1. 기초 통계학-서 론

수학정보통계학부 김 덕 기

toby123@cbnu.ac.kr

강의계획서-1

1. 교과목 정보

개설연도-학기	2024년	겨울학기	개설학과	대학(정보통계학과)
교과목번호-분반번호	0941031	01	교과목명	기초통계학
이수구분	교양선택		학점/시수	3-3-0
강의시간/강의실	월 01 ,02 ,03 [S1-2-108(41-108)] 화 01 ,02 ,03 [S1-2-108(41-108)] 수 01 ,02 ,03 [S1-2-108(41-108)] 목 01 ,02 ,03 [S1-2-108(41-108)] 금 01 ,02 ,03 [S1-2-108(41-108)]			
수업방식	대면			
강의언어			담당교수	김덕기(강사)
전화	010-3456-2055		E-mail	toby123@cbnu.ac.kr
강의정원	45		학과전화	043-273-5928
선수과목			수강대상	학부(전학년)
강의 맛보기				

2. 교과목 개요

강의개요	자료의 정리 및 요약과 자료에 바탕을 둔 의사결정능력을 키우기 위해 기초적이면서도 필수적으로 알아야 하는 통계적 이론과 방법론을 이해하고, R프로그램 실습을 통해서 예제중심의 문제해결능력(통계적 Literacy)을 배양한다.
학습목표	다양한 유형의 자료에 적절한 자료의 정리 분석 해석방법을 익히고 확률이론과 확률분포의 특징을 익히고 자료로부터 의미 있는 정보를 해석해 내고 실무적으로 통계패키지를 활용하여 다양한 자료의 형태에 맞는 적절한 통계분석 방법을 이용하여 의미 있는 결과해석 방법을 실무적으로 익힌다.

강의계획서-2

문제해결방법	실전에서 많이 사용하는 통계적 분석 방법들을 R-program을 이용하여 실제 수집한 자료에 대해 문제를 제기하고 자료를 정리, 요약하여 1차 정보를 해석, 추가 분석 및 결과 해석을 통해 2차 정보를 해석하여 현실적인 문제해결 능력을 다양한 자료로부터 학습.					
수업진행방법	강의	토의/토론	실험/실습	현장학습	개별/팀별 발표	기타
	60%	10%	30%	0%	0%	0%
	상세정보	주 별 이론 강의 중심으로 진행하고, 필요에 따라 R-program 실습을 통해 다양한 자료 분석 및 결과 해석 방법을 익히고, 개인별 실제 자료에 대한 의미 있는 통계 기법 적용 및 분석 결과를 토대로 종합 보고서 작성 방법을 익히고 전반적인 통계적 literacy 능력을 키우는데 초점을 맞춰 진행합니다.				
평가방법	중간고사	기말고사	출석	퀴즈	과제	기타
	35%	35%	10%	0%	20%	0%
	상세정보	중간시험-35%, 기말시험-35%, 출석-10%, 과제-20%				
프로그램 학습성과의 평가	1. 이론적 이해도 평가 2. R분석 및 결과 해석 능력평가 3. 과제 및 종합 보고서를 통한 실전 데이터에서 정보를 해석하는 능력평가 4. 문제 해결을 위해 올바른 통계적 방법론에 대한 이해와 활용 능력평가					
교재 및 참고문헌	1. 참고문헌 : 기초통계학 R을 이용한 통계분석3판, 노맹석 외 5인공저, 자유아카데미, 2020 2. 주교재 : R을 이용한 최신통계학, 김정연, 김주성, 나종화, 이성덕, 정대한, 조중재, 허태영, 자유아카데미, 2020					
핵심역량과 연계성	주역량:E역량(전문성) C역량:15% H역량:5% A역량:10% N역량:5% G역량:5% E역량:60%					

강의계획서-3

3. 주별 강의계획

주차	수업내용	교재범위 및 과제물	비고
1	통계학의 기초 및 서론(과목오리엔테이션+이론)	1장	대면강의
2	자료의 정리와 시각화(이론)	2장	대면강의
3	자료의 요약(수치적요약방법) (이론)	3장 (1~3장 과제)	대면강의
4	자료의 정리요약(R실습)	1~3장(R실습)	대면강의
5	확률, 확률분포(이론)	4-5장	대면강의
6	여러가지 확률분포(이론)	6장 (4~6장 과제)	대면강의
7	확률분포관련(R실습)+기출문제풀이	기출문제풀이+R실습	대면강의
8	중간시험	1장~6장(중간시험)	대면시험
9	표본추출방법 및 표본분포(이론)	7장	대면강의
10	통계적추론-추정 (이론)	8장 (7~8장 과제)	대면강의
11	통계적추론-검정(이론)	9장	대면강의
12	통계적추론(R실습)	7~9장(R실습)	대면강의
13	2집단에 관한 통계적추론 (이론)	10장(이론)+과제(9~10장)	대면강의
14	2집단에 관한 통계적추론(실습)+종합정리	10장(R분석실습)+기말기출문제	대면강의
15	기말시험	7장~10장(기말시험)	대면시험

강의계획서-4

기초통계학 평가방법: 출석(10%), 과제(20%), 중간시험(35%), 기말시험(35%)

주의 : 출석 기준 ~ 지각2번(결석1번), 결석1시간(-1점)

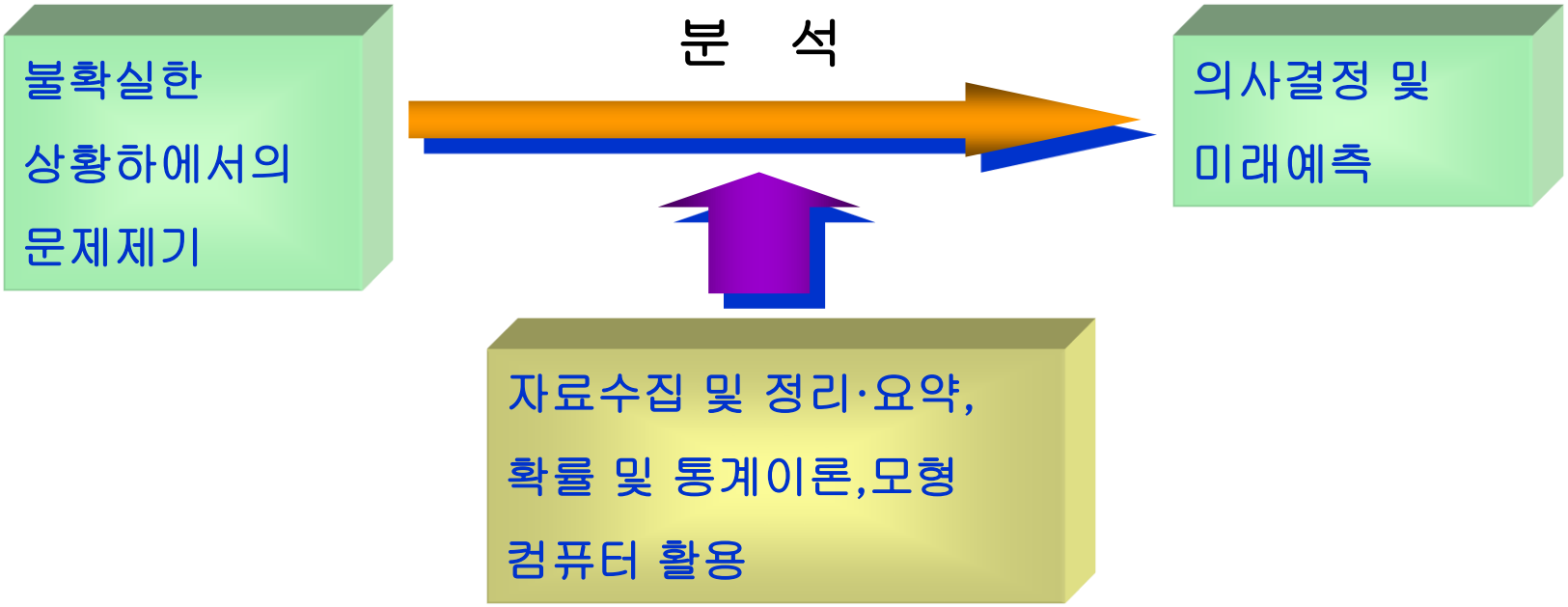
- 중간고사: 8주차 수업(1월 7일(화)), 10:00~11:10 (70분)
- 기말고사: 15주차 수업(1월 16일(목)), 10:00~11:10 (70분)
- 대면 강의와 대면 시험 원칙

