오리엔테이션+서 론

담당교수 : 김 덕 기

toby123@cbnu.ac.kr

강의계획서-1

1. 교과목 정보

E-mail : toby123@cbnu.ac.kr , HP : 010-3456-2055

강의실 : S4-1-106(21동 106호 , 강의시간 : 수요일 5-7교시

2. 교과목 개요

강의개요	This course is for an introductory course in statistics or in probability and statistics for students in engineering and computer sciences. we will cover the fundamental methodology for applying probability theory to gain insight into real, everyday statistical problems and situations.
학습목표	1. 데이터사이언스에서 빅데이터분석, 기계학습, 인공지능의 다양한 모델 학습이 대표적인 영역이다. 이러한 영역에서 아주 중요한 모델이 불확실성을 모형에 반영한 확률적(통계학적) 모델이며, 데이터사이언스 영역에서 확률과 통계가 아주 중요한 역할을 담당한다. 2. 한 학기 학습을 통해 확률과 통계의 기초적 개념을 이해하고, 다양한 자료로부터 자료를 요약(수치적, 시각적)하여 정보를 해석하는 방법과, 불확실성에 대한 확률을 계량화하는 방법을 이해한다. 3. 확률변수와 확률분포를 이해하고 이산형 확률분포의 특징과 연속형 확률분포의 특징을 이해하고 확률을 계산하는 방법을 이해한다. 4. 결합확률분포와 조건부확률분포를 이해하고, 확률분포로 많이 사용되는 대표적인 이론적 확률분포를 이해한다. 5. 표본분포로 많이 사용되는 대표적인 이론적 분포를 이해한다. 6. 모수추정 방법과 가설검정의 원리를 이해하여 합리적 의사결정방법을 익힌다. 7. 두 집단 및 다집단의 비교분석 방법을 이해한다. 8. 상관 및 회귀모형 분석에 대한 방법을 이해한다.

강의계획서-2

문제해결방법	.					
수업진행방법	강의	토의/토론	실험/실습	현장학습	개별/팀별 발표	기타
	80%	0%	10%	0%	0%	10%
	상세정보	이론 강의 중심으로 진행하며 필요에 따라 이론을 좀 더 구체적으로 이해하기 위해 R 또는 Python실습을 몇 차례 진행할 수 있다.				
평가방법	중간고사	기말고사	출석	퀴즈	과제	기타
	35%	35%	10%	0%	20%	0%
	상세정보	midterm 35% final term 35% homework 20% attenance 10%				
프로그램 학습성과의 평가	.					
교재 및 참고문헌	1. 주교재 : R과 Python을 이용한 확률 및 통계, 박진호 외 7인, 자유아카데미, 2023 2. 부교재 : Introduction to Probability and Statistics for Engineers and Scientists, Sheldon M. Ross, Elsevier,,					
핵심역량과 연계성	주역량:E역량(전문성)					

강의계획서-3

3. 주별 강의계획

주차	수업내용	교재범위 및 과제물	비고
1	Probability & Statistics	Chapter 1	
2	Descriptive Statistics	Chapter 2	
3	Elements of Probability	Chapter 3 H.W. chap1-chap3(Exercise)	
4	Discrete random variable and distribution	Chapter 4	
5	Continuous random variable and distribution	Chapter 5	
6	Joint probability distribution & Conditional probability distribution	Chapter 6 H.W. chap4-chap6(Exercise)	
7	Distributions of Sampling Statistics	Chapter 7	
8	Mid exam	Chapter 1~Chapter 7(mid exam)	
9	Parameter estimation : point estimation, interval estimation, moment method, MLE	Chapter 8	
10	One sample inference : Estimation and Hypothesis testing	Chapter 9	
11	Two sample inference : Estimation and Hypothesis testing I	Chapter 10-1	
12	Two sample inference : Estimation and Hypothesis testing II	Chapter 10-2 H.W. chap8-chap10(Exercise)	
13	Correlation and Regression analysis I	Chapter 11-1	
14	Regression analysis II	Chapter 11-2	
15	Final exam	Chapter8~Chapter 11(final exam)	

강의계획서-4

확률 및 통계 평가방법: 출석(10%), 과제(20%), 중간시험(35%), 기말시험(35%)

주의 : 출석 기준 ~ 지각2번(결석1번), 결석1시간(-1점)

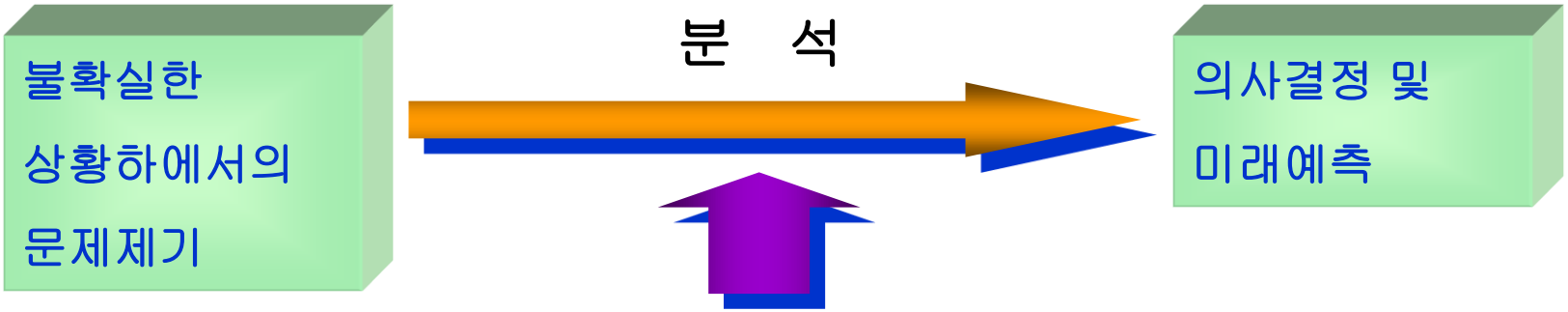
- 중간고사: 8주차 (10월 23일(수)), 13:40~15:00 (80분)
- 기말고사: 15주차 (12월 11일(수)), 13:40~15:00 (80분)
- 대면 강의와 대면 시험 원칙
- 공휴일 : 9월 3째 주(추석연휴)-수,목(2반 비대면 강의 진행)
- 공휴일 : 10월 9일(한글날)-수(비대면 강의 진행)

→ **비대면 동영상 강의 : LMS의 주차 강의 3개의 동영상을 모두 시청.**

통계학이란?

