세상은 지금 정보기술(IT)의 시대에서 데이터기술(DT)의 시대로 가고 있다.

빅데이터(Big data) 시대

빅데이터 : 기존의 관리 및 분석 체계로는 감당할 수 없을 정도의 거대한 데이터의 집합

빅데이터의 특성 :

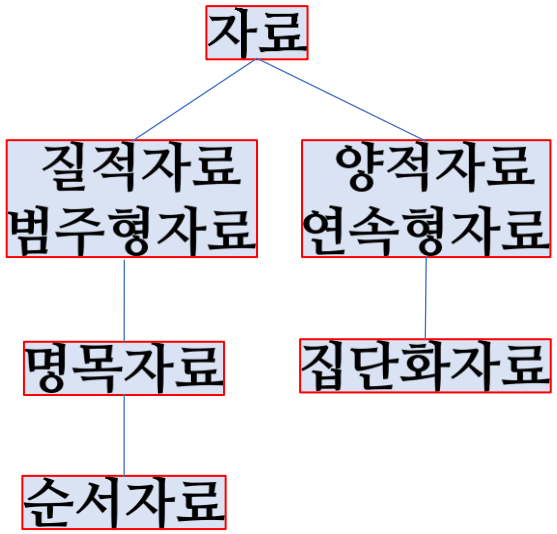
(1) 데이터의 규모(Volume) (2) 데이터 종류의 다양성(Variety) (3) 속도(Velocity)

빅데이터 기술 : 서로 다른 엄청난 양의 데이터가 실시간으로 발생될 때 무질서한 것처럼 보이는 데이터 속에서 특정 또는 일정한 패턴을 찾아내는 기술

통계학(Statistics) : 데이터에서 의미를 찾아내는 방법을 다루는 학문이다. 통계학은 수집한 자료(data)를 이용하여 연구가설(hypothesis)의 참 거짓을 판정하는 수학적 또는 확률적 논리를 제공한다.

빅데이터 기술의 기본은 통계학이다.

자료의 정리



통계실험에 의하여 얻은 자료는 숫자에 의하여 표현되며, 그 숫자가 의미를 갖는가 하면 그렇지 못한 경우도 있다. 예를 들어, 피부색이나 혈액형 또는 지역명과 같은 자료는 숫자에 의하여 표현되지 않으며, 이러한 자료를 **질적자료(qualititative data)** 또는 **범주형자료(categorical data)**라 한다. 또한 키, 몸무게 또는 강수량과 같이 자료가 숫자로 표현되며, 그 숫자가 의미를 갖는 자료를 **양적자료(quantitative data)**라 한다. 다시 말해서, 대소관계 또는 크기관계 등에 의하여 구별되는 자료를 양적자료라 한다.

한편, 각 지역에 우편번호를 부여하거나 지역별 전화번호를 부여함으로써 지역이라는 범주를 숫자로 대처할 수 있으나, 이 경우에는 숫자 고유의 특성을 갖지 못한다. 이와 같이 숫자 그 자체로는 아무런 의미가 없고 단지 범주를 사용하기 편하도록 숫자로 대치한 자료를 **명목자료(nominal data)**라 한다. 그러나 명목자료에는 각급 학교에 숫자를 부여하여 초등학교는 1, 중학교는 2, 고등학교는 3 그리고 대학 이상은 4라는 숫자로 표현할 수 있는 경우가 있으며, 이 경우에 부여된 숫자는 순서의 개념을 갖는다. 이와 같이 순서의 개념을 갖는 질적자료를 **순서자료(ordinal data)**라 한다. 또한 양적자료인 시험성적을 90점 이상 A, 80-89는 B, 70-79는 C, 60-69는 D 그리고 59점 이하는 F라는 범주로 묶어서 나타낼 수 있으며, 이러한 자료를 **집단화자료(grouped data)**라 한다.

통계분석을 위해 자료(data)를 수집하려고 한다. 무턱대고 자료를 수집하면 시간과 돈을 낭비하고 낭패를 보는 수가 있다. 제대로된 자료를 얻기 위해서는 **개체(Item)**, **요인(Factor)**, 그리고 **변수(Variable)**에 대한 개념을 이해해야 합니다.

개체(Item)는 연구자 또는 관찰자가 관심을 갖는 대상, 연구 대상, 알고 싶은 대상
요인(Factor)은 개체에 관한 특성 중 연구자가 특별히 관심을 갖는 특성
변수(Variable)는 요인을 구성하고 있는 요소

통계 데이터를 제대로 다루기 위해서는 **척도(Scale)**의 의미와 처리 방식에 대해서 숙지해야 한다. 왜냐하면, **척도의 종류에 따라서 데이터 처리 방식이 서로 다르기 때문이다.** 척도는 어떠한 대상의 특성을 "단위"를 사용하여 정량화한 것을 말한다. 그 종류는 명목척도, 순위척도, 등간척도, 비율척도 이렇게 네 가지로 구분한다.

