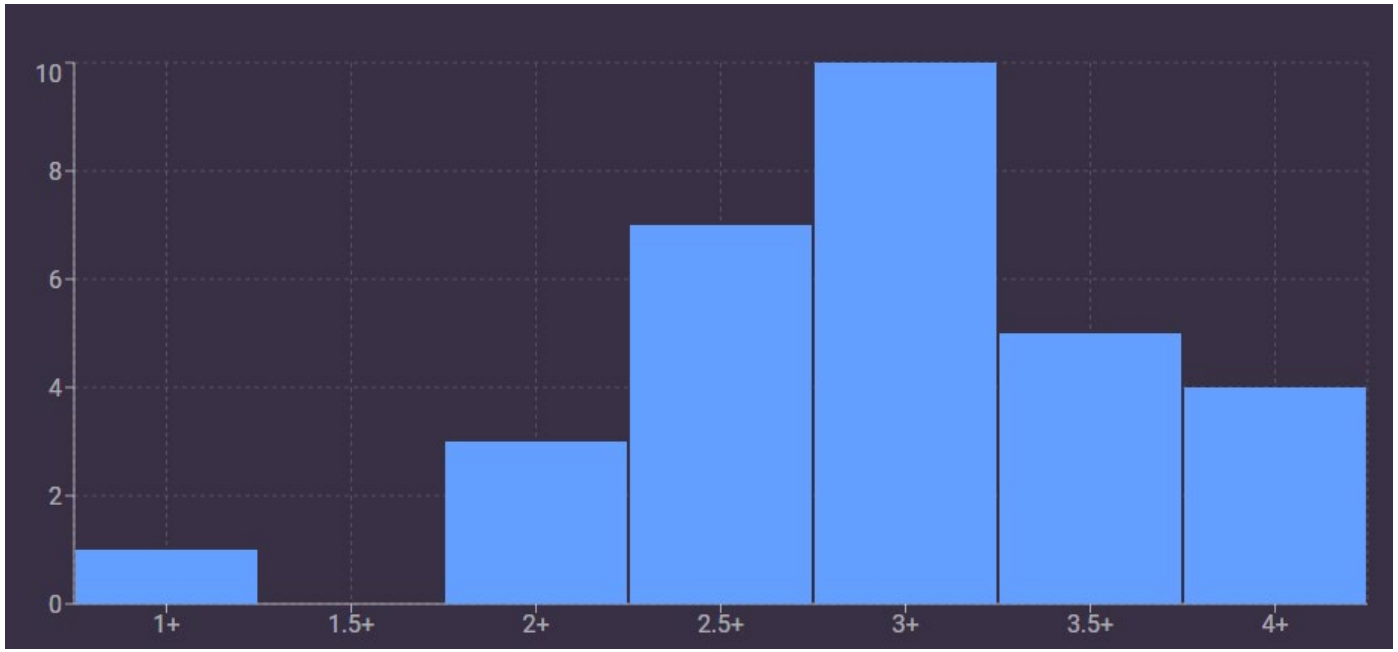
모집단과 표본

- 모집단과 표본과의 관계



자료의 축약

3.5 2.4 2.5 3.0 3.0 4.3 3.8 3.2 3.0 2.8
2.5 1.4 2.0 2.5 3.3 4.0 2.8 3.0 3.7 4.1
3.2 3.0 2.8 3.1 3.5 2.9 4.0 2.4 3.2 3.7



변동

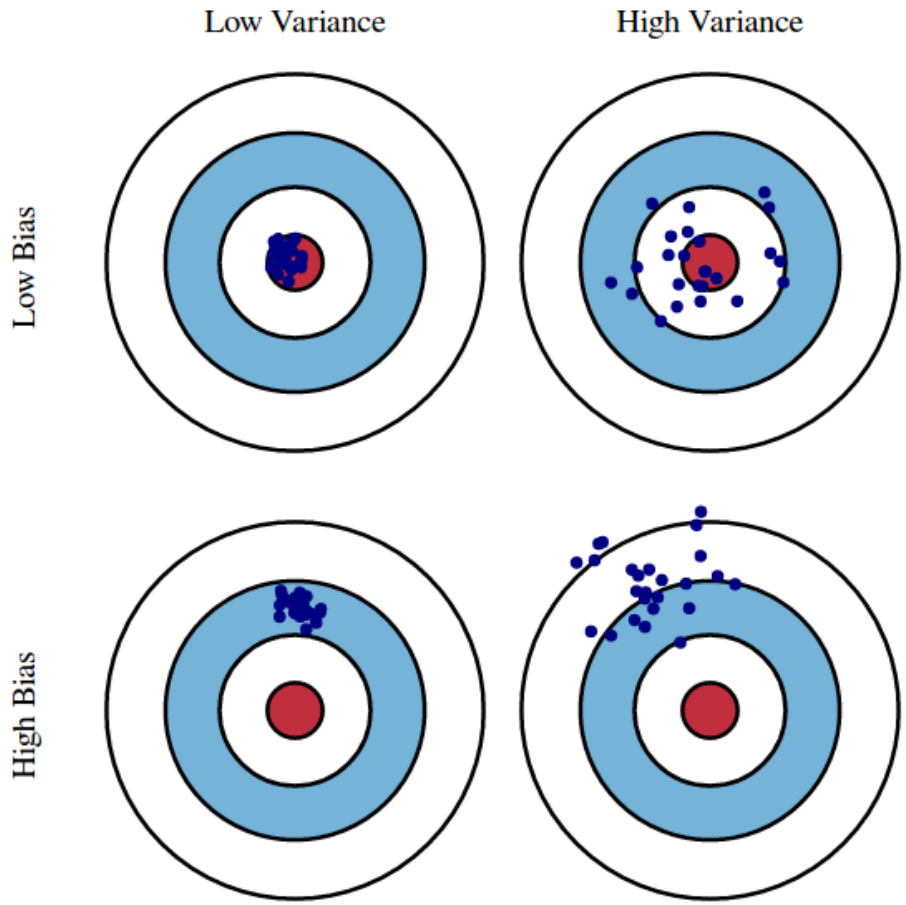


Fig. 1 Graphical illustration of bias and variance.

