자료(변수)의 두 가지 형태

1) categorical(범주형) : 명목/순서
2) quantitative(양적) : 연속/이산

- 명목(Nominal) 변수 : 순서 없는 범주를 가지는 변수
  예) 성별(남, 여), 지역(서울, 부산, 광주 …)
- 순서(Ordinal) 변수 : 순서가 있는 범주를 가지는 변수
  예) 자동차 크기(소형, 중형, 대형), 계층 (상, 중, 하)
- 연속(Continuous) 변수 : 무수히 많은 다른 값을 가짐
  예) 키, 몸무게, 온도
- 이산(Discrete) 변수 : 몇 개의 다른 값만 가짐
  예) 고장 횟수, 가족 구성원의 수

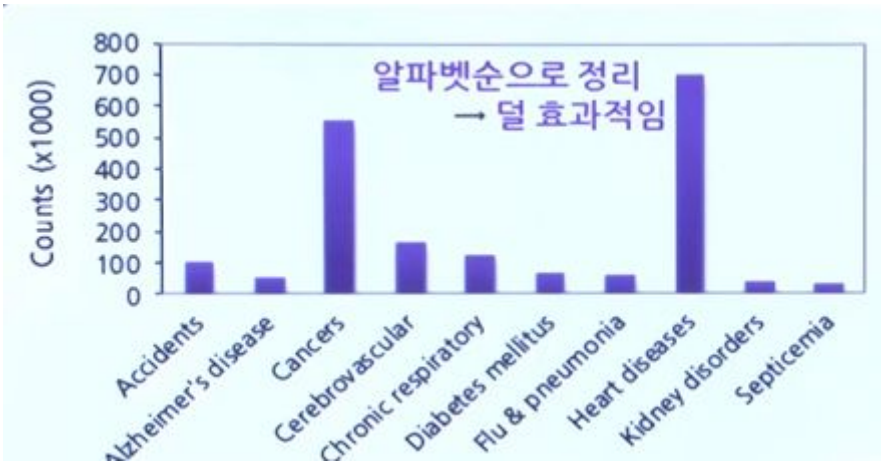
-범주형 자료형의 표현

- 도수분포표(Frequency table)
- 막대그래프(Bar graph)
  - 각 범주가 하나의 막대로 표현됨
- 파이 챠트(Pie chart)
  - 각 범주는 파이의 한 Slice로 표현됨
  - 보통 %를 사용하여 모두 더해서 1이 되도록 함

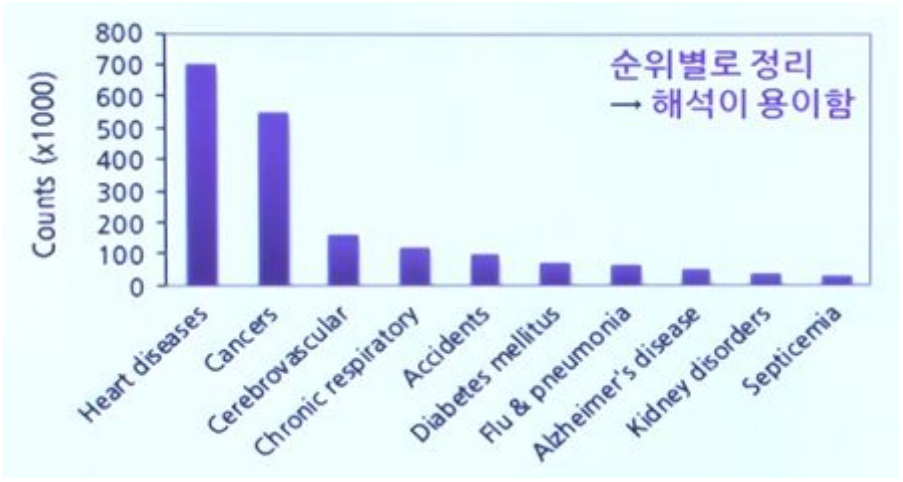
ex)2001년 미국의 상위 10개 사망원인
1.도수분포표

Rank	Causes of death	Counts	% of top 10s	% of total deaths
1	Heart disease	700,142	37%	29%
2	Cancer	553,768	29%	23%
3	Cerebrovascular	163,538	9%	7%
4	Chronic respiratory	123,013	6%	5%
5	Accidents	101,537	5%	4%
6	Diabetes mellitus	71,372	4%	3%
7	Flu and pneumonia	62,034	3%	3%
8	Alzheimer's disease	53,852	3%	2%
9	Kidney disorders	39,480	2%	2%
10	Septicemia	32,238	2%	1%
	All other causes	629,967		26%

