P/A/R/T 01

CONTENTS

기술통계학

Descriptive Statistics

기술통계학은 자료를 수집하고, 이를 정리하고 요약하는 데 필요한 자료 기술 기법이다. 이를 통해 수많은 통계자료 중에 필요한 자료를 찾아서 보기 좋게 정리하고, 정리한 자료를 여러 방법으로 분석하고 해석할 수 있다. 또한 자료를 수학적인 모델로 표현하고, 집단의 분포를 살펴봄으로써 미래를 예측하거나 의사결정을 내릴 수 있다. 따라서 기술통계학은 인문사회와 자연과학뿐만 아니라 공학에서도 중요하게 쓰인다.

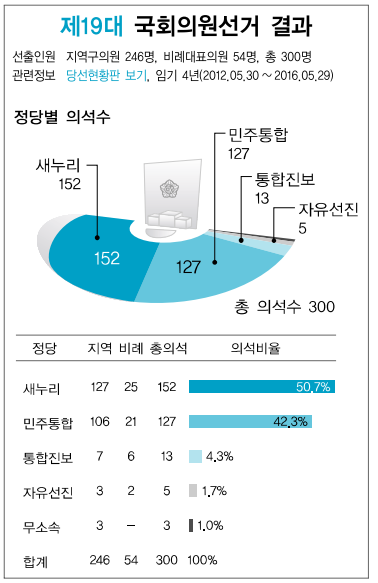
PART 01에서는 기술통계학에서 가장 기본인 자료의 정리와 요약 기법을 살펴보고, 자료를 수학적인 모델로 표현하여 집단의 분포를 수학화한 확률변수와 확률분포에 대하여 알아본다. 마지막으로 표본으로부터 얻은 통계량의 확률분포인 표본분포에 대해 살펴볼 것이다. 이는 PART 02에서 모집단의 특성을 추정하고 검증하는 데 기본이 된다.

Chapter 01 자료의 정리와 요약

학/습/목/표

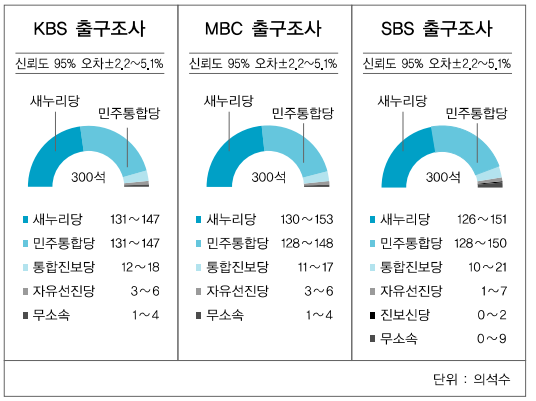
- 자료를 도표나 그림을 이용하여 정리하고, 이를 통해 자료의 특성을 파악할 수 있다.
- 자료의 대푯값을 구하고, 이를 통해 자료의 중심 위치를 알 수 있다.
- 자료의 산포도를 구하고, 이를 통해 자료의 산포 정도를 파악할 수 있다.
- 자료를 상자그림으로 나타내고, 이를 통해 자료의 집중 및 산포 정도를 파악할 수 있다.

우리가 매일 접하는 신문이나 방송을 볼 때 빠지지 않고 나오는 것이 바로 통계적 결과인데, 이러한 통계적 결과에는 두 가지 종류가 있다. 하나는 오른쪽 그림과 같이 선거 실시 후에 선거 결과를 정리한 것이고, 다른 하나는 아래 그림과 같이 선거 실시 전에 각 정당별 예상 의석수를 예측해 보는 것이다.[1]



즉 통계적 결과는 어느 집단이 가시적으로 나타내는 특성을 수량적으로 기술하는 것과, 각 집단으로부터 얻은 자료를 분석하여 그 집단이 갖고 있는 불확실한 특성을 추론하는 것으로 구분된다. 이처럼 통계적 결과를 표나 그래프로 정리하는 것을 **기술통계**라고 한다. 어떤 형태의 자료든지 목적에 따라 가공하지 않으면 그 자료가 가진 특성을 한눈에 파악하기 어렵다. 따라서 자료가 가진 정보를 쉽게 파악하기 위해서는 자료를 잘 가공해야 하는데, 이러한 과정을 **자료 정리**라고 한다.

또한, 자료의 분포 상태는 도표와 그래프로 알 수 있지만, 유사한 그래프를 비교하기 위해서는 분포의 특성을 하나의 수치로 나타내는 것이 필요하다. 이처럼 자료의 특성을 수치로 표현하는 것을 **자료 요약**이라고 한다. 자료 요약을 통해 자료의 중심과, 자료가 그 중심으로부터 얼마나 흩어져서 분포하고 있는지를 수치로 표현할 수 있다. 이 장에서는 자료를 정리하고 요약하는 다양한 기법을 살펴볼 것이다.



1 출처 : 중앙선거관리 국회의원선거 결과 및 방송 3사 출구조사 자료

Section 01

도표를 이용한 자료 정리

자료 정리의 결과는 도표로 나타낼 수 있는데, 대표적인 도표에는 도수분포표, 상대도수분포표, 누적도수분포표, 누적상대도수분포표 등이 있다.

도수분포표

수량화되어 있는 전체 자료를 그 값의 크기에 따라 일정한 계급class으로 나누고, 각 계급에 속하는 자료의 도수frequency를 대응시켜 작성한 표를 **도수분포표** frequency table라고 한다. 일반적으로 도수분포표는 다음과 같은 순서로 작성한다.

정리 1-1 도수분포표 작성 순서

❶ 자료에서 최댓값 $x_{\max}$와 최솟값 $x_{\min}$을 찾아서 범위 $R = x_{\max} - x_{\min}$을 구한다.

❷ 계급의 수 k를 정한다.

❸ 계급구간 $c = \dfrac{R}{k}$을 결정한다.

❹ 계급경계를 결정한다.

❺ 계급값 $x_i = \dfrac{(\text{계급의 양 끝값의 합})}{2}$을 구한다.

❻ 계급도수 f_i를 구한다.

이때 계급의 수는 자료의 성질, 통계의 이용 목적 등을 고려해서 정해야 한다. 계급의 수를 너무 크게 하면 얻고자 하는 자료의 전체적인 윤곽을 파악하는 데 어려움이 발생한다. 반대로, 계급의 수가 너무 작으면 분류의 의미가 없어지게 된다. 따라서 [표 1-1]과 같이 자료의 수에 따라 계급의 수를 정하는 것이 적절하다.

[표 1-1] 자료의 수에 따른 계급의 수

자료의 수	계급의 수 k
50이하	5 ~ 7
40 ~ 45	6 ~ 10
45 ~ 50	7 ~ 12
250이상	10 ~ 20

또한 계급구간은 모두 일정해야 하며, 중복되는 부분이 없어야 한다. 계급값이 복잡한 소수 형태가 나온다면, 계급구간을 분명하게 하기 위해서 그 값에 가까운 자연수로 계급값을 근사하여 사용할 수 있다.

계급경계class boundary는 자료가 이산형인지 연속형인지에 따라 그 표현법이 다르다. 특히 이산형 자료일 경우 각 계급이 중첩되는 부분이 없도록 계급상한class upper limit과 계급하한class lower limit을 분명하게 설정하는 것이 중요하다. 편의상 계급의 상한과 하한을 계급의 양 끝값이라고 한다.

Note 자료의 종류

자료data는 일관성이 없고 의미 없는 숫자나 기호를 조합한 집합으로, 질적자료qualitative data와 양적자료quantitative data로 구분된다. 질적자료는 범주형 자료라고도 하는데, 원래 숫자로 표시할 수 없는 자료이나 측정 대상의 특성을 분류하거나 확인할 목적으로 숫자를 부여한다. 이때 이 숫자들은 양적인 크기를 나타내지는 않지만, 관측된 수치 자료가 셀 수 있는 자료인 이산형 자료discrete data로 나타난다. 질적자료의 대표적인 예로는 성별, 직업, 학력, 혈액형의 구분 등이 있다. 한편, 양적자료는 자료의 크기나 양을 숫자로 표현할 수 있는 자료이다. 양적자료에는 불량품의 개수, 결점의 개수 등과 같이 셀 수 있는 정수 값으로 표현되는 이산형 자료와 물체의 길이, 무게, 온도 등과 같은 연속형 자료continuous data가 있다.

예제 1-1

다음 자료는 하천 유역의 수리시설을 점검하기 위해 지난 30년 동안 누적강우강도를 조사한 것이다. 이 누적강우강도에 대하여 도수분포표를 작성하라.

(단위 : 인치)

43.30	43.11	58.71	42.96	53.20	54.49
47.38	45.93	50.37	48.21	43.93	53.29
63.52	45.05	58.83	49.57	39.91	43.11
40.78	41.31	50.51	51.28	67.72	59.12
55.77	48.26	54.91	44.67	46.77	67.59

풀이

❶ 자료에서 최댓값과 최솟값을 찾으면 $x_{\max} = 67.72$, $x_{\min} = 39.91$이므로 범위 R은 다음과 같다.

$$R = x_{\max} - x_{\min} = 67.72 - 39.91 = 27.81$$

❷ 자료의 수가 30이므로 계급의 수를 $k = 6$으로 정한다.

❸ 계급구간 $c = \dfrac{27.81}{6} = 4.635$이므로 대략 5로 정한다.

❹ 자료의 최솟값과 최댓값을 포함하면서 계급이 중첩되지 않도록 다음과 같이 계급경계를 정한다.

계급(인치)
$35^{\text{이상}} \sim 40^{\text{미만}}$
40 ~ 45
45 ~ 50
50 ~ 55
55 ~ 60
60 ~ 65
65 ~ 70

❺ 각 계급별로 계급값 $x_i = \dfrac{(\text{계급의 양 끝값의 합})}{2}$ 을 구하여 다음과 같이 표에 작성한다.

계급(인치)	계급값
$35^{\text{이상}} \sim 40^{\text{미만}}$	37.5
40 ~ 45	42.5
45 ~ 50	47.5
50 ~ 55	52.5
55 ~ 60	57.5
60 ~ 65	62.5
65 ~ 70	67.5

❻ 계급도수 f_i를 구하여 다음과 같이 도수분포표를 완성한다.

계급(인치)	계급값	도수
$35^{\text{이상}} \sim 40^{\text{미만}}$	37.5	1
40 ~ 45	42.5	8
45 ~ 50	47.5	7
50 ~ 55	52.5	7
55 ~ 60	57.5	4
60 ~ 65	62.5	1
65 ~ 70	67.5	2
합계	–	30

상대도수분포표

계급의 상대도수를 각 계급에 대응시켜 작성한 표를 **상대도수분포표**relative frequency table라고 한다. 이때 **상대도수**relative frequency는 각 계급도수 f_i를 전체 도수 n으로 나눈 것이다. 즉 다음과 같이 상대도수는 전체 도수에 대한 각 계급도수의 비율이다.

$$(\text{상대도수}) = \frac{(\text{각 계급도수})}{(\text{전체 도수})} = \frac{f_i}{n}$$

상대도수분포표는 도수의 총합이 서로 다른 두 집단의 분포를 비교하는 데 유용하게 사용된다. 이때 상대도수의 총합은 항상 1이 됨에 유의한다.

예제 1-2

[예제 1−1]의 자료에 대하여 상대도수분포표를 작성하라.

풀이

[예제 1−1]에서 작성한 도수분포표를 이용하여 각 계급의 상대도수를 다음과 같이 구한다.

$$(\text{상대도수}) = \frac{(\text{각 계급도수})}{(\text{전체 도수})} = \frac{f_i}{n}$$

구한 각 계급의 상대도수를 표로 작성하여 다음과 같이 상대도수분포표를 완성한다.

계급(인치)	도수	상대도수
$35^{\text{이상}} \sim 40^{\text{미만}}$	1	$\frac{1}{30}$
40 ~ 45	8	$\frac{8}{30}$
45 ~ 50	7	$\frac{7}{30}$
50 ~ 55	7	$\frac{7}{30}$
55 ~ 60	4	$\frac{4}{30}$
60 ~ 65	1	$\frac{1}{30}$
65 ~ 70	2	$\frac{2}{30}$
합계	30	1

예제 1-3

다음은 1995년과 2005년의 우리나라 연령별 인구에 대한 비율을 나타낸 것이다. 이러한 인구 비율 변화가 지속될 때, 10년 후 우리나라의 인구 비율에 어떤 변화가 있을지를 예상하라.