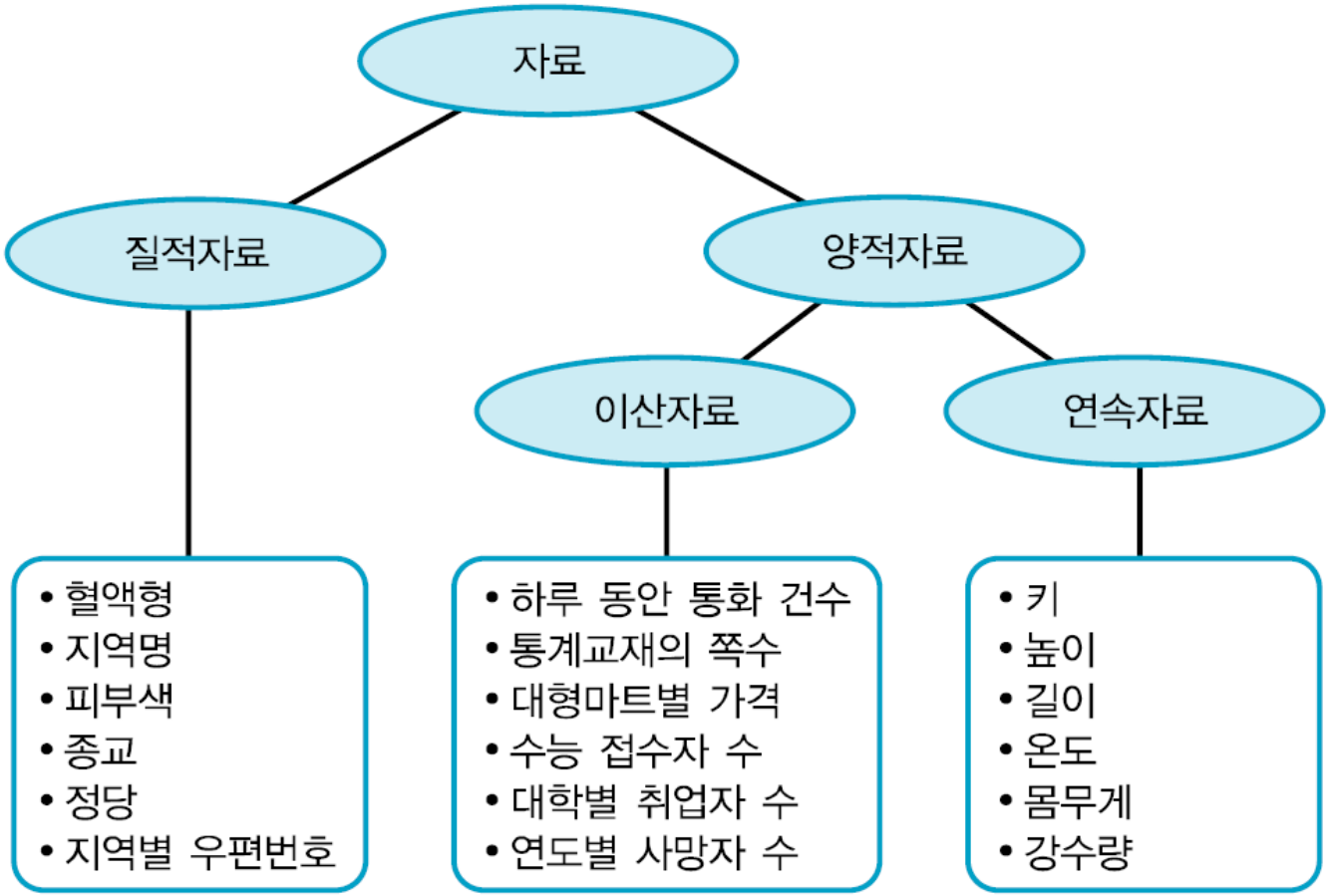
자료의 유형

- 모든 통계 자료는 수치적인 척도로 표현되거나 그렇지 않은 유형으로 분류
  - 여기서 수치적인 척도로 표현된다는 것은 자료가 숫자에 의하여 표현되며, 그 숫자 자체가 크다거나 작다 또는 많다거나 적다 등과 같이 의미를 가지는 경우를 지칭

- **양적자료**quantitative data
  : 숫자로 표현되며, 그 숫자가 의미를 갖는 자료

- **질적자료**qualitative data 또는 **범주형자료**categorical data
  : 숫자에 의하여 표현되지 않고 여러 개의 범주로 구분되는 자료

자료의 유형


자료
질적자료
양적자료
이산자료
연속자료
• 혈액형
• 지역명
• 피부색
• 종교
• 정당
• 지역별 우편번호
• 하루 동안 통화 건수
• 통계교재의 쪽수
• 대형마트별 가격
• 수능 접수자 수
• 대학별 취업자 수
• 연도별 사망자 수
• 키
• 높이
• 길이
• 온도
• 몸무게
• 강수량

자료의 정리

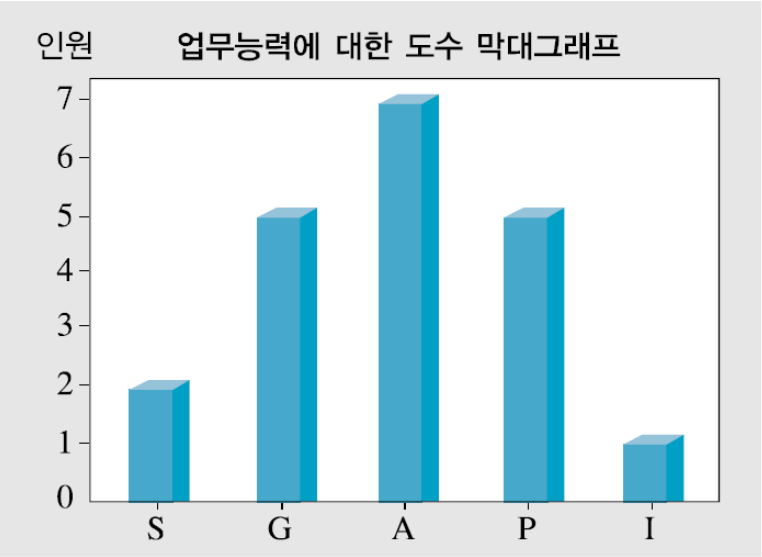
질적 자료: 도수분포표

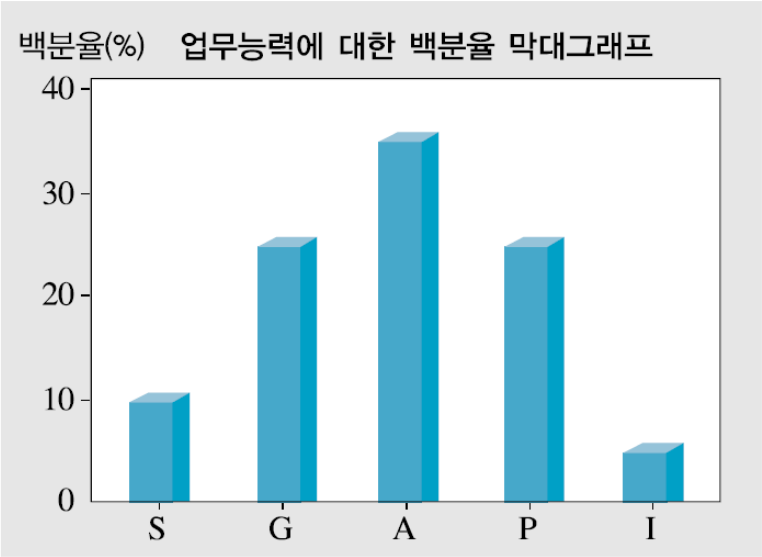
- **도수분포표**frequency distribution table : 여러 개의 범주 안에 측정된 각 범주의 도수와 상대도수 또는 범주 백분율을 기입한 표
  - 도수frequency
    : 각 범주 안에 들어가는 자료집단 안에서 관찰된 자료 수
  - 상대도수relative frequency
    : 각 범주의 도수를 자료집단 안의 전체 자료수로 나눈 값
  - 범주 백분율class percentage : 상대도수에 100을 곱한 값으로 백분율(%)