→ **강의 노트 및 실습노트 : LMS의 주차 별 업로드 ~ 다운받아 출력해 오세요**

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.



통계학의 유형(분류)

기술통계학

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

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

기본적인 통계용어 정의 1

- 자료(DATA)

  : 사건의 발생 및 사실에 대한 관찰의 결과를 표현한 것으로 숫자, 문자 기호 등으로 나타난 것을 말함.

  [예] 자료 : 청주의 A-병원을 찾는 환자의 명단

- 정보(Information)

  : 사용자의 목적에 따라서 유용한 형태로 변형된 자료를 의미함.

  [예] A-병원을 찾는 환자 중에서 60대 이상의 여자환자의 명단

- 지식(Knowledge)

  : 지식은 정보 중에서 이용자의 목적에 맞을 뿐 아니라 유용하고 부가가치를 창출할 수 있는 것일 때의 지식.

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

통계분석 과정

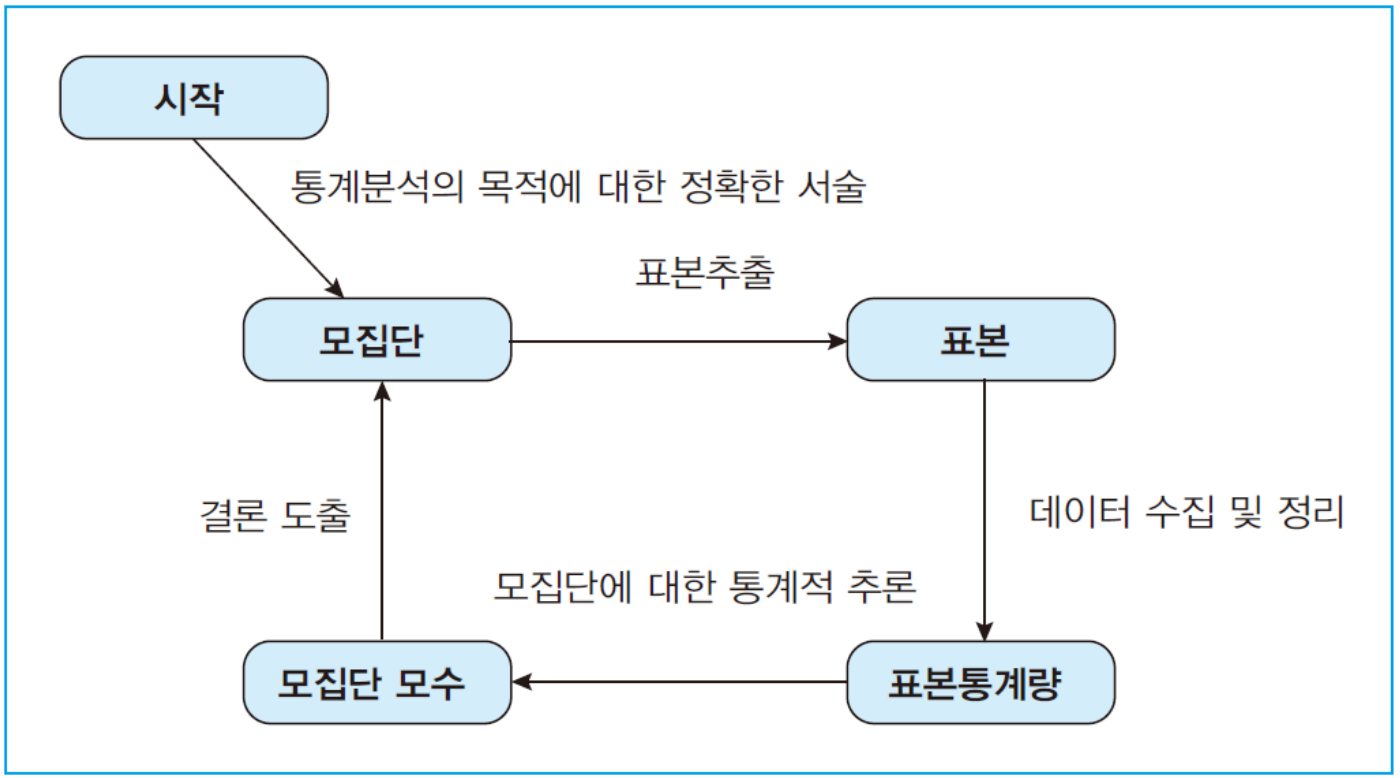
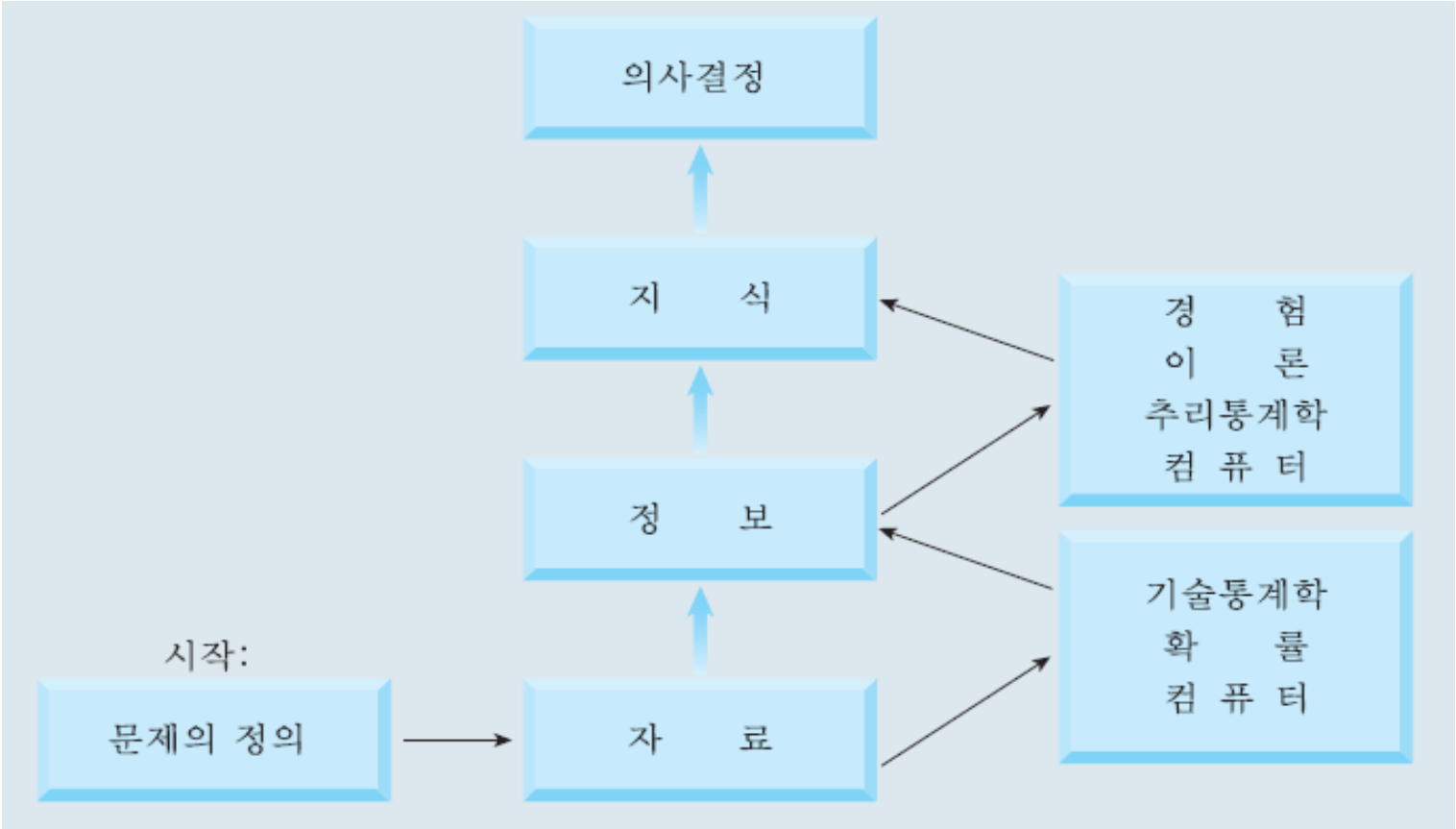