➔ Fisher : 응용수학의 한 분야로 관찰자료에 적용된 수학적 원리

➔ Khazanie : 불확실한 상황에서 관심의 대상이 되는 자료를 수집, 정리. 요약, 분석하여 그 자료에 대한 지식을 객관적이고 과학적으로 다루는 학문.



수리적 모형 :

$y = f(x) + e$

결과 원인 오차, 불확실성

⇓

풀어내는게 목표

통계학의 유형(분류)

기술통계학

1차분석

기술통계학이란 자료들의 특징을 알아보기 위하여 자료들을 수집하고 정리하여 도표 또는 표를 만들고, 분포의 형태를 알기 위하여 대푯값 또는 변동의 크기 등과 같이 수치적인 값으로 요약하는 방법을 연구하는 분야이다.

추측통계학

2차분석, 통계적 추론 (추정+검정)

추측통계학이란 모집단으로부터 추출한 일부 자료들을 이용하여 통계적 모형을 설정하고, 연구하려는 문제의 미지의 특성에 대한 결론을 유추하고 예측하는 방법을 연구하는 분야이다.

기본적인 통계 용어 정의

- 모집단(population)

  : 어떤 정보를 얻기 위하여 연구대상으로 관심을 두고 있는 집단전체

- 표본(sample)

  : 모집단 특성에 관한 정보를 얻기 위하여 모집단으로부터 추출한 또는 측정한 값들의 집합

[예] 대전에 소재한 병원을 찾는 환자 중 남녀 비율의 차이를 알고자 하는 경우

모집단 : 대전에 위치한 병원을 찾는 모든 환자, 표본 : 조사를 위해 선택된 환자

- 모수(Parameter)

  : 모집단의 특성을 나타내는 미지의 값 – ex) 모평균, 모분산, 모비율

- 통계량(Statistic)

  : 표본의 특성을 수치로 나타낸 값(특성 값) – ex) 표본평균, 표본분산, 표본비율

자료분석 단계

- 1단계: **자료수집**
  - 우리가 알고자 하는 대상(모집단)에 대한 정보를 가지고 있는 자료를 수집(관찰). 자료가 모집단의 특성을 잘 표출할 수 있도록 자료의 수집과정이 설계되어야 한다.
- 2단계: **자료의 요약 및 정리**
  - 수집된 자료가 모집단에 정보를 잘 파악할 수 있도록 그림이나 도표, 혹은 수치적인 값으로 요약 및 정리한다.
  - 예) 평균, 표준편차, 도수분포표
- 3단계: **추론(추정과 가설검정)**
  - 수집된 자료를 이용하여 통계적 기법을 통한 모집단에 대한 추측(추론)

