

**빅데이터의 이해와 활용**

Understanding and Using Big Data

07

**기술통계학**

# 학습 내용

- 01 통계학과 의사결정
- 02 자료의 정리와 도시
- 03 자료의 요약

# 학습 목표

- 통계의 개념이 의사결정에 어떻게 활용되는지 살펴볼 수 있다.
- 통계의 기본 개념들에 대해 이해하고 설명할 수 있다.
- 데이터를 활용하여 기술 통계 요약 및 도식화를 할 수 있다.

# 생각 해보기

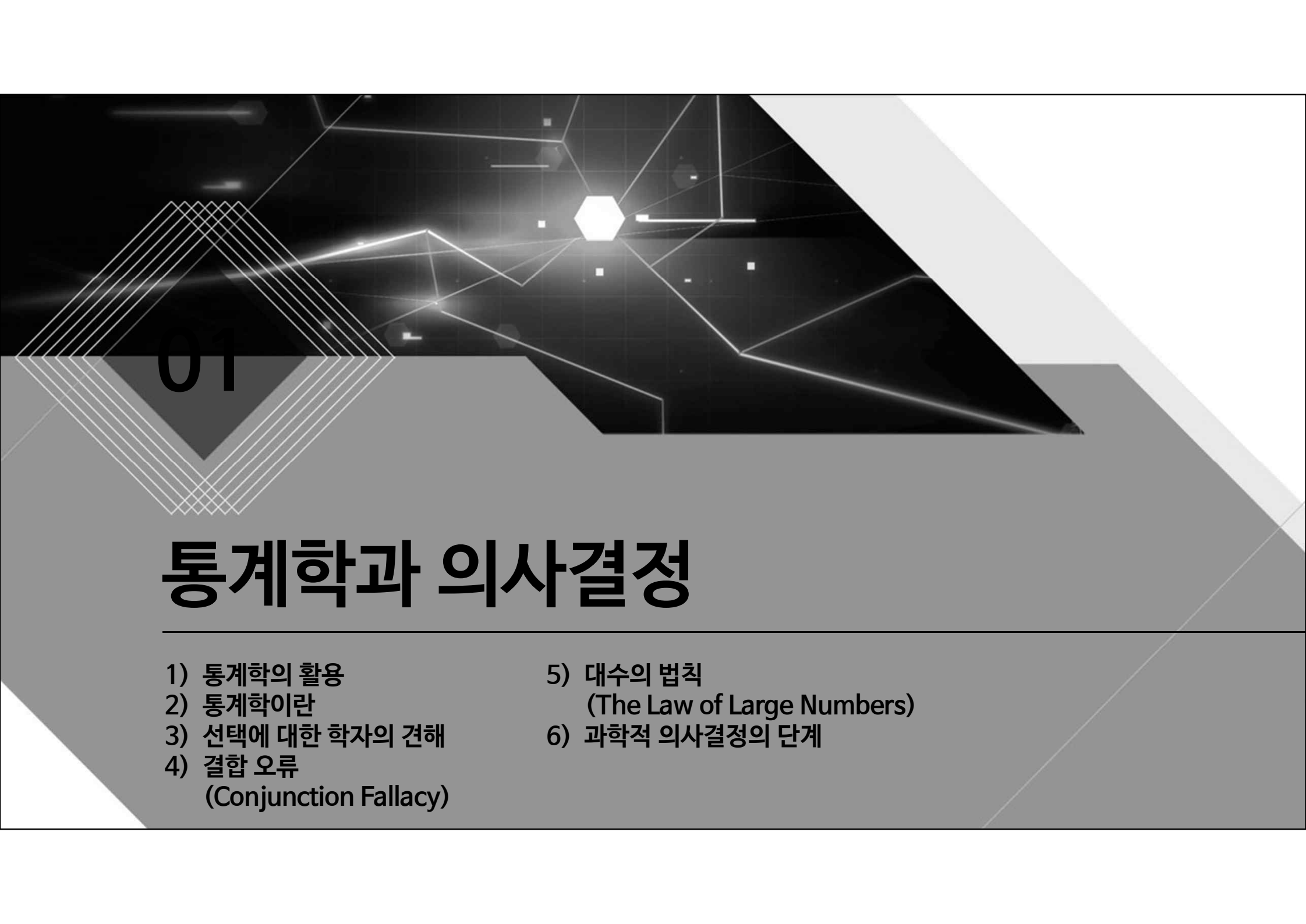
## 비행기 예약 문제

김세종씨는 일요일 저녁 상사로부터 홍콩에 있는 고객센터에서 연락이 와서 09:00 비행기를 타고 홍콩으로 오라는 전화를 받았다.

세종씨는 항공사 다섯 곳에 전화를 걸어 예약을 문의했다. 예약은 이미 완료되었고 대기 예약을 신청해 놓으면 25%, 30%, 15%, 20%, 25% 정도로 예약이 가능하다는 답변을 받았다.

이 가능성을 보고 세종씨는 내일 오전에 홍콩행 비행기를 타기는 어렵다는 생각을 했다.

세종씨의 판단은 과연 맞는 것일까?

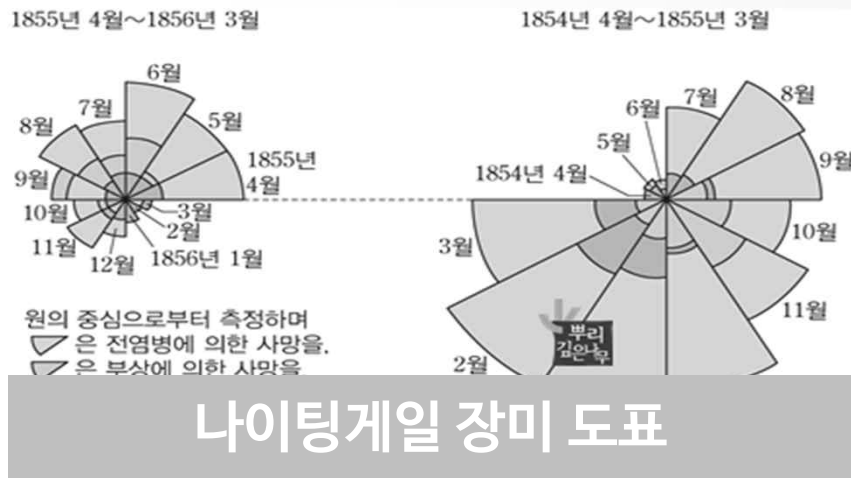


# 01 통계학과 의사결정

- 1) 통계학의 활용
- 2) 통계학이란
- 3) 선택에 대한 학자의 견해
- 4) 결합 오류  
(Conjunction Fallacy)
- 5) 대수의 법칙  
(The Law of Large Numbers)
- 6) 과학적 의사결정의 단계

# 1) 통계학의 활용

“ 통계학은 다양한 분야에서 활용되고 있음 ”



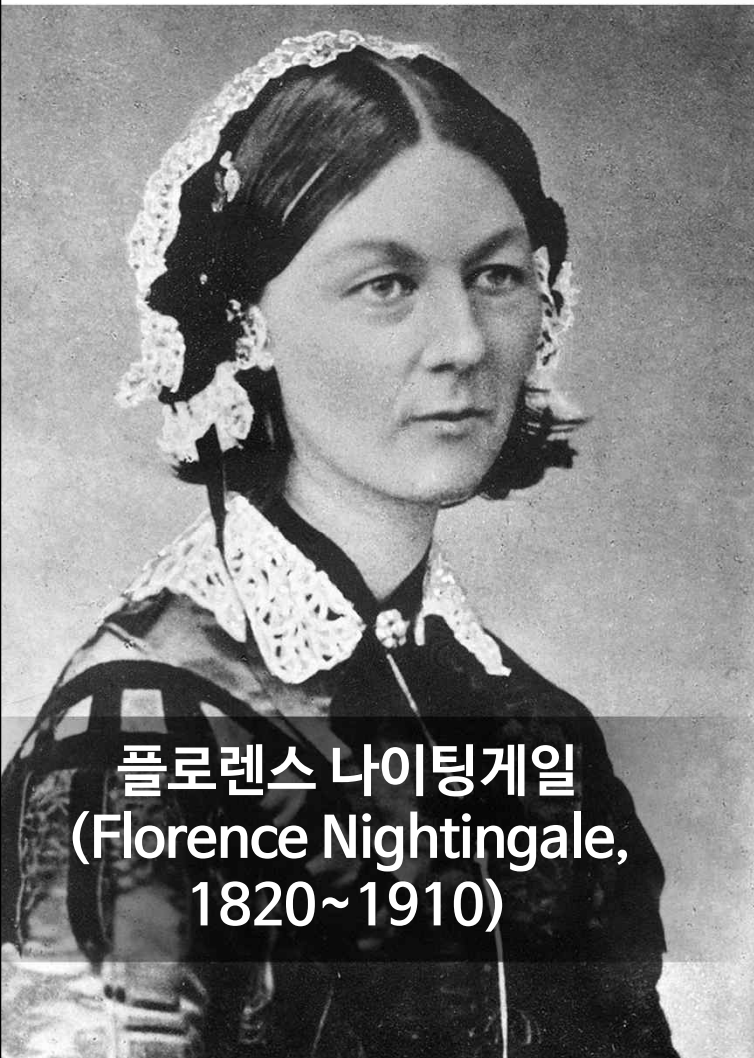
↳ 자료의 시각화



↳ 통계 분석을 활용하여 선거를 승리로 이끈 대표적인 사례

〈출처 : <https://bit.ly/2TaAStp>〉

## 1) 통계학의 활용



플로렌스 나이팅게일  
(Florence Nightingale,  
1820~1910)

“ 백의의 천사 ”

"전투에 의한 사망자보다,  
병원 환경에 의한 사망자가 더 많다."