명목척도(nominal scale): 이름 또는 범주를 나타내는 척도/ 숫자로 표현될 수 있지만 수량적인 의미를 갖지 않고, 범주(카테고리)를 구분하는 용도로 쓰이는 척도

순위척도(ordinal scale): 관찰대상이 지니는 속성에 따라 순위를 결정하는 척도

등간척도(interval scale): 속성의 차이를 양적인 차이로 측정하기 위하여 척도간 간격을 균일하게 분할하여 측정하는 척도/ 절대영점이(absolute zero) 없다.

비율척도(ratio scale): 절대영점(absolute zero)이 있는 등간 척도

독립변수(independent variable)는 연구자가 의도적으로 변화시키는 변수/ 연구자가 마음대로 조정할 수 있는 변수/ 설명변수(explanatory variable), 예측변수(predictor variable), 위험인자(risk factor)라고 부르기도 한다.

종속변수(dependent variable)는 연구자가 독립변수의 변화에 따라 어떻게 변하는지 알고 싶어하는 변수/ 반응변수(response variable), 결과변수(outcome variable), 표적변수(target variable)라고 부르기도 한다.

독립변수와 종속변수는 인과 관계를 가지고 있다.

통계는 데이터를 다루는 목적에 따라 크게 두 가지로 구분할 수 있다.

기술 통계(descriptive statistics): 수집한 데이터를 요약 묘사 설명하는 통계 기법/ 기술 통계 기법은 크게 또 두 가지로 구분할 수 있다. 하나는 데이터의 **집중화 경향(central tendency)**에 대한 기법으로 우리가 수집한 데이터를 대표하는 값이 무엇인지 또는 어떤 값에 집중되어 있는지를 다루는 기법이다. 평균(mean), 중앙값(median), 최빈값(mode) 등이 바로 집중화 경향에 속하는 것들이다. 다른 하나는 우리가 수집한 데이터가 어떻게 퍼져 있는지를 설명하는 기법이 있다. 이를 **분산도(Variation)**라고 부른다. 분산도는 말그대로 데이터가 전반적으로 어떻게 분포되어 있는지 즉, 뭉쳐 있는지 퍼져 있는지를 설명하는 방법이다. 대표적으로 표준편차(standard deviation), 사분위(quartile) 값 등이 있다.

추리 통계(inferential statistics): 수집한 데이터를 바탕으로 추론 예측하는 통계 기법

(1) 점도표(dot plot) : 질적자료/양적자료 모두 사용, 원자료의 특성을 그림으로 나타내는 가장 간단한 방법. 수평축에 각 범주 또는 자료의 측정값을 기입하고, 이 수평축 위에 각 범주 또는 측정값의 관찰 횟수를 점으로 나타낸다. 자료의 정확한 위치를 알 수 있으며, 수집한 자료가 어떠한 모양으로 흩어져 있는지 쉽게 파악할 수 있다. 그러나 자료의 수가 매우 많은 경우에는 부적당하다.

(2) 도수분포표(frequency table) : 각 범주와 그에 대응하는 도수(frequency) 그리고 상대도수(relative frequency) 등을 나열한 도표. 각 범주의 도수와 상대적인 비율을 쉽게 비교할 수 있다.

(3) 막대그래프(bar chart) : 질적자료의 각 범주를 수평축에 나타내고, 각 범주에 대응하는 도수 또는 상대도수 등을 같은 폭의 수직막대로 나타낸 그림. 도수분포표에 비하여 각 범두의; 도수 또는 상대도수를 시각적으로 쉽게 비교할 수 있다. 특히 범주의 도수가 감소하도록 범주를 재배열한 막대그래프를 **파레토 그림(Pareto chart)**이라 한다.

(4) 도수 다각형(frequency polygon) : 각 범주에 대한 막대그래프의 상단 중심부를 사선으로 연결하여 각 범주를 비교하는 그림.

(5) 원그래프(pie chart) : 질적자료의 각 범주를 상대적으로 비교할 때 많이 사용하며, 각 범주의 상대도수에 비례하는 중심각을 갖는 파이조각 모양으로 나누어진 원으로 작성한 그림.