나이(세)		$10^{미만}$	$10^{이상} \sim 20^{미만}$	20 ~ 30	30 ~ 40	40 ~ 50	50 ~ 60	60 ~ 70	70 ~ 80	$80^{이상}$
비율 (%)	1995	14.6	17.1	19.0	18.8	12.4	8.9	5.7	2.7	0.8
	2005	11.8	13.9	15.6	17.4	17.1	10.9	7.6	4.3	1.4

풀이

1995년에 비해 2005년의 40세 미만의 인구 비율(상대도수)은 감소하였으나, 40세 이상의 인구 비율은 증가했다. 또한 40세 이상의 인구 비율의 합을 보면, 1995년 30.5%에서 2005년 41.3%로 크게 증가했다. 이런 인구 비율 추세가 지속된다면 10년 후에는 40세 이상의 인구 비율이 50%를 넘을 것으로 예상할 수 있다.

누적도수분포표

도수분포표에서 첫 번째 계급부터 어떤 계급까지의 각 도수를 차례로 더한 값을 각각의 **누적도수**cumulative frequency라 한다. 다음과 같이 i 번째$(1 \le i \le k)$ 계급까지의 도수를 합한 F_i를 i 번째 계급의 누적도수라 한다.

$$F_i = f_1 + f_2 + \cdots + f_i = \sum_{v=1}^{i} f_v$$

이때 계급의 누적도수를 각 계급에 대응시켜 작성한 표를 **누적도수분포표**cumulative frequency table라고 한다. 누적도수분포표는 작은 쪽에서 몇 번째로 큰 자료의 값이 대략 얼마인가 알아보고자 할 때 유용하게 사용된다.

예제 1-4

다음 자료는 한강, 낙동강, 금강, 영산강에서 각각 10개 장소를 선택하여 조사한 생물학적 산소요구량(BOD)을 나타낸 것이다. 이 자료에 대한 계급의 수가 8인 도수분포표를 구하고, 이를 이용하여 누적도수분포표를 작성하라.

(단위 : mg/l)

1.2	1.1	0.8	0.8	1.8	1.5	2.9	2.7	3.2	4.7
1.0	1.1	5.7	7.3	2.8	3.3	2.4	2.1	2.5	3.1
1.0	1.7	0.8	1.0	3.2	4.3	3.0	5.0	3.2	5.5
2.8	5.5	3.9	6.8	6.7	7.9	3.2	1.3	0.9	1.3

풀이

자료의 최댓값과 최솟값은 각각 $x_{\max}=7.9$, $x_{\min}=0.8$이므로 범위 R은 다음과 같다.

$$R=x_{\max}-x_{\min}=7.9-0.8=7.1$$

이때 계급의 수는 $k=8$이므로, 계급구간을 구하면

$$c=\frac{7.1}{8}=0.8875$$

이다. 따라서 계급구간을 대략 1.0으로 정하여 도수분포표를 작성하면 다음과 같다.

계급(mg/l)	도수
0.0$^{\text{이상}}$ ~ 1.0$^{\text{미만}}$	4
1.0 ~ 2.0	11
2.0 ~ 3.0	7
3.0 ~ 4.0	8
4.0 ~ 5.0	2
5.0 ~ 6.0	4
6.0 ~ 7.0	2
7.0 ~ 8.0	2
합계	40

위의 도수분포표를 이용하여 누적도수를 구하면 다음과 같이 누적도수분포표를 얻을 수 있다.

계급(mg/l)	도수	누적도수
0.0$^{\text{이상}}$ ~ 1.0$^{\text{미만}}$	4	4
1.0 ~ 2.0	11	$4+11=15$
2.0 ~ 3.0	7	$15+7=22$
3.0 ~ 4.0	8	$22+8=30$
4.0 ~ 5.0	2	$30+2=32$
5.0 ~ 6.0	4	$32+4=36$
6.0 ~ 7.0	2	$36+2=38$
7.0 ~ 8.0	2	$38+2=40$
합계	40	–

누적상대도수분포표

누적도수 F_i를 전체 도수 n으로 나눈 것을 **누적상대도수**relative cumulative frequency라 한다.

$$(\text{누적상대도수}) = \frac{(\text{각 계급의 누적도수})}{(\text{전체 도수})} = \frac{F_i}{n}$$

이러한 누적상대도수를 각 계급에 대응시켜 작성한 표를 **누적상대도수분포표**relative cumulative frequency table라고 한다. 누적상대도수분포표는 각 계급의 상대도수를 누적한 것으로, 첫 번째 계급부터 어떤 계급까지의 자료가 전체 자료 중에서 얼마나 차지하는지 그 비율을 알아보고자 할 때 유용하게 사용된다.

예제 1-5

[예제 1−4]의 자료에 대하여 누적상대도수분포표를 작성하라.

풀이

[예제 1−4]에서 작성한 누적도수분포표를 이용하여 누적상대도수를 찾으면 다음과 같이 누적상대도수분포표를 얻을 수 있다. 이때 누적상대도수를 살펴보면 상대도수를 누적하여 구한 값과 같음을 알 수 있다.

계급(mg/l)	도수	상대도수	누적도수	누적상대도수
$0.0^{\text{이상}} \sim 1.0^{\text{미만}}$	4	$\frac{4}{40}(=0.1)$	4	$\frac{4}{40}(=0.1)$
1.0 ~ 2.0	11	$\frac{11}{40}(=0.275)$	15	$\frac{15}{40}(=0.375)$
2.0 ~ 3.0	7	$\frac{7}{40}(=0.175)$	22	$\frac{22}{40}(=0.55)$
3.0 ~ 4.0	8	$\frac{8}{40}(=0.2)$	30	$\frac{30}{40}(=0.75)$
4.0 ~ 5.0	2	$\frac{2}{40}(=0.05)$	32	$\frac{32}{40}(=0.8)$
5.0 ~ 6.0	4	$\frac{4}{40}(=0.1)$	36	$\frac{36}{40}(=0.9)$
6.0 ~ 7.0	2	$\frac{2}{40}(=0.05)$	38	$\frac{38}{40}(=0.95)$
7.0 ~ 8.0	2	$\frac{2}{40}(=0.05)$	40	$\frac{40}{40}(=1.0)$
합계	40	1	–	–

지금까지 살펴본 도수, 상대도수, 누적도수, 누적상대도수의 관계를 정리하면 [표 1-2]와 같다. 이때 계급구간이 중복되지 않아야 하고, 계급의 크기는 일정해야 한다. 계급을 읽을 때는 a_i 이상 a_{i+1} 미만과 같이 '이상', '미만'을 붙이면 된다.

[표 1-2] 도수, 상대도수, 누적도수, 누적상대도수의 관계

계급	계급값	도수	상대도수	누적도수	누적상대도수
a_0 이상 ~ a_1 미만 $a_1 \sim a_2$ ⋮ $a_{k-1} \sim a_k$	x_1 x_2 ⋮ x_k	f_1 f_2 ⋮ f_k	f_1/n f_2/n ⋮ f_k/n	$F_1=f_1$ $F_2=f_1+f_2$ ⋮ $F_k=f_1+f_2+\cdots+f_k$	F_1/n F_2/n ⋮ $F_k/n=1$
합계	-	$n=\sum f_i$	$\sum f_i/n=1$	-	-

예제 1-6

다음 자료는 컴퓨터 상점 100곳의 전기사용량을 나타낸 것이다. 전기사용량에 대한 도수, 상대도수, 누적도수, 누적상대도수를 표로 작성하라.

(단위 : kWh)

186	158	176	141	180	140	163	170	183	194
165	175	176	172	175	180	158	180	176	178
175	163	160	169	180	186	173	182	178	166
196	160	172	166	177	173	185	158	174	150
140	178	187	172	174	174	156	186	178	196
184	162	168	157	177	176	177	187	182	184
134	180	163	152	160	168	167	188	135	179
192	166	171	184	172	155	137	135	177	159
189	178	177	166	158	157	168	166	175	178
148	157	179	178	184	187	158	154	143	155

풀이

자료의 최댓값과 최솟값이 각각 $x_{\max}=196$, $x_{\min}=134$이므로 범위 R은 다음과 같다.

$$R=x_{\max}-x_{\min}=196-134=62$$

이때 자료의 수가 100이므로, 계급의 수를 $k=7$로 하여 계급구간을 구하면

$$c=\frac{196-134}{7}\approx 8.86$$

이다. 따라서 계급구간이 애매하지 않도록 10으로 결정하고 각 계급도수를 구한다. 이를 이용하여 상대도수, 누적도수, 누적상대도수를 구하면 다음과 같다.

계급(kWh)	도수	상대도수	누적도수	누적상대도수
$130^{\text{이상}} \sim 140^{\text{미만}}$	4	$\frac{4}{100} = 0.04$	4	$\frac{4}{100} = 0.04$
140 ~ 150	5	$\frac{5}{100} = 0.05$	$4+5=9$	$\frac{9}{100} = 0.09$
150 ~ 160	15	$\frac{15}{100} = 0.15$	$9+15=24$	$\frac{24}{100} = 0.24$
160 ~ 170	18	$\frac{18}{100} = 0.18$	$24+18=42$	$\frac{42}{100} = 0.42$
170 ~ 180	33	$\frac{33}{100} = 0.33$	$42+33=75$	$\frac{75}{100} = 0.75$
180 ~ 190	21	$\frac{21}{100} = 0.21$	$75+21=96$	$\frac{96}{100} = 0.96$
190 ~ 200	4	$\frac{4}{100} = 0.04$	$96+4=100$	1.00
합계	100	1.00	–	–

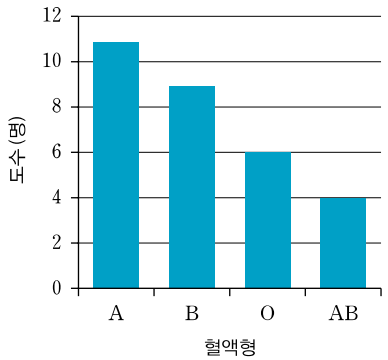
이 표를 보면 170kWh ~ 180kWh를 사용하는 상점의 수가 가장 많은 비율을 차지함을 알 수 있다. 또한 대체로 상점들이 전기를 150kWh ~ 190kWh만큼 사용한다는 것을 쉽게 파악할 수 있다.

Section
02 그림을 이용한 자료 정리

자료의 쏠림 현상은 도표보다 그림을 이용하면 더욱 쉽게 파악할 수 있다. 이러한 방법에는 막대그래프, 히스토그램, 도수분포다각형, 원그래프 등이 있다.

막대그래프

질적자료에서 각 범주의 도수를 막대 모양으로 나타낸 그림을 **막대그래프**bar chart라 한다. 막대그래프에서는 각 범주에 속하는 자료의 도수를 그림으로 표현하므로 자료의 분포를 쉽게 파악할 수 있다. [그림 1-1]은 혈액형 조사 결과를 막대그래프로 나타낸 것이다.



혈액형	도수(명)
A	11
B	9
O	6
AB	4

[그림 1-1] **혈액형 조사 결과에 대한 막대그래프**

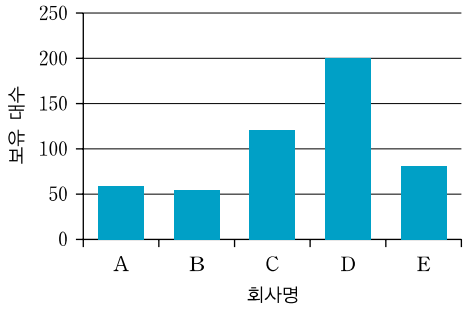
예제 1-7

다음은 자동차 부품 제조회사들이 사용하고 있는 산업용 로봇의 보유 대수를 나타낸다. 이를 이용하여 각 자동차 부품 제조회사의 산업용 로봇의 보유 대수를 막대그래프로 그려라.

회사명	산업용 로봇의 보유 대수
A	60
B	55
C	120
D	200
E	80

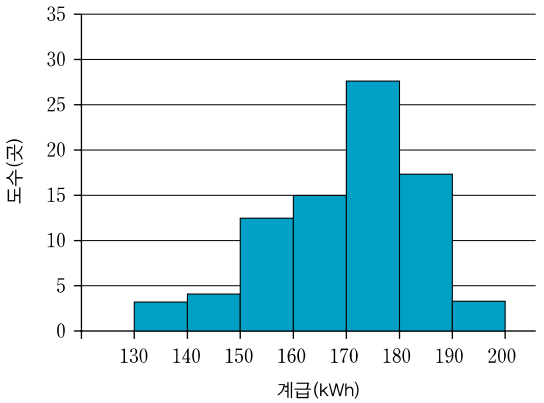
풀이

가로축에는 회사명, 세로축에는 보유 대수의 도수를 기입하여 막대그래프를 그리면 다음과 같다.