통계분석 과정

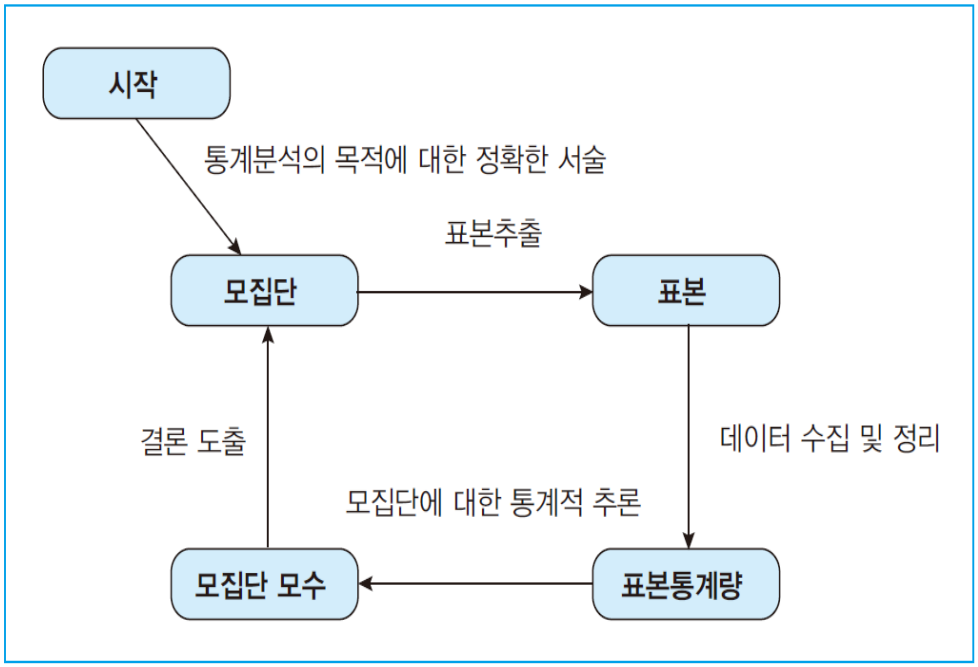


그림 1.1 통계분석 과정

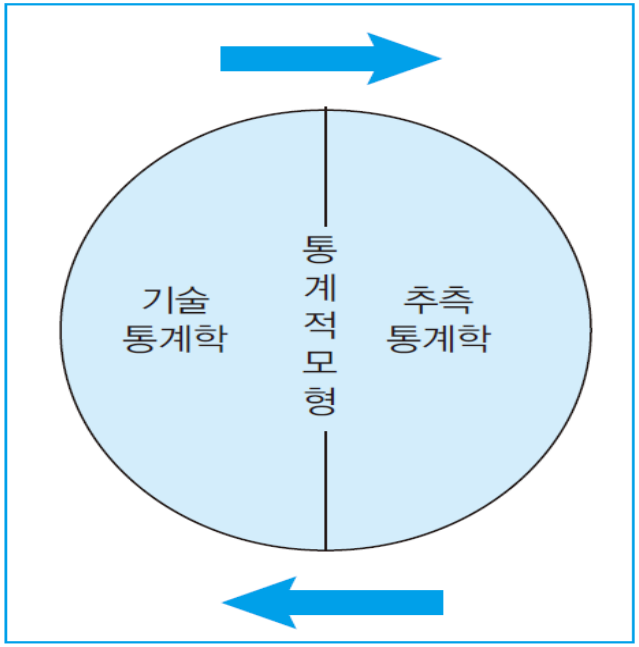
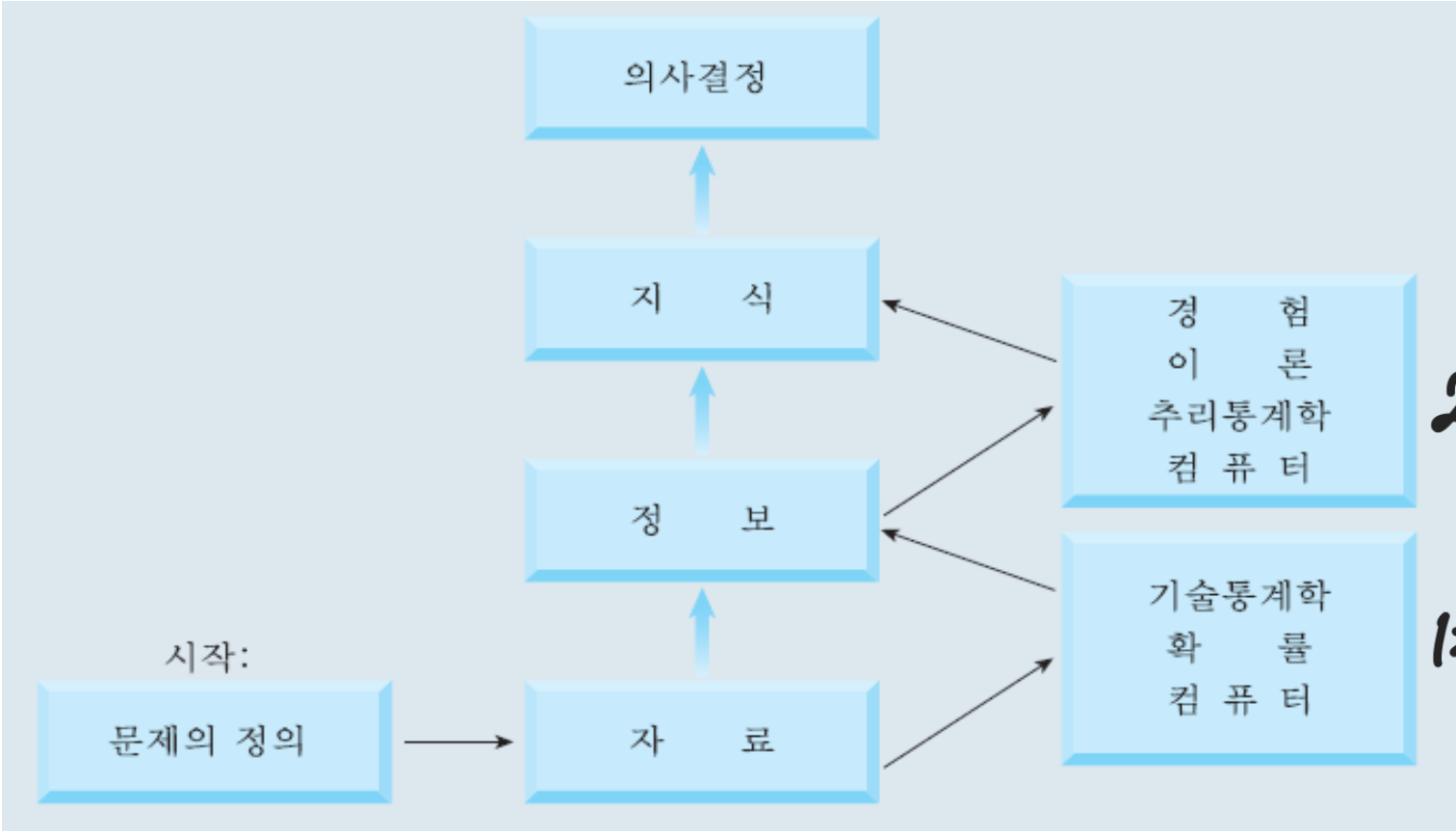


그림 1.2 기술통계학과 추측통계학의 관계

➔ **통계학은 자료의 일부만을 이용하여 모집단 전체에 대한 추론(추측)을 하므로 항상 오류의 가능성이 있다. 기본적으로 통계학은 이러한 오류의 가능성을 최소화하려고 한다.**