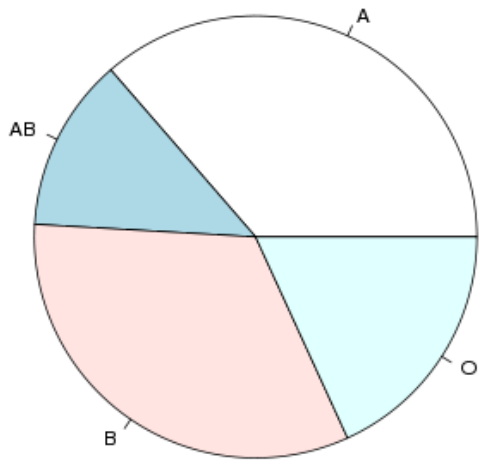
예제 1 : 어느 대학에서 통계학 수업을 수강하는 55명의 학생을 대상으로 혈액형을 조사한 결과는 다음과 같다. 이 자료를 **도수분포표**, **원그래프**, **막대그래프**로 요약하여라.

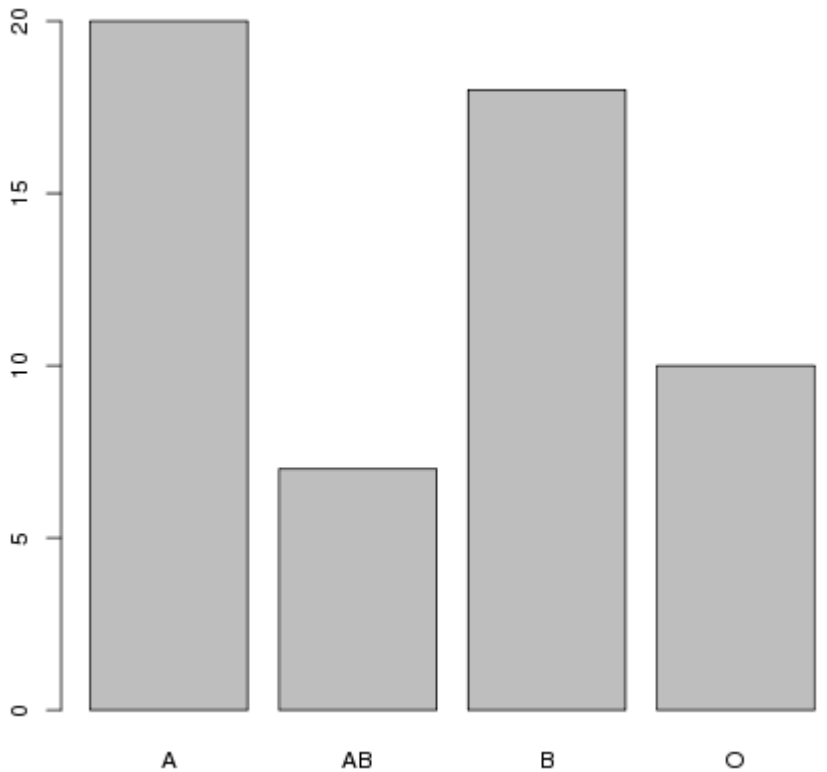
B	A	B	A	A	B	O	A	A	A	O
B	AB	B	AB	AB	A	A	O	AB	O	A
B	O	B	B	A	A	O	A	A	AB	B
B	O	B	B	B	A	AB	A	A	B	O
B	B	O	B	O	B	A	A	AB	A	A

```
blood = c("B","A","B","A","A","B","O","A","A","A","O","B","AB","B","AB",
"AB","A","A","O","AB","O","A","B","O","B","B","A","A","O","A",
"A","AB","B","B","O","B","B","B","A","AB","A","A","B","O","B",
"B","O","B","O","B","A","A","AB","A","A")
cnt = table(blood)
prop = prop.table(cnt)
cbind(cnt,prop)        # 도수분포표
pie(cnt)        # 원도표
barplot(cnt)    # 막대도표
dev.off()
```

```
null device
          1
```

```
   cnt      prop
A   20 0.3636364
AB   7 0.1272727
B   18 0.3272727
O   10 0.1818182
```





(6) 집단화 자료의 도수분포표 : 양적자료를 적당한 크기로 집단화하여 도수분포표를 만들면 전체 자료가 갖는 특성을 좀 더 쉽게 이해할 수 있으며, 이러한 방법에 의하여 자료를 정리하는 것을 집단화 자료의 도수분포표라 한다. 이때 적당한 간격으로 집단화하여 나타낸 범주들을 **계급(class)**이라 하며, 각 계급은 중복을 피하여 한 측정값이 두 계급에 동시에 포함되지 않도록 하며, 이웃하는 두 계급의 위쪽 경계에서 아래쪽 경계를 뺀 값을 **계급간격(class width)**이라 한다. 그러면 수집한 양적자료에 대한 도수분포표를 작성하기 위하여 다음과 같은 방법을 사용한다.