히스토그램

도수분포표에서 계급구간을 밑변으로 하고, 계급도수를 높이로 하는 직사각형을 좌표평면에 차례로 나타낸 그래프를 **히스토그램** histogram이라 한다. [그림 1-2]는 [예제 1-6]에서 구한 전기사용량에 대한 히스토그램이다. 이때 x축은 계급, y축은 도수를 나타낸다. 만일 히스토그램의 y축을 상대도수로 나타내면 상대도수히스토그램이 된다.



[그림 1-2] **전기사용량의 히스토그램**

히스토그램을 보면 각 계급도수를 바로 알 수 있으며, 특히 전체적인 자료의 분포 형태를 더욱 쉽게 확인할 수 있다.

예제 1-8

다음 자료는 어느 벤처기업에 근무하는 직원 30명의 연령을 조사한 것이다. 계급구간이 5인 도수분포표와 히스토그램을 구하라.

(단위 : 세)

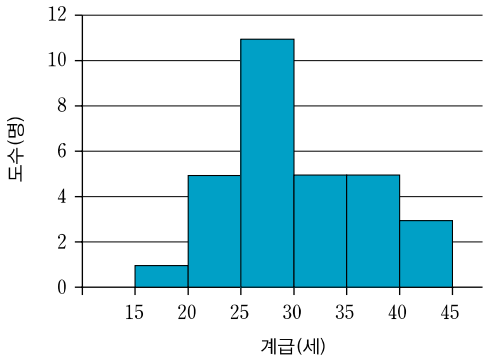
31	27	25	24	21	44	36	29	35	28
16	26	21	37	26	22	30	29	40	33
35	26	22	42	25	28	39	30	27	30

풀이

자료의 최댓값은 44, 최솟값은 16이고 계급구간이 5이므로 계급의 수는 $(44-16)/5=5.6$이다. 그리고 계급의 수를 대략 6으로 정하고, 각 계급도수와 상대도수를 구하면 다음과 같다.

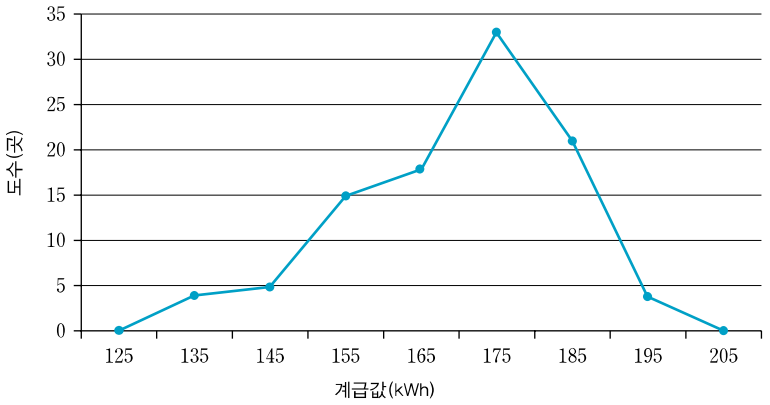
계급(세)	계급값	도수	상대도수
$15^{\text{이상}} \sim 20^{\text{미만}}$	17.5	1	$\frac{1}{30}$
20 ~ 25	22.5	5	$\frac{5}{30}$
25 ~ 30	27.5	11	$\frac{11}{30}$
30 ~ 35	32.5	5	$\frac{5}{30}$
35 ~ 40	37.5	5	$\frac{5}{30}$
40 ~ 45	42.5	3	$\frac{3}{30}$
합계	–	30	1

도수분포표에 따라 가로축에는 계급을, 세로축에는 도수를 기입하여 그리면, 다음과 같은 히스토그램을 얻는다. 한편, 이 그래프에서 세로축을 상대도수로 바꿔도 그래프의 모양은 변함이 없다.



도수분포다각형

히스토그램에서 각 계급구간의 계급값의 빈도수를 직선으로 연결하여 그린 그림을 **도수분포다각형**frequency polygon이라 한다. [그림 1-3]은 [예제 1-6]에서 구한 전기사용량에 대한 도수분포다각형을 보여준다. 도수분포다각형은 히스토그램에서 각 직사각형의 윗변의 중점, 즉 각 계급의 계급값에 해당하는 도수를 선분으로 연결하고, 양 끝은 도수가 0인 계급을 하나씩 추가하여 그 중점을 연결하여 그린 그래프이다.

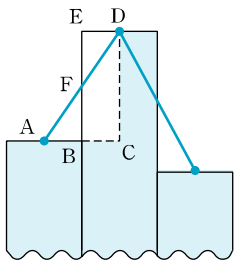


[그림 1-3] **전기사용량의 도수분포다각형**

히스토그램과 달리 도수분포다각형은 두 가지 이상의 자료를 겹쳐 그릴 수 있어, 자료들을 서로 비교하기 편리하다. 또한 다음과 같은 특징이 있다.

(도수분포다각형과 가로축으로 둘러싸인 부분의 넓이)
= (히스토그램에서 직사각형들 넓이의 합)

[그림 1-4]와 같이 히스토그램과 도수분포다각형을 함께 나타냈을 때, 두 직각삼각형 ABF와 DEF가 합동이므로, 사각형 BCDE와 삼각형 ACD의 넓이는 같다. 따라서 도수분포다각형과 가로축으로 둘러싸인 부분의 넓이가 히스토그램의 직사각형들의 넓이의 합과 같음을 확인할 수 있다.



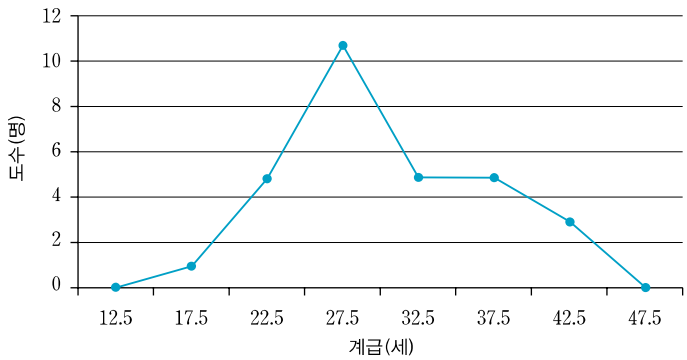
[그림 1-4] **히스토그램과 도수분포다각형**

예제 1-9

[예제 1-8]에서 구한 도수분포표를 이용하여 도수분포다각형을 그려라.

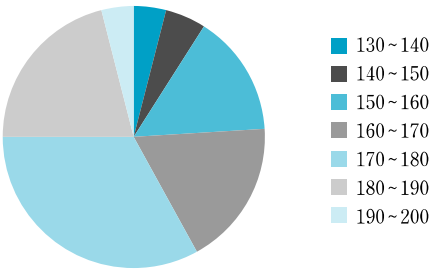
풀이

[예제 1-8]에서 구한 도수분포표의 계급값을 이용하여 도수분포다각형을 그린다. 먼저 히스토그램을 그린 후, 각 계급의 계급값에 해당하는 도수를 선으로 연결하면 다음과 같은 도수분포다각형을 얻는다. 이때 양 끝점은 세로축에 닿도록 한다.



원그래프

원을 계급의 수나 자료의 범주 수만큼 파이 모양의 여러 조각으로 나누어 표현한 그림을 **원그래프** pie chart라고 한다. [그림 1-5]는 [예제 1-6]에서 구한 전기사용량에 대한 원그래프를 보여준다.



[그림 1-5] **전기사용량의 원그래프**

원그래프의 조각 크기는 해당하는 자료의 상대도수에 비례하므로, 자료의 분포 형태를 쉽게 파악할 수 있다. 이때 파이의 중심각은 다음과 같이 구한다.

$$(\text{파이의 중심각}) = 360^\circ \times (\text{상대도수})$$

예제 1-10

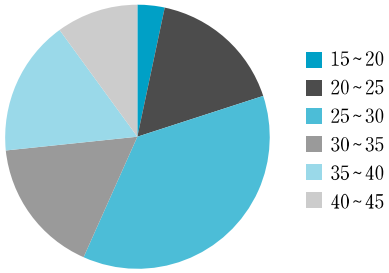
[예제 1-8]에서 구한 도수분포표를 이용하여 원그래프를 그려라.

풀이

먼저 상대도수를 이용하여 각 파이 조각의 중심각을 다음과 같이 구한다.

계급	상대도수	중심각
$15^{\text{이상}} \sim 20^{\text{미만}}$	$\frac{1}{30}$	$360° \times \frac{1}{30} = 12°$
$20 \sim 25$	$\frac{5}{30}$	$360° \times \frac{5}{30} = 60°$
$25 \sim 30$	$\frac{11}{30}$	$360° \times \frac{11}{30} = 132°$
$30 \sim 35$	$\frac{5}{30}$	$360° \times \frac{5}{30} = 60°$
$35 \sim 40$	$\frac{5}{30}$	$360° \times \frac{5}{30} = 60°$
$40 \sim 45$	$\frac{3}{30}$	$360° \times \frac{3}{30} = 36°$
합계	1	-

그 다음에 원을 그리고 각 파이 조각을 중심각에 맞게 나누면 다음과 같은 원그래프를 얻는다.



줄기-잎 그림

도수분포표나 히스토그램에서는 자료의 분포는 쉽게 파악할 수 있지만, 개개의 자료의 관측 값에 대한 정보는 알 수 없다. 자료의 분포 형태를 쉽게 파악하면서도 각 관측값을 알 수 있는 그림으로 **줄기-잎 그림**stem-and-leaf plot이 있다. 이 줄기-잎 그림은 어떤 자료에 대해 큰 수의 자릿값은 줄기에, 작은 수의 자릿값은 잎에 써서 나타낸 그림이다.

줄기-잎 그림은 다음과 같은 순서로 그린다.

정리 1-2 **줄기-잎 그림 작성 순서**

❶ 줄기와 잎을 구분한다. 이때 변동이 적은 부분을 줄기, 변동이 많은 부분을 잎으로 지정한다.
❷ 줄기 부분은 작은 수부터 차례로 나열하고, 잎 부분은 원래 자료의 관측 순서대로 나열한다.
❸ 잎 부분의 관측값을 작은 수부터 순서대로 정리한다.
❹ 전체 자료의 중앙에 놓이는 관측값이 있는 행의 맨 앞에 괄호를 만들고, 괄호 안에 그 행의 잎의 수(도수)를 기입한다.
❺ 괄호가 있는 행을 중심으로 괄호와 동일한 열에 누적도수를 위와 아래 방향에 각각 기입하고, 최소 단위와 전체 자료의 수를 기입한다.

예제 1-11

다음 50명의 통계학 성적에 대한 자료로 줄기-잎 그림을 작성하라.

(단위 : 점)

83	90	60	25	50	94	60	62	97	43	67	84	79
62	78	48	85	52	77	90	25	84	41	65	58	75
83	71	74	68	89	88	76	69	77	89	73	98	75
58	77	69	77	69	65	67	69	79	85	45		

풀이

❶ 십의 자리 수는 변동이 적고 일의 자리 수는 변동이 많으므로 십의 자리 수를 줄기, 일의 자리 수를 잎으로 정한다.

❷ 가장 작은 십의 자리 수인 '2'에서부터 가장 큰 '9'를 줄기 부분에 작성하고, 관측 순서대로 잎 부분에 기입한다.

줄기	잎
2	5 5
3	
4	3 8 1 5
5	0 2 8 8
6	0 0 2 7 2 5 8 9 9 9 5 7 9
7	9 8 7 5 1 4 6 7 3 5 7 7 9
8	3 4 5 4 3 9 8 9 5
9	0 4 7 0 8

❸ 잎 부분의 관측값을 작은 수부터 순서대로 정리한다. 이때 잎에 동일한 수가 있으면, 반복하여 모두 기입한다.

줄기	잎
2	5 5
3	
4	1 3 5 8
5	0 2 8 8
6	0 0 2 2 5 5 7 7 8 9 9 9 9
7	1 3 4 5 5 6 7 7 7 7 8 9 9
8	3 3 4 4 5 5 8 9 9
9	0 0 4 7 8

❹ 전체 자료의 수가 50이므로, 중앙에 놓이는 관측값은 크기순으로 나열하여 25번째와 26번째 자료이다. 따라서 25번째와 26번째 자료가 속해 있는 줄기 '7'행의 맨 앞에 그 행의 잎의 수(도수)인 '13'을 괄호 안에 작성한다.