# 1) 통계학의 활용

## ● 크림전쟁에서 나이팅게일



정부의 요청으로 야전병원으로 파견



병원 행정의 정상화



환자에 대한 정확한 기록과 관리

- 위생의 중요성 강조와 설득(어떻게?)
- 병원 내 사망률을 획기적으로 줄임
- 위생의 중요성을 사회에 알림



# 1) 통계학의 활용

- 학자로서의 나이팅게일

자료를 수집하고 요약하여  
설득의 도구로 자료와 통계를 사용함

자료를 숫자로만 제시하는 것이 아니라  
도표를 활용하여 변화를  
누구나 쉽게 알 수 있도록 함

“ 시각화의 선구자 ”

# 1) 통계학의 활용

- 학자로서의 나이팅게일

- ▶ 나이팅게일의 장미도표(Rose Diagram, Coxombs)

1단계

원을 중심각 30도씩 12조각으로 나눈 후  
각 조각을 월로 나타냄

2단계

각 조각별로 사망원인을 나타내는  
세 개의 쐐기(Wedge)를 겹쳐 놓음

- 사망원인
  - 파란색 : 질병
  - 빨간색 : 부상
  - 검은색 : 기타 이유

# 1) 통계학의 활용

- 학자로서의 나이팅게일

- ▶ 나이팅게일의 장미도표(Rose Diagram, Coxombs)

3단계

해당 월별 사망자 수를 각 썰기의 넓이로 함

- 면적이 넓은 썰기를 뒤에 배치하여  
작은 면적을 갖는 썰기를 가리지 않도록 함

4단계

나이팅게일의 활동(위생환경 개선) 이전 1년과  
이후 1년으로 두 장을 그린 후, 함께 배치하여  
비교할 수 있도록 함

## 2) 통계학이란

- 통계학(Statistics, 統計學)

### 통계학(사전적 정의)

수량적 비교를 기초로 하여, 많은 사실을 통계적으로 관찰하고 처리하는 방법을 연구하는 학문

〈출처 : 위키피디아〉



## 2) 통계학이란

- 통계학(Statistics, 統計學)

R. A. Fisher의  
통계학

관찰자료에 수학적 원리를 적용하는  
응용수학의 한 분야



다양한 사회 현상에 대해 자료를 바탕으로 신뢰할 만한  
정보를 제공하여 사회 현상을 파악하게 하는 학문



보다 효율적인 정보 전달 방법을 연구하는 분야이자  
중요한 통계학의 분야

## 2) 통계학이란

- 통계학(Statistics, 統計學)

- ▶ 연구대상 : R. A. Fisher

모집단

변동량

자료추약  
방법



## 2) 통계학이란

- 모집단과 표본, 그리고 기본원리

- ▶ 모집단

모집단

우리가 알고자 하는 대상 전체

↳ 조사 대상의 범위

전수조사

모집단 전체를 조사하는 방법

## 2) 통계학이란

- 모집단과 표본, 그리고 기본원리

- ▶ 표본

### 표본

모집단으로부터 조사하기 위해 선택된 조사 대상





## 2) 통계학이란

- 모집단과 표본, 그리고 기본원리

- ▶ 표본

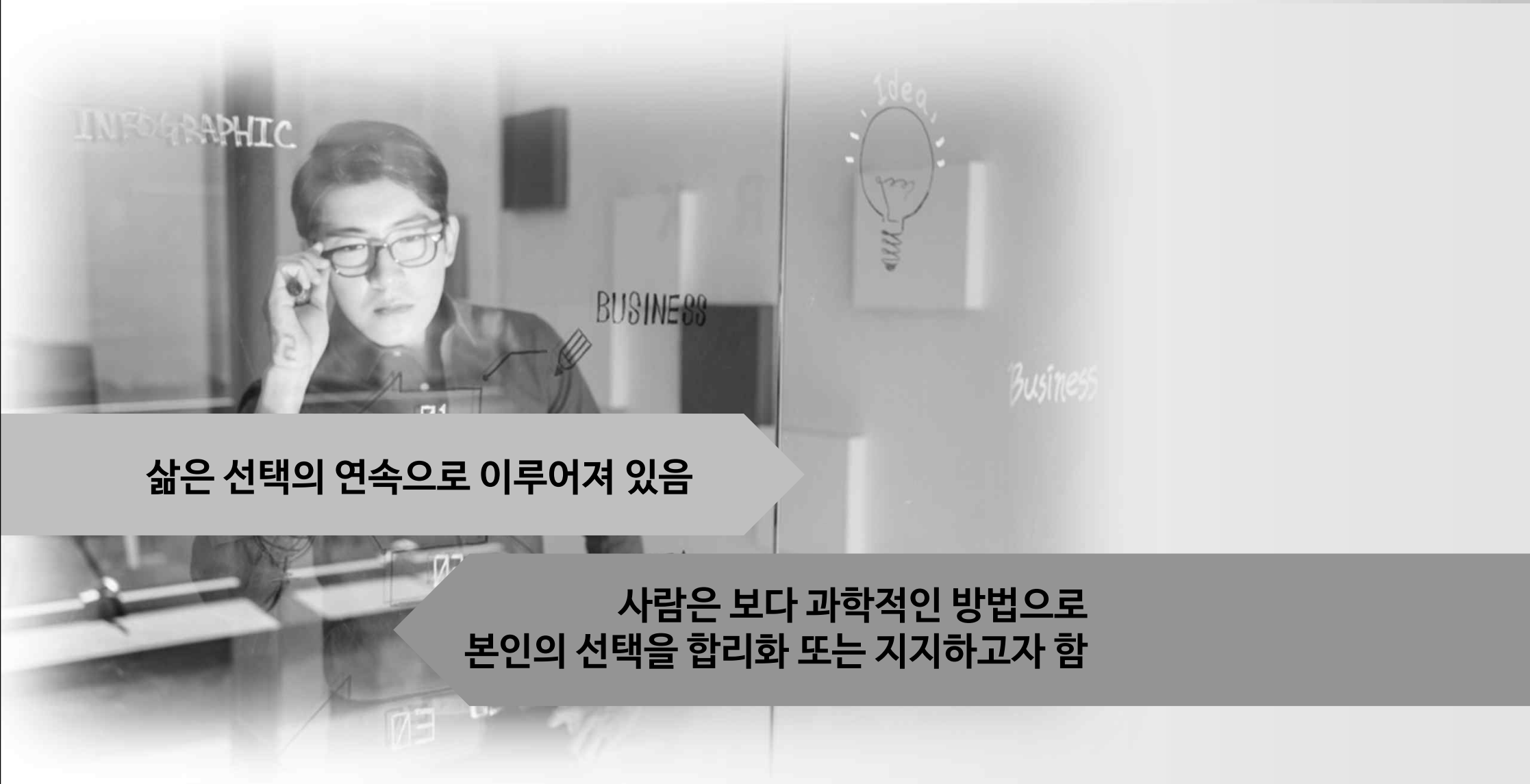
### 실제 조사 대상

- 모집단 전체를 조사하는 것이 불가능한 경우
- 수류탄과 같이 조사하면 사라지는 특성을 가진 조사 대상
- 시간적 · 공간적 제약이 있을 시 모집단을 잘 대표할 수 있는 조사 대상

### 표본조사

표본을 조사 대상으로 조사하는 방법

### 3) 선택에 대한 학자의 견해



삶은 선택의 연속으로 이루어져 있음

사람은 보다 과학적인 방법으로  
본인의 선택을 합리화 또는 지지하고자 함

### 3) 선택에 대한 학자의 견해



“훗날에 훗날에  
나는 어디선가 한숨을 쉬며 이야기할 것입니다.  
숲 속에 두 갈래 길이 있었다고,  
나는 사람이 적게 간 길을 택하였다고,  
그리고 그것 때문에 모든 것이 달라졌다고.”

- 로버트 프로스트

### 3) 선택에 대한 학자의 견해

“진정한 발견은 새로운 장소를 찾는 것이 아니라,  
새로운 관점을 갖는 것이다.”

- 마르셀 프루스트



### 3) 선택에 대한 학자의 견해



“탐색적 데이터 분석은 우리가 존재한다고 믿는 것들은 물론이고 존재하지 않는다고 믿는 것들을 발견하려는 태도, 유연성, 그리고 자발성이다.”