구분	도수	상대도수	백분율(%)
S	2	0.10	10
G	5	0.25	25
A	7	0.35	35
P	5	0.25	25
I	1	0.05	5

질적 자료: 막대그래프

- **막대그래프**bar chart : 질적자료의 각 범주를 수평축에 나타내고, 각 범주에 대응하는 도수, 상대도수, 백분율 등을 같은 폭의 수직막대로 나타낸 그림
  - 예) 업무능력을 5개의 그룹 S(Superior), G(Good), A(Average), P(Poor), I(Inferior)로 구분하여 조사하여 S 2명, G 5명, A 7명, P 5명, I 1명



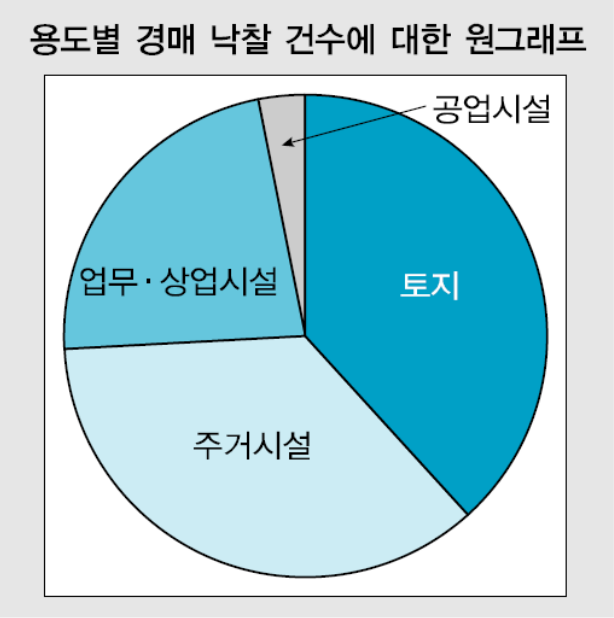


[장점] 여러 개의 범주(요인) 중에서 문제 해결에 도움을 주는 중요한 소수의 범주를 찾는데 도움을 주고, 어떤 범주가 중요한지 쉽게 파악할 수 있다.

[단점] 범주를 크기 순서로 재배열하므로 범주가 순서형인 순서자료에는 적합하지 않다.

질적 자료: 원그래프

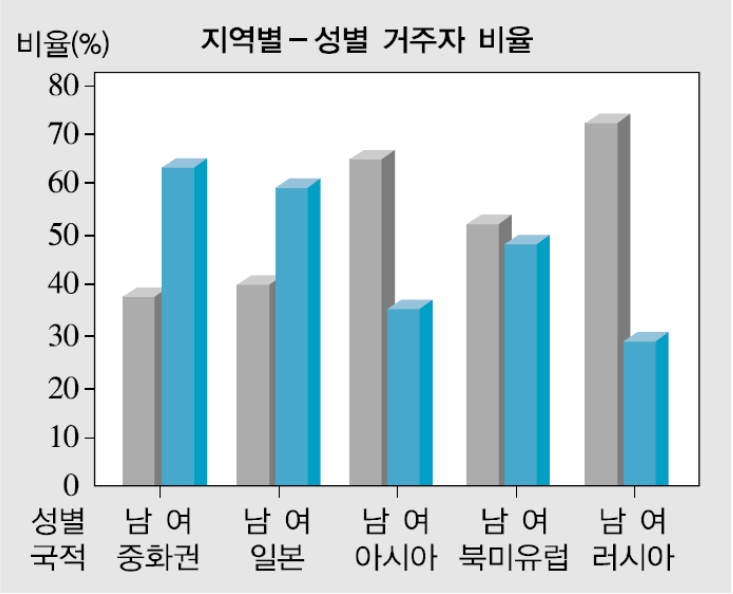
- **원그래프**pie chart : 질적 자료의 각 범주에 대한 비율 관계를 이용하여 각 범주를 상대적으로 나타낸 비율 그래프로, 각 범주의 백분율에 해당하는 중심각을 갖는 부채꼴 모양으로 나타낸 그림

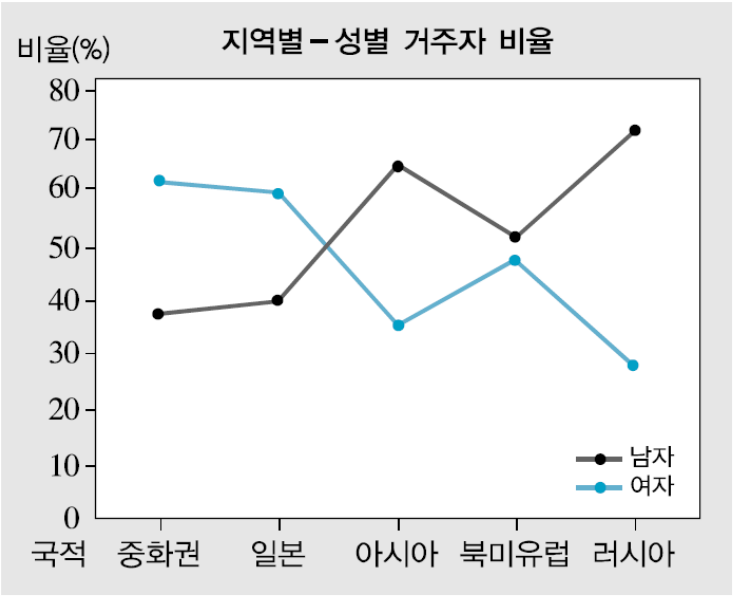


질적 자료: 꺾은선 그래프

- **꺾은선 그래프**graph of broken line : 막대그래프의 상단 중심부를 직선으로 연결하여 각 범주를 비교하는 그림

[단점] 수직축의 척도 간격에 따라 달리 해석할 수 있다.

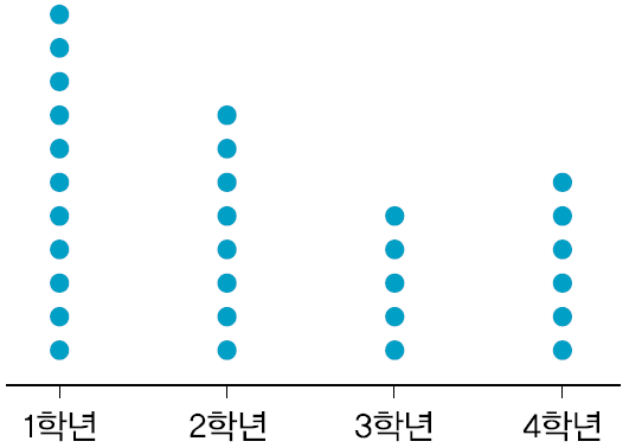




질적 자료: 점도표

- **점도표**dot plot : 수평축에 각 범주를 기입하고, 수평축 위에 각 범주 또는 측정값의 관찰 횟수를 점으로 나타낸 그림
  - 양적 자료에서 사용 가능

예) 동아리 회원은 1학년 11명, 2학년 8명, 3학년 5명, 4학년 6명



[장점] 각 범주 사이의 관찰값을 쉽게 비교할 수 있다.

