

4장 표/측도를 이용한 데이터 정리


4.1 도수분포표와 교차표

질적 데이터의 도수분포표

- 2장에서 질적 데이터의 시각화에 대해 살펴보았다. 예를 들면 표 4.1과 같은 성별 (1:남자, 2:여자) 데이터의 막대, 원, 띠그래프 등을 그려보았는데 이와 같은 그래프는 남자와 여자의 빈도수, 즉, 도수분포를 이용하여 그린 것이다. 『eStat』을 이용하여 이 성별 범주형 원시 데이터의 도수분포표를 작성하여 보자.

표 4.1 성별 데이터

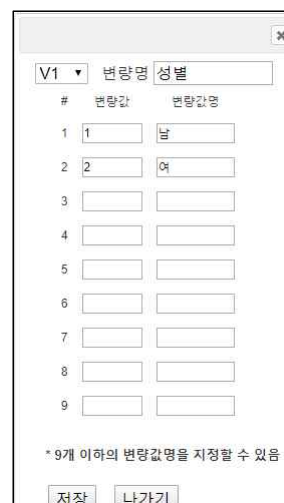
성별
1
2
1
2
1
1
1
2
1
2

- 『eStat』에 성별 데이터를 [그림 4.1]과 같이 입력한다. ‘변량편집’을 이용하여 변량명 ‘성별’을 입력하고 변량값 1과 2에 대한 변량값명 ‘남’ ‘여’를 입력한다([그림 4.2]). 이와 같이 변량값명에 대한 편집을 한 데이터는 JSON 형식으로 저장(하여야 정보를 잃어버리지 않게 된다. 다시 불러올 때도 JSON 형식으로 불러오는 아이콘 을 클릭하여야 한다.



	성별	V2	V3	V4
1	1			
2	2			
3	1			
4	2			
5	1			
6	1			
7	1			
8	2			
9	1			
10	2			

[그림 4.1] 데이터 입력




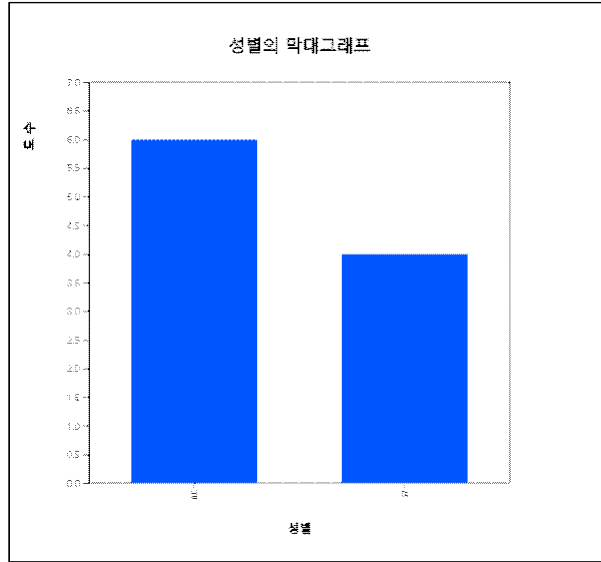
#	변량값	변량값명
1	1	남
2	2	여
3		
4		
5		
6		
7		
8		
9		

* 9개 이하의 변량값명을 지정할 수 있음

저장 나가기

[그림 4.2] 변량편집

- [그림 4.1]과 같이 변량선택박스에서 ‘분석변량’으로 성별을 선택하면 [그림 4.3]과 같은 성별의 막대그래프가 그려지고 여기서 마우스로 도수분포표 아이콘 을 선택하면 [그림 4.4]와 같은 남녀별 학생수에 대한 도수분포표가 결과저장창에 나타난다.




[그림 4.3] 성별의 막대그래프


도수분포표	분석변량	(성별)	
변량값	변량값명	도수	백분율(%)
1	남	6	60.0
2	여	4	40.0
합계		10	100
	결측수	0	

[그림 4.4] 성별 도수분포표

- 막대그래프나 원그래프는 이 도수분포를 이용하여 그린 것이다.

