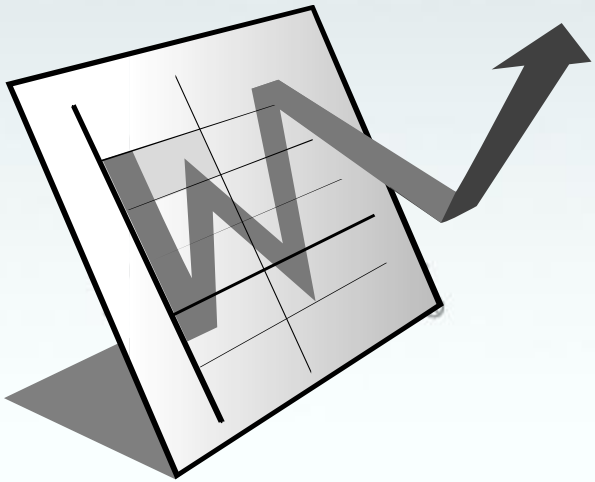
2장



표와 그림을 통한 자료의 요약

용어

- 모집단(population): 알고자 하는 전체 대상. 모든 관측 가능한 값들의 전체 집합.

- 표본(sample): 모집단의 일부분에 해당하는 관측값들의 집합

자료의 요약

- 자료의 특성을 한눈에 파악할 수 있도록 자료를 요약
- 자료를 요약하는 방법은 자료의 형태에 따라 달라진다.

- 자료의 형태별 분류
  - 수치자료(numerical data) or 양적자료(quantitative data):
    관측값이 수치로 측정되는 자료
  - 범주형 자료(categorical data) or 질적자료(qualitative data):
    관측값이 몇 개의 범주 또는 항목의 형태로 측정되는 자료

수치자료 vs 범주형자료

- 수치 자료
  - 연속형 자료(continuous data): 관측 가능한 값이 연속적인 자료
    예) 키, 몸무게
  - 이산형 자료(discrete data): 관측 가능한 값이 비연속적인 자료
    예) 교통사고 발생 건수

- 범주형 자료
  - 순위형 자료(ordinal data): 범주에 순위가 있는 자료
    예) 선호도('매우 좋다', '좋다', '그저 그렇다')
  - 명목형 자료(nominal data): 범주에 순위의 의미가 없는 자료
    예) 혈액형('A', 'B', 'O', 'AB')

- 변수(variable): 관측되는 특성을 나타낸 것
  - 수치변수(키, 몸무게), 범주형 변수(혈액형)

범주형자료의 요약: 도수분포표

- 전체 자료 중에서 각 범주에 속하는 자료의 횟수(도수)를 요약하여 나타낸다.

- 도수분포표(frequency table)
  - 도수(frequency): 각 범주에 속하는 관측값의 개수
  - 상대도수(relative frequency): 각 범주의 도수를 전체 도수로 나눈 값
  - 각 범주에서 범주와 이에 대응하는 도수(상대도수)를 나열하여 표로 작성한 것

도수분포표

- 예) 150명을 대상으로 자동차의 외형에 대한 조사
  - 범주: '좋다', '그저 그렇다', '싫다', '무응답'

답변	도수	상대도수
좋다	71	0.473
그저 그렇다	42	0.280
싫다	28	0.187
무응답	9	0.060
합계	150	1.000

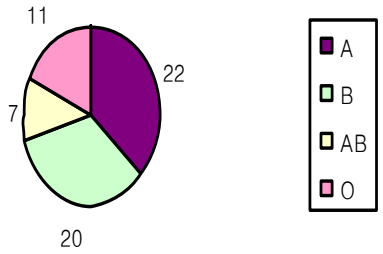
원형그래프(pie chart)

- 원을 각 범주의 상대도수에 비례하도록 중심각을 나누어 파이의 조각처럼 나타낸 것
  - 장점: 각 범주가 전체에서 차지하는 비율을 파악하기 쉽다.
  - 단점: 각 범주간 도수를 비교하는 것이 쉽지 않다. 범주의 수가 많은 경우에는 그리기가 쉽지 않다.