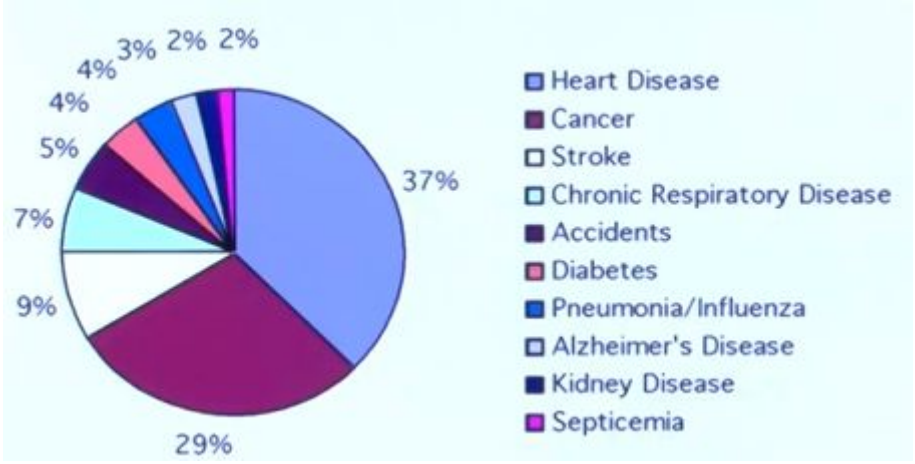
2.막대그래프



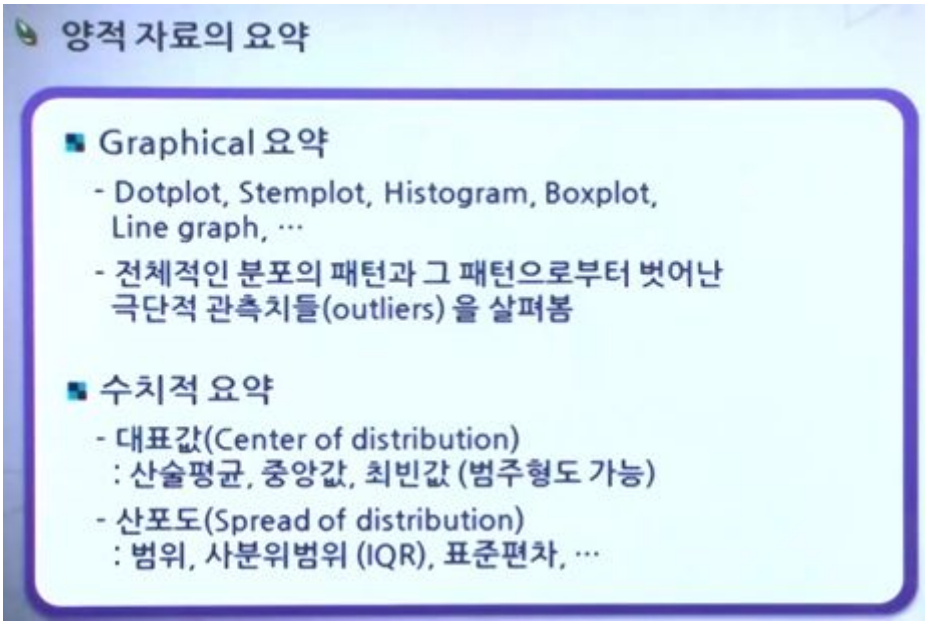
그래프로 표현하는 이유가 시각적으로 빠르게 데이터를 이해하기 위함인데, 이렇게 정렬을 해서 표현하면 하위 순서를 파악하기에 용이하지 않다.