	줄기	잎
	2	5 5
	3	
	4	1 3 5 8
	5	0 2 8 8
	6	0 0 2 2 5 5 7 7 8 9 9 9 9
(13)	7	1 3 4 5 5 6 7 7 7 7 8 9 9
	8	3 3 4 4 5 5 8 9 9
	9	0 0 4 7 8

❺ 괄호가 있는 행을 중심으로 괄호와 동일한 열에 누적도수를 위와 아래 방향에 각각 기입하고, 최소 단위 '1'과 전체 자료의 수 '50'을 기입한다.

		줄기	잎	
↓	2	2	5 5	(최소 단위) = 1
↓	2	3		(자료의 수) = 50
↓	6	4	1 3 5 8	
↓	10	5	0 2 8 8	
↓	23	6	0 0 2 2 5 5 7 7 8 9 9 9 9	
	(13)	7	1 3 4 5 5 6 7 7 7 7 8 9 9	
↑	14	8	3 3 4 4 5 5 8 9 9	
↑	5	9	0 0 4 7 8	

최종 완성된 줄기-잎 그림을 왼쪽으로 90° 회전하면, 계급간격이 10인 히스토그램의 모양이 된다. 또한 각각의 자료 값도 제시되어 있어 최솟값과 최댓값을 바로 알 수 있다. 줄기-잎 그림은 자료가 크기순으로 나열되어 있으므로 중앙에 있는 자료도 쉽게 파악할 수 있다. 이 경우 25번째와 26번째 자료가 중앙에 있는 자료이므로 그 값은 각각 73과 74이다.

[예제 1-11]에서 자료의 집중 현상을 좀 더 세분화하여 볼 수 있도록, 잎의 자료가 0 ~ 4인 경우와 5 ~ 9인 경우의 줄기를 각각 '·'와 '*'으로 구분하면 [표 1-3]과 같이 계급간격이 5인 세분화된 줄기-잎 그림을 얻을 수 있다.

[표 1-3] **계급간격이 5인 줄기-잎 그림**

	줄기	잎	
2	2*	5 5	(최소 단위) = 1
2	3·		(자료의 수) = 50
2	3*		
4	4·	1 3	
6	4*	5 8	
8	5·	0 2	
10	5*	8 8	
14	6·	0 0 2 2	
23	6*	5 5 7 7 8 9 9 9 9	
(3)	7·	1 3 4	
24	7*	5 5 6 7 7 7 7 8 9 9	
14	8·	3 3 4 4	
10	8*	5 5 8 9 9	
5	9·	0 0 4	
2	9*	7 8	

예제 1-12

다음 자료는 여러 정유회사에서 생산하는 자동차용 휘발유의 옥탄가(octane rating)를 나타낸 것이다. 이 자료에 대한 줄기-잎 그림을 작성하라.

88.5	87.8	83.4	86.7	87.5	91.5	88.6	100.3	95.6	93.3
94.7	91.1	91.0	94.2	87.8	89.9	88.3	87.6	84.3	86.7
88.2	90.8	88.3	98.8	94.2	92.7	93.2	94.5	91.0	89.8
84.5	83.8	87.6	92.5	96.5	89.9	87.3	89.2	90.4	91.0
91.6	95.4	89.5	89.4	94.4	92.2	91.0	90.7	90.4	89.8
87.0	86.5	85.1	84.3	89.1	91.3	92.4	97.5	85.6	86.0
92.3	91.5	84.2	90.8	92.5	91.1	93.3	94.5	88.6	87.5

풀이

❶ 자료에서 변동이 적은 정수 부분을 줄기로, 변동이 많은 소수점 아랫부분을 잎으로 정한다.

❷ 정수 부분의 가장 작은 수 '83'부터 가장 큰 수 '100'까지를 줄기 부분에 작성하고, 관측 순서대로 소수점 아랫부분을 잎에 기입한다.

❸ 잎 부분의 관측값을 작은 수부터 순서대로 정리한다. 이때 잎에 동일한 수가 있으면 반복하여 모두 기입한다.

❹ 전체 자료의 수가 70이므로, 중앙에 놓이는 관측값은 크기순으로 나열하여 35번째와 36번째 사이의 자료이다. 따라서 35번째와 36번째 자료가 속해 있는 줄기 '90' 행의 맨 앞에 그 행의 도수인 '5'를 괄호 안에 작성한다.

❺ 괄호가 있는 행을 중심으로 괄호와 동일한 열에 누적도수를 위와 아래 방향에 각각 기입하고, 최소 단위 '1'과 전체 자료의 수 '70'을 기입한다.

	줄기	잎	
2	83	4 8	(최소 단위) = 1
6	84	2 3 3 5	(자료의 수) = 70
8	85	1 6	
12	86	0 5 7 7	
20	87	0 3 5 5 6 6 8 8	
26	88	2 3 3 5 6 6	
34	89	1 2 4 5 8 8 9 9	
(5)	90	4 4 7 8 8	
31	91	0 0 0 0 1 1 3 5 5 6	
21	92	2 3 4 5 5 7	
15	93	2 3 3	
12	94	2 2 4 5 5 7	
6	95	4 6	
4	96	5	
3	97	5	
2	98	8	
1	99		
1	100	3	

Section 03

대푯값

주어진 자료의 분포 상태를 수치화할 때는 가장 먼저 그 분포의 중심이 어디에 있는지를 살펴봐야 한다. 이때 분포의 중심 위치를 나타내는 값을 대푯값이라고 하며, 이러한 대푯값에는 산술평균, 중앙값, 최빈값, 백분위수, 절사평균 등이 있다.

산술평균

평균 mean에는 산술평균, 기하평균, 조화평균 등이 있는데, 일반적으로 평균이라 하면 **산술평균** arithmetic mean을 의미한다. 산술평균은 가장 보편적으로 사용하는 자료의 중심 측도이다. 변량 X에 대한 n개의 자료가 $x_1, x_2, \cdots, x_n$으로 주어질 때, 변량 X의 산술평균 $\overline{x}$는 다음과 같이 정의된다.

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$$

그리고 변량 X가 모집단[2]에서 얻은 관측값이 $x_1, x_2, \cdots, x_N$으로 주어질 때, 변량 X의 산술평균을 **모평균** population mean이라 하고, μ로 나타낸다. 변량 X의 모평균 μ는 다음과 같이 정의된다.

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1}{N}(x_1 + x_2 + \cdots + x_N)$$

변량 X의 자료가 도수분포로 주어졌을 때는 **가중산술평균** weighted arithmetic mean을 정의하는데, 이는 다음과 같이 구한다.

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{k} f_i x_i = \frac{1}{n}(f_1 x_1 + f_2 x_2 + \cdots + f_k x_k), \quad \sum_{i=1}^{k} f_i = n$$

이러한 산술평균의 성질을 정리하면 다음과 같다.

2 모집단은 통계적인 관측 대상이 되는 집단 전체를 뜻한다.

정리 1-3 산술평균의 성질

(1) 산술평균에 대한 편차[3]의 합은 0 이다.

$$\sum_{i=1}^{n}(x_i-\overline{x})=(x_1-\overline{x})+(x_2-\overline{x})+\cdots+(x_n-\overline{x})=0$$

(2) 산술평균은 편차의 제곱의 합을 최소로 한다. 즉 산술평균에 대한 편차의 제곱의 합은 임의의 수에 대한 편차의 제곱의 합보다 크지 않다.

$$\sum_{i=1}^{n}(x_i-\overline{x})^2 \le \sum_{i=1}^{n}(x_i-a)^2 \quad (\text{단, } a\text{는 상수})$$

(3) 산술평균은 주어진 자료를 모두 사용하므로 정보 손실이 없고, 특히 표본들의 평균인 표본평균은 모집단을 추론할 때 유용하게 사용된다.

(4) 산술평균은 양적자료에 대해서만 구할 수 있으며, 대다수의 자료와 멀리 떨어져 있는 값인 극단값outlier(이상점)에 매우 민감하게 작용한다.

중앙값

변량 X의 n개의 자료 $x_1, x_2, \cdots, x_n$을 작은 값부터 크기순으로 배열했을 때, 한가운데에 위치한 값을 **중앙값**median(중위수)이라고 하고, Me로 나타낸다. 이 값은 자료의 수가 홀수 또는 짝수이냐에 따라 [표 1-4]와 같이 구한다.

[표 1-4] 중앙값 계산

n	중앙값	설명
홀수	$Me=x_{\left(\frac{n+1}{2}\right)}$	$\frac{n+1}{2}$번째 값
짝수	$Me=\frac{x_{\left(\frac{n}{2}\right)}+x_{\left(\frac{n}{2}+1\right)}}{2}$	$\frac{n}{2}$번째 값과 $\frac{n}{2}+1$번째 값의 평균

중앙값은 다음과 같이 편차의 절댓값의 합을 최소로 하는 성질이 있다.

$$\sum_{i=1}^{n}|x_i-Me| \le \sum_{i=1}^{n}|x_i-a| \quad (\text{단, } a\text{는 상수})$$

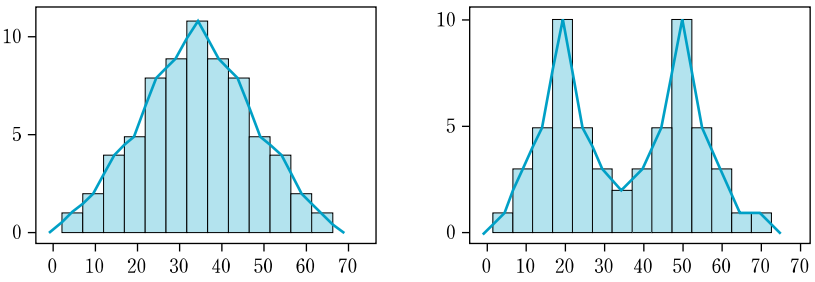
자료의 수가 적으면 중앙값은 구하기 쉽기 때문에 산업 현장에서 사용하기 편리하지만, 자료의 수가 너무 많으면 중앙값을 구하기가 쉽지 않다. 그리고 중앙값은 극단값에 영향을

3 편차는 자료의 변량에서 평균을 뺀 값이다.

받지 않고 자료의 특징을 잘 표현하는 장점이 있는 반면, 자료 전체의 정보를 충분히 이용하지 못하는 단점이 있다. 따라서 극단값이 있는 경우에는 중앙값이 산술평균보다 자료 전체의 특징을 더 잘 나타낸다고 할 수 있다.

최빈값

변량 X의 자료 중에서 가장 많이 나타나는 값을 **최빈값**mode(최빈수)이라 하고, Mo로 나타낸다. 최빈값은 [그림 1-6]과 같이 도수분포곡선에서 최고봉에 해당하는 변량의 값이 된다. [그림 1-6(a)]의 단봉형 도수분포곡선에서는 최빈값이 35로 한 개이고, [그림 1-6(b)]의 쌍봉형 도수분포곡선에서는 최빈값이 20과 50으로 두 개가 존재함을 알 수 있다.



(a) 단봉형 (b) 쌍봉형

[그림 1-6] **도수분포곡선에서의 최빈값**

따라서 최빈값은 가장 이해하기 쉬운 대푯값이며, 반드시 하나만 존재하는 것도 아니다. 최빈값은 주로 의류 업계에서 기성복의 치수를 정하거나 선호도를 조사할 때 이용한다.

예제 1-13

다음 50명의 통계학 성적에 대해 산술평균, 중앙값, 최빈값을 각각 구하라.

(단위 : 점)

83	90	60	25	50	94	60	62	97	43	67	84	79
62	78	48	85	52	77	90	25	84	41	65	58	75
83	71	74	68	89	88	76	69	77	89	73	98	77
58	77	69	75	69	65	67	69	79	85	45		

풀이

먼저 이 자료의 산술평균을 구하면 다음과 같다.

$$\bar{x} = \frac{1}{50}(83 + 90 + 60 + 25 + \cdots + 85 + 45) = 70.48$$

중앙값과 최빈값을 구하기 위해 이 자료의 줄기-잎 그림을 그리면 다음과 같다.