- 예) 학생 60명의 혈액형 조사

혈액형	도수	상대도수 (%)	각도
A	22	36.67	132
B	20	33.33	120
AB	7	11.67	42
O	11	18.33	66
합계	60	100	360

그림 : 혈액형에 대한 원형그래프
(예제2 p.9)



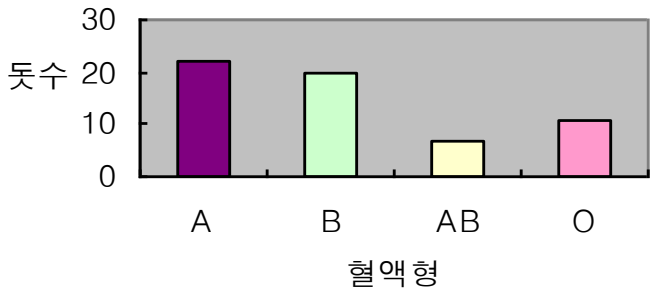
막대그래프(bar chart)

- 각 범주에서 도수의 크기를 막대의 높이로 나타낸 그림
  - 장점: 각 범주간 도수를 비교하기 쉽다.
  - 단점: 각 범주가 전체에서 차지하는 비율을 파악하기 쉽지 않다.

- 예) 학생 60명의 혈액형 조사

혈액형	도수	상대도수 (%)	각도
A	22	36.67	132
B	20	33.33	120
AB	7	11.67	42
O	11	18.33	66
합계	60	100	360

그림 : 혈액형에 대한 막대그래프
(예제2 p.9)



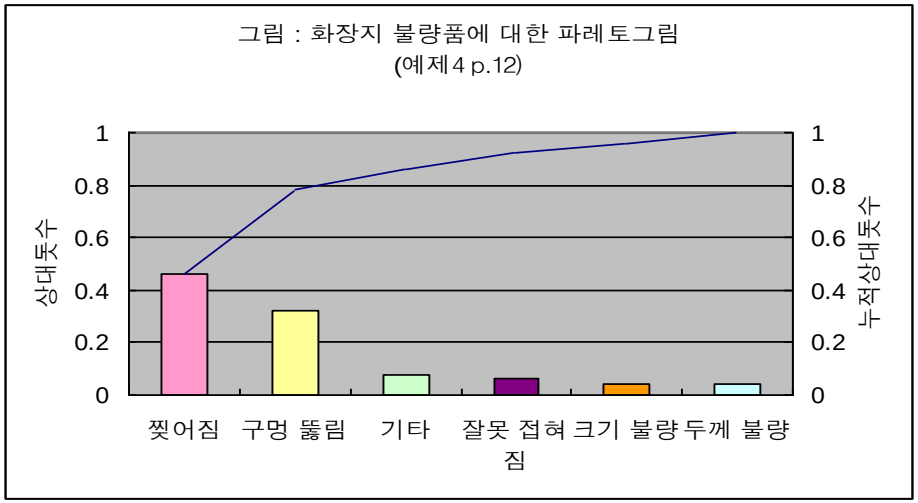
파레토그림(Pareto diagram)

- 막대그래프의 일종으로 상대도수가 큰 순서대로 범주를 왼쪽부터 차례로 배열한 후, 누적상대도수를 각 범주의 막대 위 중앙에 표시하고 그 점을 연결한 그림

- 예) 화장지 불량의 종류에 대한 조사
  - 범주: '구멍 뚫림', '잘못 접혀짐', '크기 불량', '두께 불량', '찢어짐', '기타'

불량의 종류	도수	상대도수	누적 상대도수
찢어짐	23	0.46	0.46
구멍 뚫림	16	0.32	0.78
기타	4	0.08	0.86
잘못 접혀짐	3	0.06	0.92
크기 불량	2	0.04	0.96
두께 불량	2	0.04	1.00
합계	50	1.00	

파레토그림(Pareto diagram)



- Pareto: 이탈리아 경제학자
  - '전제 부의 80% 정도를 약 20%의 사람이 소유하고 있다.'
  - 중요한 소수(vital few): 전체 부의 80%를 점유하는 20%
  - 사소한 다수(trivial many): 나머지 80%

파레토그림(Pareto diagram)

- 파레토그림은 여러 개의 범주 중에서 문제의 해석이나 해결에 도움을 주는 중요한 소수의 범주를 찾는데 도움을 준다.

- 장점: 각 범주들이 차지하는 비율과 상대도수가 증가하는 비율을 파악할 수 있으므로 어느 범주가 중요한지 쉽게 파악할 수 있다.

- 단점: 순위형자료에는 유용하지 않다. (why?)