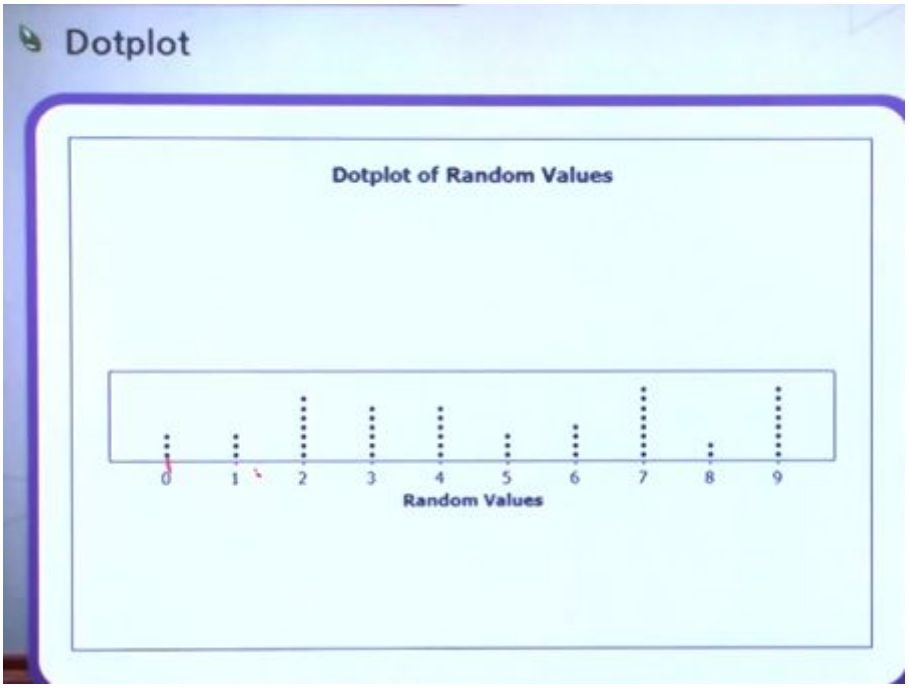
따라서 다음과 같이 순위별로 정리하는 것이 좋다.



-양적 자료형의 표현



ex)



0에 3개 1에 3개...의 데이터가 존재하는 것이다.

줄기-잎 그림(Stemplot)

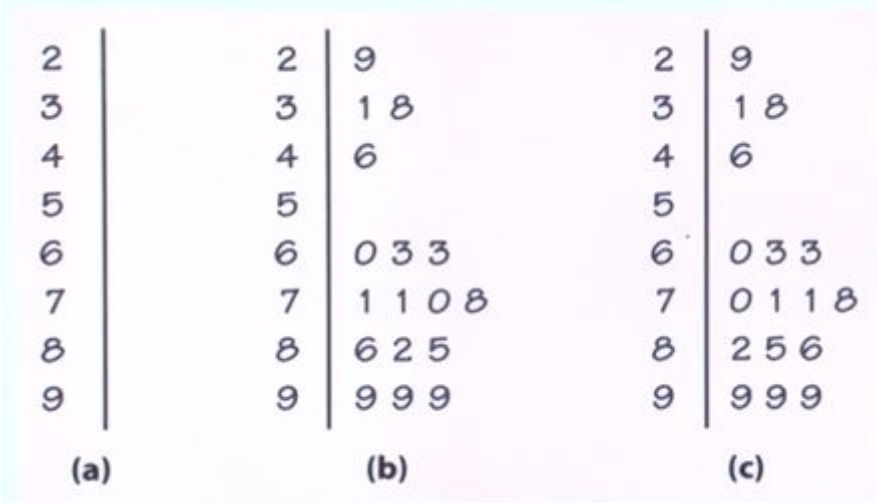
- 실제 자료의 수치를 그대로 사용하면서 분포의 형태를 보여주는 그림
- 쉽고 빠르게 그릴 수 있으며 정보의 손실이 없음
- 모든 값이 양수면서 데이터의 양이 많지 않을 경우 좋음
- 두 개의 연관된 분포를 비교하고 싶을 때, 같은 줄기를 공유하는 Back-to-back stemplot이 유용함