[단점] 관찰값의 수에 해당하는 점을 찍어서 나타내므로 그 수가 매우 많은 경우에는 부적당하다.

양적 자료: 도수분포표

- **도수분포표**frequency distribution table : 양적 자료를 적당한 간격으로 집단화하여 계급, 도수, 상대도수, 누적도수, 누적상대도수, 계급값 등을 기입한 표
  - 계급class
    : 양적자료를 적당한 간격으로 집단화하여 나타낸 범주
  - 계급폭class width
    : 이웃하는 두 계급의 위쪽 경계에서 아래쪽 경계를 뺀 값

$$\text{계급의 폭} = \frac{\text{최댓값} - \text{최솟값}}{\text{계급의 수}}$$

- **상대도수분포표**relative frequency distribution table : 각 계급구간과 각 계급구간의 상대도수를 기록하여 자료들을 집단화시킨 표
  - 계급 상대도수class relative frequency
    : 계급의 도수를 전체 자료수로 나눈 값

$$\text{계급상대도수} = \frac{\text{계급의 도수}}{\text{전체 도수}}$$

양적 자료: 도수분포표

- **누적상대도수분포표**cumulative relative frequency distribution table : 각 계급구간과 해당 계급구간을 포함하여 이전 계급구간에 있는 상대도수들을 모두 합한 도수분포표
  - 누적도수cumulative frequency : 이전 계급까지의 모든 도수를 합한 도수
  - 누적상대도수cumulative relative frequency : 이전 계급까지의 모든 상대도수를 합한 상대도수

[장점] 자료의 대략적인 중심의 위치(50% 위치)를 알 수 있다.
전체 자료의 흩어진 정도를 파악할 수 있다.

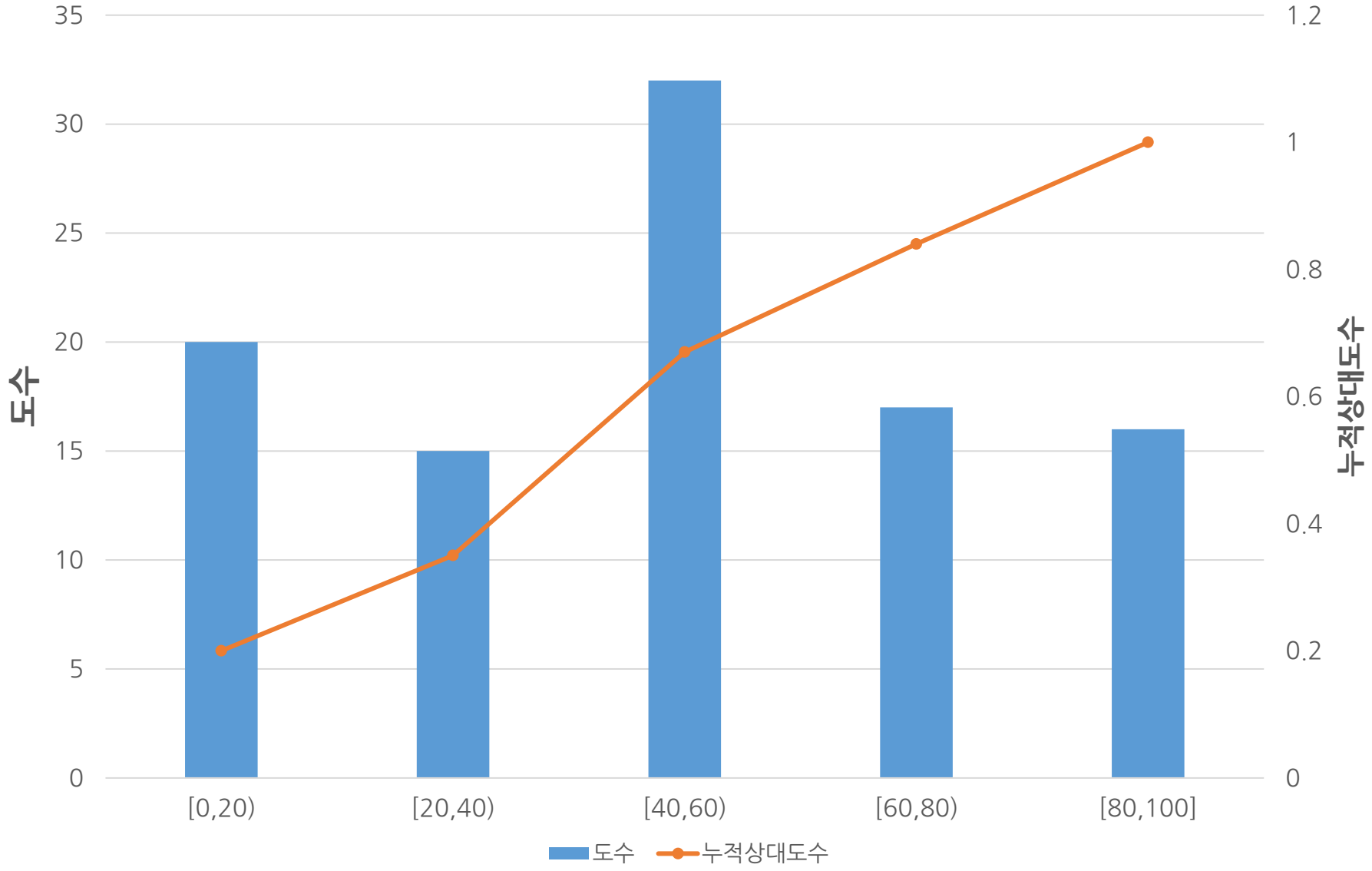
[단점] 각 계급 안에 들어 있는 자료의 정확한 값을 알 수 없다.

양적 자료: 도수분포표

- 도수분포표의 예시

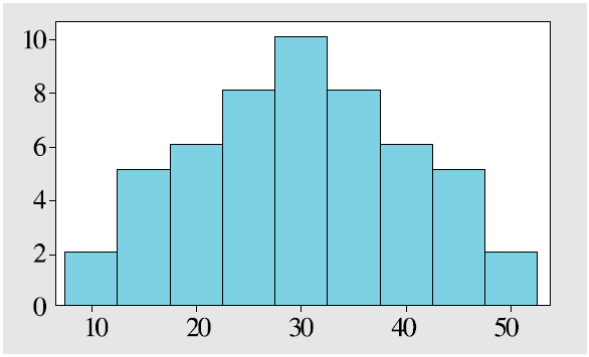
계급	도수	상대도수	누적상대도수
$0 \le x < 20$	20	0.20	0.20
$20 \le x < 40$	15	0.15	0.35
$40 \le x < 60$	32	0.32	0.67
$60 \le x < 80$	17	0.17	0.84
$80 \le x \le 100$	16	0.16	1.00
전체	100	1	