그림 1.1 통계분석 과정

불확실한 상황에서 의사결정과정

- 의사결정과정



통계학 – 변동(Variation)의 중요성 -1

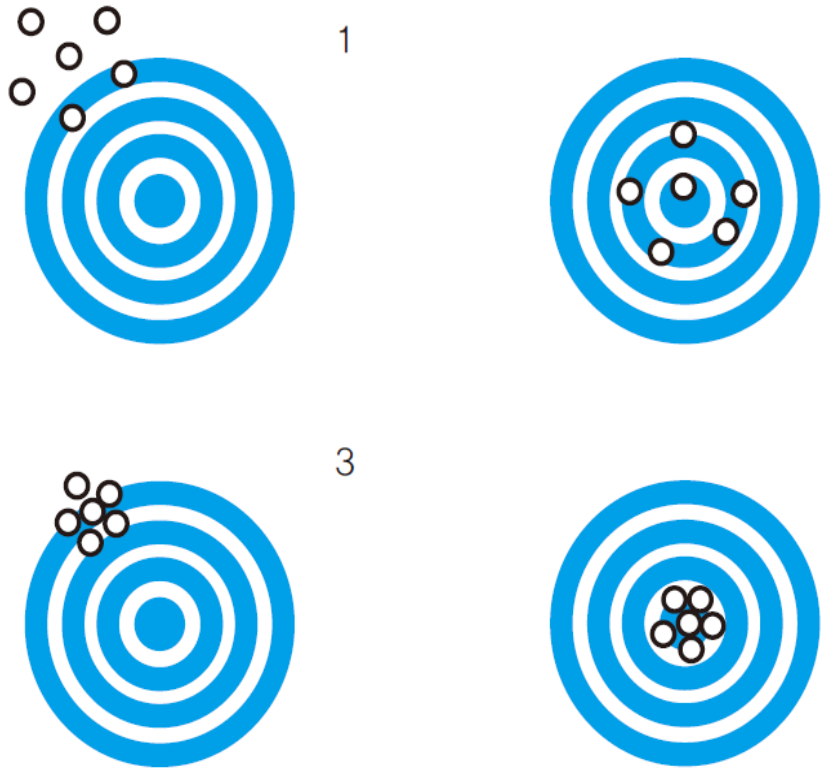


그림 1.5 정확성과 정밀성

6-시그마 운동 ?

품질공학-정확성, 통계학-unbiased

품질공학-정밀성, 통계학-유효성

-적은 변동

2→4 ,3→4 로 개선하기 수월한 것은?

➔ 변동의 중요성.