- 존 튜키

## 4) 결합 오류 (Conjunction Fallacy)

강세종씨는 33세이며 미혼이며, 건장한 체격을 가지고 있으며, 매우 활발하고, 지적인 이미지를 가지고 있다. 대학에서는 컴퓨터 공학과 심리학을 전공하였으며 소수자 차별, 비핵화 문제 등 사회적 이슈에 대해서도 관심이 많아 시민 운동에도 참여하였다.

상기에 기술된 강세종씨의 10개의 특징으로 부터 강세종씨라고 특징 지을 수 있는 가능성이 가장 높은 것부터 순서를 매겨보세요.

## 4) 결합 오류 (Conjunction Fallacy)

☐ a

강세종씨는 컴퓨터 프로그래머이다.

☐ b

강세종씨는 사회운동가이다.

☐ c

강세종씨는 컴퓨터 프로그래밍을 학원에서 가르치며,  
방사선 피해자들을 상담하는 심리 상담가이다.

☐ d

강세종씨는 건강을 위해 헬스를 하며, 인권운동을 하고 있다.

☐ e

강세종씨는 결혼을 앞두고 있는 대학강사이다.

## 5) 대수의 법칙(The Law of Large Numbers)

저출산 문제를 해결할 수 있는 정책을 마련하고자 고심하고 있는 복지부의 김세종 과장은 출산 현황을 파악하기 위해 서울에 있는 대형 산부인과 A와 소도시에 있는 소형 산부인과 B의 출산 자료를 확보하여 분석을 했다.

A 산부인과에서는 매일 45명의 신생아가 태어났고, B 산부인과에서는 매일 15명의 신생아가 태어났다. 통계적으로 신생아의 성별은 50%는 남아, 50%는 여아라고 생각할 수 있다. 그러나 두 병원에서 매일 태어나는 남녀의 비율은 어떤 날은 남아 30%, 여아 70%, 어떤 날은 남아 60%, 여아 40% 등으로 다르다. 각 병원에서는 신생아의 성별의 차이가 5% 이상 나는 날을 기록을 해 두었다. 그리고 김과장은 성별의 차이가 5% 이상 나는 날을 세어 보았다. 어느 병원에서 성별의 차이가 5% 이상 나는 날이 더 많을까?

1 A 병원

2 B 병원

3 대략 같다.  
(5% 오차 이내)

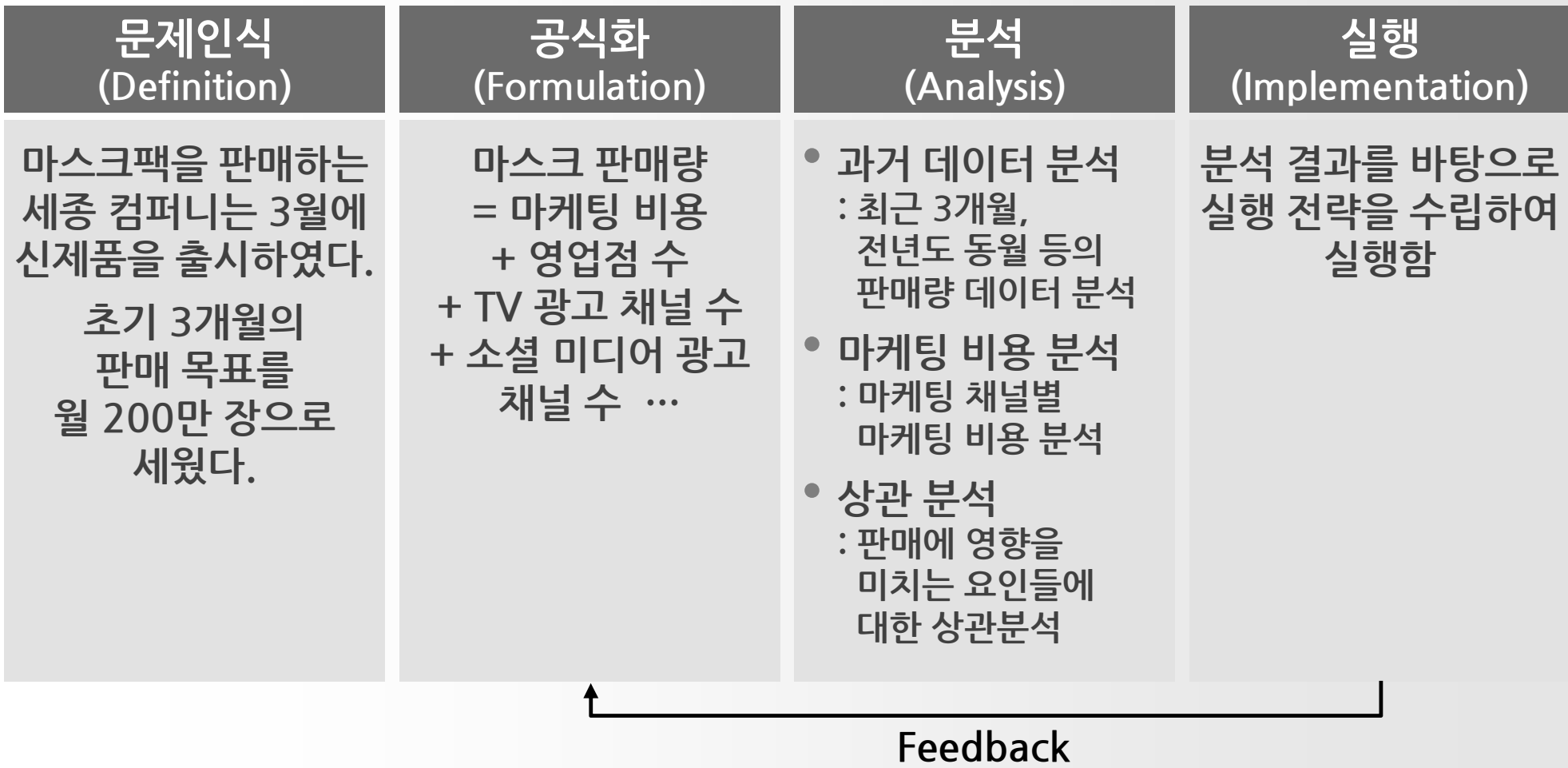


## 6) 과학적 의사결정의 단계



{ 통계학은 공식화와 분석 분야에 활용되어  
과학적 의사결정을 지원함 }

## 6) 과학적 의사결정의 단계





02

# 자료의 요약과 도시

---

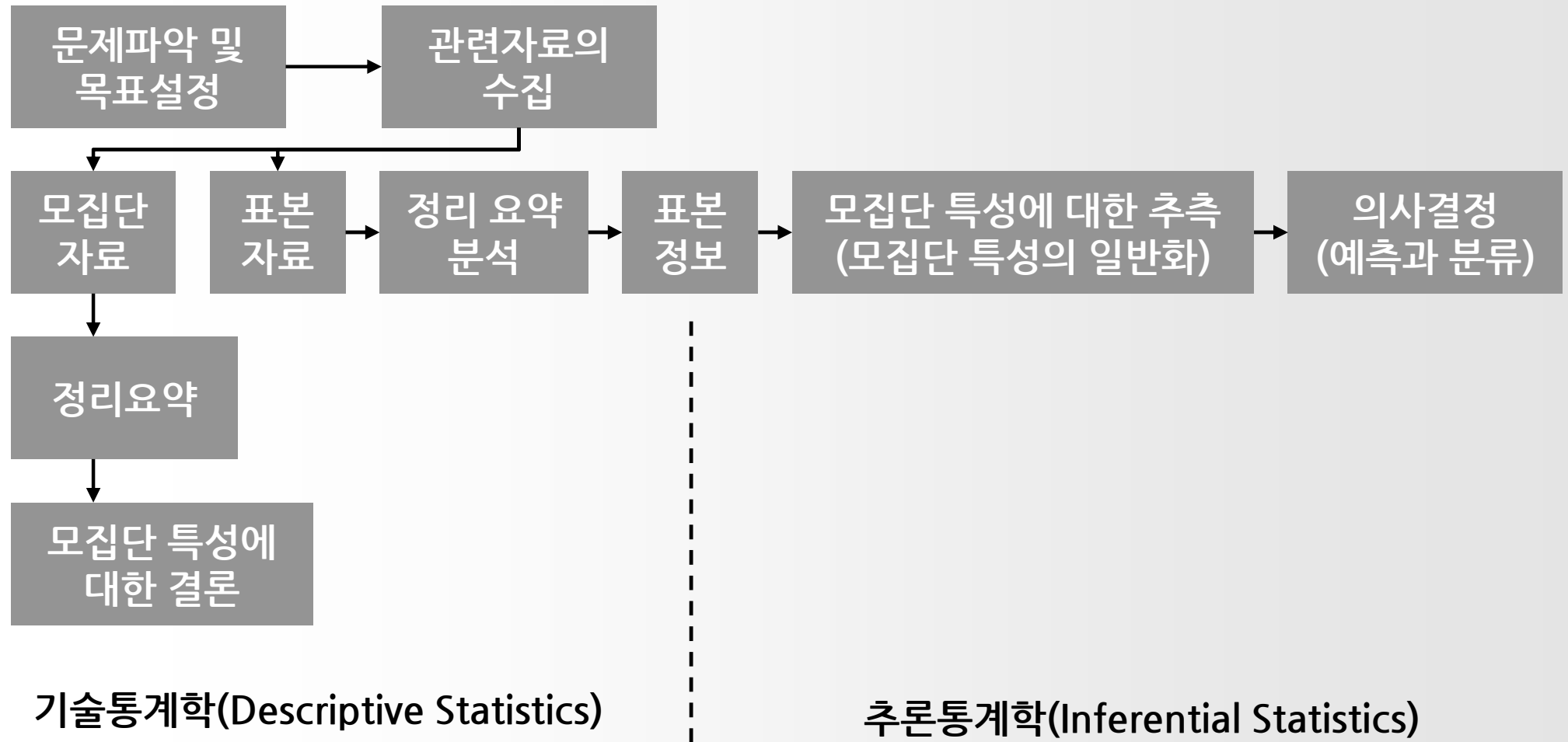
- 1) 통계학과 의사결정
- 2) 자료의 요약과 도시
- 3) 자료의 정리와 도시