(a) 계급의 수(k)를 결정한다. 자료의 수가 n일 때, **Sturges 공식**이라 불리는 $k = 1 + 3.3 \log_{10} n$에 가까운 정수를 택한다.

자료의 수		30	50	120	250	500	1000
계급수	Sturges 방법	6	7	8	9	10	11

(b) 계급의 간격(w)을 적당히 구한다.

$$w = \frac{\text{자료의 최대 관찰값} - \text{자료의 최소 관찰값}}{k}$$

(c) 제1계급의 하한을 결정한다. 이웃하는 계급간의 중복을 피하기 위하여 제1계급의 하한으로 다음을 택한다.

$$\text{최소 관찰값} - \frac{\text{최소단위}}{2}$$

예제 2 : 머리의 직경이 50mm인 볼트를 제조하는 회사로부터 100개의 볼트를 임의로 수집하여 측정한 결과는 다음과 같다고 하자. 집단화 자료에 대한 **도수분포표**를 완성하여라.

49.6	50.5	49.9	51.6	49.6	48.7	49.7	49.1	48.7	51
50.1	48.7	50.4	50.6	51.5	49.4	51.1	49.8	49.8	49
47.2	50.4	49.1	50.5	50.9	49.8	49.6	49.3	50.5	50.2
52	50.7	50.4	48.6	50.9	51.2	50.7	48.5	50	51.3
47.6	49.1	51	51.9	49.5	49.7	48.6	49.7	48.5	48.3
50.5	48.7	50.5	49.1	50.4	51.2	50.4	49.9	50	50.4
50.7	49.3	50.8	49.8	48.9	49	49.5	49.9	49.7	51.3
51	49.5	49.9	49.6	50.5	50.3	48.9	49.2	51.2	48
49.8	49.1	48.8	51.7	49.7	50.3	50.6	50	49.6	51.2
47.6	50.8	49.7	49.9	50.6	49.7	49.9	49.7	51.8	55.1

최소 단위 : 0.1

계급수 : $k = 1 + 3.3\log_{10} 100 = 7.6 \approx 8$

계급 간격 : $\frac{55.1-47.2}{8} = 0.9875 \approx 1$

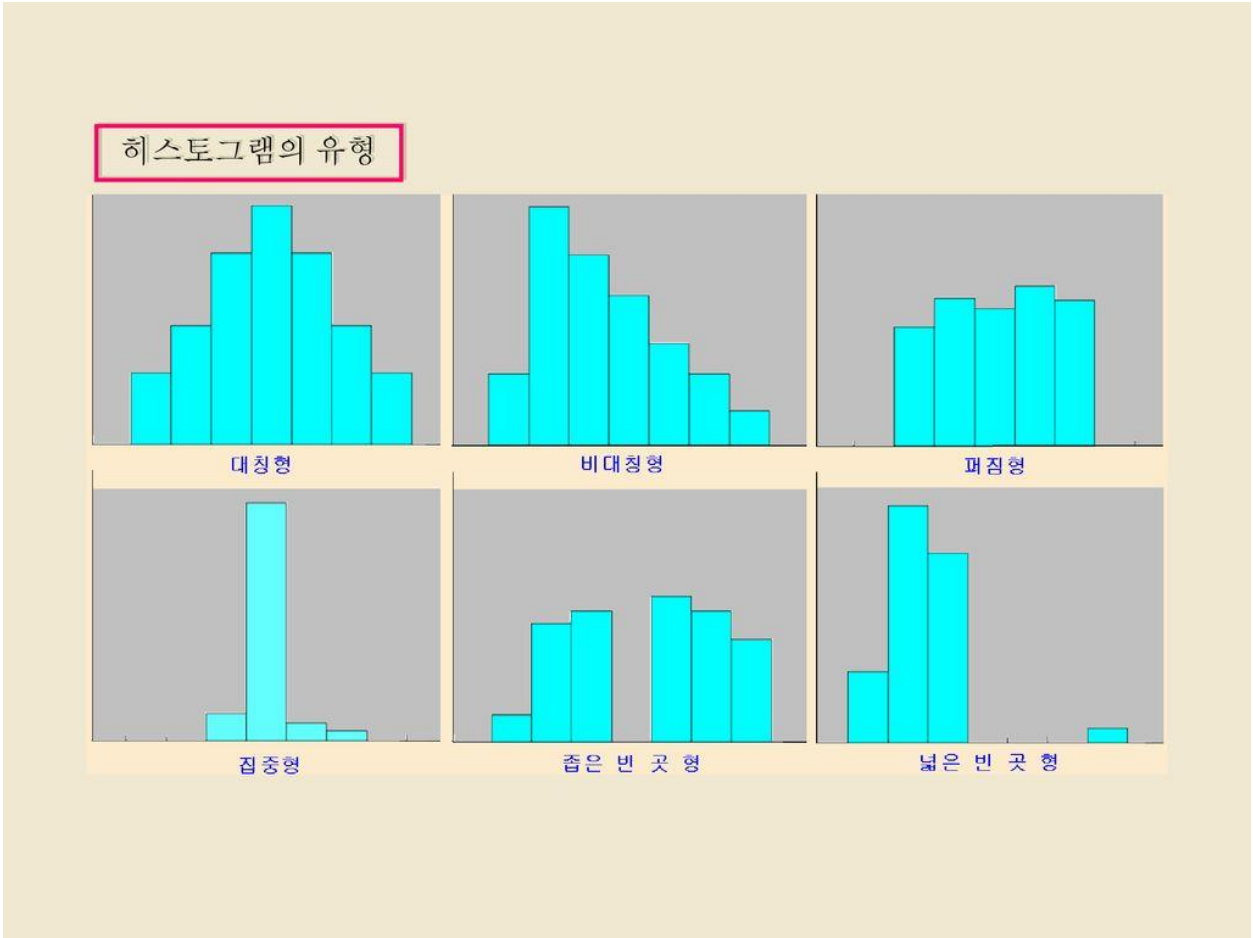
제1계급의 하한 : $47.2 - \frac{0.1}{2} = 47.15$