통계학 – 변동(Variation)의 중요성 -2

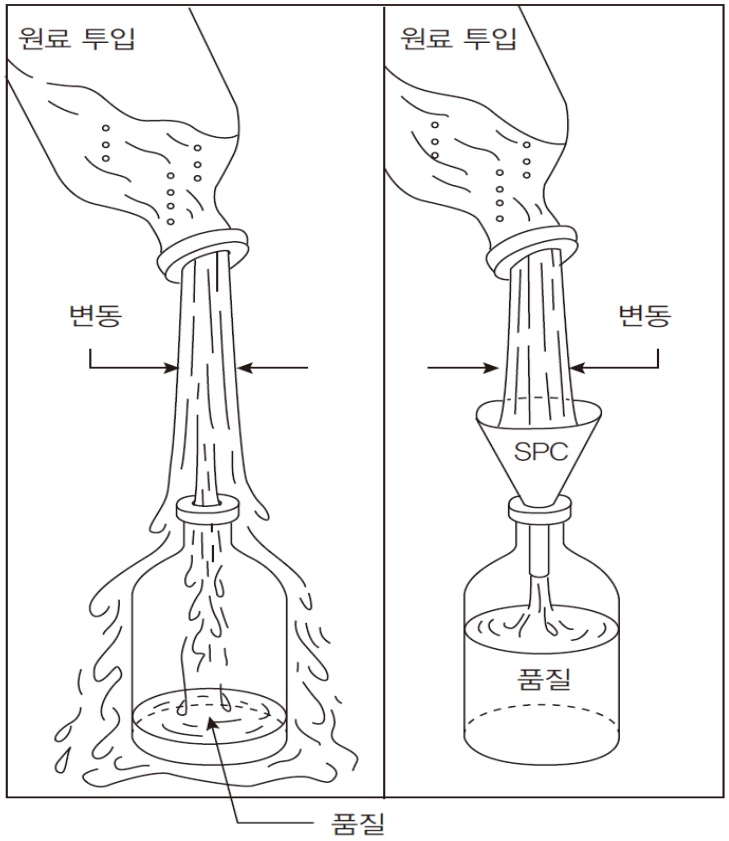


그림 1.6 변동의 의미

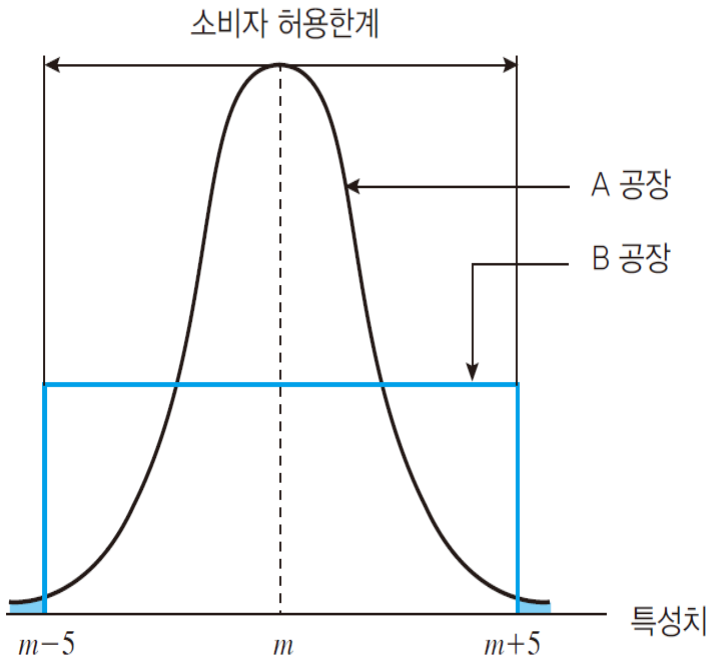
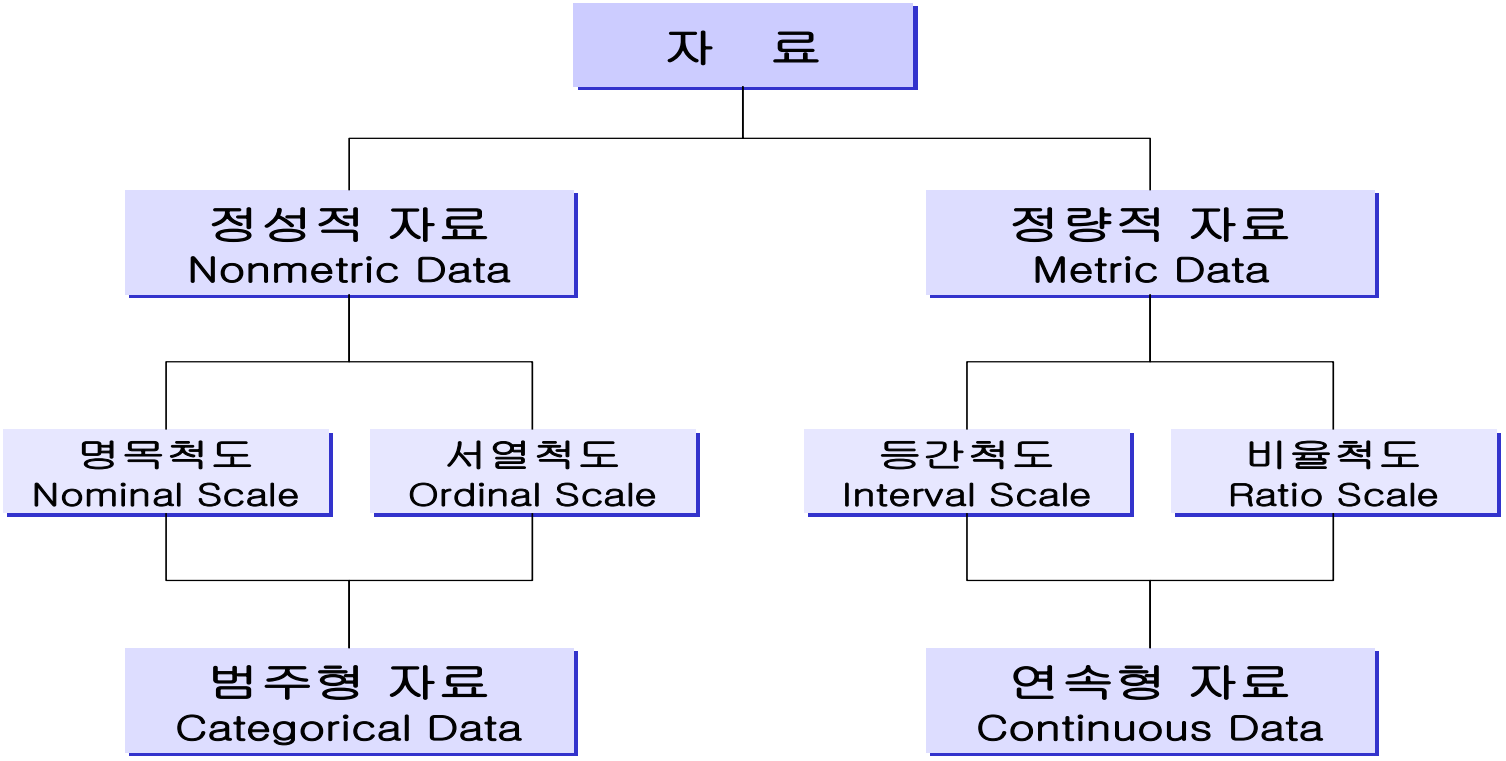


그림 1.7 텔레비전 색상밀도 분포

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. (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 비율계산이 가능한 수치를 부여한 척도. (예) 판매량, 매출액, 무게, 소득 등	기술통계분석 집단 평균분석 회귀분석 다변량 분석

자료의 분류



주1) 범주형 자료(= 정성적 자료, 비계량적 자료, 질적 자료)

주2) 연속형 자료(= 정량적 자료, 계량적 자료, 양적 자료)

자료척도-분석방법1

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

자료의 척도	범주형(명목, 서열)	양적 자료(등간, 비율)
범주형(명목, 서열)	범주형자료(빈도, 교차)분석, 카이 제곱분석 등	독립T-검정, 대응T-검정 분산분석 F-검정 등
양적 자료(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 ( )분석
3. 성별에 따라 좋아하는 색에 차이가 있는가 ? 위 분석카테고리 중 ( )분석