이산형 자료의 요약

- 관측값의 종류가 적은 경우: 범주형 자료를 요약하는 방법을 사용
- 관측값의 종류가 많은 경우: 연속형 자료를 요약하는 방법을 사용

- 예) 60개의 콩깍지를 대상으로 각 깍지의 콩의 수를 조사한 자료
  - 관측값의 종류: 1, 2, 3, 4, 5, 6
  - 관측값의 종류가 6가지 많지 않으므로 범주형 자료를 요약하는 방법을 사용
  - 도수분포표, 원형그래프, 막대그래프 등

연속형 자료의 요약: 점도표

- 연속형 자료는 연속적인 값을 가지므로 범주형 자료처럼 몇 개의 범주로 나뉘어 있지 않음.

- 점도표(dot diagram)
  - 수평선 위에 각 관측값에 해당하는 위치에 점을 찍어 표시한 그림
  - 관측값의 수가 적은 경우 (20~25개 이하)에 주로 사용

연속형 자료의 요약: 점도표

- 예) 국가별 교육대비 GNP 비율

〈표 5〉 GNP에 대한 교육비 지출 비율

국가	GNP대비 교육비 지출 비율
한국	4.4
일본	4.7
미국	5.3
타이완	5.7
캐나다	7.6
영국	5.2
이탈리아	5.4
말레이시아	5.5

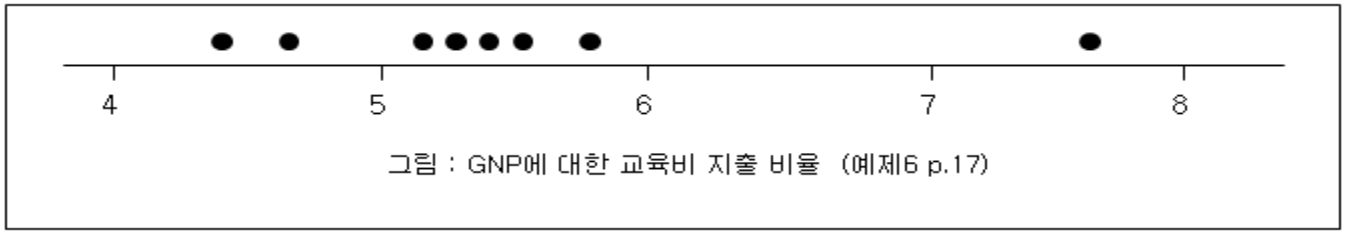


그림 : GNP에 대한 교육비 지출 비율 (예제6 p.17)

- 주로 5~6% 사이에서 교육비 지출
- 한 국가(캐나다)는 다른 나라에 비해 많은 교육비 지출

점도표

- 점도표는 자료의 분포 특징을 쉽게 파악할 수 있도록 한다.

- 자료의 수가 많은 경우에는 점도표가 적절하지 않고, 이 경우에는 자료를 몇 개의 그룹으로 나누어 표시

도수분포표(frequency table)

- 관측값을 몇 개의 구간(계급, class)으로 나누고, 이 계급에 속하는 관측값의 수(도수)를 세어 작성

- 계급구간(class interval): 각 계급에 포함되는 값의 범위

- 도수분포표의 작성방법
  1. 자료의 범위(range=최댓값-최솟값)를 구한다.
  2. 계급구간의 폭(or 계급의 수): 계급의 수가 5~15개가 되도록 자료의 범위를 계급의 수로 나누어 계급구간의 폭으로 정한다.
  3. 계급구간: 관측값이 계급의 경계에 놓이지 않도록 계급구간(or 계급의 시작값)을 정한다.
  4. 각 계급에서 도수와 상대도수를 구한다.

- 시작값을 정하는 일반적인 방법
  - 관측값이 경계에 오지 않도록
  - 최솟값, 최댓값이 계급의 중간에 오도록

도수분포표(frequency table)

- 계급구간의 수(or 계급구간의 폭)를 정하는 방법
  - 계급의 수가 적으면(계급구간의 폭이 크면) 자료가 너무 간략히 요약되어 많은 정보를 잃어버린다.
  - 계급의 수가 크면(계급구간의 폭이 작으면) 각 계급별로 어떤 경향을 가지는지 파악하기 힘들다.
  - 자료 전체에 대한 분포 경향을 잘 나타내도록 계급의 수를 정한다. 보통 5~15개 정도로 한다.