## 1) 통계학과 의사결정



{ 기술통계 분석을 거쳐 가설을 세우고  
의사결정을 위한 추론통계의 절차로 진행됨 }

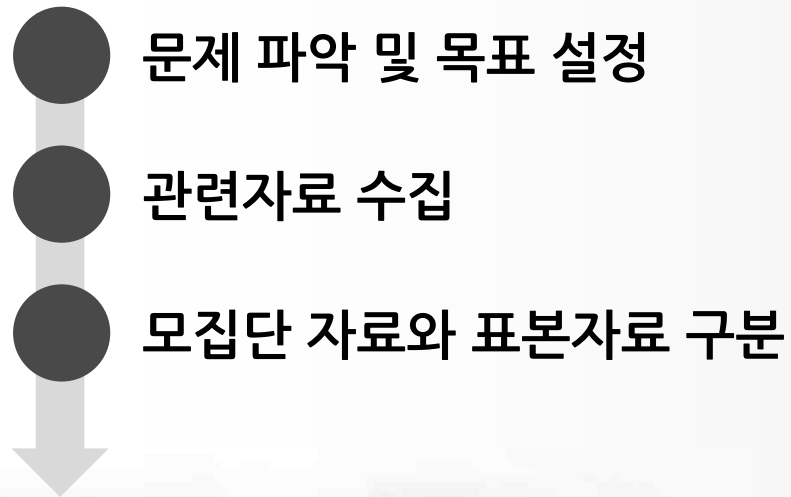
# 1) 통계학과 의사결정



# 1) 통계학과 의사결정

- 통계분석 절차

EDA(Exploratory Data Analysis)



# 1) 통계학과 의사결정

## ● 통계분석 절차

### 기술통계학

- 모집단 자료를 정리 요약하여 모집단 특성에 대한 결론을 도출함
- 자료의 특성을 나타냄

### 추론통계학

- 표본 정보를 바탕으로 모집단 특성에 대한 추측과 의사결정을 위한 예측 및 분류
- 기술통계의 과정을 통해 획득한 정보를 분석주제의 검증 과정을 거쳐 결론을 도출함으로 의사결정에 기여함

# 1) 통계학과 의사결정

## ● 기술통계학(Descriptive Statistics)

### 기술통계학

자료를 수집하고 정리하여 도표나 표를 만들거나 자료를 요약하여 대표값이나 변동의 크기를 구하는 방법을 다루는 분야



평균, 중앙값, 최빈값 등을 활용하여 중심화 경향 확인



분산, 표준편차, 백분위수, 최대값, 최소값을 활용하여 퍼짐 정도 확인



왜도와 첨도 등을 활용하여 분포형태와 대칭 여부 확인



# 1) 통계학과 의사결정

- 추론통계학(Inferential Statistics)

## 추론통계학

자료에 내포되어 있는 정보를 분석하여 불확실에 대한 추론을 다루는 분야



통계적 모형을 설정하고, 설정된 모형이 합리적인지 평가함



자료로부터 얻어지는 정보를 근거로  
미지의 특성에 대한 결론을 내리고  
미래에 일어날 현상에 대한 예측을 함

# 1) 통계학과 의사결정

- 추론통계학(Inferential Statistics)

모집단으로부터 얻은 표본자료를 사용하여  
모집단 전체에 대한 특징을 추측함

모집단에 대한 일련의 의사결정 방법을  
연구하는 분야로 현대 통계학의 핵심 분야



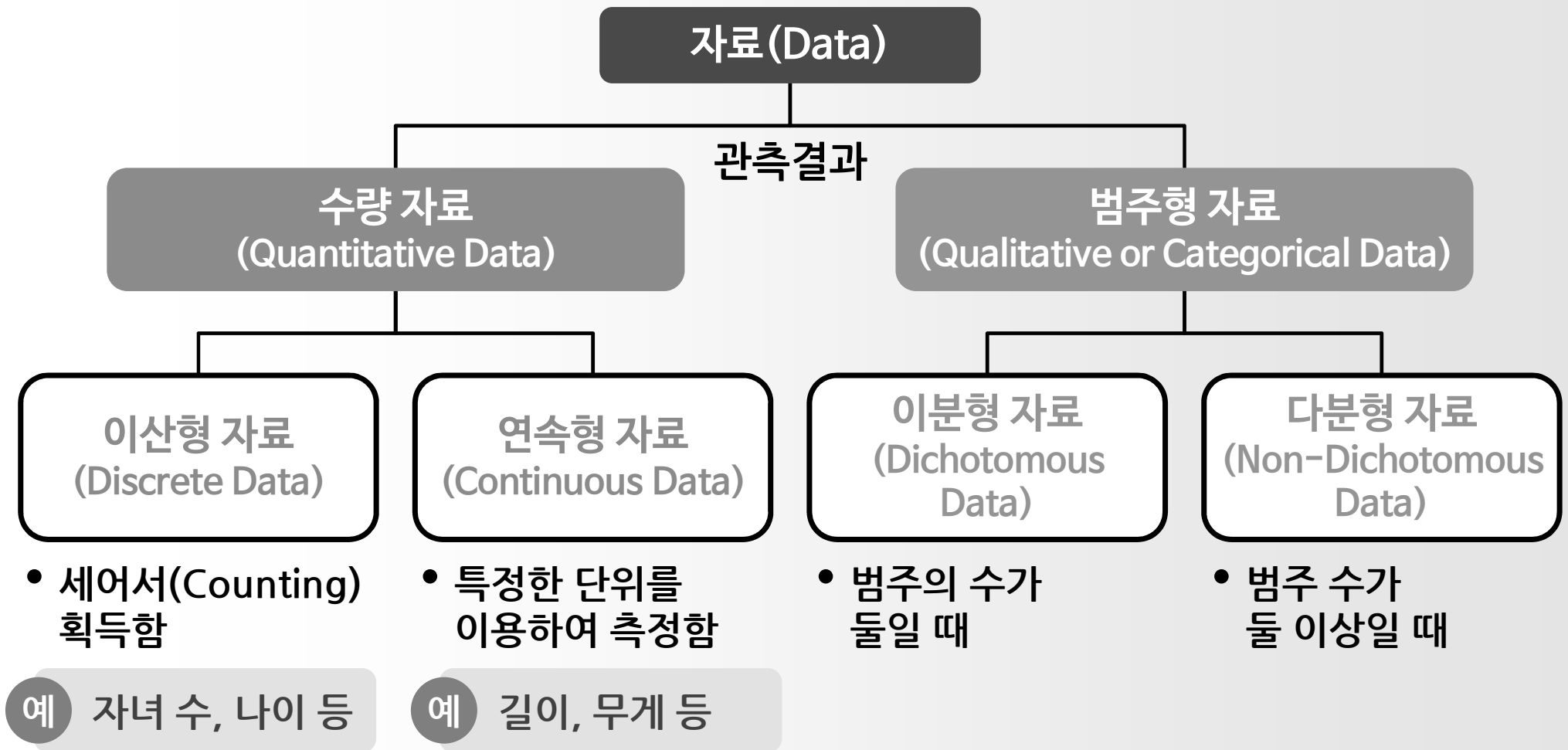
## 2) 자료의 요약과 도시

자료(Data)

관심의 대상이 되는 개체의 성격이나 속성을  
관측한 결과(Outcomes of Observation)