TABLE 1.2

Literacy rates (percent) in Islamic nations

Country	Female percent	Male percent	Country	Female percent	Male percent
Algeria	60	78	Morocco	38	68
Bangladesh	31	50	Saudi Arabia	70	84
Egypt	46	68	Syria	63	89
Iran	71	85	Tajikistan	99	100
Jordan	86	96	Tunisia	63	83
Kazakhstan	99	100	Turkey	78	94
Lebanon	82	95	Uzbekistan	99	100
Libya	71	92	Yemen	29	70
Malaysia	85	92			

Table 1-2
Introduction to the Practice of Statistics, Fifth Edition
© 2005 W. H. Freeman and Company

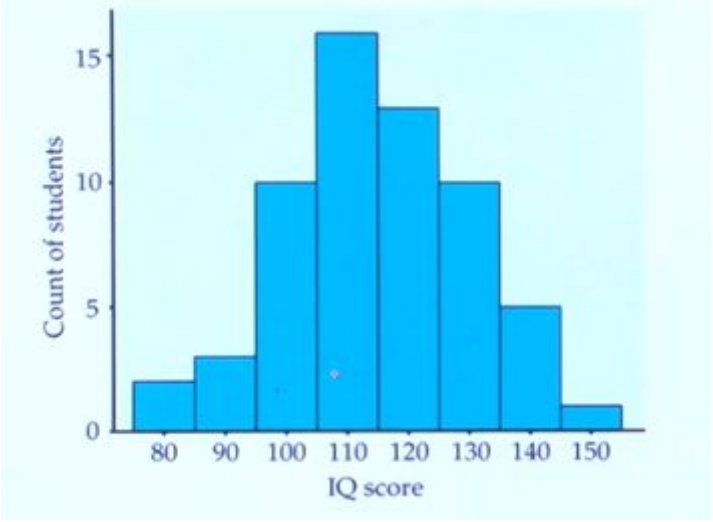


->

정보의 손실이 없이 데이터를 파악할 수 있고, 분포도 알 수가 있다.
50개 이하의 데이터일 때 많이 쓴다.
그 이상일 때는 보통 히스토그램을 쓴다.

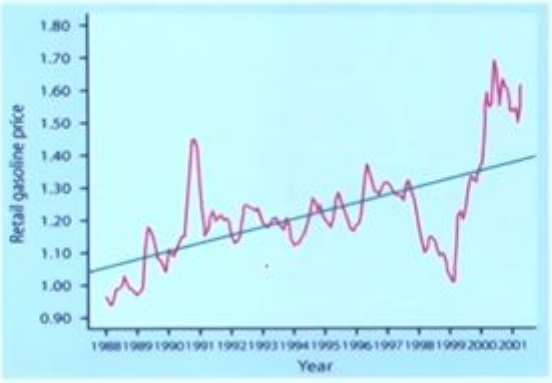
히스토그램(Histogram)

- 자료의 범위(range)를 몇 개의 구간(class)으로 나누고 각 구간에 들어가는 관측치의 빈도(frequency) 또는 상대빈도(relative frequency)만을 나타내는 그림
- Dataset이 큰 경우 좋음
- 히스토그램의 각 막대는 그 class의 빈도에 비례함
- 그리는 방법 :
  - 자료의 범위를 구간(class)으로 나눔 (class의 개수는 5-10개 정도가 적당하며 그리면서 조절)
  - 각 class에 들어가는 관측치의 개수(frequency)를 계산
  - 각 class별로 빈도 또는 상대빈도를 표현