양적 자료: 도수분포그래프

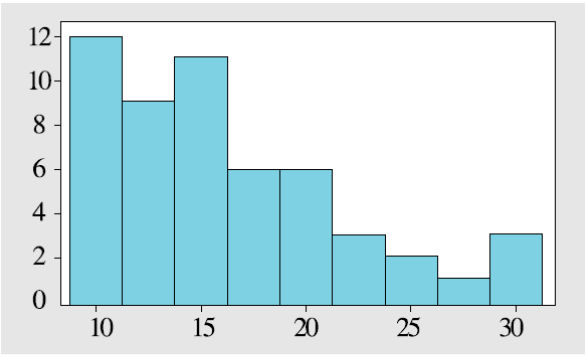


양적 자료: 히스토그램

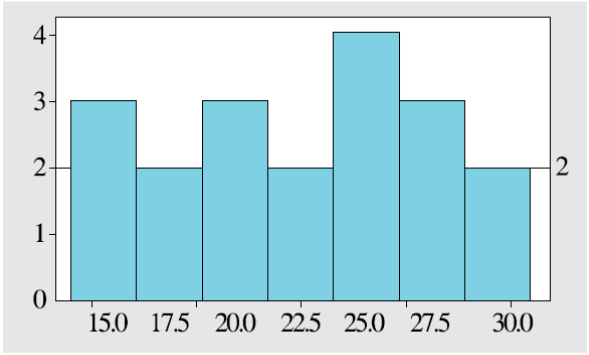
- **히스토그램**histogram : 수평축에 도수분포표의 계급간격, 수직축에 각 계급의 도수를 높이로 갖는 사각형으로 작성한 그림
  - 누적도수히스토그램 : 수직축에 누적도수를 나타낸 히스토그램
  - 상대도수히스토그램 : 수직축에 상대도수를 나타낸 히스토그램
  - 누적상대도수히스토그램 : 수직축에 누적상대도수를 나타낸 히스토그램



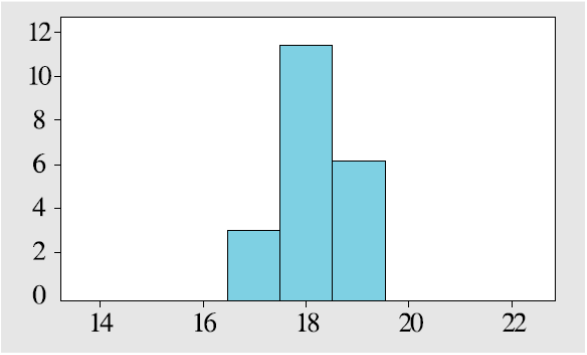
(a) 대칭형



(b) 비대칭형



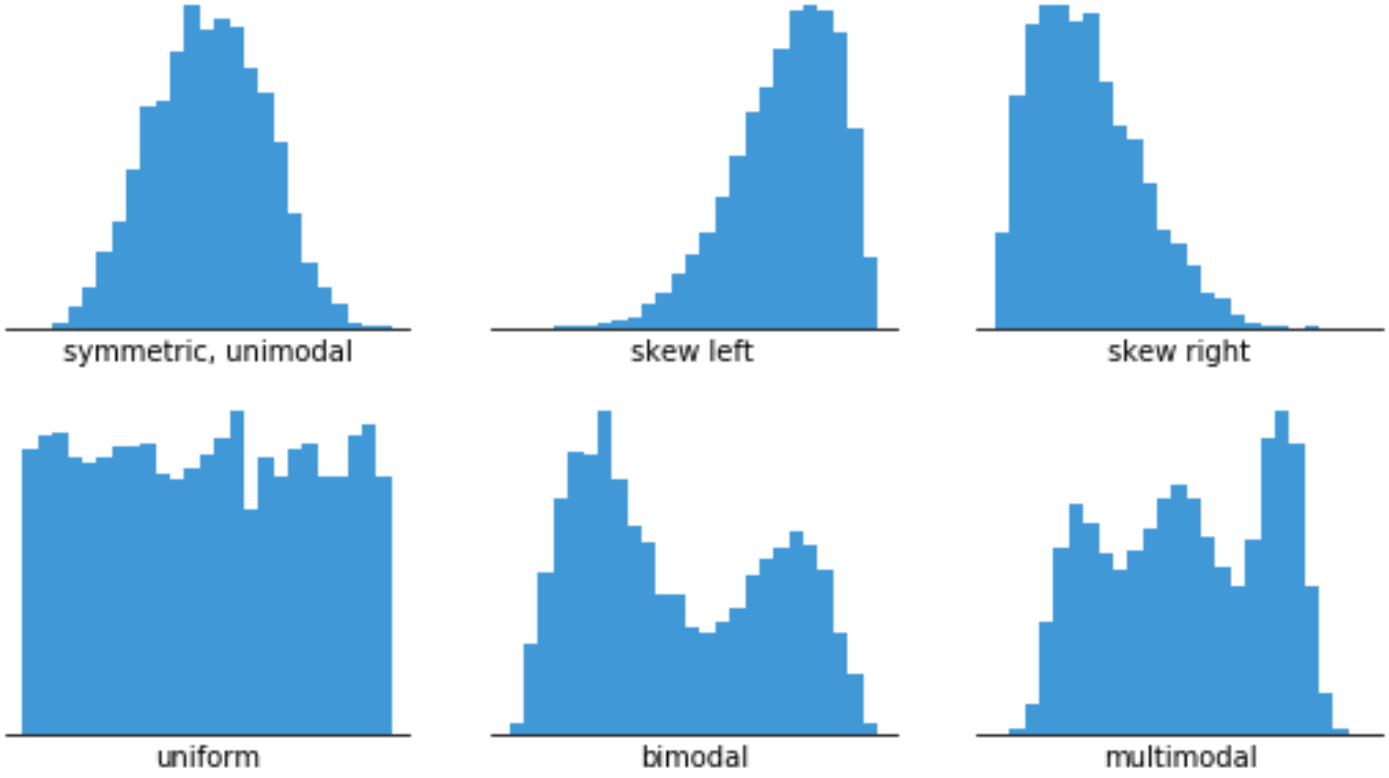
(c) 퍼짐형



(d) 집중형

양적 자료: 히스토그램

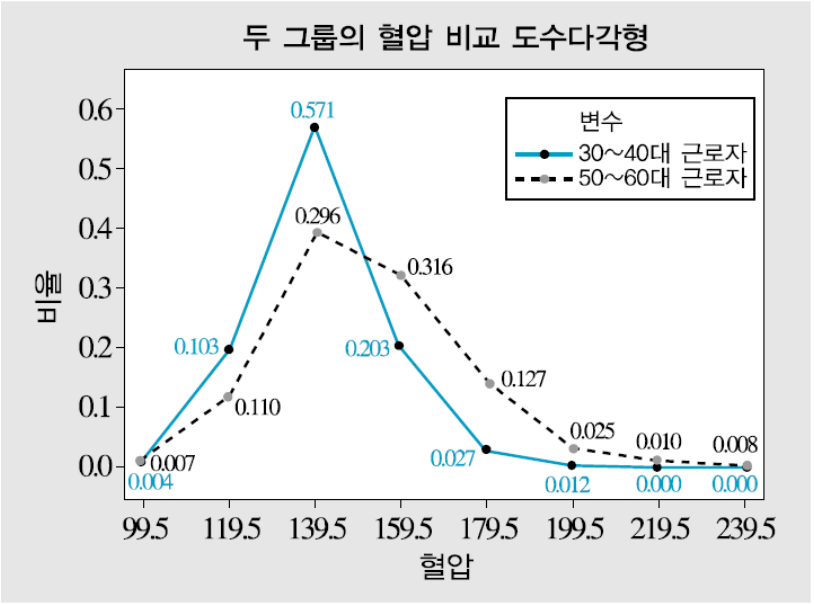
- 히스토그램의 여러가지 모양



양적 자료: 도수다각형

- **도수다각형**frequency polygon : 히스토그램에서 연속적인 막대의 상단중심부를 선분으로 연결하여 다각형으로 표현한 그림

[장점] 두 개 이상의 자료집단을 비교하는데 널리 사용한다.