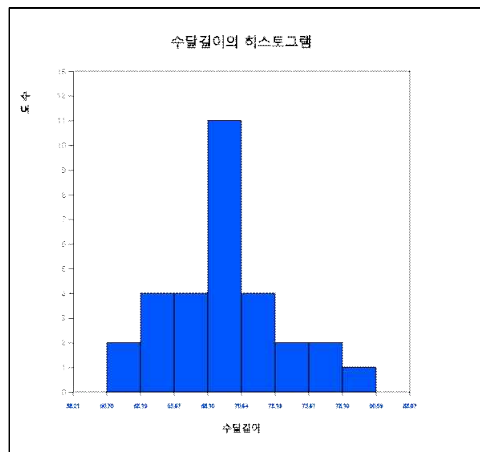
양적 데이터의 도수분포표

- 양적 데이터에 대한 도수분포표는 구간을 나누어 각 구간에 속하는 데이터의 빈도수를 조사하여 작성한다. 일반적으로 동일한 간격을 갖고, 서로 중복되지 않는 계급구간(class interval)을 여러 개 설정해 각 구간에 속하는 데이터의 개수를 도수분포표에 나타낸다. 이를 위해 먼저 최댓값과 최솟값을 구하여 데이터의 범위를 알아본 다음 구간의 개수를 결정한다. ‘몇 개의 구간을 할 것인가?’는 분석자의 선택인데 일반적으로 데이터의 수에 따라 5개에서 10개 사이의 구간의 수가 많이 이용된다. 구간의 개수가 정해지면 데이터값의 범위(=최댓값-최솟값)를 구간의 개수로 나누어 구간의 너비를 계산한다. 각 구간의 시작점과 끝점은 대개 ‘~ 이상(≥)에서 ~ 미만(<)’으로 정한다.
- 예를 들어 수달의 길이(『eStat』의 ) ⇨ 01Korean ⇨ 031연속_수달의길이.csv)

데이터의 히스토그램과 도수분포표를 『eStat』을 이용하여 구해보자. 『eStat』에서 **Ex** ⇨ 01Korean ⇨ 031연속_수달의 길이.csv를 불러온다([그림 4.5]). 마우스로 히스토그램 아이콘 을 클릭하고 변량명 '수달의길이'를 선택하면 [그림 4.6]과 같은 히스토그램이 그려진다.

수달길이	V2	V3	V4
1	63.2		
2	65.3		
3	67.6		
4	68.7		
5	69.7		
6	60.7		
7	72.4		
8	75.2		
9	64.4		
10	76.5		
11	68.3		
12	69.3		
13	70.2		
14	71.3		
15	74.2		
16	63.6		
17	66.1		
18	67.9		
19	68.7		
20	70.5		
21	72.3		
22	72.8		
23	77.6		
24	78.1		

[그림 4.5] 수달길이 데이터



[그림 4.6] 수달길이의 히스토그램

- 그래프 밑의 선택 대화상자에서([그림 4.7]) '도수분포표'를 클릭하면 [그림 4.8]과 같은 구간별 도수분포표가 결과저장창에 나타난다.

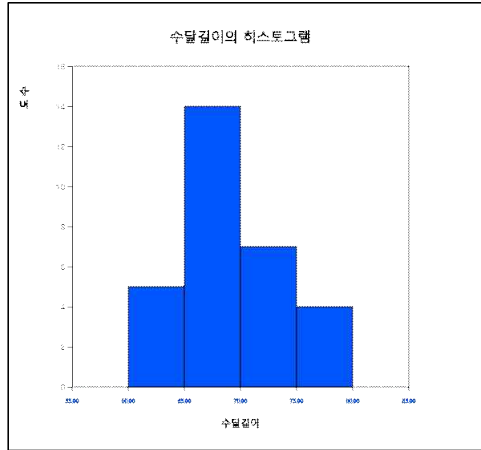
<input type="checkbox"/> 평균	<input type="checkbox"/> 도수표시	<input type="checkbox"/> 도수분포다각형	<input checked="" type="checkbox"/> 도수분포표
<input checked="" type="checkbox"/> 새 구간으로 실행		구간시작 0	구간너비 10