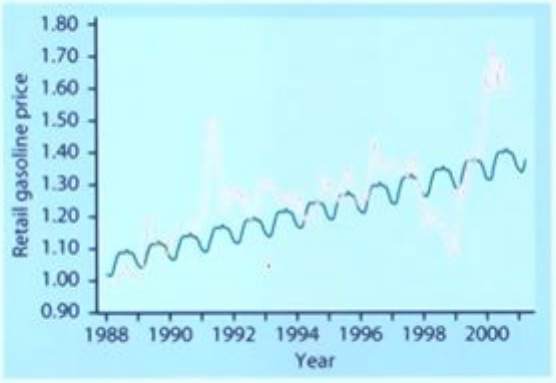
Line graph(time plot)

- 시계열 자료인 경우 x축을 시간으로 한 time plot에서 trend와 seasonal variation 등을 찾을 수 있음



Trend : a rise or fall that persists over time, despite small irregularities

- 시계열 자료인 경우 x축을 시간으로 한 time plot에서 trend와 seasonal variation 등을 찾을 수 있음



Seasonal variation : a pattern that repeats itself at regular intervals of time

TABLE 1.4

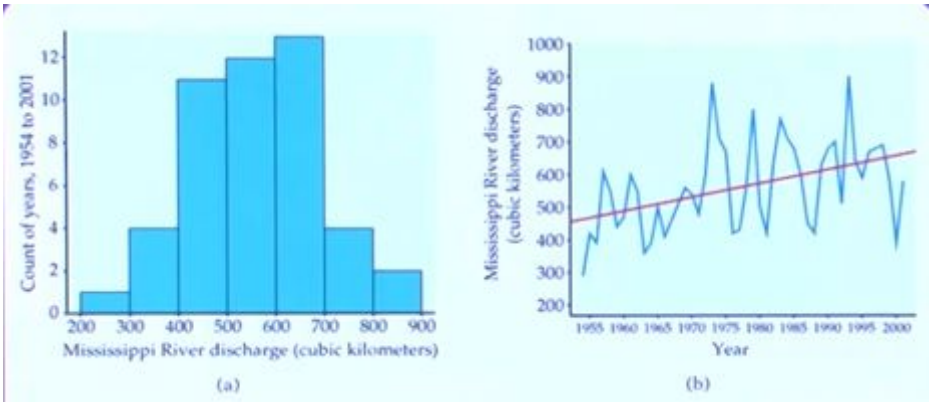
Yearly discharge of the Mississippi River (in cubic kilometers of water)

Year	Discharge	Year	Discharge	Year	Discharge	Year	Discharge
1954	290	1966	410	1978	560	1990	680
1955	420	1967	460	1979	800	1991	700
1956	390	1968	510	1980	500	1992	510
1957	610	1969	560	1981	420	1993	900
1958	550	1970	540	1982	640	1994	640
1959	440	1971	480	1983	770	1995	590
1960	470	1972	600	1984	710	1996	670
1961	600	1973	880	1985	680	1997	680
1962	550	1974	710	1986	600	1998	690
1963	360	1975	670	1987	450	1999	580
1964	390	1976	420	1988	420	2000	390
1965	500	1977	430	1989	630	2001	580

Table 1.4

Introduction to the Practice of Statistics, Sixth Edition

->



문제

다음은 어떤 기업의 최근 2년간 월간 매출액을 나타낸 데이터이다.

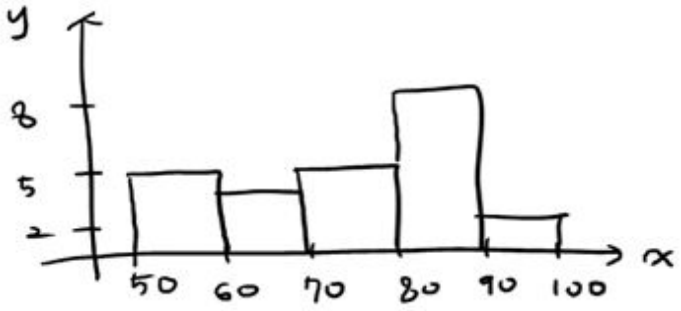
65	75	78	50	60	90	84	82
81	54	92	72	70	66	79	80
55	95	85	87	90	57	64	90

도수분포표, 히스토그램, 줄기-잎 그림을 그려보시오.