양적 자료: 줄기-잎 그림

- **줄기-잎 그림**stem-leaf plot : 실제 측정값을 이용하여 변동이 적은 부분은 줄기로 하고, 변동이 많은 부분은 잎모양으로 나타낸 그림

[장점] 도수분포표나 히스토그램이 갖고 있는 특성을 그대로 보존한다. 각 계급 안에 들어있는 개개의 측정값을 제공한다.

48	40	46	47	45	47	40	48	48	49
41	40	41	43	49	41	32	31	31	32
31	38	34	31	37	33	33	33	39	34
36	37	39	52	55	54	53	53	57	57
51	54	53	21	25	22	20	29	27	20

2	1520970
3	21121841733394679
4	8067570889101391
5	2543377143

How to Spot a Misleading Graph

https://www.youtube.com/watch?v=E91bGT9BjYk&vl=en

R Programming

프로그래밍 언어란?

- 프로그래밍 언어는 컴퓨터 시스템을 구동시키는 소프트웨어를 작성하기 위한 형식언어
  - 프로그래밍언어는 기계어(machine language)만을 이해하는 컴퓨터와 자연어(natural language)를 구사하는 인간 사이의 의사소통 수단으로 작용

- R
  - R은 통계 계산과 그래픽을 위한 프로그래밍 언어

기본 연산자: 산술 연산자

[표 1-3] R의 산술연산자

연산자	설명	예	결과
+	더하기	3 + 2	5
–	빼기	3 – 2	1
*	곱하기	3 * 2	6
/	나누기	3 / 2	1.5
^ 혹은 **	승수	3 ^ 2	9
x %% y	X를 y로 나눈 나머지 값 반환	3 %% 2	1
x %/% y	나누기의 결과를 정수로	3 %/% 2	1

기본 연산자: 논리 연산자

[표 1-4] R의 논리 연산자

연산자	설명	예	결과
<	좌변이 보다 작은	5 < 5	FALSE
<=	좌변 이하	5 <= 5	TRUE
>	좌변이 보다 큰	5 > 5	FALSE
>=	좌변 이상	5 >= 5	TRUE
==	값이 같은	5 == 5	TRUE
!=	값이 다른	5 != 5	FALSE
!x	부정형 연산	!TRUE	FALSE
x | y	x OR y (논리합)	TRUE | FALSE	TRUE
x & y	x AND y (논리곱)	TRUE & FALSE	FALSE

상수와 변수

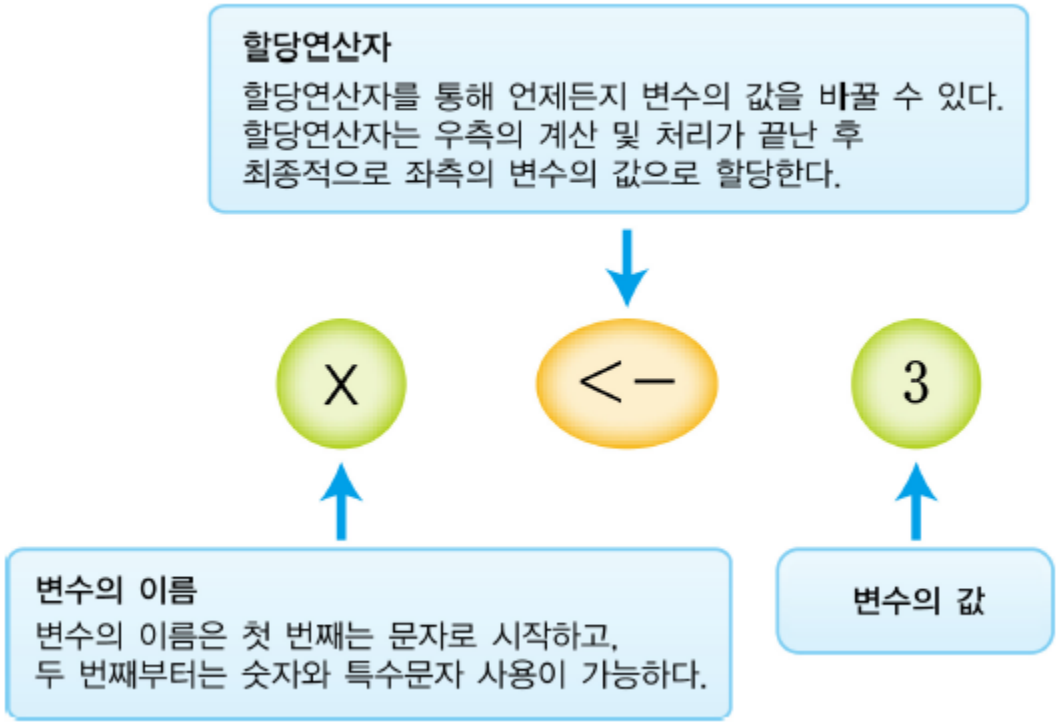
- 상수
  - 나타내는 값을 바꿀 수 없는 자료
  - 상수는 프로그래밍을 좀 더 유연하고 편하게 하기 위해 미리 정의되어 있는 자료들

- 변수
  - 상황에 따라 값을 바꿀 수 있는 자료들

변수

- 변수 사용하기
  - 이름 : 저장하고자 하는 값을 가장 잘 나타낼 이름
    - 문자, 숫자, 특수문자(점(.), 밑줄(_)) 사용 가능
    - 변수의 이름은 숫자로 시작할 수 없음
    - 점(.)로도 시작할 수 있으나 바로 뒤에 숫자가 나올 수 없음
    - R에서 사용하는 예약어들은 변수명으로 사용할 수 없음(for, function 등)
    - 변수명은 대소문자를 구분
  - 대입연산자(혹은 할당연산자)
    - <-
    - 할당연산자를 이용하여 변수에 원하는 값을 저장
  - 변수의 초기화
    - 변수를 미리 만들어 놓고 값으로 결측상태를 나타내는 NA, 혹은 값이 정해지지 않은 상태를 나타내는 NULL을 이용하거나 연산에 따라 항등원을 이용하여 값을 초기화하고 필요한 시점에 원하는 값을 저장하여 사용

변수


할당연산자
할당연산자를 통해 언제든지 변수의 값을 바꿀 수 있다.
할당연산자는 우측의 계산 및 처리가 끝난 후
최종적으로 좌측의 변수의 값으로 할당한다.
X
<-
3
변수의 이름
변수의 이름은 첫 번째는 문자로 시작하고,
두 번째부터는 숫자와 특수문자 사용이 가능하다.
변수의 값