## 2) 자료의 요약과 도시



## 2) 자료의 요약과 도시

{ 자료(Data)를 나타내는 수치에 사용된 척도의 구분 }

01 수의 순서(Order) 유무

02 수 사이 거리(Distance) 개념 유무

03 수에 의미 있는 원점(Natural Zero Point) 유무



명목척도

서열척도

구간척도

비율척도

## 2) 자료의 요약과 도시

### 명목척도

단지 관측 결과의 특성을 분류하고자 할 때  
사용함

예 VVIP 고객은 1번, VIP 고객은 2번,  
일반 고객은 3번을 부여할 때 사용하는 숫자

### 서열척도

수의 순서만 있으며 거리나 의미 있는  
원점은 없는 경우

예 품질 수준에 따라 고급 1번, 중급 2번,  
저급 3번을 부여하는 경우

## 2) 자료의 요약과 도시

### 구간척도

순서와 거리의 개념은 있으나 의미 있는 원점이 없는 경우

예 온도, 시간

### 비율척도

순서, 거리 뿐 아니라 의미 있는 원점도 있는 경우

예 무게, 길이, 학생 수, 자녀 수 등

## 2) 자료의 요약과 도시

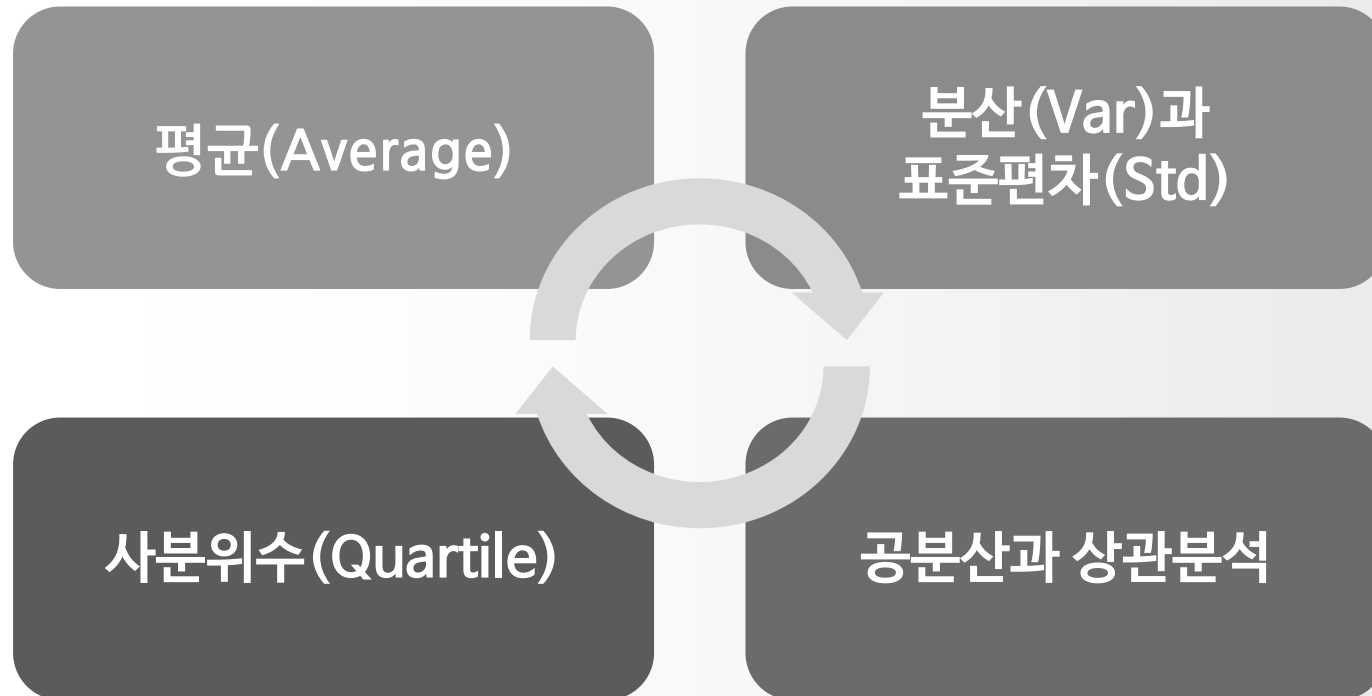
구분	관측결과의 특성	순서	거리	의미 있는 원점 (비율)
명목척도	O	X	X	X
서열척도	O	O	X	X
구간척도	O	O	O	X
비율척도	O	O	O	O

〈출처 : 통계학의 이해 노부호 외, 법문사〉



## 2) 자료의 요약과 도시

- 기술통계학의 주요 지표



## 2) 자료의 요약과 도시

### ● 기술통계학의 주요 지표

평균

분산과  
표준편차

사분위수

공분산과  
상관분석

- 평균에 관심을 갖는 이유는 예측을 하기 위함

예 어떤 학생의 3개월 간 수학 평균 점수가 90점이라면 4번째 시험에서도 90점 정도의 점수를 받게 될 것이라고 예측할 수 있음

- 평균의 함정에 유의해야 함

➡ [https://dbr.donga.com/article/view/1303/article\\_no/6962](https://dbr.donga.com/article/view/1303/article_no/6962)

## ▶ 강 평균 깊이 150cm, 우리 군사들 건너라?

군치'로 뭉뚱그려 설명한 것이 무리였다"고 말했다. **1** 여기서 말하는 우화는 다음과 같다. "100명의 군인들이 강을 건넌다. 군인들의 평균 키는 180cm, 강의 평균 깊이는 150cm다. 보고를 받은 장군은 도강을 명령했다. 강 언저리를 지나면서 물이 갑자기 깊어졌고 병사들이 한 명, 두 명 빠져죽기 시작했다. 겁이 난 병사들은 뒤를 흘깃흘깃 쳐다봤지만 장군은 '돌격 앞으로'만 외쳤다. 물에 빠져죽는 병사가 속출하자 장군은 당황했다. 그제야 장군은 회군을 명령했다. 하지만 이미 많은 군사를 잃은 뒤였다. 알고 보니 이 강의 최대 수심은 2m였고, 군사 중 2m가 넘는 사람은 30명이 채 안 됐다."이번 연말정산 사태

## 2) 자료의 요약과 도시

### ● 기술통계학의 주요 지표

평균

분산과  
표준편차

사분위수

공분산과  
상관분석

- 분산은 관측 값에서 평균을 뺀 값을 제곱하고, 그것을 모두 더한 후 전체의 개수로 나누어 구함
- 표준편차는 분산 값에 루트를 씌워서 평균과 단위가 맞는 편차의 평균 값을 나타냄
- 분산과 표준편차에 관심을 갖는 이유는 예측의 정확성과 집단의 동질성을 파악하기 위함

## 2) 자료의 요약과 도시

### ● 기술통계학의 주요 지표

평균

분산과  
표준편차

사분위수

공분산과  
상관분석

- 사분위수(0, 25%, 50%, 75% 등)는 데이터 표본을 4개의 동일한 부분으로 나눈 값
- 사분위수를 활용하여 데이터 집합의 범위와 중심 위치를 직관적으로 파악이 가능함

## 2) 자료의 요약과 도시

### ● 기술통계학의 주요 지표

평균

분산과  
표준편차

사분위수

공분산과  
상관분석

- 공분산은 두 개의 확률변수의 관계를 보여주는 값, 즉 확률 변수 X에 대해 Y가 변하는 정도를 나타내는 값
- 상관분석은 두 확률변수 간에 어떤 선형적 관계를 갖고 있는지 분석하는 방법
- 두 변수는 서로 독립적이거나 상관된 관계일 수 있으며, 두 변수 간의 관계의 강도를 상관 관계라고 함

## 2) 자료의 요약과 도시

R은...

각 패키지의 Cheatsheet(간략참조자료)를  
제공하고 있음

이 자료를 참고하면  
보다 쉽게 활용법을 이용할 수 있음



## 2) 자료의 요약과 도시

{ 다음의 메소드는 dplyr과 ggplot Cheetsheet에서 일부 인용 }

구분	절차	메소드
탐색 (Exploration)	Summarize	summarise( ), count( ), summarise_all( ), summarise_at( ), summarise_if( ), mean( ), sum( ), first( ), last( ), quantile( )
	NA Handling	isna(), sum(isna()), na.omit( ), complete.cases( ), sapply(dataset, function(x) ifelse(is.na(x), mean(x, na.rm=TRUE), x))
	Subset Observation	filter(.data...), distinct, sample_frac, sample_n, slice, top_n, select(data.), contain, ends_with, matchs, starts_with,



## 2) 자료의 요약과 도시

{ 다음의 메소드는 dplyr과 ggplot Cheetsheet에서 일부 인용 }

구분	절차	메소드
재구조화 (Restructuring) 분석(Analysis)	Reshaping	mutate, transmute, mutate_all, add_column, rename
	Combine	bind_cols( ), left_join, right_join, inner_join, full_join, rbind, cbind, merge
	Group Data	group_by, ungroup,
시각화 (Visualization)	Data Visualization	geom_area(stat ='bin), geom_bar( ), geom_blank( ), geom_boxplot( ), geom_violin( )
분석 (Analysis)	Statistical Test	correlation test, t-test, Regression, Anova, Chisquer-terst, Wilconxon test, Anderson test