1. 도수 분포표

구분	도수
60 이하	5
61~70	4
71~80	5
81~90	8
91~100	2
합계	24

2. 히스토그램



3. 줄기-잎 그림

5	0 4 5 7
6	0 4 5 6
7	0 2 5 8 9
8	0 1 2 4 5 7
9	0 0 0 2 5

대표값

- 산술평균(mean)
  - 계산이 쉽고 수학적으로 다루기 쉬움
  - 모든 관측치를 사용하므로 특이값에 영향을 많이 받음
- 중앙값(median)
  - 관측한 자료를 순서대로 배열하여 가장 중앙에 있는 값
  - 순위를 사용해 중앙에 있는 값만 사용하므로 특이값에 영향을 받지 않음

대표값

- 최빈값(mode)
  - 관측치 가운데 가장 여러 번 나타난 값
  - 여러 개 존재하거나 존재하지 않을 수 있고 중심을 잘 대변하지 못하는 경우가 많음
  - 이산변수에 주로 사용하고, 범주형 자료에도 사용가능
- 분포가 한쪽으로 치우쳐있는 경우나 특이값들이 있는 경우 중앙값이 더 적합하고
  그렇지 않은 경우 대부분 산술평균이 적합함

중앙값의 계산

1	1	0.6
2	2	1.2
3	3	1.6
4	4	1.9
5	5	1.5
6	6	2.1
7	7	2.3
8	8	2.3
9	9	2.5
10	10	2.8
11	11	2.9
12	12	3.3
13		3.4
14	1	3.6
15	2	3.7
16	3	3.8
17	4	3.9
18	5	4.1
19	6	4.2
20	7	4.5
21	8	4.7
22	9	4.9
23	10	5.3
24	11	5.6
25	12	6.1