	줄기	잎	
2	2	5 5	(최소 단위) = 1
2	3		(자료의 수) = 50
6	4	1 3 5 8	
10	5	0 2 8 8	
23	6	0 0 2 2 5 5 7 7 8 9 9 9 9	
(13)	7	1 3 4 5 5 6 7 7 7 7 8 9 9	
14	8	3 3 4 4 5 5 8 9 9	
5	9	0 0 4 7 8	

자료의 수가 50이므로, 중앙값은 자료를 작은 값부터 크기순으로 나열하여 25번째와 26번째의 평균값이다. 25번째와 26번째 자료가 $x_{25} = 73$, $x_{26} = 74$이므로 중앙값은 다음과 같다.

$$Me = \frac{(73 + 74)}{2} = 73.5$$

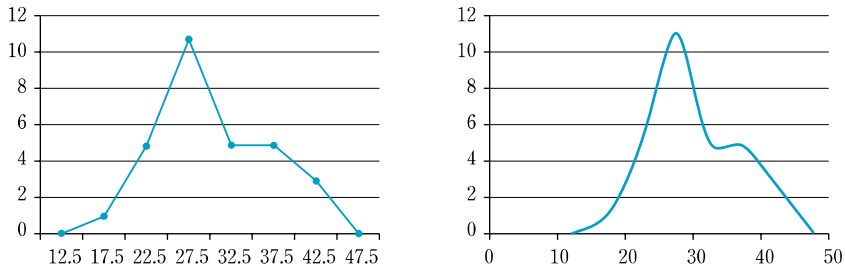
또한 자료에서 빈도가 가장 많은 값은 69가 4번, 77이 4번씩이다. 따라서 최빈값은

$$Mo = 69,\ 77$$

이다. 이를 통해 최빈값이 항상 하나만 존재하는 것은 아님을 확인할 수 있다.

Note 도수분포곡선

도수분포다각형에서 자료의 수를 늘리고 계급의 폭을 좁게 하여 구간의 수를 늘려 나가면 도수분포다각형이 결국 곡선 형태로 수렴하는데, 이 곡선을 도수분포곡선이라 한다. [예제 1-8]의 도수분포곡선은 다음과 같이 그릴 수 있다.



(a) 도수분포다각형 (b) 도수분포곡선

[그림 1-7] 도수분포다각형과 도수분포곡선

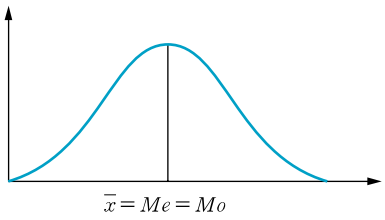
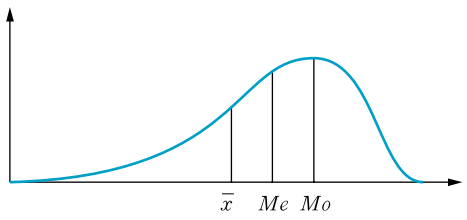
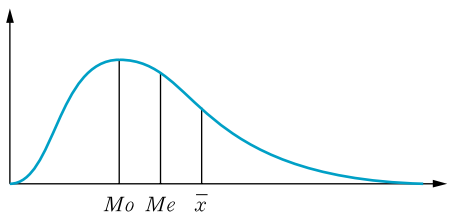
산술평균, 중앙값, 최빈값 사이의 관계

도수분포곡선이 단봉형으로 나타나고, 극히 비대칭이 아닌 자료 집단의 산술평균($\overline{x}$), 중앙값(Me), 최빈값(Mo) 사이에는 도수분포도가 봉우리가 하나인 단봉형으로 극히 비대칭이 아닐 때, 다음과 같은 '피어슨의 실험 공식'이 성립한다.

$$\overline{x} - Mo \approx 3(\overline{x} - Me)$$

또한 도수분포곡선의 모양에 따른 $\overline{x}$와 Me 및 Mo 사이의 관계를 살펴보면 다음과 같다.

[표 1-5] 도수분포곡선 모양에 따른 산술평균, 중앙값, 최빈값 사이의 관계

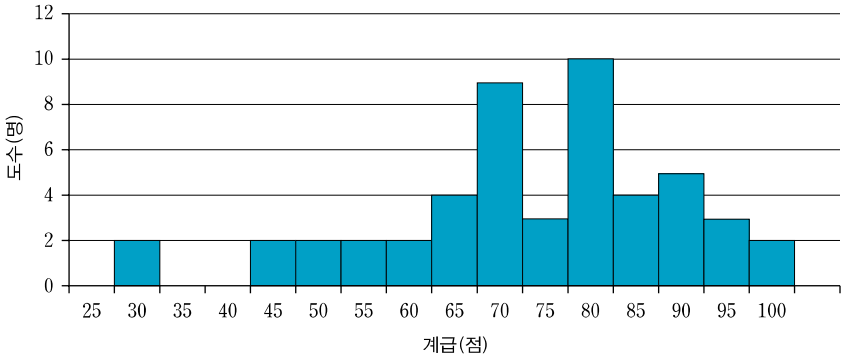
분류	도수분포곡선	관계
도수분포가 완전히 대칭인 경우		$\overline{x} = Me = Mo$
도수분포가 오른쪽으로 치우친 경우		$\overline{x} < Me < Mo$
도수분포가 왼쪽으로 치우친 경우		$Mo < Me < \overline{x}$

예제 1-14

[예제 1-13]의 통계학 성적에 대한 산술평균, 중앙값, 최빈값 사이의 관계를 설명하라.

풀이

통계학 성적에 대한 히스토그램을 그리면 다음과 같다.



이 히스토그램을 보면 약간 오른쪽으로 치우쳤음을 알 수 있다. 이때 산술평균($\overline{x}$), 중앙값(Me), 최빈값(Mo)를 구하여 비교하면 다음과 같다.

$$\overline{x}\,(=70.48) < Me\,(=73.5) < Mo\,(=77)$$

이 경우에는 $(\overline{x}-Me) \approx 2.16(\overline{x}-Mo)$으로 피어슨의 실험 공식을 완전하게 따르지는 않는데, 이는 이 그래프가 완벽한 단봉형이 아니기 때문이다.[4]

백분위수와 사분위수

변량 X의 n개의 자료를 작은 값부터 크기순으로 배열했을 때, $0 \le p \le 1$인 p에 대하여 전체 자료를 $100p\%$와 $100(1-p)\%$로 나누는 값을 제$100p$ **백분위수**percentile라 한다. 자료의 수가 n개일 때, 제$100p$ 백분위수는 그 값보다 작거나 같은 자료의 수가 np개 이상이고, 그 값보다 크거나 같은 자료의 수가 $n(1-p)$개 이상인 값이다.

자료 1, 2, 3, 4, 5를 생각해보자. 이 자료의 제 30 백분위수는 전체 자료를 $(100\times0.3)\%$와 $(100\times0.7)\%$로 나누는 값으로, 그 값보다 작거나 같은 자료의 수가 $5\times0.3=1.5$개 이상이어야 하고, 그 값보다 크거나 같은 자료의 수가 $5\times0.7=3.5$개 이상이어야 한다. 이때 2보다 작거나 같은 자료가 2개이고, 2보다 크거나 같은 자료가 4개이므로 제 30 백분위수는 2가 된다. 그러나 제 40 백분위수는 그 값보다 작거나 같은 자료의 수가 $5\times0.4=2$개 이상이어야 하고, 그 값보다 크거나 같은 자료의 수가 $5\times0.6=3$개 이상이어야 한다. 이때 2와 3사이의 값은 그 값보다 작은 자료의 수가 2이고 크거나 같은 값이 3개이므로, 2와 3사이의 모든 값이 제 40 백분위수가 될 수 있다. 따라서 백분위수는 어떻게 정의하느냐에 따라 다양하게 표현될 수 있다. 여기서는 다음과 같은 방법으로 백분위수를 구하기로 한다.

4 하지만 피어슨의 실험 공식에 어느 정도 의미가 있다고 볼 수 있다.

정리 1-4 **제$100p$ 백분위수 구하는 방법**

❶ 자료를 작은 값부터 크기순으로 배열한다.

❷ 자료의 수 n에 p를 곱하여 다음과 같은 기준으로 제$100p$ 백분위수를 결정한다.

- 만일 np가 정수이면, np번째로 큰 자료와 $(np+1)$번째로 큰 자료의 평균을 택한다.
- 만일 np가 정수가 아니면, np의 정수 부분에 1을 더한 값 m을 구하고 m번째로 큰 자료를 택한다. 자료와 멀리 떨어져 있는 값인 극단값에 매우 민감하게 작용한다.

특히 제 25, 50, 75 백분위수는 자료를 4등분하는 위치에 있는 값으로, 이 값을 **사분위수** quartile라고 한다. 이를 각각 Q_1, Q_2, Q_3로 표시하면, Q_1을 제 1 사분위수, Q_2를 제 2 사분위수(중앙값), Q_3을 제 3 사분위수라 한다.

예제 1-15

다음 자료에서 제 50 백분위수와 제 25 백분위수를 구하라.

16, 25, 4, 18, 11, 13, 20, 8, 11, 9

풀이

자료를 다음과 같이 작은 값부터 크기순으로 배열한다.

$$4,\ 8,\ 9,\ 11,\ 11,\ 13,\ 16,\ 18,\ 20,\ 25$$

제 50 백분위수는 $n=10,\ p=0.5$이므로 $np=5$이다. np가 정수이므로, 다음과 같이 제 50 백분위수는 5번째로 큰 값과 6번째로 큰 값의 평균이다.

$$\frac{(11+13)}{2}=12$$

또한 제 25 백분위수는 $np=10\times 0.25=2.5$이므로, 정수 부분인 2에 1을 더한다. 따라서 제 25 백분위수는 $(2+1)$번째로 큰 값, 즉 3번째로 큰 값인 9이다.

절사평균

산술평균은 극단값의 유무에 따라 많은 영향을 받으므로, 수집한 자료에 극단값이 있을 때 이 극단값을 제거하면 좀 더 바람직한 평균을 산출할 수 있다. 이러한 척도를 얻기 위해 수집한 자료를 작은 값부터 크기순으로 나열하고 $0 \le \alpha \le 0.5$인 α에 대하여 양 끝에서 $100\alpha\,\%$에 해당하는 자료를 제거하고 남은 나머지 자료들의 평균을 구한다. 이와 같이 얻은 평균을 $100\alpha\,\%$ **절사평균**trimmed mean이라 한다.

정리 1-5 절사평균 구하는 방법

❶ 자료를 작은 값부터 크기순으로 배열한다.

❷ $0 \le \alpha \le 0.5$인 α에 대하여 자료의 수 n에 α를 곱하여 다음과 같은 기준으로 자료의 수를 제거한다.

- 만일 αn이 정수이면, 이 정수에 해당하는 자료의 수만큼 양 끝에서 제거한다.
- 만일 αn이 정수가 아니면, αn을 넘지 않는 최대 정수에 해당하는 자료의 수만큼 양 끝에서 제거한다.

❸ 제거하고 남은 자료에 대하여 산술평균을 구한다.

예제 1-16

다음 자료에서 15% 절사평균을 구하라.

68, 70, 67, 10, 72, 68, 70, 71

풀이

자료를 다음과 같이 작은 값부터 크기순으로 배열한다.

$$10,\ 67,\ 68,\ 68,\ 70,\ 70,\ 71,\ 72$$

자료의 수가 $n=8$, $\alpha=0.15$이므로 $\alpha n=0.15\times 8=1.2$이다. 이때 αn이 정수가 아니므로, 1.2를 넘지 않는 최대 정수에 해당하는 1만큼 자료의 수를 양 끝에서 제거한다.

$$67,\ 68,\ 68,\ 70,\ 70,\ 71$$

이때 이들의 산술평균을 구하면

$$\frac{67+68+68+70+70+71}{6}=69$$

이므로, 주어진 자료에 대한 15% 절사평균은 69이다.

Section
04 산포도

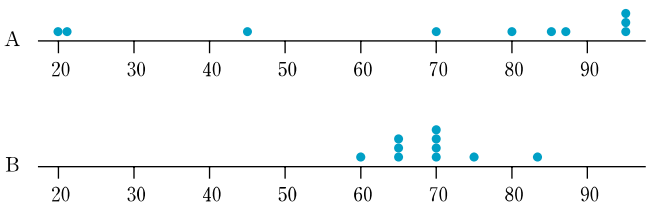
자료의 분포 상태를 알기 위해서는 자료의 중심 위치 이외에도 자료의 흩어진 정도를 함께 고려해야 한다. 이와 같이 자료의 흩어진 정도를 **산포도**measure of dispersion라 한다.