2. 자료의 정리와 시각화

수학정보통계학부 김 덕 기

toby123@cbnu.ac.kr

자료의 정리 및 시각화-범주형 자료

(1) 범주형 자료에 적합한 그래프

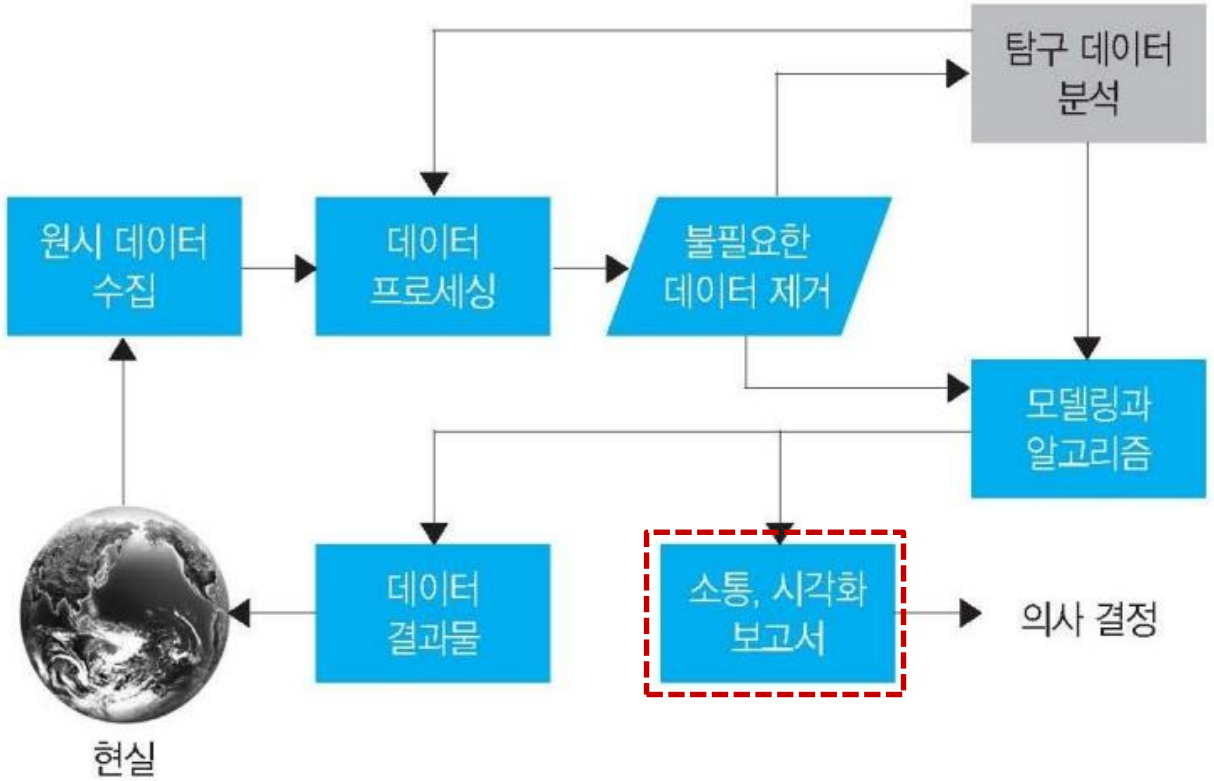
<table>
<tr><td rowspan="3">막대도표</td><td colspan="2">단순 막대도표 : 하나의 범주형 자료의 경우</td></tr>
<tr><td>수평누적 막대도표</td><td rowspan="2">둘 이상의 범주형 변수에 의해 2차원 이상의 분할표로 표현되는 자료의 경우</td></tr>
<tr><td>수직누적 막대도표</td></tr>
<tr><td>선도표</td><td colspan="2">단순 선도표 : 단순 막대도표를 선도표로 표현한 것
다중 선도표 : 누적 막대도표를 선도표로 표현한 그림
하락-선 선도표</td></tr>
<tr><td>면적도표</td><td colspan="2">단순 면적도표
수직누적 면적도표</td></tr>
<tr><td>원도표</td><td colspan="2"></td></tr>
<tr><td>히스토그램</td><td colspan="2"></td></tr>
</table>

자료의 정리 및 시각화-연속형 자료

(2) 연속형 자료에 적합한 그래프

히스토그램	범주형 자료의 형태로 변환 후 히스토그램을 그림
산점도	단순산점도 : 두 변수의 관계를 그림으로 표현 3-차원산점도 : 세 변수 사이의 관계를 나타내고자 할 때 산점도행렬 : 세 개 이상의 변수들 사이의 관계를 알아보고자 할 때 겹쳐그리기
P-P도표	귀무가설 하의 확률분포의 누적비율과 자료들의 누적비율의 산점도
Q-Q도표	귀무가설 하의 확률분포의 분위수와 자료들의 경험적분포의 분위수의 산점도
상자도표	한 범주형 변수들의 각 수준별로 상자그림을 작성
오차막대도표	한 범주형 변수들의 각 수준별로 신뢰구간, 표준편차, 표준오차 등이 수직선의 형태로 평균과 함께 출력
순차도표	시간을 X-축으로 하고 관측값을 Y-축으로 하는 산점도
시계열도표	시계열 자료의 패턴을 알아보기 위한 도표

Data Science Process –시각화



데이터 사이언스 프로세스 흐름도

자료의 도표화–연속형 자료–히스토그램

- 자료의 정리 및 시각화 :

  자료의 특성을 파악함에 있어, 자료를 도표나 그래프에 의해 표현하면 보다 많은 정보를 보다 빨리 시각적으로 파악할 수 있다.

> 자료 : (P.12) 어떤 여중생 35명에 대한 신장(단위:cm)

147	153	155	152	146	160	155
155	155	146	155	163	151	159
158	148	155	146	152	155	151
140	150	160	152	153	160	155
156	154	146	153	155	156	156

자료의 정리 및 시각화(계속)

도수분포표

- 전체 자료를 동일한 간격을 가지는 서로 중복되지 않는 몇 개의 계급구간으로 나누어 각 구간에 속하는 도수를 세어 나타낸 표

일반적인 작성절차

1. 자료 중 최대값과 최소값을 찾아 범위(=최대값-최소값)을 구함.
2. 자료의 크기에 따라 5~15정도의 계급의 수를 정하고, **계급의 폭 (=범위÷계급의 수)**을 정함.
3. 첫 계급구간의 **시작값(=최소값-자료값의 최소단위×½)**을 정하고, 2에서 구한 계급의 폭에 따라 나머지 계급을 설정.
4. 각 계급의 도수(상대도수, 누적상대도수 등)을 구함.