- 도수분포표는 작성하는 방법에 따라 달라진다. (주관적)

- 주의) 자료의 특성에 따라서는 계급구간의 폭을 다르게 할 수 있다. (예, 소득 자료)

도수분포표(frequency table)

- 예) 통계학과 신입생 51명의 키
  - 최솟값: 152, 최댓값: 183
  - 자료의 범위 = 183 – 152 = 31
  - 계급의 수를 7로 정하면, 계급의 폭 = 31/7 = 4.4 -> 5
  - 첫 번째 계급의 시작값을 149.5로 정한다. (관측값이 경계에 오지 않고, 최솟값과 최댓값이 계급의 중간에 오도록), 첫 번째 계급은 149.5 ~ 154.5

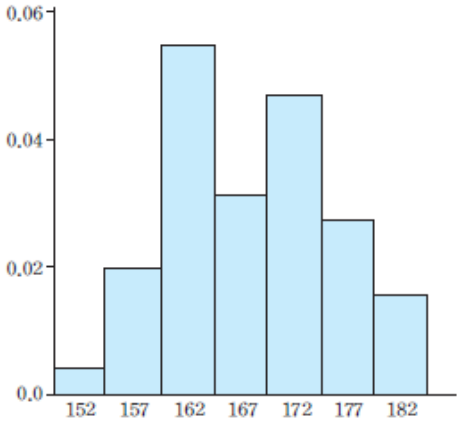
계급구간	도수	상대도수
149.5 ~ 154.5	1	0.02
154.5 ~ 159.5	5	0.098
…	…	…
179.5 ~ 184.5	4	0.078
합계	51	1.000

히스토그램(histogram)

- 도수분포포를 바탕으로 각 계급에서 도수의 크기를 막대로 나타낸 그림 (이산형 자료의 막대그래프에 대응)

- 막대의 높이 = 상대도수/(계급구간의 폭)
- 히스토그램의 전체 면적은 1이 된다. 계급구간의 폭이 모두 같은 경우에는 각 계급의 막대의 높이를 이용하여 비교하고, 계급구간의 폭이 다른 경우에는 막대의 넓이를 이용하여 비교

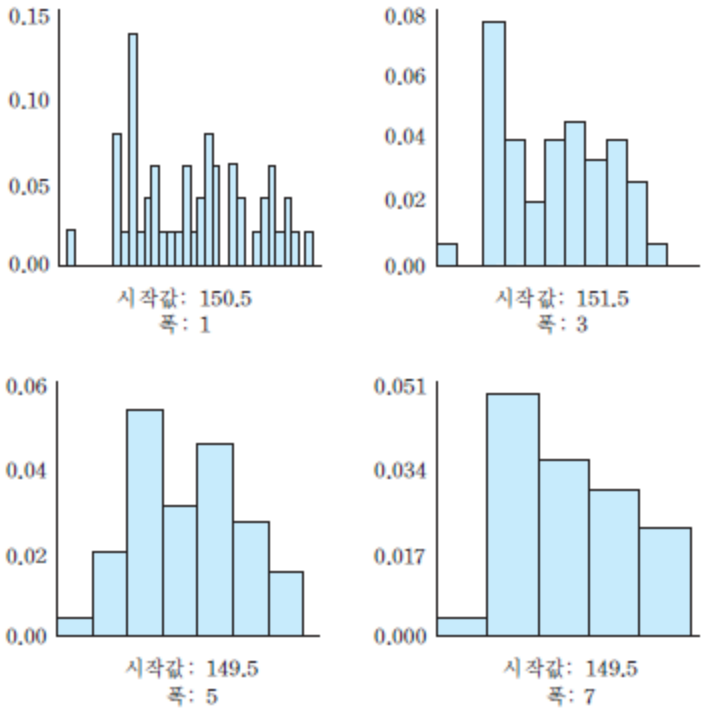
- 예) 통계학과 신입생 51명의 키
  - 두 개의 봉우리



| 그림 7 | 통계학과 신입생의 키에 대한 히스토그램

히스토그램

- 계급구간의 수(폭)와 시작값의 변화에 따라 히스토그램은 달라진다.