[표 1-6]과 같이 통계학 성적의 평균이 69.3점으로 동일한 두 그룹 A, B를 살펴보자. 이 두 그룹의 점수의 중앙값은 각각 82.5점과 70점으로 약 12점 정도의 차이가 나지만, 최댓값과 최솟값의 차는 B 그룹(23점)보다 A 그룹(75점)이 3배 이상 크게 나타난다.

[표 1-6] **평균이 같은 두 그룹**

그룹	통계학 성적	총점	평균
A	20, 45, 95, 80, 70, 85, 95, 87, 21, 95	693	69.3
B	60, 65, 70, 75, 70, 70, 70, 65, 65, 83	693	69.3

또한 [그림 1-8]과 같이 수평축에 점수에 해당하는 점을 찍어 나타내는 점도표로 점수 분포를 비교하면, 두 그룹의 점수 분포에는 분명히 큰 차이가 있음을 알 수 있다.



[그림 1-8] A **그룹과** B **그룹의 점도표**

즉 B 그룹의 점수는 평균 69.3점을 중심으로 밀집되어 있는 반면, A 그룹의 점수는 최고 점수와 최저 점수의 차가 매우 크고 점수들도 다양하게 나타난다. 따라서 두 그룹 점수의 중심 위치만 알고 있다면, 두 그룹이 유사한 점수 분포를 갖는다고 결론을 내릴 수 있으나, 실제 두 그룹의 점수 분포는 명확하게 차이가 있다. 즉 대푯값만으로는 각각의 자료에 대한 정보를 충분히 제공하지 못한다는 것이다. 따라서 대푯값으로 반영할 수 없는 분포 특성을 살펴보기 위해서는 산포도를 고려해야 한다.

이제 범위, 사분위수 범위, 평균편차, 분산, 표준편차, 변동계수 등과 같이 널리 이용되는 산포도의 척도들을 살펴보자.

범위

변량 X의 자료가 $x_1, x_2, \cdots, x_n$일 때, X의 **범위**range는 이들 자료의 최댓값($x_{\max}$)과 최솟값($x_{\min}$)의 차를 말하며, 보통 R로 표시한다.

$$R = x_{\max} - x_{\min}$$

범위는 가장 쉽게 구할 수 있는 산포도로, 자료의 수가 많지 않을 때에는 산포에 대한 유용한 추정치가 되므로 품질관리 등 산업 현장에서 많이 이용된다. 그러나 범위는 극단값의 영향을 많이 받으므로 매우 불안정한 산포도이다.

사분위수 범위

범위는 자료의 두 극단값의 차이만을 나타내기 때문에 자료의 산포를 나타내기에 불충분하다. 이러한 단점을 일부 보완한 산포도가 **사분위수 범위**interquartile range이다. 사분위수 범위는 다음과 같이 제 3 사분위수와 제 1 사분위수의 차이로 정의된다.

$$(\text{사분위수 범위}) = Q_3 - Q_1$$

사분위수 범위를 응용한 자료 표현의 방법으로는 상자그림box plot이 있는데, 이는 Section 05에서 살펴 볼 것이다.

분산과 표준편차

편차의 합 대신에 편차의 제곱의 합을 이용하는 **분산**variance과 **표준편차**standard deviation는 산포의 척도로 가장 많이 쓰이며, 자료가 모집단이냐 표본이냐에 따라 다르게 정의된다.

■ 분산

변량 X의 자료가 N개의 원소 $x_1, x_2, \cdots, x_N$으로 이루어진 모집단일 때, 모집단의 분산인 **모분산**population variance은 σ^2으로 나타내며, 다음과 같이 정의한다.

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

여기에서 μ는 모평균이다. 또한 변량 X의 자료가 n개의 원소 $x_1, x_2, \cdots, x_n$으로 이루어진 모집단의 한 표본일 때, X의 **표본분산**sample variance은 s^2으로 나타내며, 다음과 같이 정의한다.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

그리고 X의 자료가 도수분포로 주어졌을 때의 표본분산은 다음과 같이 구한다.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{k} f_i(x_i - \overline{x})^2, \quad \sum_{i=1}^{k} f_i = n$$

■ 표준편차

모분산의 양의 제곱근인 σ를 **모표준편차**population standard deviation, s를 **표준편차**sample standard deviation라 하는데, 각각 다음과 같이 정의한다.

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

또한 X의 자료가 도수분포로 주어졌을 때의 표준편차는 다음과 같다.

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{k} f_i(x_i - \overline{x})^2}, \quad \sum_{i=1}^{k} f_i = n$$

분산은 자료의 값을 제곱하여 얻어지므로 그 값이 큰 반면에, 표준편차는 분산에 제곱근을 취한 것이므로 그 값이 분산보다는 작다. 그러므로 표준편차가 분산보다 다루기 쉽다. 또한 분산의 단위는 자료의 측정 단위의 제곱인 반면에, 표준편차는 자료의 단위와 같으므로 산포도를 측정할 때는 분산보다 표준편차를 주로 사용한다. 일반적으로 표준편차의 값이 클수록 산포도가 커지는데, 이는 자료가 넓게 흩어져 있음을 의미한다.

Note 표본표준편차

표본표준편차 s는 크기가 n인 하나의 표본의 자료 값들의 표준편차이다. 따라서 표본의 크기 n이 충분히 커지면 모집단의 크기와 비슷해지기 때문에 표본표준편차를 모표준편차(표준편차) 대신에 쓸 수 있게 된다. 여기서는 모표준편차를 대신하여 표본표준편차를 쓰기로 한다.

예제 1-17

다음 자료에 대하여 물음에 답하라.

10, 11, 12, 13, 12, 14, 13, 11, 13, 12, 12, 11

(a) 범위를 구하라.
(b) 사분위수 범위를 구하라.
(c) 표본분산과 표준편차를 각각 구하라.

풀이

(a) 최댓값 $x_{\max}=14$, 최솟값 $x_{\min}=10$이므로 범위 R은 다음과 같다.

$$R = x_{\max} - x_{\min} = 14 - 4 = 10$$

(b) 12개의 자료를 작은 값부터 크기순으로 나열하면 다음과 같다.

$$10,\ 11,\ 11,\ 11,\ 12,\ 12,\ 12,\ 12,\ 13,\ 13,\ 13,\ 14$$

이때 제 1 사분위수 Q_1은 $np=12\times0.25=3$이므로, 3번째로 큰 자료와 4번째로 큰 자료의 평균인 11이다. 그리고 제 3 사분위수 Q_3는 $np=12\times0.75=9$이므로, 9번째로 큰 자료와 10번째로 큰 자료의 평균인 13이다. 따라서 사분위수 범위는 다음과 같다.

$$(\text{사분위수 범위}) = Q_3 - Q_1 = 13 - 11 = 2$$

(c) 산술평균을 구하면

$$\overline{x} = \frac{1}{12}(10+11+\cdots+11) = 12$$

이므로, 표본분산과 표준편차를 구하면 다음과 같다.

$$s^2 = \frac{1}{12-1}\{(10-12)^2+(11-12)^2+\cdots+(11-12)^2\} = \frac{14}{11} \fallingdotseq 1.273$$

$$s \fallingdotseq \sqrt{1.273} \fallingdotseq 1.128$$

변동계수

변량 X의 산술평균을 $\overline{x}$, 표준편차를 s라 할 때 X의 **변동계수**coefficient of variation를 CV라 하고, 다음과 같이 정의한다.

$$CV = \frac{s}{\overline{x}} \times 100\,(\%)$$

변동계수는 산술평균에 대한 표준편차의 상대적 크기를 나타내는 척도로서, 여러 다른 종류의 통계 집단이나 동종의 집단일지라도 평균이 크게 다를 때 산포를 비교하기 위해 쓰인다. 변동계수가 크다는 것은 변동 폭이 크다는 것을 의미하며, 변동계수는 보통 백분율로 나타낸다. 변동계수의 제곱은 **상대분산**relative variance이라 한다.

예제 1-18

고소득층과 저소득층의 하루 일당에 대한 변동계수를 구하고, 상대적으로 두 자료 집단의 흩어진 정도를 분석하라.

(단위 : 천 원)

저소득층	11.5	12.2	12.0	12.4	13.6	10.5
고소득층	171	164	167	156	159	164

풀이

저소득층과 고소득층의 산술평균을 각각 $\overline{x_l}$, $\overline{x_h}$, 표준편차를 각각 s_l, s_h라 하자. 두 집단의 산술평균을 구하면 다음과 같다.

$$\overline{x_l} = \frac{11.5+12.2+12+12.4+13.6+10.5}{6} \fallingdotseq 12$$

$$\overline{x_h} = \frac{171+164+167+156+159+164}{6} = 163.5$$

그리고 저소득층과 고소득층의 분산과 표준편차를 구하면

$$s_l^2 = \frac{1}{5}\sum_{i=1}^{n}(x_i-12)^2 = 1.052,\quad s_l = \sqrt{1.052} \fallingdotseq 1.026$$

$$s_h^2 = \frac{1}{5}\sum_{i=1}^{n}(x_i-163.5)^2 = 29.1,\quad s_h = \sqrt{29.1} \fallingdotseq 5.39$$

이므로, 저소득층과 고소득층의 변동계수는 다음과 같다.

$$CV_l = \frac{s_l}{\overline{x_l}} \times 100 = \frac{1.026}{12} \times 100 = 8.55(\%)$$

$$CV_h = \frac{s_h}{\overline{x_h}} \times 100 = \frac{5.39}{163.5} \times 100 = 3.3(\%)$$

두 집단의 소득 분포를 표준편차로 비교하면 $s_l < s_h$이므로 고소득층의 소득이 저소득층의 소득보다 더 넓게 분포한다고 할 수 있다. 그러나 이는 고소득층과 저소득층의 절대적인 비교이다. 따라서 고소득층과 저소득층의 상대적인 비교를 위해 변동계수 척도로 비교하면 $CV_h < CV_l$이므로 고소득층의 소득이 저소득층 소득보다 변동 폭이 적다고 할 수 있다.

5점 요약 표시

주어진 자료의 중앙값 Me, 제 1 사분위수 Q_1, 제 3 사분위수 Q_3, 최댓값 $x_{\max}$, 최솟값 $x_{\min}$을 구하여, 다음과 같이 5개의 통계량 조합으로 나타내는 방법을 **5점 요약 표시** 5-number summary라 하며, 줄기-잎 그림과 더불어 자료를 요약할 때 유용하게 사용된다.

$$[x_{\min},\ Q_1,\ Me,\ Q_3,\ x_{\max}]$$

예제 1-19

다음 자료를 5점 요약 표시로 나타내라.

60, 64, 72, 80, 92, 64, 68, 72, 76, 80, 84, 84, 76, 88, 88, 92, 96, 88, 92, 76

풀이

20개의 자료를 작은 값부터 크기순으로 나열하면 다음과 같다.

$$60,\ 64,\ 64,\ 68,\ 72,\ 72,\ 76, 76,\ 76,\ 80,\ 80,\ 84,\ 84,\ 88, 88,\ 88,\ 92,\ 92,\ 92,\ 96$$

최댓값과 최솟값은 각각 $x_{\max} = 96$, $x_{\min} = 60$이고, 중앙값은 자료의 수 20이 짝수이므로 $\frac{20}{2}$ 번째 값과 $\frac{21}{2}$ 번째 값의 평균, 즉 10번째 값과 11번째 값의 평균이다.

$$Me = \frac{1}{2}(80 + 80) = 80$$

또한 제 1 사분위수 Q_1은 $np = 20 \times 0.25 = 5$이므로, 5번째로 큰 자료와 6번째 큰 자료의 평균인 72이다. 제 3 사분위수 Q_3는 $np = 20 \times 0.75 = 15$이므로, 15번째로 큰 자료와 16번째로 큰 자료의 평균인 88이다.

따라서 5점 요약 표시를 나타내면 다음과 같다.

$$[x_{\min},\ Q_1,\ Me,\ Q_3,\ x_{\max}] = [60, 72, 80, 88, 96]$$

왜도와 첨도

변량 X의 산술평균 $\overline{x}$나 표준편차 s의 값만으로는 X의 분포의 형태를 완전하게 결정할 수 없다. 왜냐하면 평균이 같고 표준편차가 같아도 분포의 형태가 현저하게 다를 수 있기 때문이다. 이러한 부분을 보완하기 위해 다음과 같이 왜도와 첨도를 사용한다.

