Box Plot

Box Plot(상자그림 또는 상자도표)은 데이터를 시각화하는 도구 중 하나로 Boxplot이라 칭하기도 하며, Whisker를 포함하므로 Box and whisker plot 또는 Box and whisker diagram으로 불리기도 한다.

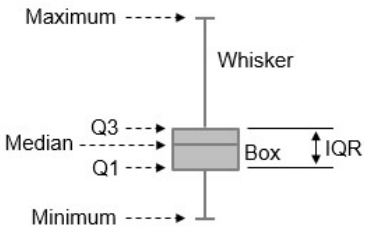
상자도표는 데이터로부터 얻어진 5개의 통계치(Five-number summary)인 최대 변수, 최소 변수 및 사분위수(Quartile)를 수직 방향 또는 수평 방향으로 도식화 한 것이다.

사분위수란 변수들을 4개의 그룹으로 나누는 3개의 수를 의미한다. 아래 그림은 9개의 측정치를 이용해 상자도표가 그려지는 절차를 보여주는 예이다. 먼저 변수들을 순서대로 나열한 후 25%, 50%, 75%에 해당하는 값을 정한다. 이때 50%에 해당하는 2사분위수가 중앙값(Median)이다.

25
44
51
41
98
14
46
53

14
25
41
44
46
48
56
98

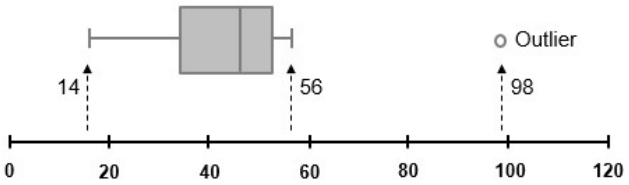
Minimum	14
Quartile1(Q1)	33 = (25+41)/2
Median	45 = (44+46)/2
Quartile3(Q3)	52 = (48+56)/2
Maximum	98



1사분위수(Q1)는 중앙값(Q2)보다 작은 변수들의 중앙값이며, 3사분위수(Q3)는 중앙값보다 큰 변수들의 중앙값이다. 이 두 사분위수의 거리를 Interquartile Range(IQR)라 칭한다.

위의 상자도표에서 Whisker의 양단이 최대 변수 및 최소 변수에 위치한지만 동일한 데이터로부터 작성된 아래의 상자도표에서 Whisker는 다른 값들을 나타낸다.

IQR	19 = 52-33
1.5 IQR	28.5 = 19x1.5
Q1-1.5 IQR	4.5 = 33-28.5
Q3+1.5 IQR	80.5 = 52+28.5

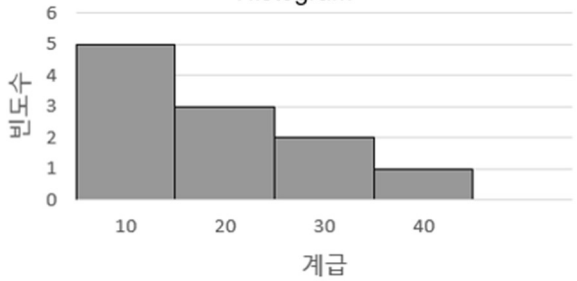


위의 상자도표는 Whisker는 물론 작은 원(o)도 가진다. 여기서 Whisker의 양단은 Q1으로부터 아래로 IQR의 1.5배 이내에 존재하는 변수 중 최소 변수인 14와 Q3로부터 위로 IQR의 1.5배 이내에 존재하는 변수 중 최대 변수인 56에 각각 위치한다.

Q1으로부터 아래로 IQR의 1.5배, Q3로부터 위로 IQR의 1.5배 범위를 벗어나는 변수 즉, 비정상적으로 분포된 변수들을 극단값(Outlier)이라 칭하는데, IQR의 3배 범위 내에 있는 극단값을 Mild Outlier, 그 밖에 존재하는 극단값을 Extreme Outlier로 구분하기도 한다. 위의 예에서 변수 98은 Mild Outlier이므로 작은 원(o)으로 표시되었다. 단, Extreme Outlier의 경우 별표(*)를 이용하여 표시한다.

상자도표는 히스토그램(Histogram)이나 줄기잎도표(Stem and Leaf Plot)에 비해 데이터의 분포를 상세히 보여주지는 못하지만, 분포의 비대칭 여부와 Outlier의 존재여부 및 그 개수를 보여주는데 유용하다.

<Histogram>



<Stem and Leaf Plot>

Stem	Leaf			
0	4	5	7	7
1	0	2	5	
2	0	2	6	
3	4			