## 2) 자료의 요약과 도시



검색창에 패키지명과 Cheatsheet를 입력하면 각 패키지별 Cheatsheet를 볼 수 있음

➡ <https://rstudio.com/resources/cheatsheets/>



## 2) 자료의 요약과 도시

Bar Plot은  
그룹 또는 범주 데이터의 빈도수를 보여줌

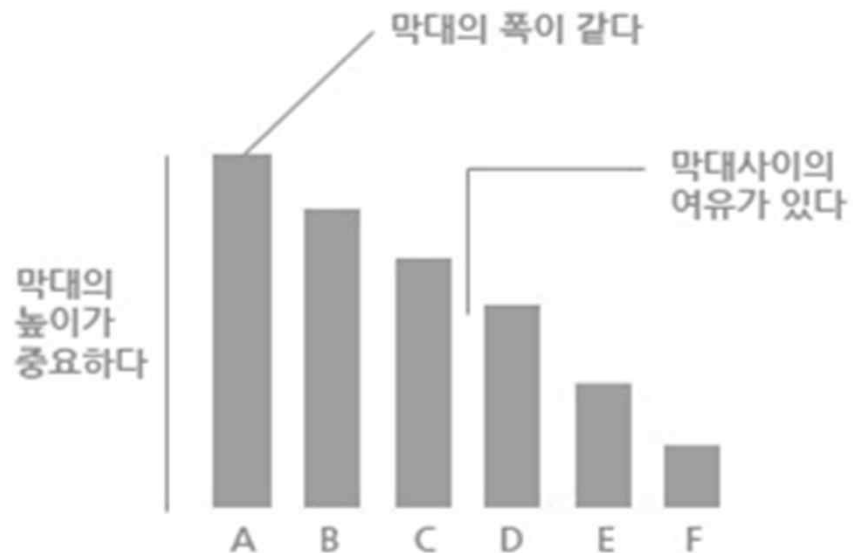
Histogram은 연속적인 변수의  
구간별 빈도수를 보여줌



## ▶ Bar Plot vs. Histogram

### Bar Plot

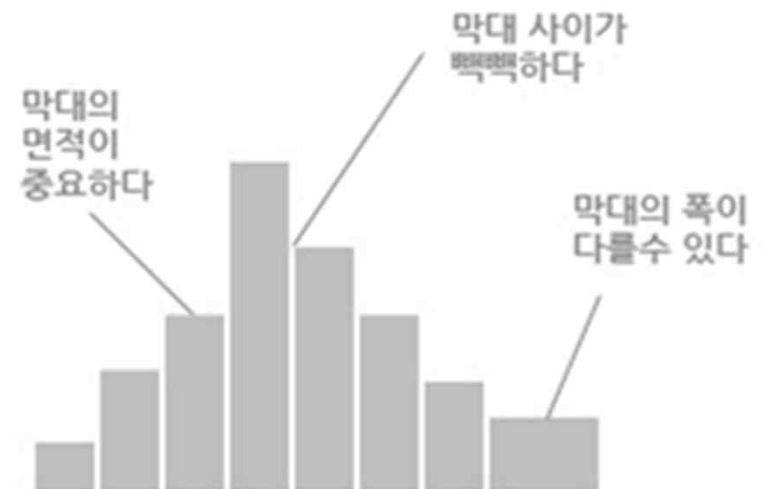
- 범주 데이터의 빈도를 나타낼 때 사용
- X값은 꼭 연속형이 아니어도 됨
- Y값은 X값의 크기



Vs.

### Histogram

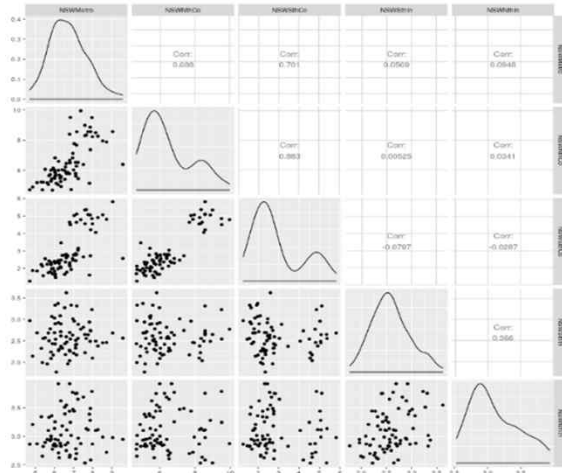
- 도수분포를 막대그래프로 나타낸 그래프
- X값은 연속적이고 분할 가능한 수
- Y값은 X의 빈도수를 표시



## ▶ Scatter Plot vs. Box Plot

### Scatter Plot

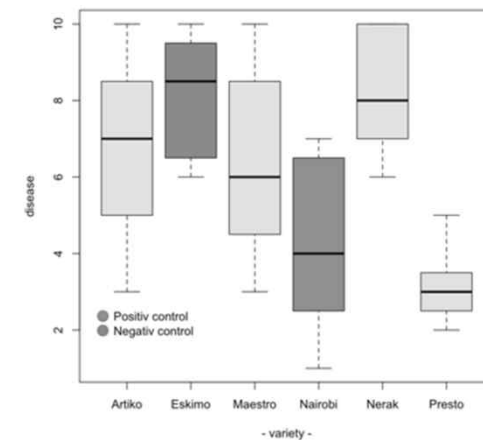
- 직교 좌표계를 이용해 두 개 변수 간의 관계 표현
- 두 개 변수 간의 선형 또는 비선형 형태와 같은 수학적 모델을 확인하여 방향성과 강도 확인



Vs.

### Box Plot

- 데이터 집합의 범위와 중앙값을 빠르게 확인
- 통계적 이상치 확인이 용이함
- 최소값, 최대값 및 제 1, 2, 3 분위수의 값 확인



〈참조 : 데이터 시각화 예, <https://www.tableau.com/ko-kr/learn/articles/best-beautiful-data-visualization-examples>〉



03

# 자료의 요약(실습)

---

1) 코로나19 현황 데이터를 활용한 EDA 및 시각화

## 2) 자료의 요약과 도시

```
install.packages('dplyr')
```

- ▶ # EDA를 위한 dplyr 패키지 설치

```
install.packages('readxl')
```

- ▶ # 엑셀 파일 불러오기 위한 패키지 설치

## 2) 자료의 요약과 도시

```
require(dplyr)
```

- ▶ # 패키지 불러오기,
- ▶ 설치된 패키지라도 사용할 때마다 불러와야 함

```
require(readxl)
```

- ▶ # 패키지 불러오기



## 2) 자료의 요약과 도시

```
setwd('C:/Users/tkpeo/Documents/r_statistics')
```

- ▶ # working directory 세팅, 파일을 저장해 놓은 경로를 설정함

## 2) 자료의 요약과 도시

### #0. 데이터 불러오기

```
covid <- read_excel('covid19.xlsx')
```

▶ # 파일 불러오기

```
is.data.frame(covid)
```

▶ # 데이터 형태 확인

## 2) 자료의 요약과 도시

### #1. 데이터 구조 확인

`dim(covid)`

▶ # 데이터 프레임의 행과 열의 개수 확인 360행, 5개 열

`nrow(covid)`

▶ # 행의 개수 확인 360행

`ncol(covid)`

▶ # 열의 개수 확인 5컬럼

`head(covid)`

▶ # 상위 6개의 값 확인

## 2) 자료의 요약과 도시

### #1. 데이터 구조 확인

```
tail(covid)
```

▶ # 하위 6개 값 확인

```
names(covid)
```

▶ # 컬럼명 확인

```
names(covid) <- c('time', 'location',  
                  'state1', 'state2', 'count')
```

▶ # 컬럼명 바꾸기

## 2) 자료의 요약과 도시

### #2. 결측치 확인

```
is.na(covid)
```

▶ # 전체 확인

```
sum(is.na(covid))
```

▶ # 전체 na값 합계 확인

```
colSums(is.na(covid))
```

▶ # 컬럼별 na값 확인

## 2) 자료의 요약과 도시

### #3. 날짜 데이터 다루기

```
install.packages('lubridate')
```

▶ # 날짜 다루기 위한 패키지 설치

```
require(lubridate)
```

▶ # 패키지 불러오기

```
covid$newdate <- date(covid$time)
```

▶ # 일자별 발표 시간은 매일 동일하여 시간 제외한 날짜 데이터 생성

## 2) 자료의 요약과 도시

### #4. 2월 29일 지역별 검사현황 및 확진자 현황

```
covid %>% filter(newdate == "2020-02-29") %>%
```

```
group_by(location, state1) %>%
```

## 2) 자료의 요약과 도시

### #4. 2월 29일 지역별 검사현황 및 확진자 현황

```
summarise(total = sum(count), ► # 검사 현황 및 확진자 현황
           mean = mean(count), ► # 검사 현황 및 확진자 평균
           var = var(count), ► # 검사 현황 및 확진자 분산
           std = sd(count), ► # 검사 현황 및 확진자 표준편차
           '25%' = quantile(count, probs = 0.25), ► # 검사 현황 및 확진자 4분위수
           '50%' = quantile(count, probs = 0.5),
           '75%' = quantile(count, probs = 0.75),
           iqr = IQR(count))
```