계급값= $\frac{\text{위쪽 경계}+\text{아래쪽 경계}}{2}$

계급	계급간격	도수	상대도수	누적도수	누적상대도수	계급값
제1계급	47.15~48.15	4	0.04	4	0.04	47.65
제2계급	48.15~49.15	18	0.18	22	0.22	48.65
제3계급	49.15~50.15	36	0.36	58	0.58	49.65
제4계급	50.15~51.15	29	0.29	87	0.87	50.65
제5계급	51.15~52.15	12	0.12	99	0.99	51.65
제6계급	52.15~53.15	0	0.00	99	0.99	52.65
제7계급	53.15~54.15	0	0.00	99	0.99	53.65
제8계급	54.15~55.15	1	0.01	100	1.00	54.65
합 계		100	1.00	100	1.00	

위 표로부터 전체 자료를 크기순으로 나열하여 가장 가운데 놓이는 자료값을 나타내는 누적상대도수가 0.5인 위가 대략적으로 제3계급의 끝부분에 있다는 사실과 전체 자료의 흩어진 정도를 알 수 있다.
55.1은 대다수의 자료로부터 멀리떨어져 있는 측정값이다.
이러한 자료를 **이상점(outlier)**이라 한다.
위 표로부터는 원자료의 정확한 측정값을 알 수 없다.

(7) 히스토그램(histogram) : 집단화 자료에 대한 도수분포표의 계급간격을 수평축에 작성하고, 수직축에 도수(누적도수) 또는 상대도수(누적상대도수)에 해당하는 막대모양으로 작성한 그림. 도수분포표에 비하여 보다 더 시각적으로 중심의 위치와 자료가 어떠한 모양으로 흩어져 있는가에 대하여 쉬게 파악할 수 있으며, 자료의 분포가 대칭성을 갖는 경우와 넓게 흩어지거나 집중되는 경우 그리고 중간에 빈 곳이 있는 정도 등을 쉽게 알 수 있다.



(8) 도수분포다각형(frequency polygon) : 양적자료에 대하여 시각적인 효과를 주는 또 다른 방법으로, 히스토그램의 연속적인 막대의 상단중심부를 직선으로 연결하여 다각형으로 표현한 그림.