■ 왜도

분포의 대칭이나 비대칭의 정도를 표시하는 척도를 **왜도**skewness(비대칭도)라고 한다. 변량 X의 분포에 대한 왜도는 α로 나타내며, 다음과 같이 정의한다.

$$\alpha = \frac{\sum_{i=1}^{n}[(x_i - \overline{x})/s]^3}{n-1} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^3}{s^3} = \frac{\mu_3}{s^3}$$

여기서 s는 표준편차이다. 이때 α값에 따라 분포의 형태를 알 수 있으며, α의 절댓값이 클수록 비대칭의 정도가 심하다는 것을 의미한다.[5]

- $\alpha = 0$이면 대칭분포이다.[6]
- $\alpha > 0$이면 왼쪽으로 치우친 분포이다.
- $\alpha < 0$이면 오른쪽으로 치우친 분포이다.

■ 첨도

뾰족함의 정도를 나타내는 척도를 **첨도**kurtosis라 한다. 변량 X의 분포에 대한 첨도는 β로 나타내며, 다음과 같이 정의한다.

$$\beta = \frac{\sum_{i=1}^{n}[(x_i - \overline{x})/s]^4}{n-1} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^4}{s^4} = \frac{\mu_4}{s^4}$$

여기서 s는 표준편차이다. 이때 β값에 따라 분포의 형태를 알 수 있다.

- $\beta = 3$이면 뾰족한 정도가 표준정규분포와 같다.
- $\beta > 3$이면 표준정규분포[7]보다 정점이 높고 뾰족하다.
- $\beta < 3$이면 표준정규분포보다 정점이 낮고 완만하다.

---

5 $n-1$ 대신 n을 사용하는 경우도 있으나 여기서는 $n-1$을 사용하기로 한다.
6 분포가 평균에 의해서 대칭인 분포로서, 4장에서 배울 정규분포와 $t-$ 분포가 이에 해당된다.
7 평균이 0, 분산이 1인 정규분포를 표준정규분포라 한다. 이에 대해서는 4장에서 자세하게 다룬다.

예제 1-20

다음 자료에 대하여 왜도와 첨도를 각각 구하고, 이를 통해 자료의 분포 형태를 파악하라.

1, 3, 2, 0, 1, 1, 2, 3, 2, 4, 3

풀이

왜도를 구하기 위해 먼저 산술평균 $\overline{x}$와 분산 s^2을 구하면

$$\overline{x} = \frac{1}{11}(1+3+2+\cdots+4+3) = 2$$

$$s^2 = \frac{1}{10}\{(1-2)^2+(3-2)^2+\cdots+(3-2)^2\} = 1.4$$

이다. μ_3을 구하면

$$\mu_3 = \frac{1}{10}\{(1-2)^3+(3-2)^3+\cdots+(3-2)^3\} = 0$$

이므로, 왜도는 다음과 같다.

$$\alpha = \frac{0}{s^3} = 0$$

또한 첨도를 구하기 위해 μ_4를 구하면

$$\mu_4 = \frac{1}{10}\{(1-2)^4+(3-2)^4+\cdots+(3-2)^4\} = 3.8$$

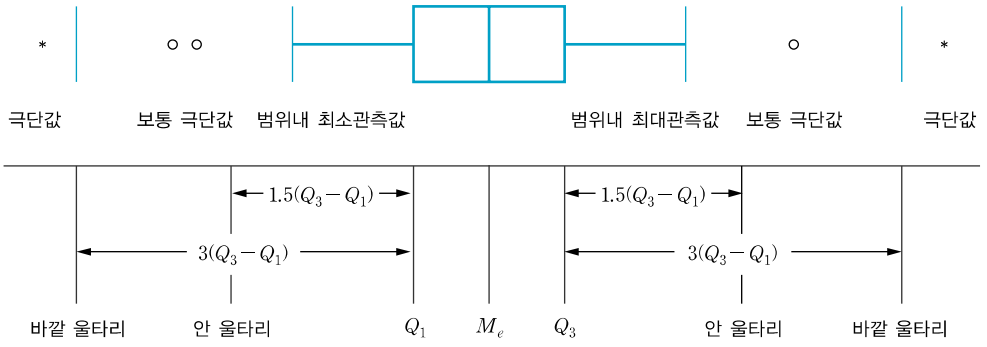
이므로, 첨도는 다음과 같다.

$$\beta = \frac{3.8}{(1.4)^2} \fallingdotseq 1.9388$$

왜도가 $\alpha = 0$이므로 분포는 대칭을 이루고 있으며, 첨도가 $\beta < 3$이므로 표준정규분포보다 정점이 낮고 완만한 형태를 이룬다.

Section
05 상자그림

줄기-잎-그림은 자료의 분포 형태나 범위, 자료의 집중 현상 등의 정보를 주는 반면, [그림 1-9]와 같은 **상자그림**box plot은 분포의 대칭성, 자료의 중심 위치, 산포도(또는 흩어진 정도), 분포의 꼬리 부분에서의 집중 정도 등을 파악하는 데 유용하다.



[그림 1-9] **상자그림**

상자그림을 작성하는 순서는 다음과 같다.

정리 1-6 상자그림의 작성 순서

❶ 사분위수의 값 Q_1, Q_3와 중앙값 Me를 결정한다.

❷ Q_1과 Q_3를 상자 형태로 연결하고, 중앙값의 위치에 선을 표시한다.

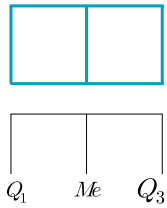
❸ 사분위수 범위 $(Q_3 - Q_1)$을 계산하고, Q_1과 Q_3로부터 각각 오른쪽, 왼쪽으로 $1.5(Q_3 - Q_1)$ 크기의 범위 내의 인접 값을 실선으로 연결하여 표시한다.

❹ 안 울타리로부터 $1.5(Q_3 - Q_1)$ 크기의 범위를 바깥 울타리로 표시하고 보통 극단값이 존재하면 ○로 표시한다.

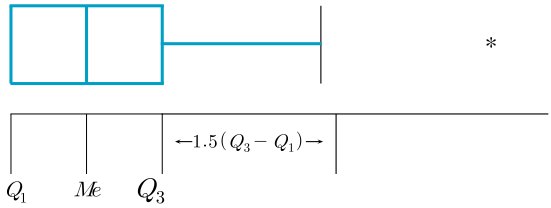
❺ 바깥 울타리 경계를 벗어난 값을 *로 표시하고, 이 점을 극단값으로 판정한다.

이제 [정리 1-6]에 따라 [그림 1-9]의 상자그림을 작성해보자.

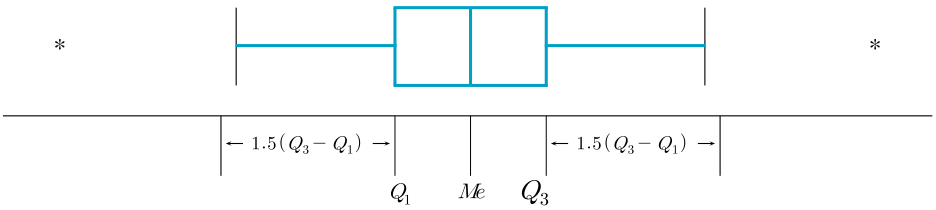
❶ 제 1 사분위수 Q_1과 제 3 사분위수 Q_3 사이를 중앙 부분(동체)이라 생각하여 상자 형태로 나타내고, 중앙값 Me의 위치에 선을 그려 넣는다.



❷ Q_3보다 큰 수는 Q_3를 나타내는 선으로부터 $Q_3 + 1.5(Q_3 - Q_1)$을 넘지 않는 최대의 수(인접 값)까지 실선으로 연결한다. 그리고 이보다 큰 수는 중심으로부터 멀리 떨어진 값(이상점)이라고 보고, 값을 *로 나타낸다.



❸ Q_1보다 작은 수도 Q_1을 나타내는 선으로부터 $Q_1 - 1.5(Q_3 - Q_1)$보다 작지 않은 최소의 수(인접 값)까지 실선으로 연결한다. 그리고 이보다 작은 수는 이상점이라고 보고, 값을 *로 나타낸다.



이때 사분위수 Q_1과 Q_3에서 $1.5(Q_3 - Q_1)$만큼 떨어져 있는 값을 **안 울타리**inner fence라고 하고, 왼쪽과 오른쪽 안 울타리는 다음과 같이 구한다.

- 왼쪽 안 울타리 : $f_l = Q_1 - 1.5(Q_3 - Q_1)$
- 오른쪽 안 울타리 : $f_u = Q_3 + 1.5(Q_3 - Q_1)$

또한 사분위수 Q_1과 Q_3에서 $3(Q_3 - Q_1)$ 만큼 떨어져 있는 값을 **바깥 울타리**outer fence 라고 하고, 왼쪽과 오른쪽 바깥 울타리는 다음과 같이 구한다.

- 왼쪽 바깥 울타리 : $f_L = Q_1 - 3(Q_3 - Q_1)$
- 오른쪽 바깥 울타리 : $f_U = Q_3 + 3(Q_3 - Q_1)$

❹ 왼쪽 안 울타리보다 작거나 오른쪽 안 울타리보다 큰 수는 각각 ○로 나타낸다.

❺ 왼쪽 바깥 울타리보다 작거나 오른쪽 바깥 울타리보다 큰 수는 각각 *로 나타낸다. 이때 안 울타리와 바깥 울타리 사이에 놓이는 값을 **보통 극단값**이라 하고, 바깥 울타리 외부에 놓이는 값을 **극단값**이라고 한다. 이와 같은 방법으로 상자그림을 완성하면 [그림 1-9]와 같다.

예제 1-21

다음 자료에 대하여 상자그림을 그려라.

83, 90, 60, 20, 50, 94, 60, 62, 97, 43, 67, 84, 79, 62, 78

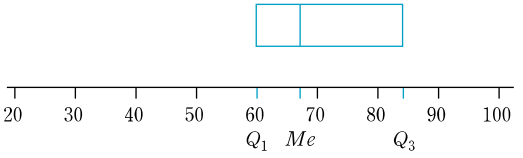
풀이

주어진 자료를 작은 값부터 크기순으로 배열하면 다음과 같다.

20, 43, 50, 60, 60, 62, 62, 67, 78, 79, 83, 84, 90, 94, 97

❶ 제 1 사분위수 Q_1은 $np = 15 \times 0.25 = 3.75$이므로, 4번째 자료인 60이다. 그리고 제 3 사분위수 Q_3는 $np = 15 \times 0.75 = 11.25$이므로, 12번째 자료인 84이다. 또한 중앙값 Me를 구하면 자료가 15개이므로 8번째 자료인 67이다.

❷ Q_1과 Q_3를 상자 형태로 연결하고, 중앙값의 위치에 선을 표시한다.

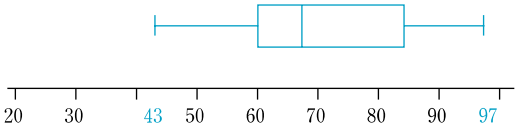


❸ 사분위수 범위를 구하면 $Q_3 - Q_1 = 84 - 60 = 24$이므로, 안 울타리를 구하면 다음과 같다.

$$f_l = Q_1 - 1.5(Q_3 - Q_1) = 60 - 36 = 24$$
$$f_u = Q_3 + 1.5(Q_3 - Q_1) = 84 + 36 = 120$$

안 울타리와 Q_1, Q_3, 인접 값을 연결한다. 이때 인접 값은 각각 43과 97이다. 인접 값까지

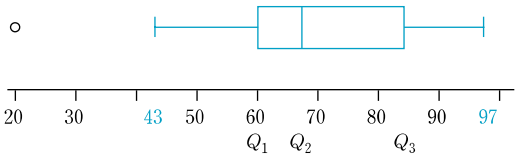
실선으로 표시한다.



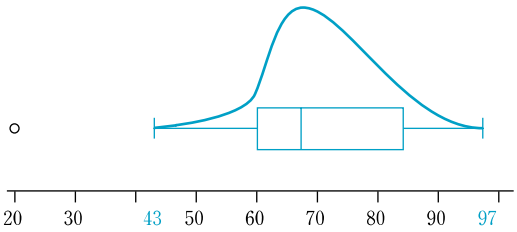
❹ 바깥 울타리를 구한다. 최솟값이 20, 최댓값이 97이므로 극단값은 없다. 따라서 이 경우에는 왼쪽, 오른쪽 바깥 울타리가 모두 존재하지 않는다.

$$f_L = Q_1 - 3\times(Q_3 - Q_1) = 60 - 72 = -12$$
$$f_U = Q_1 + 3\times(Q_3 - Q_1) = 60 + 72 = 132$$

❺ 마지막으로 보통 극단값과 극단값을 표시한다. 극단값은 없지만 보통 극단값은 20으로 하나 존재한다.