## 2) 자료의 요약과 도시

```
descriptive <- covid %>% filter(newdate  
  == '2020-02-29') %>%
```

▶ # descriptive 변수로 저장하기

```
group_by(location, state1) %>%
```

## 2) 자료의 요약과 도시

```
summarise(total = sum(count),  
           mean = mean(count),  
           var = var(count),  
           std = sd(count),  
           '25%' = quantile(count, probs = 0.25),  
           '50%' = quantile(count, probs = 0.5),  
           '75%' = quantile(count, probs = 0.75),  
           iqr = IQR(count))
```

## 2) 자료의 요약과 도시

### #5. 지역별 확진 및 검사 현황 그래프 그리기

```
require(ggplot2)
ggplot(data = descriptive, aes(x = location, y = total, fill = state1))+
  geom_bar(stat = 'identity')+
  ggtitle('코로나19 현황')+
  theme(plot.title = element_text(hjust = 0.5))
```

## 2) 자료의 요약과 도시

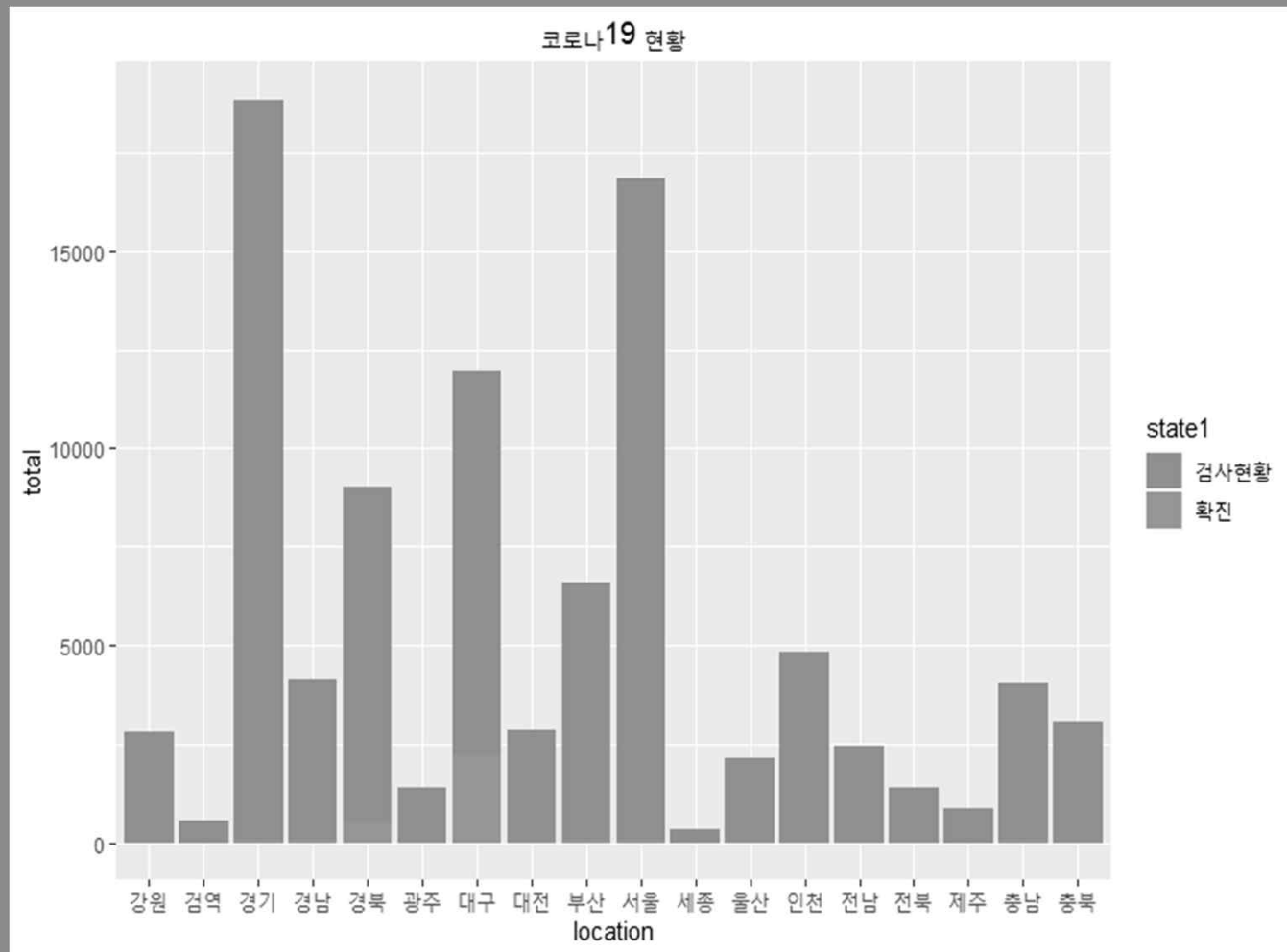
### #5. 지역별 확진 및 검사 현황 그래프 그리기

```
ggplot(data = descriptive %>% filter(state1 == '확진'), aes(x = location, y = total))+  
  geom_bar(stat = 'identity')+  
  ggtitle('코로나19 확진현황')+  
  theme(plot.title = element_text(hjust = 0.5))
```

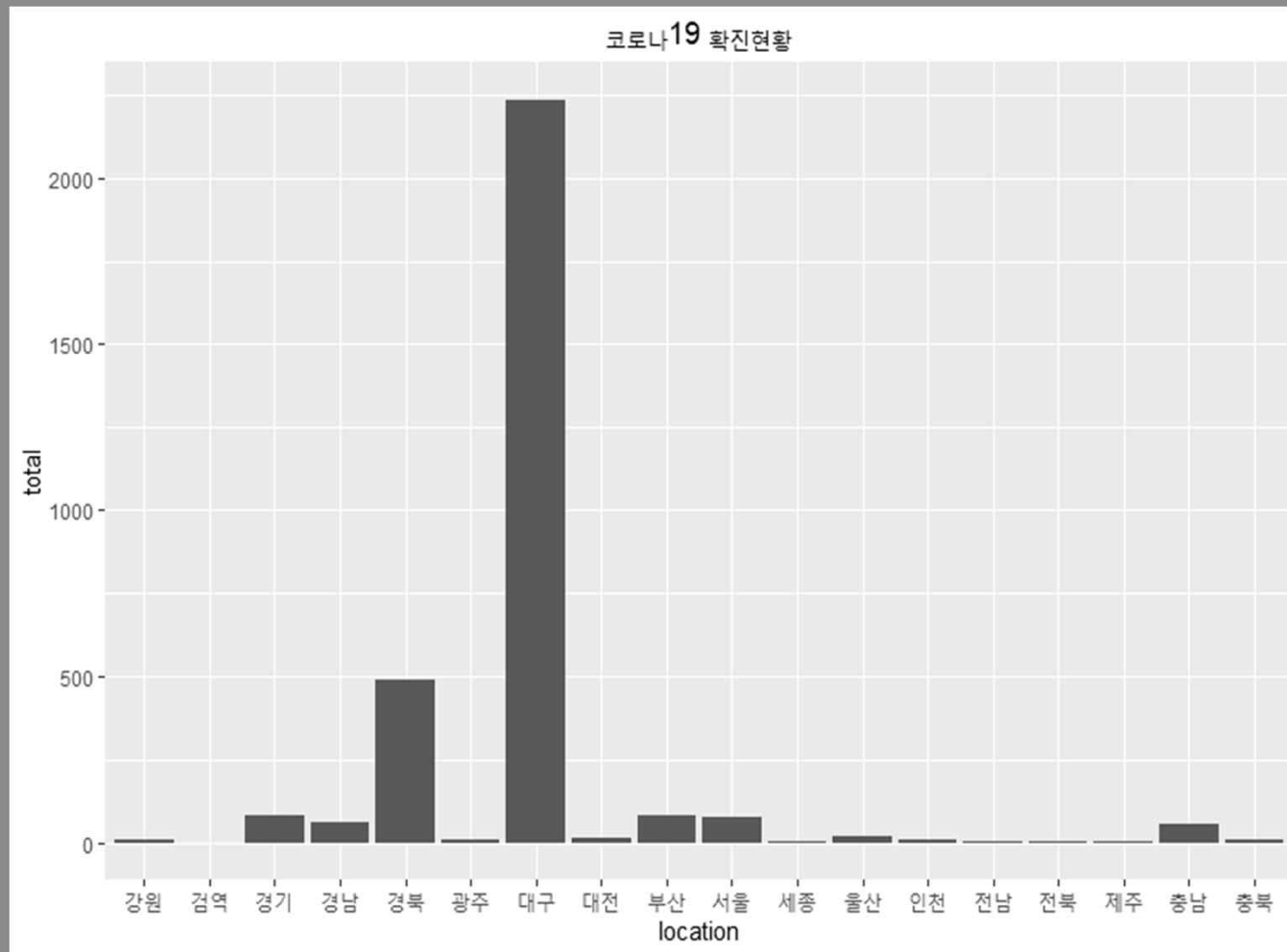
```
ggplot(data = descriptive %>% filter(state1 == '검사현황'), aes(x = location, y =  
total))+  
  geom_bar(stat = 'identity')+  
  ggtitle('코로나19 검진현황')+  
  theme(plot.title = element_text(hjust = 0.5))
```