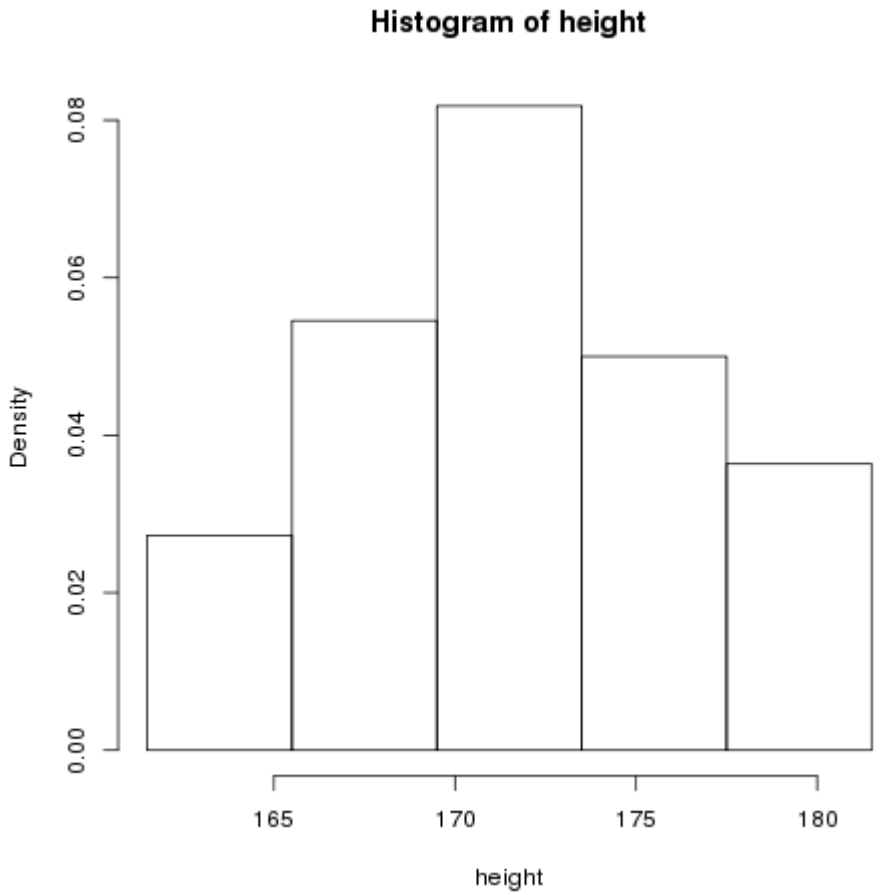
예제 2: 다음의 자료는 어느 대학에서 임의로 선정한 남학생 55명의 키를 기록한 것으로 단위는 센티미터(cm)이다. 이 자료에 대한 **도수분포표**와 **히스토그램**을 그려라.

170	178	171	168	173	178	171	174	170	170	175
170	169	166	162	170	171	175	175	171	171	170
172	179	164	170	181	178	180	177	166	169	168
165	163	175	166	178	165	168	167	177	168	177
174	174	176	179	169	173	167	170	173	170	162

```
height = c(170,178,171,168,173,178,171,174,170,170,175,
170,169,166,162,170,171,175,175,171,171,170,
172,179,164,170,181,178,180,177,166,169,168,
165,163,175,166,178,165,168,167,177,168,177,
174,174,176,179,169,173,167,170,173,170,162)
boundaries = seq(161.5, 181.5, by=4)          # 계급 구간
cnt=table(cut(height, boundaries))            # 도수분포표
prop=prop.table(cnt)
cbind(Freq=cnt, relative=prop)                # 상대 도수분포표
hist(height,breaks=boundaries,probability=T) # 히스토그램
dev.off()
```

```
          Freq  relative
(162,166]    6 0.1090909
(166,170]   12 0.2181818
(170,174]   18 0.3272727
(174,178]   11 0.2000000
(178,182]    8 0.1454545

null device
          1
```



(9) 줄기-잎 그림(stem-leaf-display) : 히스토그램 또는 도수분포다각형 등은 수집한 자료에 대한 중심의 위치와 흩어진 모양을 대략적으로 제공하지만, 각 계급의 자료값에 대한 정확한 정보는 제공하지 못한다. 이러한 단점을 보완하기 위하여 고안된 그림으로 줄기-잎 그림이 있으며, 이 그림을 도수분포표나 히스토그램이 갖고 있는 성질을 그대로 보존하면서 각 계급 안에 들어있는 개개의 측정값을 제공한다는 장점이 있다. 그림을 작성하기 위하여

(a) 줄기와 잎을 구분한다. 이때, 변동이 작은 부분을 줄기, 변동이 많은 부분을 잎으로 지정한다.

(b) 줄기 부분을 작은 수부터 순차적으로 나열하고, 잎 부분을 원자료의 관찰 순서대로 나열한다.

(c) 이제 잎 부분의 관찰값을 순서대로 나열하고 전체 자료의 중앙에 놓이는 관찰값이 있는 행의 맨 앞에 괄호( )를 만들고, 괄호 안에 그 행의 잎의 수(도수)를 기입한다.

(d) 괄호가 있는 행을 중심으로 괄호와 동일한 열에 누적도수를 위와 아래방향에서 각각 기입하고, 최소단위와 자료의 전체 개수를 기입한다.

예제 3 : 머리의 직경이 50mm인 볼트를 제조하는 회사로부터 100개의 볼트를 임의로 수집하여 측정한 결과는 다음과 같다고 하자. 집단화 자료에 대한 **줄기-잎 그림**을 완성하여라.