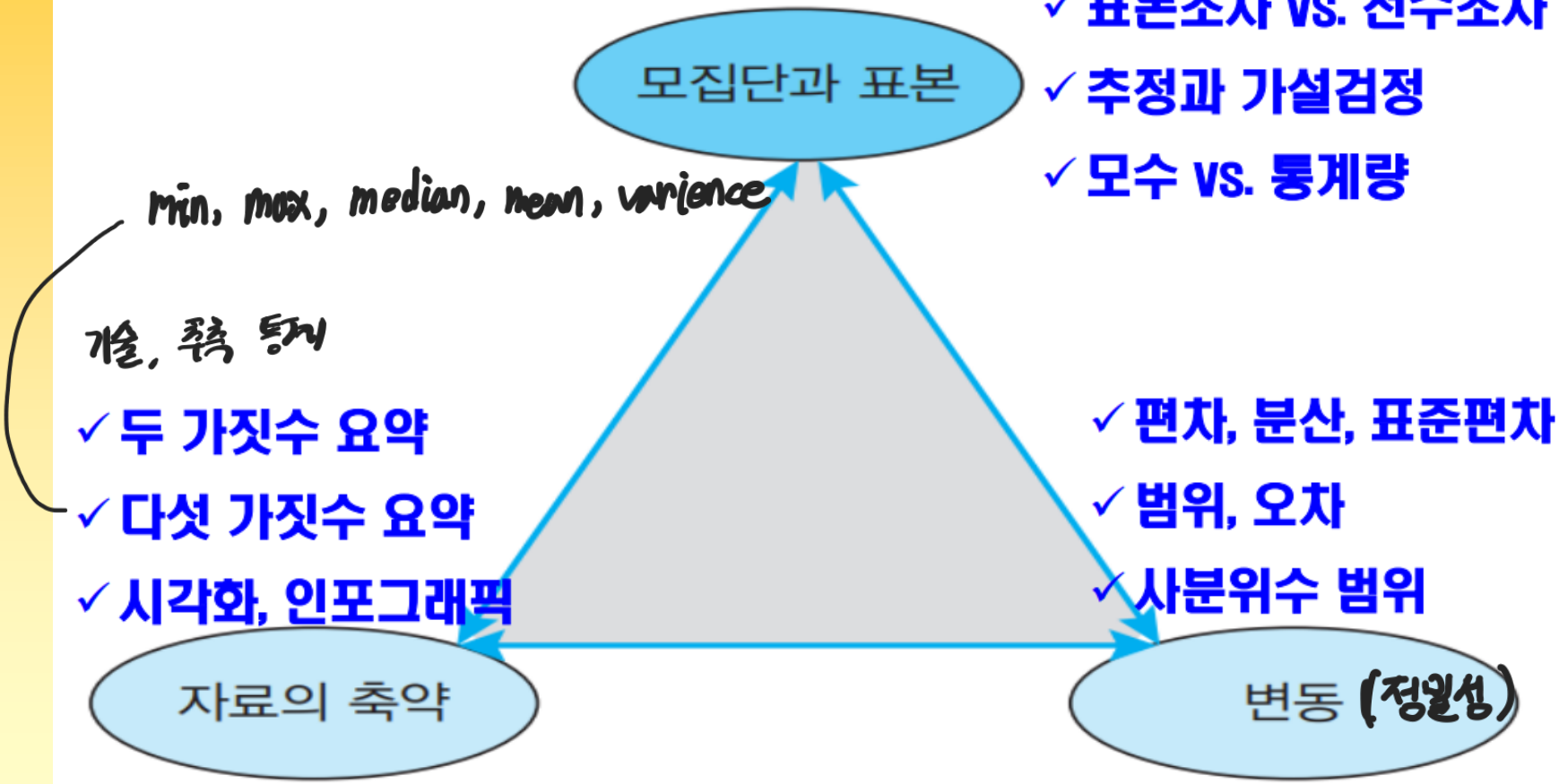
불확실한 상황에서 의사결정과정

- 의사결정과정



통계학의 세 가지 주제

Survey Census



자료의 축약 – 히스토그램

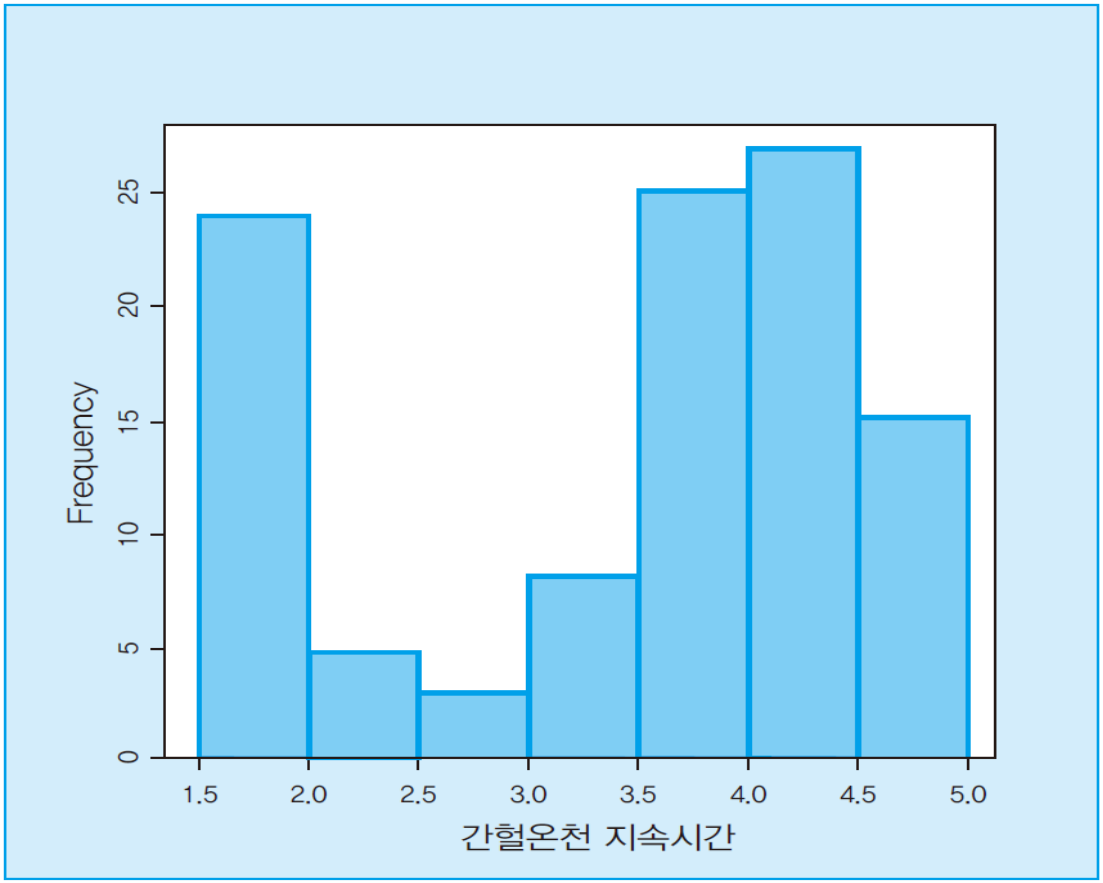


그림 1.4 간헐온천 지속시간에 대한 히스토그램

107개 자료를 도수분포표를 작성하여 7개 기둥의 히스토그램으로 축약

평균=3.43, 분산=1.51

최빈값=4.25

변동(Variation) -1

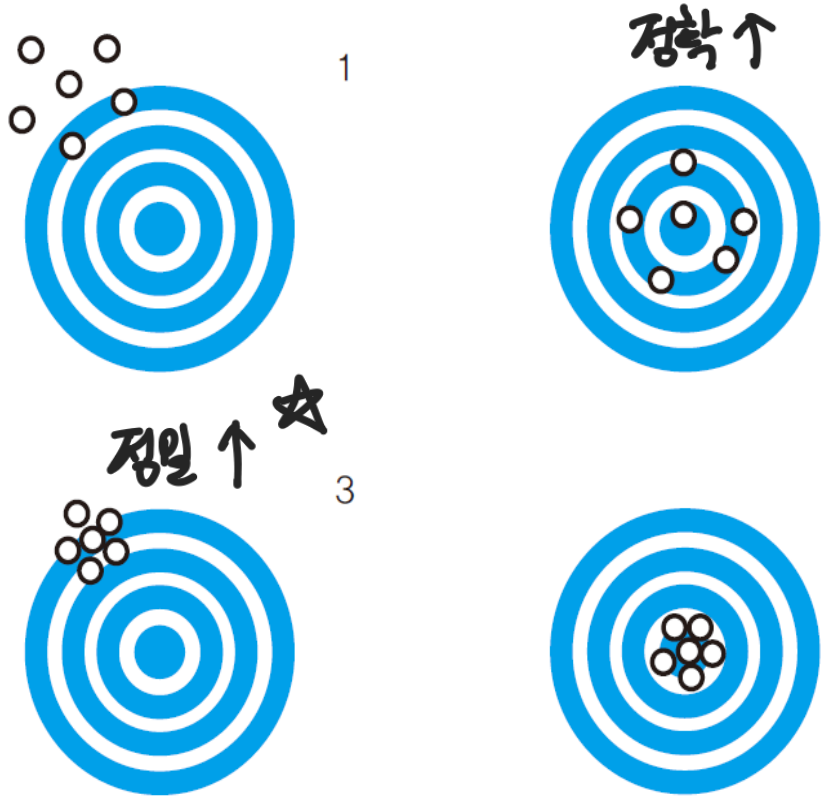