히스토그램

- 히스토그램의 변형
  - 키에 대한 히스토그램은 두 개의 봉우리가 있다. 남녀를 비교하기 위해 두 그룹으로 나누어 히스토그램을 작성하고 연결

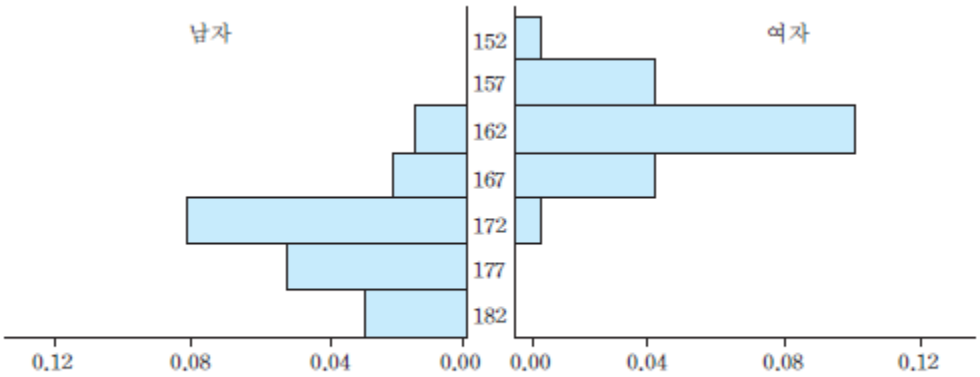


그림 9 | 신입생의 남녀간의 키 비교

도수다각형(frequency polygon)

- 히스토그램의 각 계급구간의 막대 상단의 중앙점을 연결한 그림

- 자료의 분포 특징(관측값의 변화에 따른 도수, 상대도수의 변화, 자료의 중심위치, 퍼진 정도 등)을 히스토그램보다 쉽게 파악할 수 있다. (중앙점을 선으로 연결함으로써)

- 통계학과 신입생 51명의 키

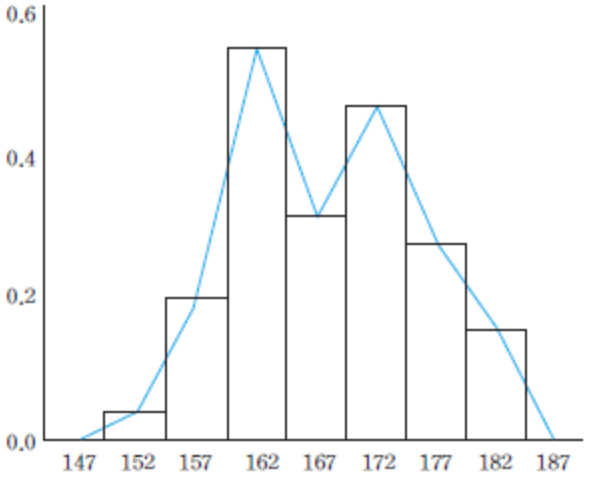


그림 10 | 통계학과 신입생의 키에 대한 도수다각형

도수다각형

- 하나의 좌표에 여러 종류의 도수다각형을 나타내어 여러 자료를 비교하기 쉽다.

- 예) 두 공장 A와 B에서 생산되는 나사의 직경에 대한 자료

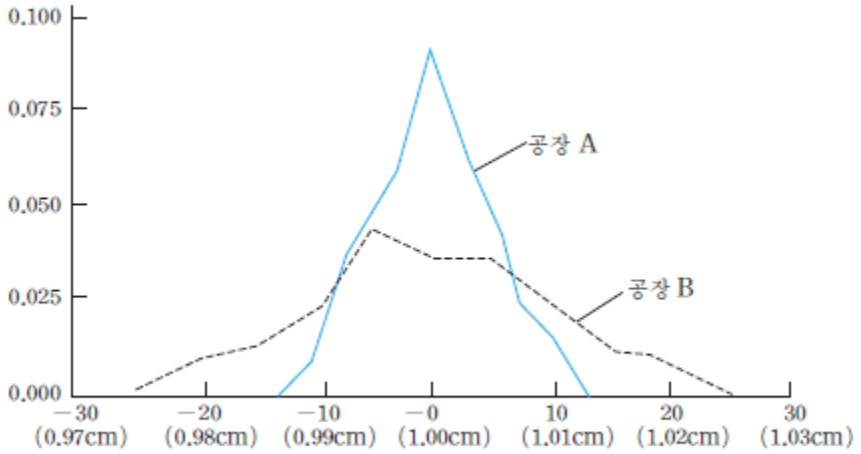


그림 11 | 나사의 직경에 대한 도수다각형

공장 A에서 생산된 나사의 직경이 공장 B에서 생산된 나사에 비해 더 좁은 구간에 분포하고 있다.