참고로 완성된 상자그림을 살펴보면 Q_1과 Q_2 사이가 Q_2와 Q_3 사이보다 좁게 나타난다. 이는 Q_1과 Q_2 사이에 자료 값이 많이 밀집되어 있음을 의미한다. 즉 자료가 좀 더 넓게 포진하고 있음을 알 수 있다. 따라서 다음 그림의 곡선 모양과 같이 주어진 자료는 중앙값을 중심으로 대칭성을 약간 벗어나는 형태임을 알 수 있다.



예제 1-22

다음은 어느 대학의 남녀 신입생의 키에 대한 자료이다. 이에 대하여 상자그림을 그리고, 각 분포를 비교하라.

(단위 : cm)

남자	181	170	179	183	178	171	177	174	163	167	163	173	178	170
	167	177	176	180	169	178	173	173	171	171	170	174	180	
여자	161	160	158	169	162	163	158	160	160	158	160	165	170	152
	160	162	167	168	166	164	158	160	160	159				

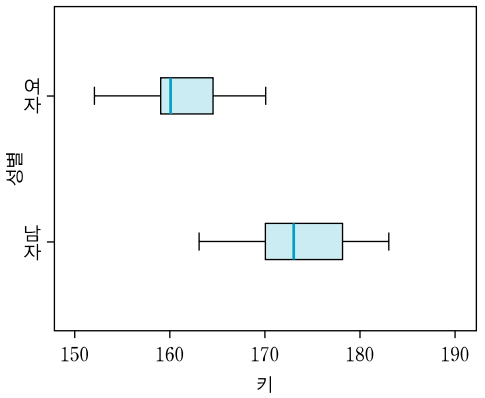
풀이

남자는 자료의 수가 27이고 $27\times0.25=6.75$, $27\times0.75=20.25$이므로, 자료를 작은 값부터 크기순으로 나열했을 때 Q_1은 7번째 자료인 170, Me는 14번째 자료인 173, Q_3는 21번째 자료인 178이다. 그러므로 사분위수 범위는 8이 되고, 안 울타리를 다음과 같이 구한다.

$$f_l = Q_1 - 1.5\times(Q_3 - Q_1) = 170 - 1.5\times 8 = 158$$
$$f_u = Q_3 + 1.5\times(Q_3 - Q_1) = 178 + 1.5\times 8 = 190$$

이때 자료의 최솟값과 최댓값이 각각 163과 183이므로 보통 극단값과 극단값이 모두 없다.

여자는 자료의 수가 24이고 $24\times0.25=6$, $24\times0.75=18$이므로, 자료를 작은 값부터 크기순으로 나열했을 때 Q_1은 6번째와 7번째 자료의 평균인 159.5, Me는 12번째와 13번째 자료의 평균인 160, Q_3는 18번째와 19번째 자료의 평균인 164.5이다. 그러므로 사분위수 범위는 5이고, $f_l=152$, $f_u=172$이다. 또한 모든 자료가 이 범위 안에 있으므로 보통 극단값과 극단값이 모두 없다. 남자와 여자의 상자그림을 그리면 다음과 같다.[8]



이 그림으로부터 남자가 여자보다 평균 10cm정도 크다는 것을 알 수 있다. 사분위수 범위는 남자가 여자보다 크지만, 전체 범위는 남자와 여자가 거의 같다는 것을 알 수 있다. 또한 남자의 키는 비교적 대칭에 가까운 반면에, 여자의 경우는 제 1 사분위수와 중앙값 사이가 좁게 몰려있는 모양으로 분포가 대칭이 아니라 중앙값을 중심으로 작은 쪽에 집중적으로 몰려 있음을 알 수 있다.

8 편의상 SPSS를 이용하여 상자그림을 그렸다.

Chapter 01 연습문제

1.1 주어진 자료에 대하여 다음을 구하라.

1, 3, 2, 2, 3, 9, 5, 5, 8, 9, 5, 8

(a) 산술평균
(b) 중앙값
(c) 최빈값
(d) 사분위수
(e) 사분위수 범위
(f) 분산과 표준편차
(g) 변동계수
(h) 왜도
(i) 첨도
(j) 5점 요약 표시

1.2 자료 $x_1, x_2, \cdots, x_n$의 산술평균을 $\overline{x}$라 할 때, 각각의 자료에 상수 a를 곱한 값 $ax_1, ax_2, \cdots, ax_n$의 산술평균을 구하라.

1.3 자료 $x_1, x_2, \cdots, x_n$에서 가평균을 U라 하자. $u_i = x_i - U$라 할 때, $u_1, u_2, \cdots, u_n$의 산술평균 $\overline{u}$를 구하라.

1.4 다음은 조선 왕조 27대 왕들의 수명을 표로 나타낸 것이다. 이 자료에 대한 줄기-잎 그림을 작성하라.

(단위 : 세)

왕	수명	왕	수명	왕	수명	왕	수명
태조	73	정종	62	태종	45	세종	53
문종	38	단종	16	세조	51	예종	28
성종	37	연산군	30	중종	56	인종	30
명종	33	선조	56	광해군	66	인조	54
효종	40	현종	33	숙종	59	경종	36
영조	82	정조	48	순조	44	헌종	22
철종	32	고종	67	순종	52	–	–

1.5 다음은 40대 여자와 10대 여자의 체중을 나타낸다. 어느 쪽의 산포도가 더 큰지 구하라.

여자	평균체중	표준편차
40대	51	5
10대	26.5	3.7

1.6 다음에 주어진 자료에 대하여 물음에 따라 상자그림을 그려라.

3.5	81.5	27.6	33.2	12.0	20.5	19.0	21.2	22.8	20.5
21.7	24.4	6.4	17.3	22.1	22.2	22.3	24.7	22.7	32.6
15.8	21.9	21.4	26.8	22.3	26.1	22.0	24.7	21.6	22.1

(a) 사분위수 Q_1, Q_2, Q_3를 구하라.
(b) 사분위수 범위를 구하라.
(c) 안 울타리 f_l, f_u를 구하라.
(d) 극단값이 있으면 그 자료 값을 구하고, 극단값의 위치를 ○ 또는 * 로 나타내라.
(e) 상자그림을 작성하라.

1.7 다음 임의의 자료에 대하여 물음에 답하라.

22	19	27	22	27	11	22	48	24	19
15	18	36	33	32	21	37	16	33	16
24	41	39	17	28	22	21	33	17	18

(a) 줄기-잎-그림을 작성하라.
(b) 줄기-잎 그림에 대응하는 상자그림을 작성하라.
(c) 평균과 중앙값과 최빈값을 각각 구하라.
(d) 사분위수를 구하고, 제20, 80 백분위수를 구하라.

1.8 다음 자료는 어느 대학교 2학년 학생 40명의 확률통계학 점수이다.

(단위 : 점)

91	85	64	45	92	82	95	89	83	78
67	67	15	79	67	85	79	76	82	57
55	99	68	72	79	80	64	76	68	81
66	81	91	64	73	74	86	67	62	97

(a) 계급의 수가 5인 도수분포표를 작성하라.
(b) 계급간격이 5인 줄기-잎 그림을 그려라.
(c) 상자그림을 그려라.
(d) 평균과 분산을 각각 구하라.

1.9 다음 표는 1992년부터 2000년까지의 서울과 경기 간 인구 이동 추이를 나타낸다. 서울에서 경기로 이주한 사람과 경기에서 서울로 이주한 사람 수에 대한 도수분포다각형을 작성하고, 순이동에 대한 원그래프를 작성하라.

(단위 : 천 명)

연도	서울 → 경기(A)	경기 → 서울(B)	순이동(A−B)
1992	428	254	174
1994	522	259	283
1996	521	291	230
1998	407	277	130
2000	436	313	123

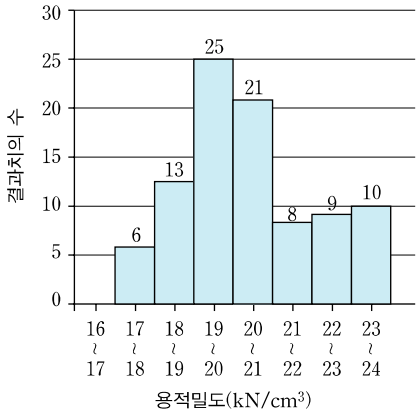
1.10 오른쪽 표는 30년 동안 연간 누적강수량을 나타낸 도수분포표이다.

(a) 연간 누적강수량의 사분위수는 각각 어떤 구간에 있는지 구하라.

(b) 연간 누적강수량의 평균과 표준편차를 각각 구하라.

계급(mm)	도수
38 ~ 42	3
42 ~ 46	7
46 ~ 50	5
50 ~ 54	6
54 ~ 58	3
58 ~ 62	3
62 ~ 66	1
66 ~ 70	2

1.11 다음은 잔류 흙의 용적밀도를 나타내는 히스토그램이다. 이때 용적밀도의 평균, 표준편차를 각각 구하라.



1.12 컴퓨터 실습 다음은 29년 동안의 강우강도를 기록한 표이다. Excel을 이용하여 이 자료에 대한 도수분포표, 히스토그램, 도수분포다각형을 그려라.

연도	강우강도(in)	연도	강우강도(in)	연도	강우강도(in)
1	43.30	11	54.49	21	58.71
2	53.02	12	47.38	22	42.96
3	63.52	13	40.78	23	55.77
4	45.98	14	45.05	24	41.31
5	48.26	15	50.37	25	58.83
6	50.51	16	54.91	26	48.21
7	49.57	17	51.28	27	44.67
8	43.93	18	39.91	28	67.72
9	46.77	19	53.29	29	43.11
10	59.12	20	67.59	–	–

1.13 컴퓨터 실습 Excel을 이용하여 [예제 1-11]의 50명의 통계학 성적에 대한 평균, 분산, 표준편차, 첨도, 왜도를 구하라. 그리고 분포 형태를 파악하라.

1.14 컴퓨터 실습 다음은 어느 정유회사에서 생산하는 자동차용 휘발유의 옥탄가 자료이다. Excel을 이용하여 평균, 분산, 표준편차, 첨도, 왜도를 구하라. 그리고 분포 형태를 파악하라.

90.7	90.0	92.2	91.0	88.5	87.8	83.4	82.1	88.6	100.2
95.6	93.3	88.2	91.0	92.7	93.2	91.0	93.4	85.3	88.6
96.1	98.9	89.9	89.8	91.1	89.7	88.2	93.7	84.3	97.9
87.9	90.1	88.3	93.3	95.4	91.6	88.9	92.6	97.4	87.4
86.7	90.4	91.1	92.6	88.8	89.3	89.8	89.2	88.6	89.0
96.1	95.6	92.2							

1.15 컴퓨터 실습 어느 전자회사의 서비스센터에서는 고객 만족도를 조사하여 서비스 개선을 위해 다음과 같은 5점 척도로 설문조사를 실시하였다.

① 매우 만족한다. ② 만족한다. ③ 보통이다.
④ 불만족이다. ⑤ 매우 불만족이다.

응답자의 응답 결과가 다음과 같을 때, Excel을 이용하여 위 설문에 대한 도수분포표를 작성하고, 원그래프와 막대그래프를 그려라.

2	2	3	2	3	4	1	2	3	2
1	1	3	3	3	4	5	4	1	3
2	2	3	3	2	2	5	1	3	3
2	1	4	3	4	2	4	3	2	3
1	4	4	3	3	3	2	3	1	3

1.16 컴퓨터 실습 다음 자료는 어느 공업도시의 공기 오염도를 측정하기 위해 대기 중의 오존농도를 측정한 것이다.

3.5	9.4	6.6	1.4	6.0	5.7	3.8	5.4	3.5	5.6
7.4	8.0	11.2	4.7	4.2	6.0	4.1	5.8	1.7	1.6
2.0	3.7	3.1	3.8	4.4	3.9	2.5	3.0	3.4	1.0
6.5	3.5	7.7	4.7	3.5	3.9	8.8	8.2	2.5	4.6
5.7	9.7	1.6	2.5	4.0	4.4	4.8	5.1	5.9	2.7

(a) Excel을 이용하여 히스토그램과 도수분포다각형을 그려라.
(b) Excel을 이용하여 자료의 평균, 중앙값, 표준편차, 왜도, 첨도를 구하라.