그림 1.5 정확성과 정밀성

6-시그마 운동 ?

품질공학-정확성, 통계학-unbiased

품질공학-정밀성, 통계학-유효성

-적은 변동

2→4 ,3→4 로 개선하기 수월한 것은?

➔ 변동의 중요성.

4>1 순으로 좋다

변동(Variation) -2

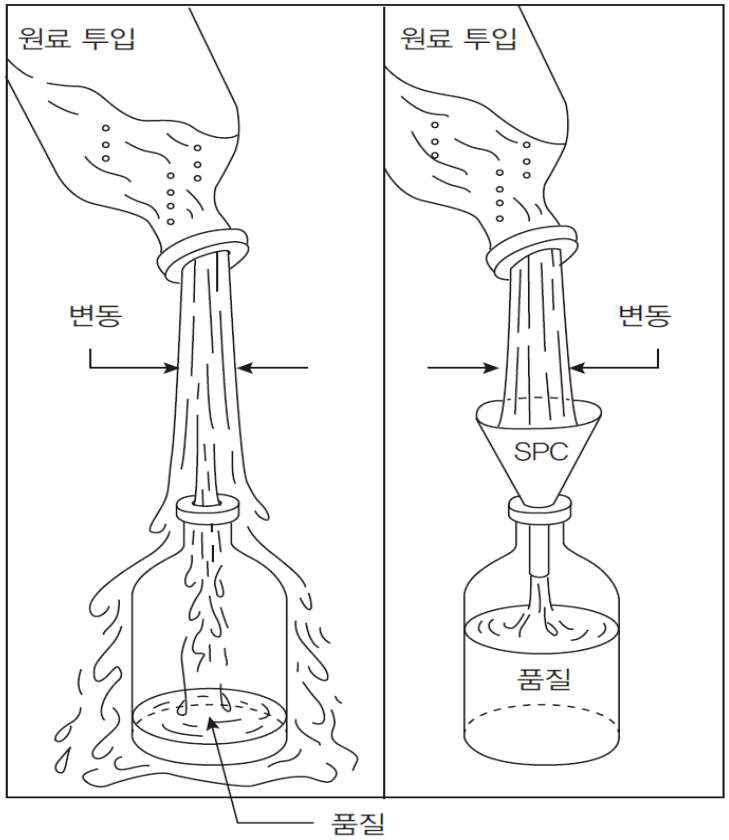


그림 1.6 변동의 의미

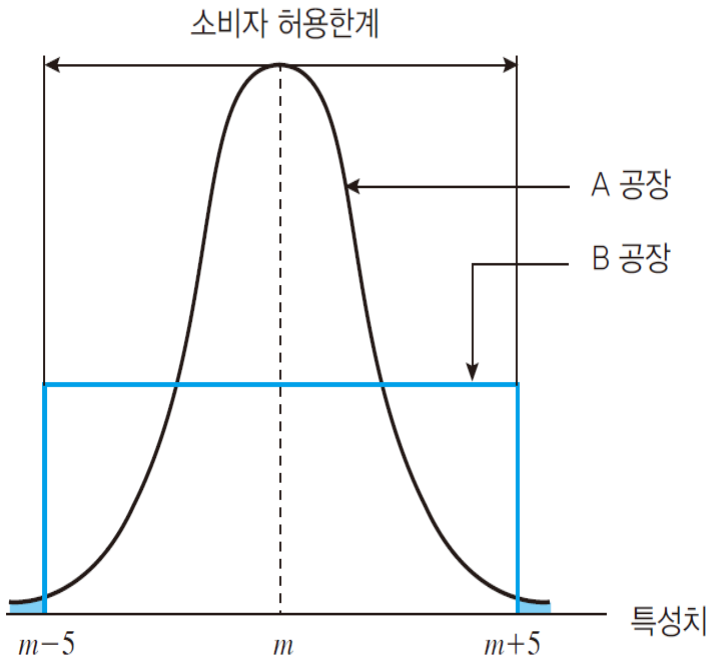


그림 1.7 텔레비전 색상밀도 분포

정형 → 형태가 정해짐

질적

양적 (평균을 낼 수 있음)