1. 관측치를 크기 순서대로 배열
   n = 관측치의 개수

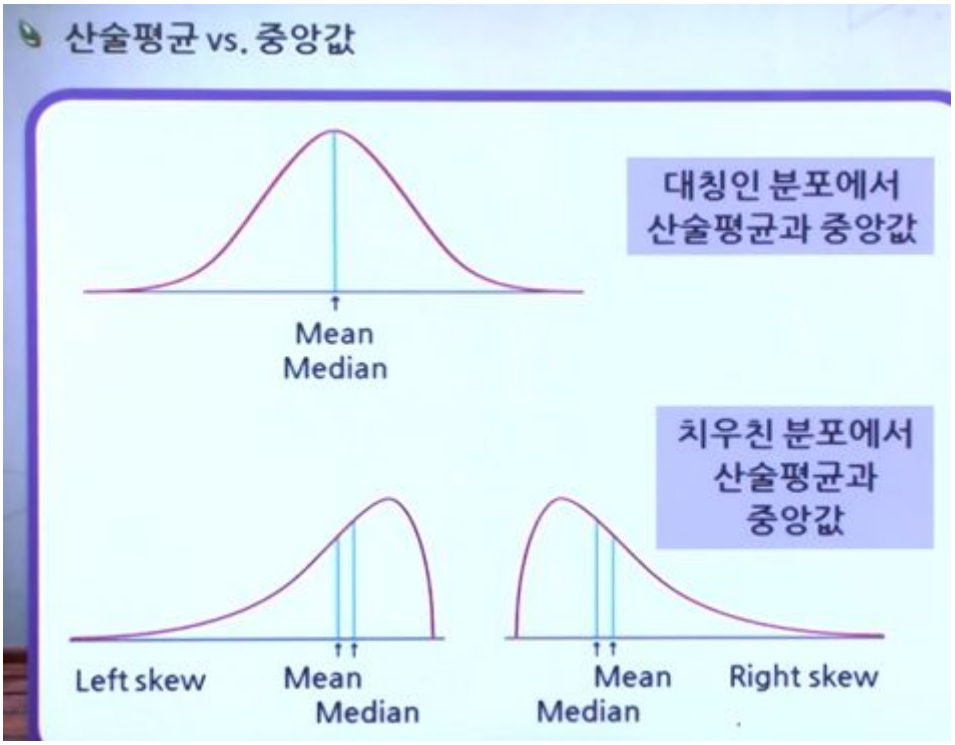
2. a. n 이 홀수이면 중앙값은
   $(n+1)/2$ 번째 관측치

← $n = 25$
$(n+1)/2 = 26/2 = 13$
Median = 3.4

2. b. n 이 짝수이면, 중앙값은
   가장 중앙에 있는
   두 개의 관측치의 평균

$n = 24$ →
$n/2 = 12$
Median = (3.3+3.4) /2 = 3.35

1	1	0.6
2	2	1.2
3	3	1.6
4	4	1.9
5	5	1.5
6	6	2.1
7	7	2.3
8	8	2.3
9	9	2.5
10	10	2.8
11	11	2.9
12		3.3
13		3.4
14	1	3.6
15	2	3.7
16	3	3.8
17	4	3.9
18	5	4.1
19	6	4.2
20	7	4.5
21	8	4.7
22	9	4.9
23	10	5.3
24	11	5.6


산술평균 vs. 중앙값
대칭인 분포에서
산술평균과 중앙값
Mean
Median
치우친 분포에서
산술평균과
중앙값
Left skew
Mean
Median
Mean
Median
Right skew

산포도(spread)

- 범위 (range)
  - 최대값-최소값
  - 간단하지만 특이값에 큰 영향을 받음
- 4분위 범위(IQR, interquartile range)
  - 특이값에 영향 받지 않음
- 표준편차
  - 가장 널리 이용되며 통계적 추론에 유용
  - 산술평균처럼 특이값에 영향을 받음

$$S = +\sqrt{S^2}, S^2 = \sum (X_i - \overline{X})^2 / (n-1)$$

사분위 범위(IQR)

- 백분위수 (percentile, quantile)
  : p백분위수란 p%의 관측치는 이 값 아래에 있고 나머지는 이 값보다 위에 있게 되는 값을 말함
- 중앙값 : 50 백분위수
- Q1=25 백분위수=제1사분위수 (first quartile)
- Q3=75 백분위수=제3사분위수(third quartile)
- IQR=Q3-Q1
- 다섯 숫자 요약(Five-number summary)
  : min Q1 median Q3 max

상자그림(Boxplot)

- 다섯 숫자 요약의 graphical result
- 상자는 중앙 50%의 자료를 표시
- 여러 개의 분포를 한 눈에 비교할 때 유용함

- 그리는 방법 :
  - Q1과 Q3로 끝나는 상자를 그린다.(상자의 길이=IQR)
  - 상자 안에 줄을 그어 중앙값을 표시한다.
  - Q3+1.5IQR보다 크거나 Q1-1.5IQR보다 작은 값은 ⋆ 또는 다른 symbol로 표시한다(outliers).
    : "1.5IQR criterion"
  - 상자의 끝에서 Outlier가 아닌 값 중에 가장 큰 값과 가장 작은 값까지 줄을 긋는다.

다섯 숫자 요약과 상자그림

25	6	6
24	5	6
23	4	5
22	3	5
21	2	5
20	1	5
19	6	4
18	5	4
17	4	4
16	3	4
15	2	4
14	1	4
13		3
12	6	3
11	5	3
10	4	3
9	3	3
8	2	2
7	1	2
6	6	2
5	5	2
4	4	2
3	3	2
2	2	1
1	1	1