자료구조: 벡터

- 벡터 : 동일한 자료형의 단일 값들이 한군데 모여있는 자료구조
  - 벡터는 R의 가장 기본적인 자료 저장 방법
  - 벡터는 동일한 자료형을 갖는 값들의 집합으로, 일반적으로 하나의 속성을 저장하는 단위로 사용

- 벡터 생성하기
  - 벡터 생성 연산자 ":"(콜론)
    - '시작값 : 종료값'의 형태로 사용하며, 시작값부터 종료값까지 1씩 더하거나 빼서 벡터를 생성

```
> 1:5
[1] 1 2 3 4 5
> 5:1
[1] 5 4 3 2 1
```

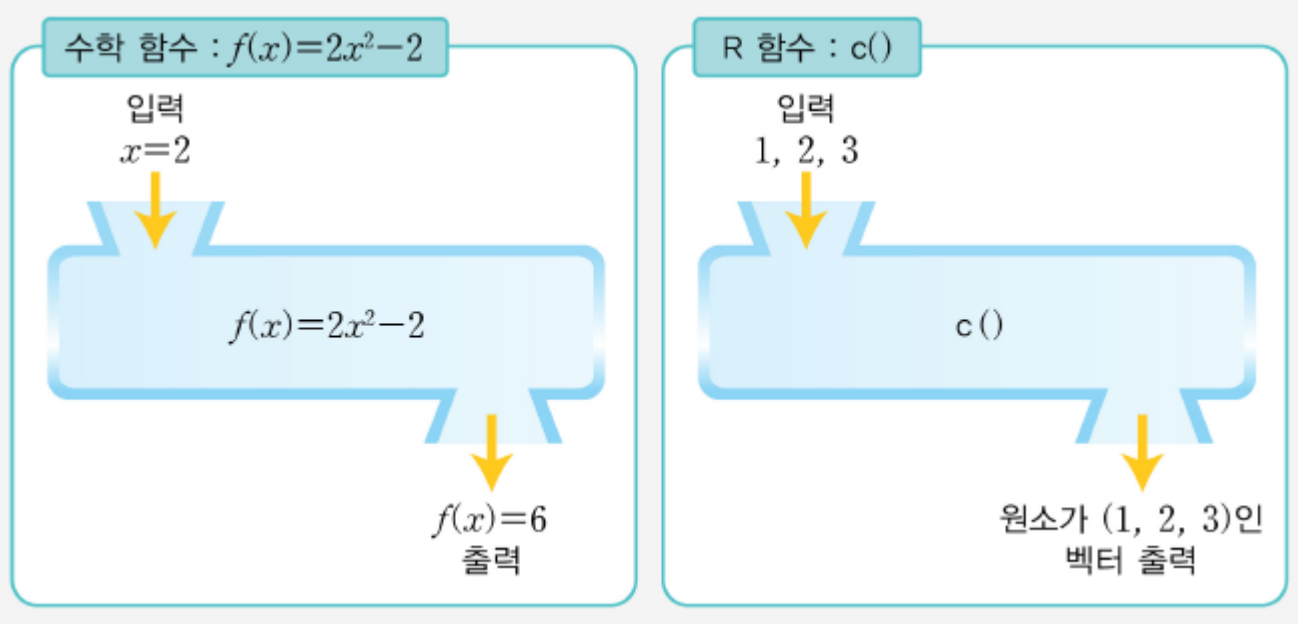
자료구조: 벡터

- 벡터생성함수 : c(), seq(), rep()

```
> c(1, 2, 3)
[1] 1 2 3
> c(1, 2, 3, c(4, 5, 6))
[1] 1 2 3 4 5 6
> x <- c(1, 2, 3)
> x
[1] 1 2 3
```

※ 함수

- 함수
  - 프로그래밍 언어에 따라 프로시져(procedure), 메소드(method) 등으로 불리며, 수학에서의 함수와 마찬가지로 함수 작동에 필요한 입력으로 함수내부에서 계산을 함으로써 생성되는 적절한 출력을 내보내는 코드들의 모임
    - c() 함수의 경우 함수명 “c” 이후 소괄호( ) 사이에 벡터를 구성하는 원소들을 콤마(,)로 구분하여 입력으로 사용하고, 출력으로는 입력에 사용된 자료들로 구성된 벡터를 반환



※ R 도움말

c {base} ❶ R Documentation

Combine Values into a Vector or List ❷

Description

This is a generic function which combines its arguments. ❸

The default method combines its arguments to form a vector. All arguments are coerced to a common type which is the type of the returned value, and all attributes except names are removed.

Usage ❹

```
c(..., recursive = FALSE)
```

Arguments ❺

`...`	objects to be concatenated.
`recursive`	logical. If `recursive = TRUE`, the function recursively descends through lists (and pairlists) combining all their elements into a vector.

Details ❻

The output type is determined from the highest type of the components in the hierarchy NULL < raw < logical < integer < double < complex < character < list < expression. Pairlists are treated as lists, but non-vector components (such names and calls) are treated as one-element lists which cannot be unlisted even if `recursive = TRUE`.

`c` is sometimes used for its side effect of removing attributes except names, for example to turn an array into a vector. `as.vector` is a more intuitive way to do this, but also drops names. Note too that methods other than the default are not required to do this (and they will almost certainly preserve a class attribute).

※ R 도움말

- R의 도움말 구조
  1. 함수의 이름과 함수가 속한 패키지
     - c() 함수의 경우 기본 패키지인 base에 속하고 있음을 나타냄
     - R에서 함수는 패키지 별로 존재
  2. 함수에 대한 간략한 설명
  3. Description : 함수에 대한 자세한 설명
  4. Usage : 함수의 사용법
     - 다른 프로그래밍 언어에서 이야기하는 함수의 원형과 유사한 형태로 함수의 원형은 해당함수가 출력으로 전달하는 자료의 유형 및 모든 전달인자를 기술하는 것을 뜻함
  5. Arguments : 함수 호출 시 전달되는 값 (전달인자) 에 대한 설명
     - 위의 예에서 recursive로 전달되는 값은 인수로 논리값(logical)을 가짐을 나타내며, 그 값이 TRUE일 경우 벡터 생성의 순서를 반대로 함을 설명
     - 이처럼 이 섹션에서는 전달인자의 유형과 값에 대해 설명
  6. Details : 함수 사용에 대한 자세한 설명

자료구조: 벡터

- seq() 함수로 벡터생성 : 전달인자 사용

```
> seq(from=1, to=5, by=2)
[1] 1 3 5
> seq(1, 5, 2)
[1] 1 3 5
```

- rep(): 기존에 있는 벡터를 반복하여 새로운 벡터를 만듬

```
> rep(c(1, 2, 3), times=2)
[1] 1 2 3 1 2 3
> rep(c(1, 2, 3), each=2)
[1] 1 1 2 2 3 3
```