자료의 유형 – 통계분석방법

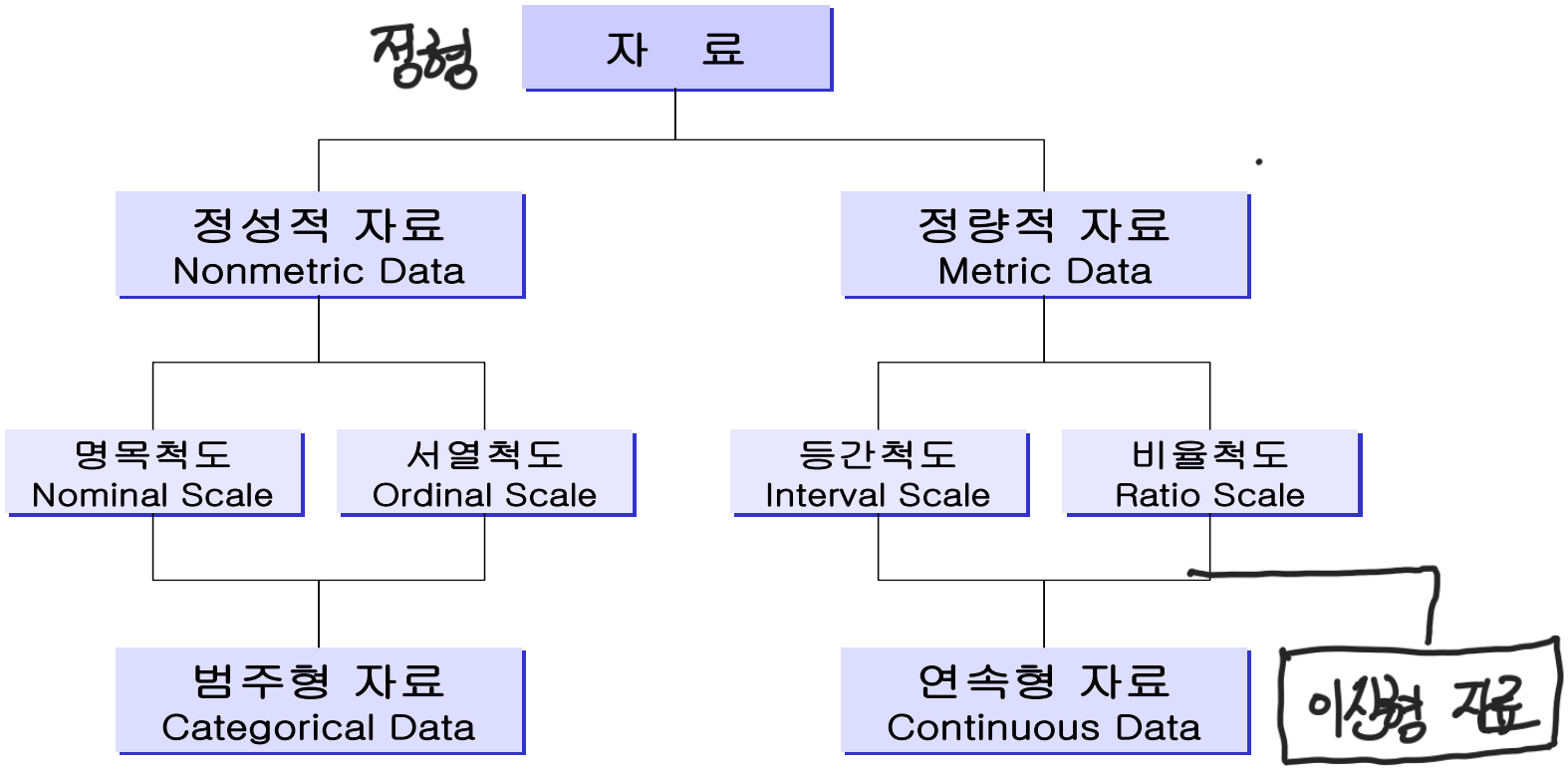
구 분	정의 및 예	통계분석방법
명목 척도 (Nominal)	명목척도는 관심대상의 특성을 범주로 분류하여 각 범주에 숫자를 부여한 척도. (예) 성별 : 남=1, 여=2 직업 : 회사원=1, 공무원=2, 자영업=3, 학생=4, 기타=5 지역 : 서울=1, 경기=2, 강원=3, 충청=4, 전라=5, 경상=6	빈도분석 교차분석 범주형 자료분석
서열 척도 (Ordinal)	관심대상의 특성을 크기 순으로 나열하고 이에 숫자를 부여한 척도. (예) 올림픽순위 : 금=1, 은=2, 동=3 군대계급 : 이등병=1, 일등병=2, 상병=3, 병장=4 게임횟수 : 1~2회/주 =1, 3~4회/주 =2, 5회 이상/주 =3	빈도분석 교차분석 범주형 자료분석 다변량 분석
등간 척도 (Interval)	관심대상의 특성을 나타내는 측정치 사이의 거리를 일정한 간격으로 표시하는 척도. 대칭관계 (예) 만족도 : 매우불만=1, 불만=2, 보통=3, 만족=4, 매우만족=5 실험온도 : 0도, 50도, 100도, 150도, 200도	기술통계분석 집단 평균분석 회귀분석 다변량 분석
비율 척도 (Ratio)	절대적 원점이 존재하며 비율 계산이 가능한 수치를 부여한 척도. nothing 몇배나 해석 가능 (예) 판매량, 매출액, 무게, 소득 등	기술통계분석 집단 평균분석 회귀분석 다변량 분석

정보량이 많다.

질적 ↔ 양적

없다가 아님.