줄기-잎 그림(stem-and -leaf plot)

- 히스토그램과 도수다각형은 자료의 분포를 쉽게 파악할 수 있는 장점이 있는 반면 개개의 관측값에 대한 정보를 잃어버린다는 단점이 있다.

- 줄기-잎 그림은 관측값을 앞단위(줄기)와 뒷단위(잎)로 나누어 나무의 줄기와 잎 모양으로 나타낸 그림

- 줄기-잎 그림의 작성방법
  1. 관측값을 앞단위와 뒷단위로 나눈다.
  2. 앞단위를 줄기로 하여 순서대로 세로로 배열하고, 그 옆에 수직선을 그린다.
  3. 뒷단위를 잎으로 하여 앞단위의 오른쪽에 가로로 기입한다.
  4. 각 줄기에서 잎 부분의 값을 작은 숫자가 왼쪽에 오도록 크기순으로 재배열한다.

줄기-잎 그림

- 통계학과 신입생 51명의 키
  - 줄기: 첫 두 자리
  - 잎: 끝자리

15	
16	
17	
18	
1단계	

15	888289
16	1092300037537009207864
17	0981743807608331104
18	1300
2단계	

15	288889
16	00000001223334567778899
17	0000111333446778889
18	0013
3단계	

줄기-잎 그림

- 히스토그램과 같이 자료의 분포 모양을 파악할 수 있으며 관측값 개개의 정보도 얻을 수 있는 장점이 있다. 반면 자료의 수가 너무 많거나 흩어져 있는 경우에는 적절하지 않다.

15^-	2
15^+	88889
16^-	00000001223334
16^+	56777899
17^-	000011133344
17^+	6778889
18^-	0013

그림 13 | 통계학과 신입생의 키에 대한 두 줄기-잎 그림

줄기를 세분화한 경우

	15^-	2
	15^+	88889
33	16^-	000000012234
977	16^+	56789
44333111000	17^-	0
9888776	17^+	
3100	18^-	

그림 14 | 통계학과 남자 신입생과 여자 신입생의 키에 대한 두 줄기-잎 그림

남녀로 구분한 경우

분포의 모양

- 대칭형 분포(symmetric distribution): 예) 종모양 분포
- 이봉형 분포(bimodal distribution): 두 개의 다른 집단의 가능성
- 균일형 분포(uniform distribution)
- 편중된 분포(skewed distribution)
  - 오른쪽으로 편중(skewed to the left)
  - 왼쪽으로 편중(skewed to the right)

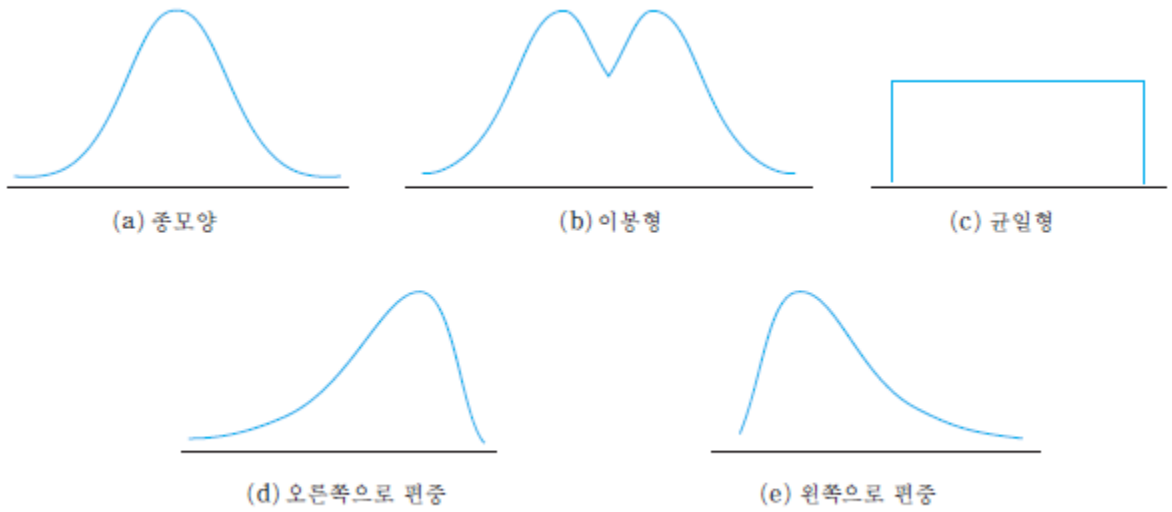


그림 15 여러 가지 분포 모양의 예