자료구조: 벡터

- 벡터 내의 원소에 접근하기
  - 위치정보로 접근하기
    - 벡터에 포함되는 자료는 인덱스(Index)라는 위치정보를 가짐
    - 인덱스는 1부터 시작하는 정수
    - 벡터 이름 뒤에 대괄호 [ ]를 써서 인덱스를 지정하여 벡터 내의 원하는 위치의 원소들을 추출

```
> rep(c(1, 2, 3), times=2)
[1] 1 2 3 1 2 3
> rep(c(1, 2, 3), each=2)
[1] 1 1 2 2 3 3
> x <- c(5, 4, 3, 2, 1)
> x[1]
[1] 5
> x[1, 2, 3]
Error in x[1, 2, 3] : incorrect number of dimensions
> x[c(1, 2, 3)]
[1] 5 4 3
> x[-c(1, 2, 3)]
[1] 2 1
> length(x)
[1] 5
> x[3:length(x)]
[1] 3 2 1
```

자료구조: 벡터

- 논리값이 TRUE인 원소 접근하기
  - 벡터 이름 뒤에 대괄호 안에 논리값 벡터를 넣어 논리값이 TRUE인 자료들을 추출

```
> ex <- c(1, 3, 7, NA, 12)
> ex < 10
[1]  TRUE  TRUE  TRUE    NA FALSE
> ex[ex < 10]
[1]  1  3  7 NA
> ex[ex %%  2 == 0]
[1] NA 12
> ex[is.na(ex)]
[1] NA
> ex[ex %%  2 == 0 & !is.na(ex)]
[1] 12
```

자료구조: factor

- factor
  - 저장 값의 크기보다 의미가 중요한 질적 자료를 위해 사용
  - 예를 들어 숫자 1, 2, 3은 산술연산을 통해 계산되는 본래의 숫자로써 기능을 하지만, factor로 지정된 1, 2, 3은 단지 세 개의 그룹 혹은 상태를 구별 짓는 의미로 사용
- factor 생성함수 : factor()

```
factor(x = character(), levels, labels = levels, ordered=FALSE)
```

▷ x : factor로 만들 벡터

▷ levels : 주어진 데이터 중 factor의 각 값(수준)으로 할 값을 벡터 형태로 지정(여기서 빠진 값은 NA로 처리).

▷ labels : 실제 값 외에 사용할 각 수준의 이름(벡터). 예를 들어 데이터에서 1이 남자를 가리킬 경우 labels를 통해 '남자' 혹은 'M' 등으로 변경.

▷ ordered : 순위형 자료 여부(TRUE/FALSE)로, levels에 입력한 순서를 가짐.

자료구조: factor

```
> x <- c(1, 2, 3, 4, 5)
> factor(x, levels=c(1, 2, 3, 4))
[1] 1    2    3    4    <NA>
Levels: 1 2 3 4
> factor(x, levels=c(1, 2, 3, 4), labels=c("a", "b", "c", "d"))
[1] a    b    c    d    <NA>
Levels: a b c d
> factor(x, levels=c(1, 2, 3, 4), ordered=TRUE)
[1] 1    2    3    4    <NA>
Levels: 1 < 2 < 3 < 4
```

자료구조: 데이터 프레임

- 데이터 프레임
  - 자료 처리를 위해 가장 많이 사용된 자료구조
  - 서로 다른 벡터로 구성된 자료들을 열로 배치한 자료구조
  - 직접 입력하는 경우보다 외부 데이터로부터 가져오는 경우가 많음
- 생성함수 : data.frame()

```
data.frame(..., row.names = NULL,
           stringsAsFactors = default.stringsAsFactors())
```

▷ ... : 데이터 프레임을 구성할 열 정의 (값 혹은 열이름 = 자료의 형태)

▷ row.names : 행의 이름으로 사용할 값 저장. 기본값은 NULL로 각 행의 번호 저장

▷ stringsAsFactors : 문자열로 구성된 자료를 factor로 변환할지 여부로 기본값은 문자열을 factor로 변환

자료구조: 데이터 프레임

```
> name <- c("철수", "영희", "길동")
> age <- c(21, 20, 31)
> gender <- factor(c("M", "F", "M"))
> character <- data.frame(name, age, gender)
> str( character )
'data.frame':   3 obs. of  3 variables:
 $ name  : Factor w/ 3 levels "길동","영희",..: 3 2 1
 $ age   : num  21 20 31
 $ gender: Factor w/ 2 levels "F","M": 2 1 2
> character
  name age gender
1 철수  21      M
2 영희  20      F
3 길동  31      M
```

외부로부터 자료 가져오기

[표 1-6] 통계청이 제공하는 다양한 서비스

서비스명	주소	설명
국가통계포털	http://kosis.kr	주제별로 다양한 통계들을 제공하며, 통계를 얻기 위한 조사들에 대한 설명들도 함께 제공하는 서비스
국가지표체계	http://www.index.go.kr	국가주요지표, e-나라지표, 국민 삶의 지표, 녹색성장 지표 등 우리나라의 각종 상황들을 지표화 및 시각화하여 한눈에 알아볼 수 있도록 한 서비스
SGIS+ 통계지리정보서비스	http://sgis.kostat.kr	각종 통계들을 지리정보와 결합하여, 지도를 통해 정보들을 탐색할 수 있도록 하였으며, 지도 제작을 위한 각종 지리정보를 제공하는 서비스
마이크로데이터 통합서비스	https://mdis.kostat.go.kr	통계자료가 아닌 원자료에서 입력오류 등을 제거한 마이크로데이터(microdata, 통계기초자료)를 제공하는 서비스로서 무료로 제공하는 공공용 마이크로데이터, 유료로 제공하는 인가된 마이크로데이터를 사용할 수 있는 서비스

외부로부터 자료 가져오기

- 데이터를 불러올 수 있는 함수
  - read.csv
    - ‘,’로 구분된 데이터
  - read.delim
    - tab으로 구분된 데이터
  - read.table
    - 구분자를 지정할 수 있음

```
> data <- read.csv("./서울시연령별성별인구_2017.csv", header=TRUE)
> str(data)
'data.frame':   66874 obs. of  6 variables:
 $ 행정구역별.읍면동.: Factor w/ 240 levels "가평군","강남구",..: 108 108 108 108 108 108 108 10
 $ 연령별           : Factor w/ 28 levels "0~4세","10~14세",..: 28 28 28 28 28 28 28 28 1 1 ...
 $ 항목             : Factor w/ 8 levels "내국인(명)","내국인_남자(명)",..: 5 6 8 7 1 2 4 3 5 6
 $ 단위             : logi  NA NA NA NA NA NA ...
 $ X2017.년         : Factor w/ 23863 levels "0","10","10.3",..: 23599 16909 17334 23319 23204
 $ X                : logi  NA NA NA NA NA NA ...
```

과제

과제 2

- 뉴스 기사에 삽입된 도표 찾아보기
  - 원그래프, 막대그래프, 히스토그램, 꺾은선 그래프 등을 사용한 뉴스 기사를 찾기
    - 3개 이상의 서로 다른 그래프를 활용한 기사를 선택
      - 한 기사가 아니라 각기 다른 기사에서 다른 종류의 그래프를 찾아도 됨
    - 사용한 도표가 적절한지, 개선할 수 있는 방향은 없는지 의견 작성
    - 올해 나온 뉴스 기사 중에서 찾기: 출처 명시