* 계급의 수: $K = 1 + \log_2 N$ [sturges formula : N(자료총수), K(계급수)]

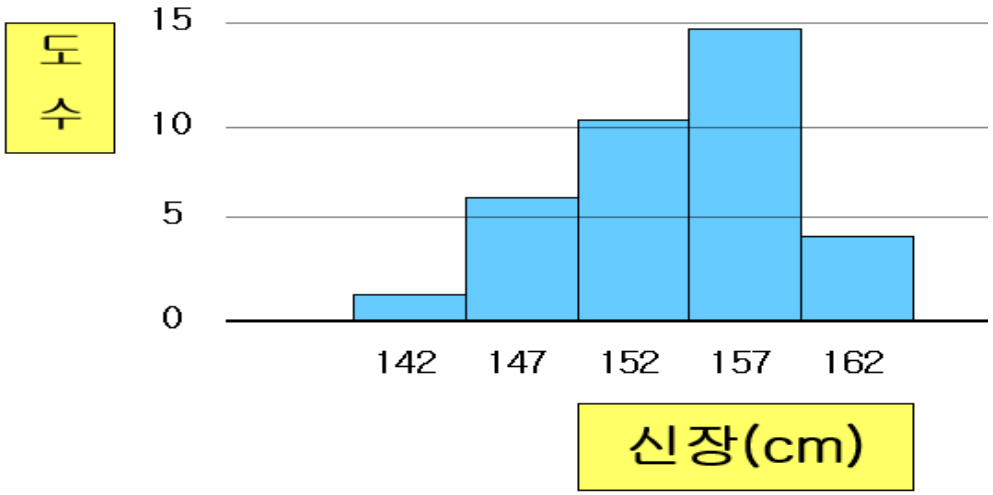
자료의 정리 및 시각화(도수분포표)

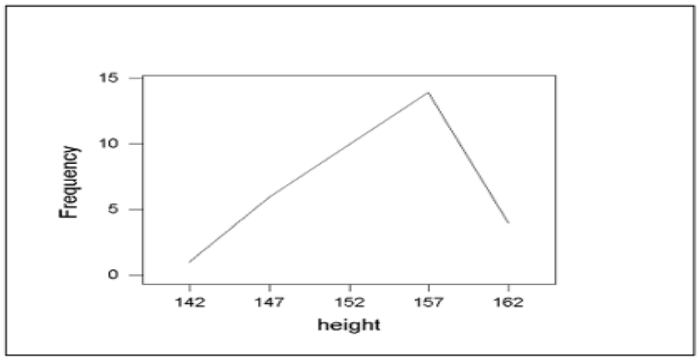
[표2-4] 신장에 대한 도수분포표

신장(계급)	학생수(도수)	상대도수	누적도수	누적상대도수
139.5~144.5	1	0.03	1	0.03
144.5~149.5	6	0.17	7	0.20
149.5~154.5	10	0.29	17	0.49
154.5~159.5	14	0.40	31	0.89
159.5~164.5	4	0.11	35	1.00
계	35	1		

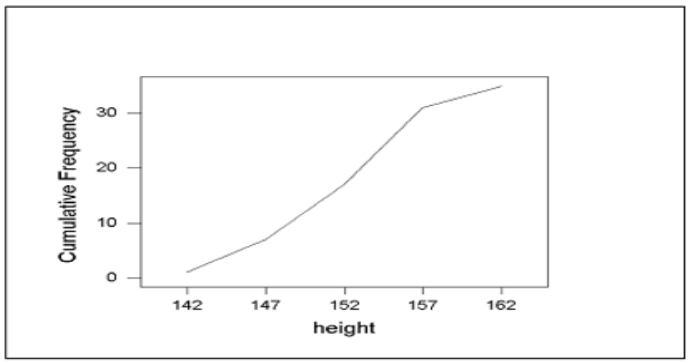
1. 최대값=163, 최소값=140 => 범위=163-140=23
2. 계급의 수=5 => 계급의 폭=23/5=4.6≒5
3. 첫 계급구간의 시작 값=140-1*½ =139.5

자료의 정리 및 시각화(히스토그램)





도수다각형



누적도수다각형

자료의 도표화–범주형–도수분포표+막대도표

진로 (범주)	학생 수 (도수)
공무원	12
기업체	25
대학원진학	5
해외연수	7
어학 및 자격증준비	1
합계	50

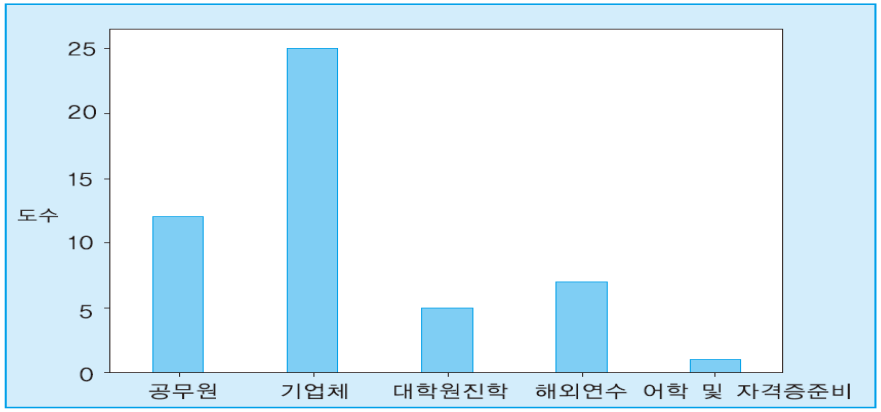


그림 2.2 졸업 후 진로에 대한 막대그래프

표 2.5 졸업 후 진로의 도수분포표

진로 (범주)	학생 수 (도수)	백분율 (%)
공무원	12	24.0
기업체	25	50.0
대학원진학	5	10.0
해외연수	7	14.0
어학 및 자격증준비	1	2.0
합 계	50	100.0

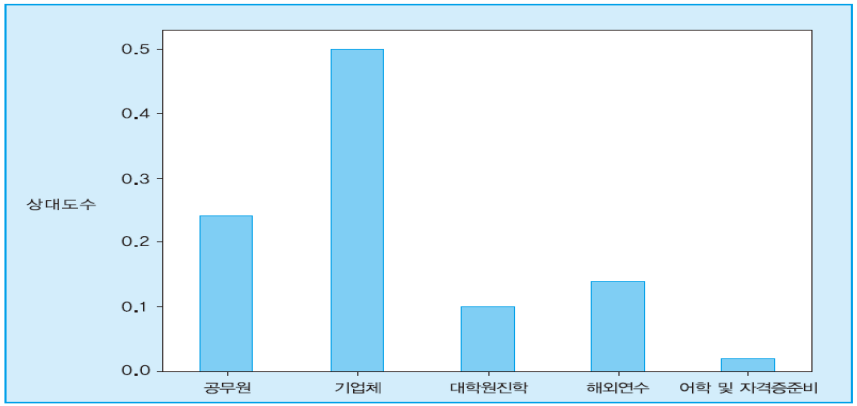


그림 2.3 졸업 후 진로에 대한 막대그래프

질적 자료의 정리 – 원그래프

표 2.5 졸업 후 진로의 도수분포표

진로 (범주)	학생 수 (도수)	백분율 (%)	원그래프 안의 범주들의 각도
공무원	12	24.0	86.4
기업체	25	50.0	180.0
대학원진학	5	10.0	36.0
해외연수	7	14.0	50.4
어학 및 자격증준비	1	2.0	7.2
합 계	50	100.0	360.0