49.6	50.5	49.9	51.6	49.6	48.7	49.7	49.1	48.7	51
50.1	48.7	50.4	50.6	51.5	49.4	51.1	49.8	49.8	49
47.2	50.4	49.1	50.5	50.9	49.8	49.6	49.3	50.5	50.2
52	50.7	50.4	48.6	50.9	51.2	50.7	48.5	50	51.3
47.6	49.1	51	51.9	49.5	49.7	48.6	49.7	48.5	48.3
50.5	48.7	50.5	49.1	50.4	51.2	50.4	49.9	50	50.4
50.7	49.3	50.8	49.8	48.9	49	49.5	49.9	49.7	51.3
51	49.5	49.9	49.6	50.5	50.3	48.9	49.2	51.2	48
49.8	49.1	48.8	51.7	49.7	50.3	50.6	50	49.6	51.2
47.6	50.8	49.7	49.9	50.6	49.7	49.9	49.7	51.8	55.1

줄기 : 변동이 많은 부분은 정수 부분)

잎 : 변동이 많은 부분(소수 부분)

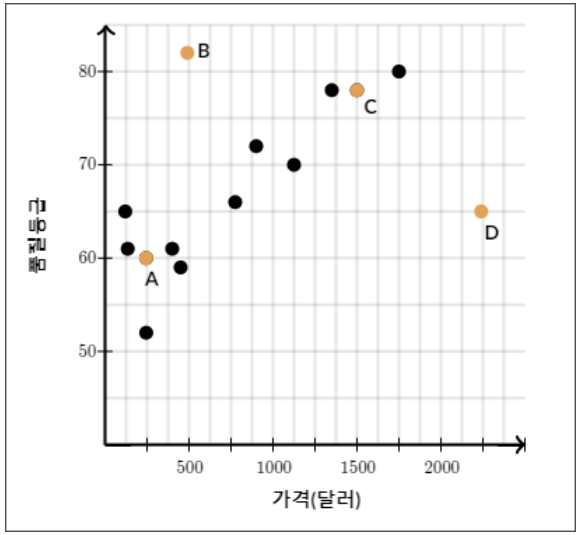
최소단위 : 0.1

누적도수 줄기부분

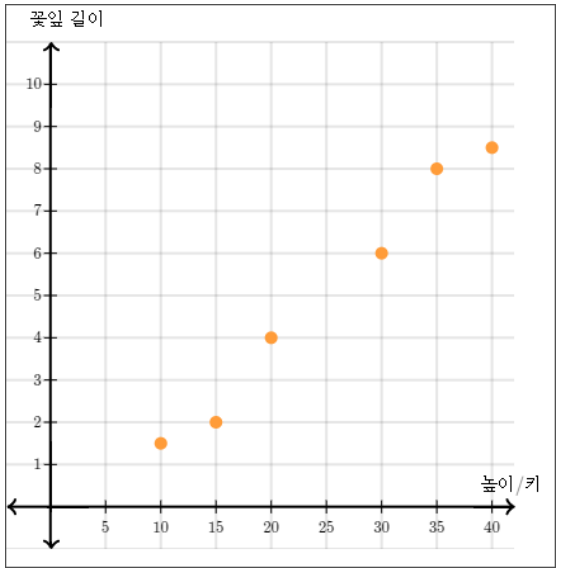
3	47	266
16	48	035566777899
(38)	49	00111112334555666667777777788888999999
46	50	00012334444445555556667778899
17	51	00012222335678 9
2	52	0
1	53	
1	54	
1	55	1

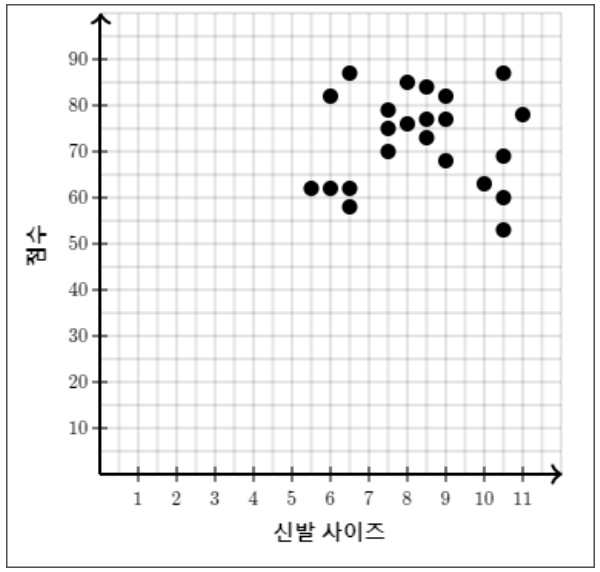
잎 부분을 0-4인 경우와 5-9인 경우로 더 세분화하여 만들 수도 있고, 90도 회전하여 만들 수도 있고, 두 자료집단을 줄기를 기준으로 좌우로 만들어 쉽게 비교할 수도 있다.