자료의 분류



주1) 범주형 자료(= 정성적 자료, 비계량적 자료, 질적 자료)

주2) 연속형 자료(= 정량적 자료, 계량적 자료, 양적 자료)

질적, 양적이 헷갈리면 평균을 붙여봐라
양적에 객관성을 붙이면 질적 자료가 될 수 있다.

등간 C 서열 ☆

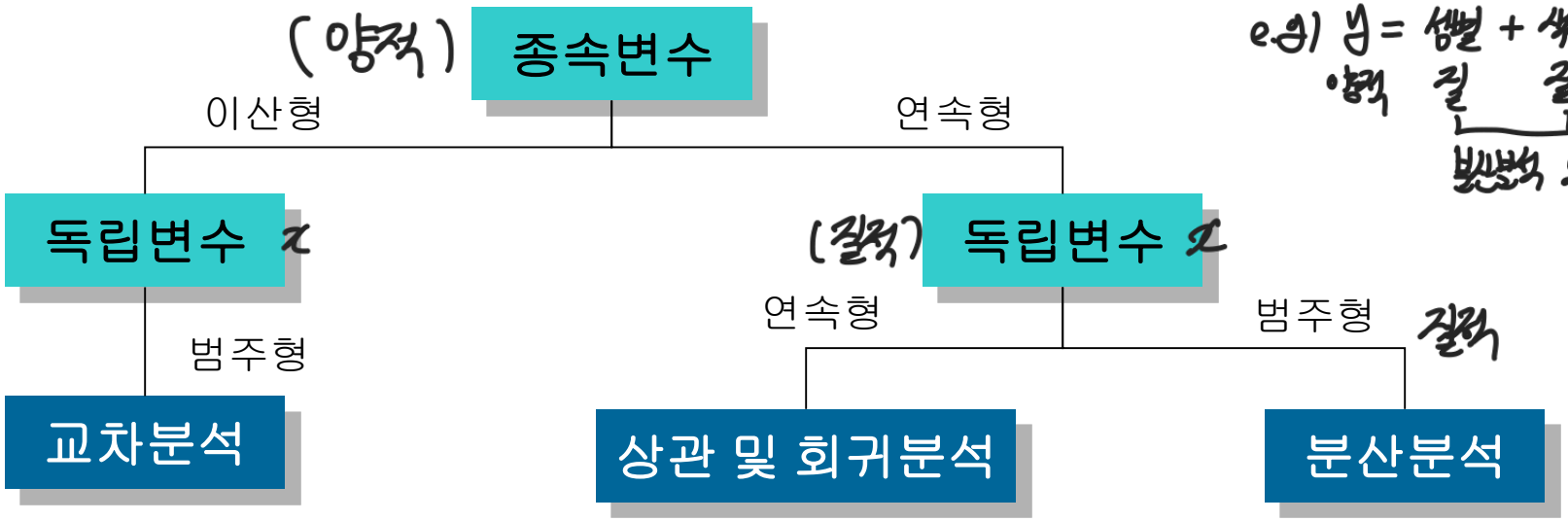
자료척도-분석방법1

(설문1) 성별을 답해 주십시오 (　　　) 명목
(설문2) 혈액형을 답해 주십시오 (　　　)형 〃
(설문3) 가장 좋아하는 색을 답해 주십시오 (　　　)색 〃
(설문4) 귀하의 연령을 답해 주십시오 (　　　)세 비율
(설문5) 이 상품을 처음으로 알게 된 계기는 다음 중 어느 것입니까? 명목
1.TV의 광고 2.라디오의 광고 3.신문의 광고
4.저널광고 5.아는 사람의 소개 6.기타 (　　　)
(설문6) 이 상품의 만족도는? 리커트 척도
1.대단히 불만 2.불만 3.약간 불만 4.약간 만족 5.만족 등간
6.대단히 만족

자료의 척도	범주형(명목, 서열)	연속형(등간, 비율)
범주형(명목, 서열)	범주형자료(빈도, 교차)분석, 카이제곱분석 등	독립T-검정, 대응T-검정 분산분석 F-검정 등 – 3집단 이상
연속형(등간, 비율)	독립T-검정, 대응T-검정 분산분석 F-검정 등	상관분석, 회귀분석, 요인분석 등

1. 설문1~설문6의 자료의 척도를 쓰시오. (　　　)
2. 성별에 따른 상품만족도에 차이가 있는가? 위 분석카테고리 중 (독립T-검정)분석
3. 성별에 따라 좋아하는 색에 차이가 있는가? 위 분석카테고리 중 (교차)분석