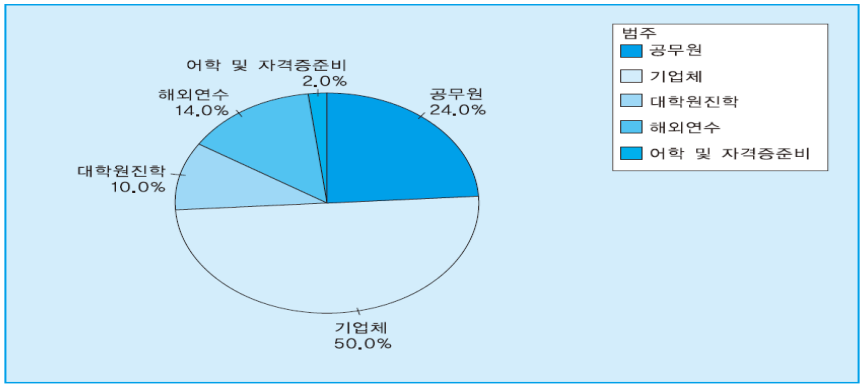


그림 2.4 졸업 후 진로에 대한 원그래프

360도 * 0.24=86.4도

양적 자료 – 범주(계급) 개수의 적절성-1

표 2.7 남학생 50명의 체중

72	74	73	76	66	86	70	71	77	71
70	72	71	72	70	72	79	74	70	74
72	77	78	72	69	68	76	67	69	73
72	73	66	67	72	68	68	67	71	67
69	75	70	68	73	70	68	69	70	71

계급의 수가 5인 경우

계급	도수
65.5 ~ 70.5	22
70.5 ~ 75.5	21
75.5 ~ 80.5	6
80.5 ~ 85.5	0
85.5 ~ 90.5	1
합 계	50

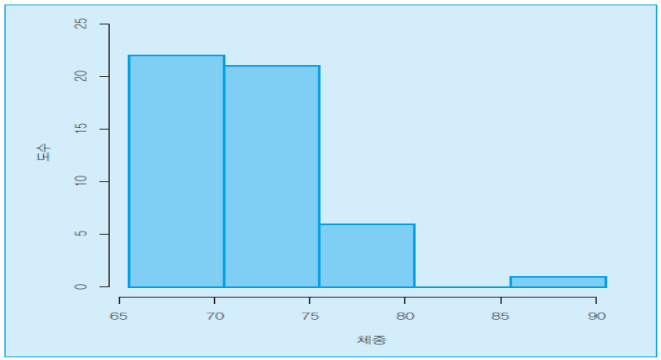
계급의 수가 7인 경우

계급	도수
65.5 ~ 68.5	11
68.5 ~ 71.5	16
71.5 ~ 74.5	15
74.5 ~ 77.5	5
77.5 ~ 80.5	2
80.5 ~ 83.5	0
83.5 ~ 86.5	1
합 계	50

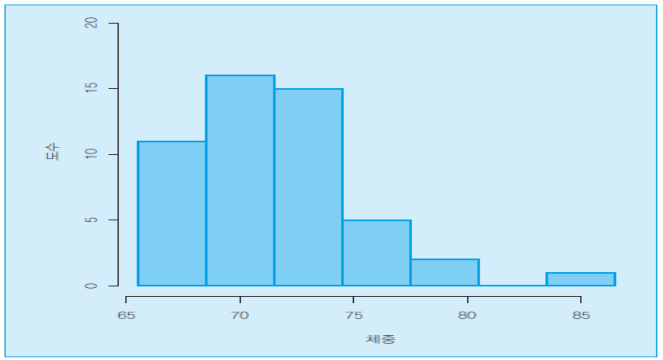
계급의 수가 9인 경우

계급	도수
63.75 ~ 66.25	2
66.25 ~ 68.75	9
68.75 ~ 71.25	16
71.25 ~ 73.75	12
73.75 ~ 76.25	6
76.25 ~ 78.75	3
78.75 ~ 81.25	1
81.25 ~ 83.75	0
83.75 ~ 86.25	1
합 계	50