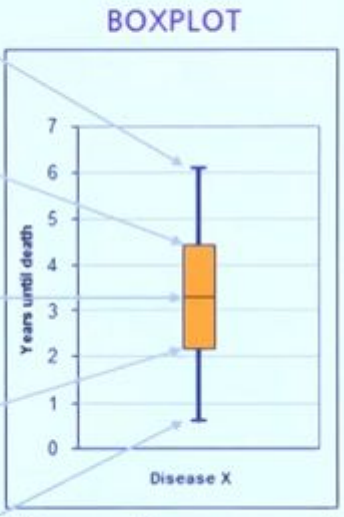
Largest = max = 6.1

Q_3 = third quartile = 4.35

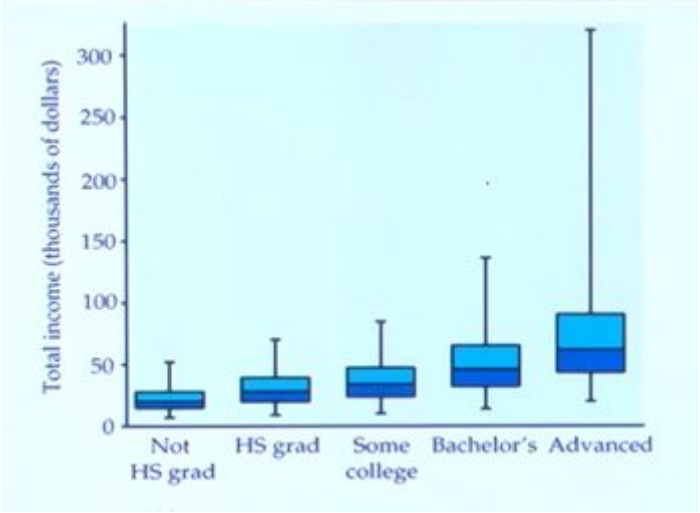
M = median = 3.4

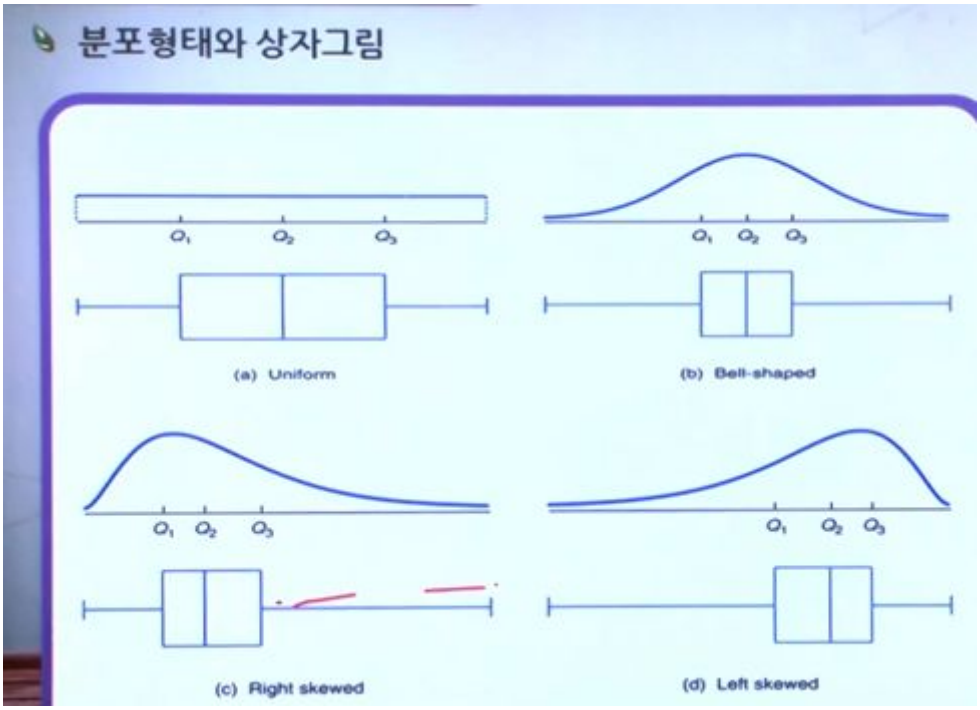
Q_1 = first quartile = 2.2

Smallest = min = 0.6



Five-number summary :
min Q_1 M Q_3 max


Total income (thousands of dollars)
300
250
200
150
100
50
0
Not HS grad
HS grad
Some college
Bachelor's
Advanced


분포형태와 상자그림
Q_1 Q_2 Q_3
(a) Uniform
(b) Bell-shaped
(c) Right skewed
(d) Left skewed