자료척도 – 분석방법2



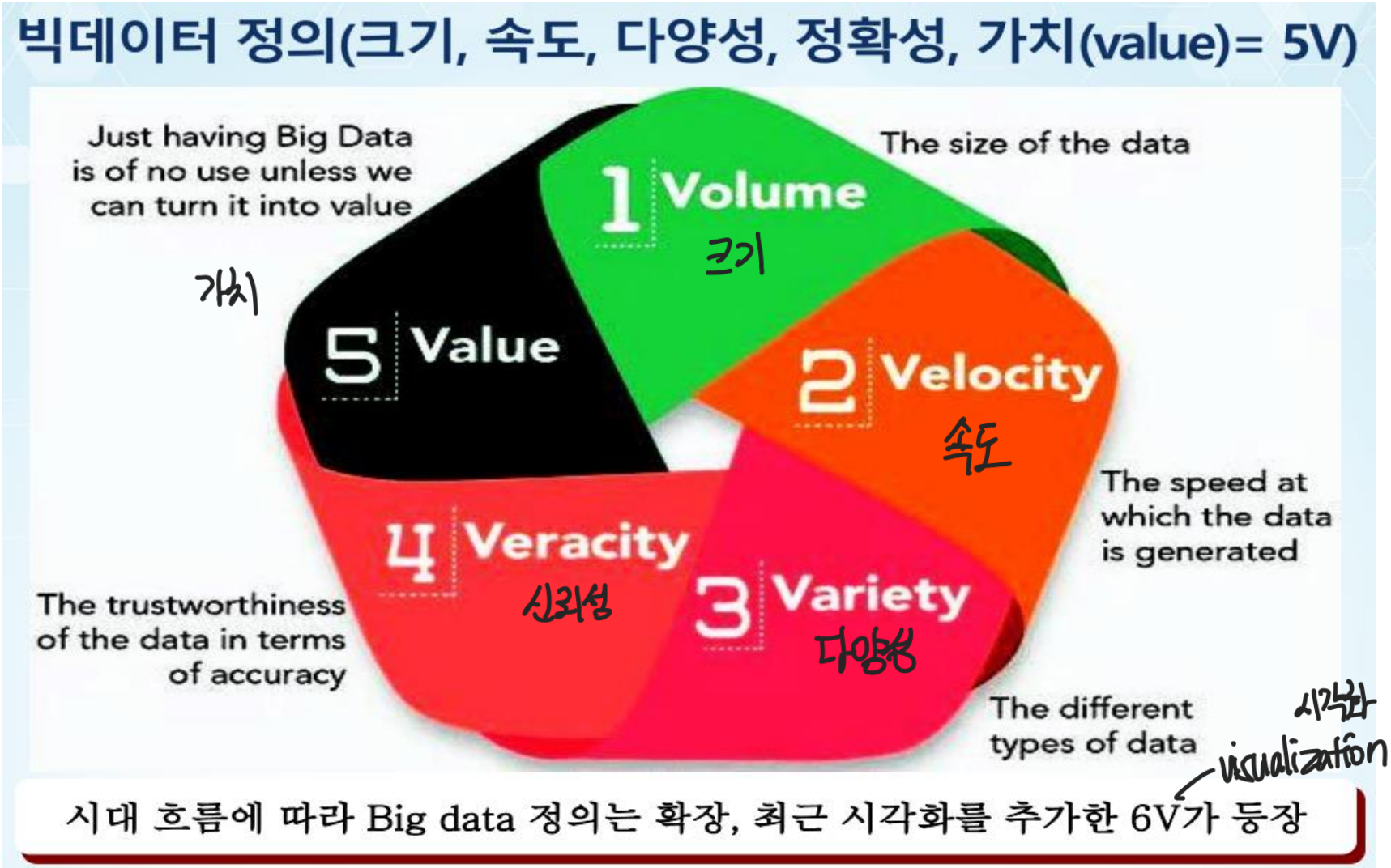
독립변수(변인) : 설명변수, 외생변수(변인)–어떤 변수에 영향을 주는 변수 또는 원인변수.

종속변수(변인) : 반응변수, 내생변수(변인)–어떤 변수에 영향을 받는 변수 또는 결과변수.

1. 연령에 따른 만족도의 관계분석 ? (상관, 회귀)
2. 성별에 따른 좋아하는 색의 관계분석 ? (교차분석)
3. 성별에 따른 만족도의 관계분석 ? (분산분석)

연령 / 성별	W	R	G
m	f_{11}	f_{12}	···
w	f_{21}	···	·

빅데이터의 정의


빅데이터 정의(크기, 속도, 다양성, 정확성, 가치(value)= 5V)
Just having Big Data is of no use unless we can turn it into value
가치
5 Value
The size of the data
1 Volume
크기
2 Velocity
속도
The speed at which the data is generated
4 Veracity
신뢰성
The trustworthiness of the data in terms of accuracy
3 Variety
다양성
The different types of data
시각화
visualization
시대 흐름에 따라 Big data 정의는 확장, 최근 시각화를 추가한 6V가 등장

반정형 = 비정형 + 정형

3종류의 데이터 : 정형, 비정형, 반정형자료

빅데이터의 성격과 속성

- 데이터는 생산 방식에 따라 구조적데이터(정형데이터, **structured data**)와 비구조적 데이터(비정형데이터, **unstructured data**), 반구조적데이터(반정형데이터, **semi-structured data**)로 나눔
- 정형데이터: 기온, 압력, 이혼율, 확진자수, 사고건수, 키, 몸무게, 불량건수 선호도, 우선순위, 혈액형, 성별, 지역, 학위, 결혼상태(미혼, 유배우), 소득 등
- 반정형데이터: 바코드, **QR**코드, **HTML(www. ..)**, 웹로그, 센서 데이터 등
- 비정형데이터: 영상, 음성, 문서, 자연어, **SNS** 댓글, 사진, 동영상 등

홍길동씨가 아침출근에 접하는 다양한 데이터:『홍길동씨는 아침에 스마트폰 *알람 소리*에 기상하여 스마트폰에 적힌 *시간, 날짜, 문자, 카톡* 등을 확인한 뒤, 세수 및 간단한 아침 식사를 하고 출근한다. *시계의 시간*을 보면서 *지하철 도착시각*에 맞춰 지하철에 탑승하기 위해 *A역 3번 출구*로 들어가 개찰구에 교통카드를 터치하여 승인 *신호*와 *소리*를 듣고 개찰구를 통과한다. 정기적으로 운행하는 지하철에 탑승한 뒤, 스마트폰으로 개인 *메일*을 확인하기 위해 *아이디*와 *비밀번호*를 입력하고 로그인하여 메일을 확인한다. 인터넷 사이트(인터넷 사이트를 구성하는 *HTML 언어*)들을 방문하면서 최신 뉴스를 보고, 무의식중에 지하철이 어디쯤을 달리고 있는지 *지하철 노선도*를 확인 한다. 지하철 노선도에는 각역의 한글·영어 *명칭*과 그 아래에 *숫자(넘버)*가 표기되어 있다. 목적지에 도착하여 출근 시간 몇 분 전, 모닝커피를 좋아하는 홍길동 씨는 회사 근처의 카페에서 카페라테 *tall 사이즈*를 포장 구매take-out한다. 이때 커피 가격으로 *4,600원*을 낸다. 회사

데이터 마이닝과 텍스트 마이닝

정형 데이터에서 진주, 비정형 데이터에서 진주 찾기

데이터(정형) 마이닝 ⊂ 텍스트(비정형) 마이닝

비정형자료를 정형화해서 분석하거나 비정형자료 자체를 분석