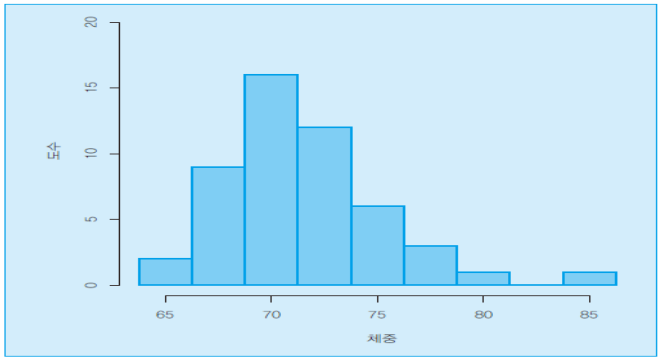
양적 자료 – 범주(계급) 개수의 적절성-2



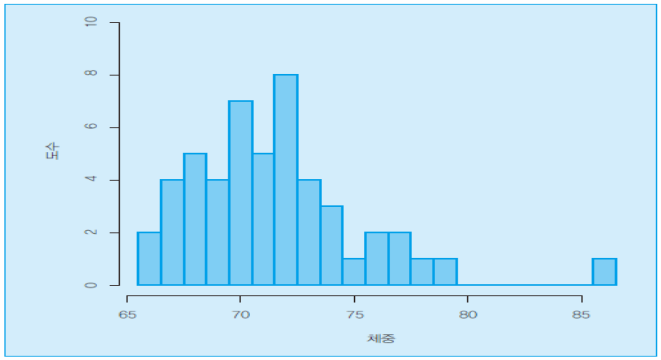
(a) 계급의 수가 5일 때



(b) 계급의 수가 7일 때



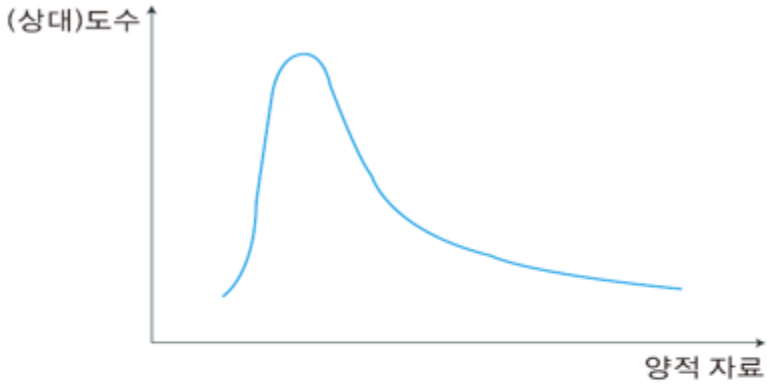
(c) 계급의 수가 9일 때



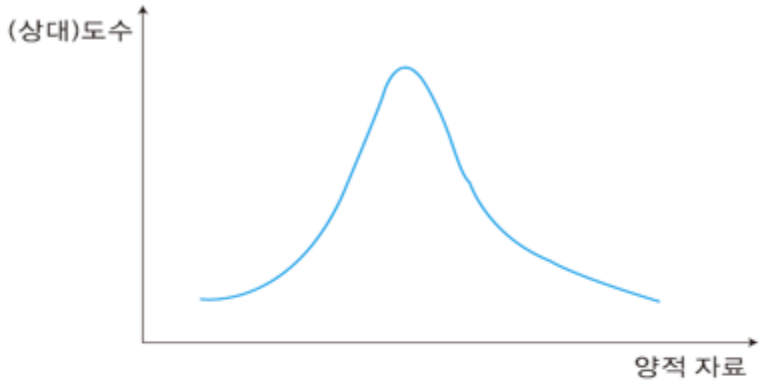
(d) 계급의 수가 21일 때

그림 2.7 각 계급의 수에 따른 히스토그램

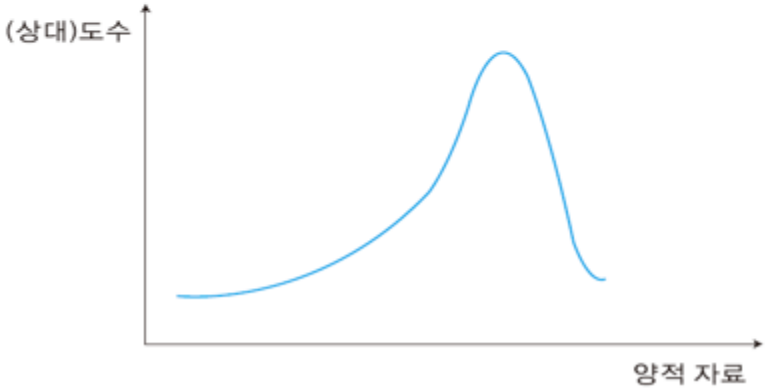
히스토그램의 형태-대칭, 비대칭, 쌍봉분포



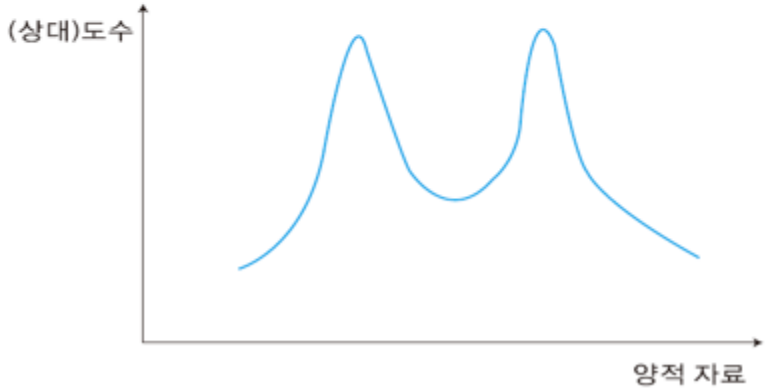
(a) 오른쪽으로 꼬리가 있는 형태



(b) 대칭인 형태



(c) 왼쪽으로 꼬리가 있는 형태

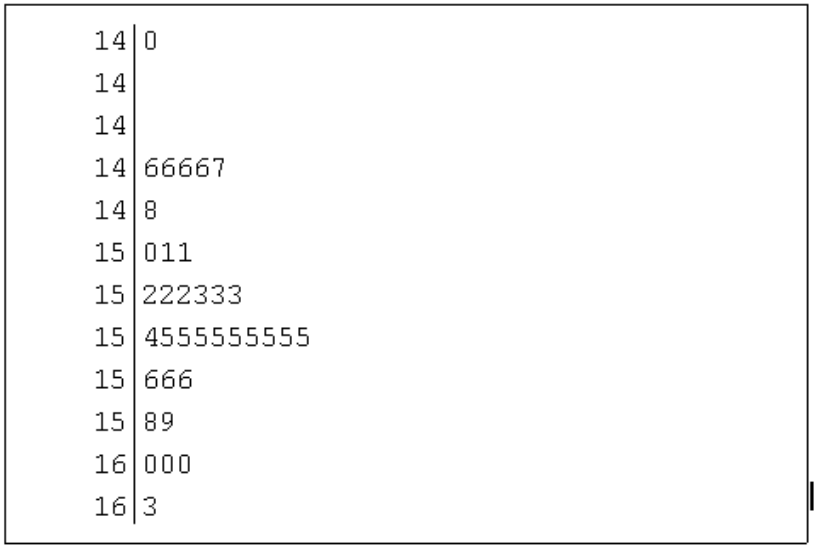


(d) 대칭이나 봉우리가 두 개인 형태

자료의 정리 및 시각화(줄기와 잎 그림)

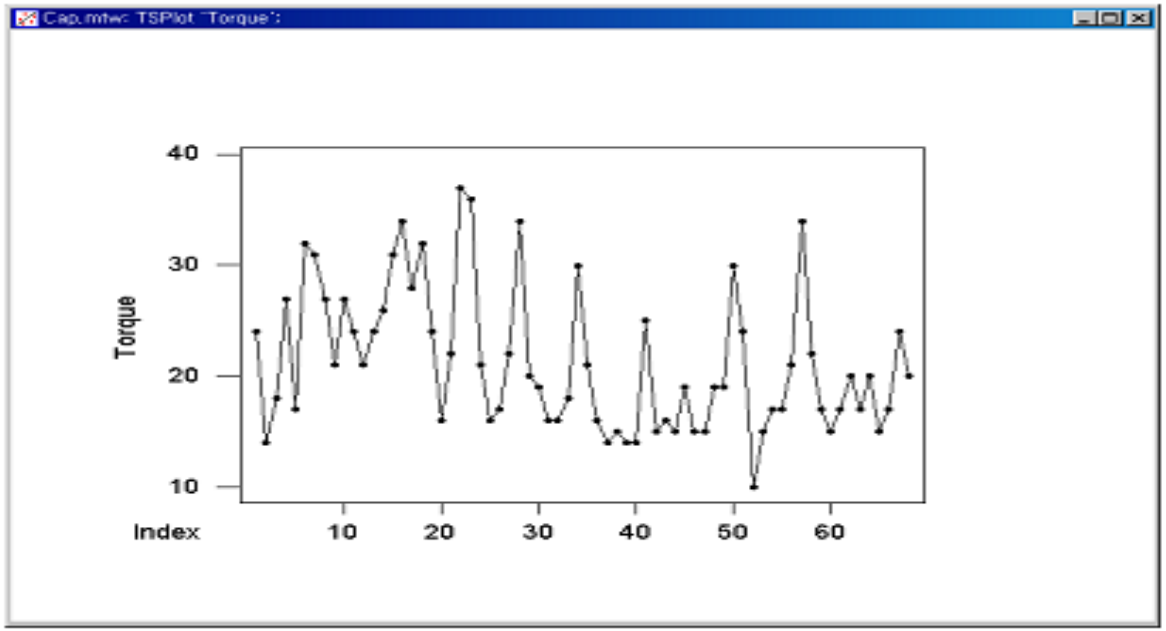
자료 : 어떤 여고생 35명에 대한 신장(단위 : cm)

147	153	155	152	146	160	155
155	155	146	155	163	151	159
158	148	155	146	152	155	151
140	150	160	152	153	160	155
156	154	146	153	155	156	156



- 히스토그램과 비교할 때 줄기와 잎 그림의 장점은 자료의 분포 형태와 자료 값을 그대로 유지하고 있다는 점.

자료의 정리 및 시각화(꺽은선 그래프)



- 시간의 흐름에 따른 자료의 변화(추세, 분포변화)를 표현하는데 유용하다.