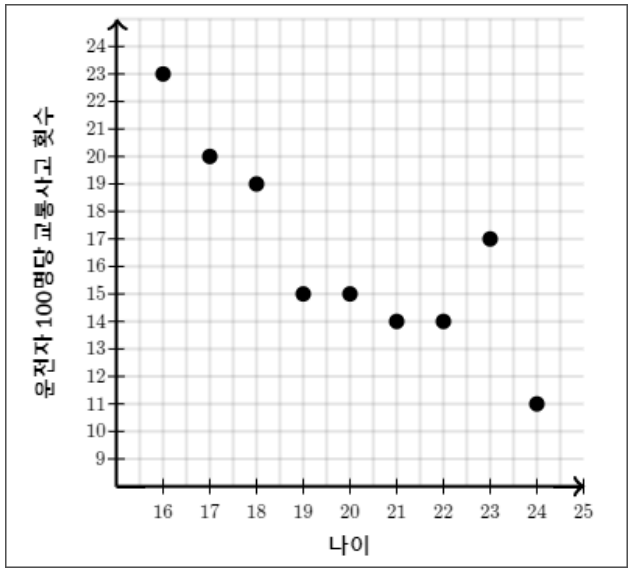
(10) 산점도(scatter diagram) : 두 종류의 자료가 독립변수와 응답변수의 관계를 가짐으로써 각각의 자료가 (x, y) 형태의 쌍으로 나타나는 경우가 있다. 이와 같이 쌍으로 주어진 자료를 나타내는 가장 좋은 방법으로 산점도를 사용한다. 이 산점도의 가로축은 독립변수 x를 기입하고, 세로축은 응답변수 y를 기입한다.

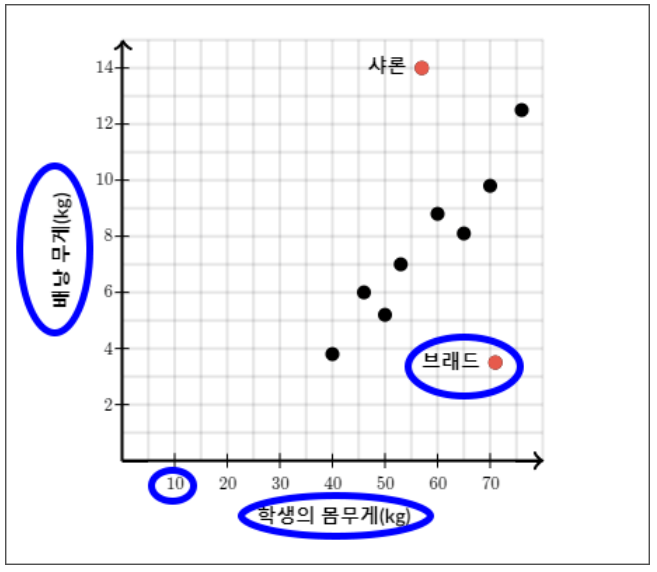


한편 산점도에 가장 적합한 직선 $y = ax + b$를 구할 수 있다면, 다음 관측값을 예측할 수 있다. 또한 이상점으로 판단되는 자료의 쌍을 쉽게 알 수 있다. 다시 말해서, 산점도를 이용하여 쌍으로 주어지는 자료를 나타냄으로써 다음 자료 쌍을 예측하거나 또는 평면 위에 점으로 표현된 자료가 의미가 있는 자료인지 아니면 오류에 의한 자료인지 명확하게 보여준다.









13주 과제

다음은 기초 통계학 기말시험성적을 나타낸 것이다.

23	60	70	32	57	74	52	70	82
36	80	77	81	95	41	65	92	85
55	76	52	10	64	75	78	25	80
98	81	67	41	71	83	54	64	72
88	62	74	43	60	78	89	76	84
48	84	90	15	79	34	67	17	82
69	74	63	80	85	61			

(1) 도수분포표, 원그래프, 막대그래프를 작성하여라. .
(2) 히스토그램을 작성하여라. 추정된 분포의 그래프를 그리고, 분포의 치우침에 대해 설염하여라.
(3) 줄기를 1,2,...,9로 하여 줄기-잎 그림을 작성하여라.
(4) 표본평균, 표본중앙값, 그리고 표본표준편차를 계산하여라.