〈소스 및 예제 파일 다운로드: [https://github.com/LEESUAJE1978/r\\_statistics](https://github.com/LEESUAJE1978/r_statistics)〉

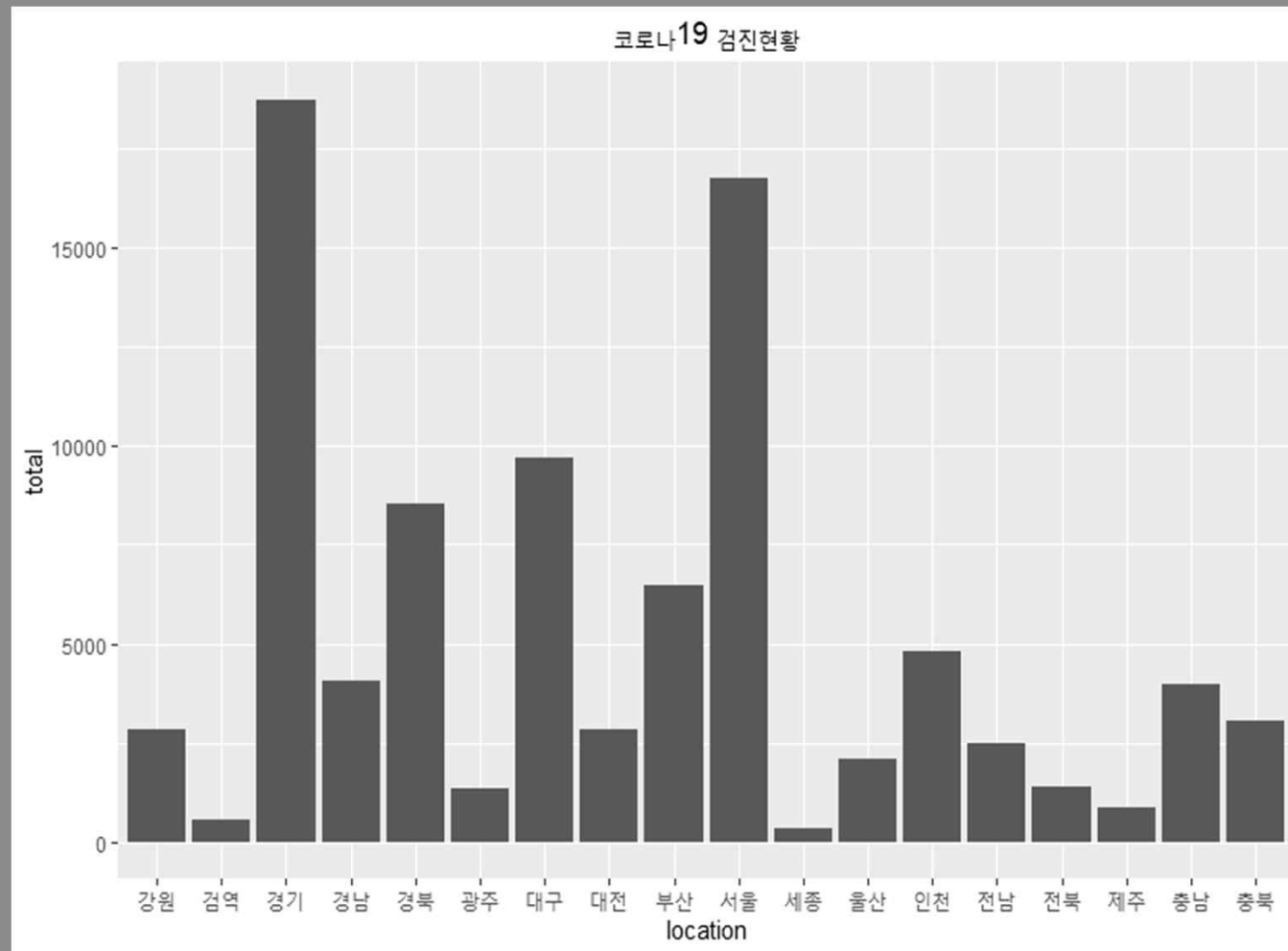
## ▶ 코로나19 현황 데이터를 활용한 EDA 및 시각화



## ▶ 코로나19 현황 데이터를 활용한 EDA 및 시각화



## ▶ 코로나19 현황 데이터를 활용한 EDA 및 시각화



〈소스 및 예제 파일 다운로드 : [https://github.com/LEESUAJE1978/r\\_statistics](https://github.com/LEESUAJE1978/r_statistics)〉

# 학습 평가

Q1

Q2

Q1

대수의 법칙(The Law Of Large Numbers)에 대해서 바르게 설명한 것은?

- 1 동일한 확률 분포를 가진 독립 확률 변수  $n$ 개의 평균의 분포는  $n$ 이 적당히 크다면 정규 분포에 가까워진다.
- 2 큰 모집단에서 무작위로 뽑은 표본의 평균이 전체 모집단의 평균과 가까울 가능성이 높다.
- 3 어떤 주어진 값들을 크기의 순서대로 정렬했을 때 가장 중앙에 위치하는 값이다.
- 4 확률변수가 기댓값으로부터 얼마나 떨어진 곳에 분포하는지를 가늠하는 숫자이다.



# 학습 평가

Q1

Q2

Q1

대수의 법칙(The Law Of Large Numbers)에 대해서 바르게 설명한 것은?

- 1 동일한 확률 분포를 가진 독립 확률 변수  $n$ 개의 평균의 분포는  $n$ 이 적당히 크다면 정규 분포에 가까워진다.
- ☒ 2 큰 모집단에서 무작위로 뽑은 표본의 평균이 전체 모집단의 평균과 가까울 가능성이 높다.
- 3 어떤 주어진 값들을 크기의 순서대로 정렬했을 때 가장 중앙에 위치하는 값이다.
- 4 확률변수가 기댓값으로부터 얼마나 떨어진 곳에 분포하는지를 가늠하는 숫자이다.

정답

2

해설

1번은 중심극한 정리에 대한 설명이고 3번은 중앙값에 대한 설명, 4번은 분산에 대한 설명입니다.

# 학습 평가

Q1

Q2

Q2

다음의 특성을 가진 그래프의 종류는 무엇인지 고르시오.

- 데이터 집합의 범위와 중앙값을 빠르게 확인
- 통계적 이상치 확인 용이
- 최소값, 최대값 및 제 1, 2, 3 분위수의 값 확인

1 히스토그램(Histogram)

3 박스플롯(Box Plot)

2 산점도(Scatter Plot)

4 바플롯(Bar Plot)

# 학습 평가

Q1

Q2

Q2

다음의 특성을 가진 그래프의 종류는 무엇인지 고르시오.

- 데이터 집합의 범위와 중앙값을 빠르게 확인
- 통계적 이상치 확인 용이
- 최소값, 최대값 및 제 1, 2, 3 분위수의 값 확인

1 히스토그램(Histogram)

☒ 박스플랏(Box Plot)

2 산점도(Scatter Plot)

4 바플랏(Bar Plot)

정답

3

해설

박스 플랏은 분위수와 이상치를 시각화한 것으로 데이터의 퍼짐 정도와 이상치 등을 직관적으로 파악할 수 있습니다.

# 정리 하기

## 기술통계학

### ✓ 통계학이란

- 수량적 비교를 기초로 하여, 많은 사실을 통계적으로 관찰하고 처리하는 방법을 연구하는 학문
- 통계분석은 기술통계학과 추론 통계학으로 구분되고 기술 통계 분석을 거쳐 가설을 세우고 의사결정을 위한 추론 통계의 절차로 진행됨
- 기술통계학의 주요 지표는 평균, 분산, 표준편차, 사분위 수 등임

# 정리 하기

## 기술통계학

- ✓ 자료의 요약과 도시
  - 막대 그래프(Bar Chart)
    - : 범주형 데이터의 빈도를 나타낼 때 활용
  - 히스토그램(Histogram)
    - : 도수 분포를 막대 그래프로 나타낸 그래프
  - 산점도(Scatter Plot)
    - : 직교 좌표계를 이용해 두 개 변수 간의 관계를 표현
  - 상자그림(Box Plot)
    - : 요약 통계량(최소값, 최대값, 제1사분위수, 제3사분위수, 중앙값)을 상자를 가지고 나타낸 그래프



- 다음 시간에 살펴 볼 내용 -

## 08강 확률의 개념

수고하셨습니다.