[그림 4.7] 히스토그램의 선택사항

구간별 도수분포표	그룹명	0
계급구간 (수달길이)	그룹1 (null)	합계
1 [60.70, 63.19)	2 (6.7%)	2 (6.7%)
2 [63.19, 65.67)	4 (13.3%)	4 (13.3%)
3 [65.67, 68.16)	4 (13.3%)	4 (13.3%)
4 [68.16, 70.64)	11 (36.7%)	11 (36.7%)
5 [70.64, 73.13)	4 (13.3%)	4 (13.3%)
6 [73.13, 75.61)	2 (6.7%)	2 (6.7%)
7 [75.61, 78.10)	2 (6.7%)	2 (6.7%)
8 [78.10, 80.59)	1 (3.3%)	1 (3.3%)
합계	30 (100%)	30 (100%)

[그림 4.8] 수달 길이의 구간별 도수분포표

- 만일 히스토그램 구간을 60kg에서 5kg간격으로 재조정하기 위해서는 그래프 선택사항에서 '구간시작'을 60, 구간너비를 5로 설정한 후 '새구간으로 실행' 버튼을 누르면 [그림 4.9]와 같은 히스토그램이 나타난다. 선택사항의 '도수분포표'를 클릭하면 [그림 4.10]의 도수분포표가 나타난다.



[그림 4.9] 구간 조정된 히스토그램

구간별 도수분포표	그룹명	0
계급구간 (수달길이)	그룹1 (null)	합계
1 [60.00, 65.00)	5 (16.7%)	5 (16.7%)
2 [65.00, 70.00)	14 (46.7%)	14 (46.7%)
3 [70.00, 75.00)	7 (23.3%)	7 (23.3%)
4 [75.00, 80.00)	4 (13.3%)	4 (13.3%)
합계	30 (100%)	30 (100%)



[그림 4.10] 구간 조정된 도수분포표

교차표

- 교차표**(cross table 또는 contingency table)는 두 개의 범주형 변량을 요약하여 그 연관된 특성을 연구하는데 매우 효과적인 표로서 한 변량의 도수분포표와 유사하다. 교차표는 두 변량의 가능한 변량값을 행과 열로 나누어 행변량의 속성과 열변량의 속성이 교차하는 부분에 셀(cell)을 만든 후, 각 데이터마다 행변량과 열변량의 데이터값을 조사하여 해당되는 셀에 속하는 데이터의 빈도수를 조사한다. 분석을 위해 각 셀의 빈도수 밑에 행의 합에 대한 백분율, 열의 합에 대한 백분율, 그리고 전체 백분율을 표시하기도 한다.
- 교차표는 범주형 데이터에 대해서 작성하는 것이지만 연속형 데이터의 경우 구간을 나누어 범주형 데이터로 만들어 교차표를 작성할 수도 있다.
- 교차표를 작성하여 분포를 살펴보면 대략 두 변량 사이의 관련성을 알 수 있다. 이를 좀 더 자세히 알아보기 위해서는 행변량과 열변량의 독립성검정, 또는 동질성검정 등의 통계분석을 할 수 있는데, 11장에서 자세히 알아보기로 한다.
- 한 여론조사에서 성별(1:남자, 2:여자)과 함께 결혼여부(1:미혼, 2:결혼, 3:기타)를 조사한 데이터가 표 4.2와 같다. 성별 결혼여부에 대한 교차표를 구해 보자.


표 4.2 성별 결혼여부 데이터

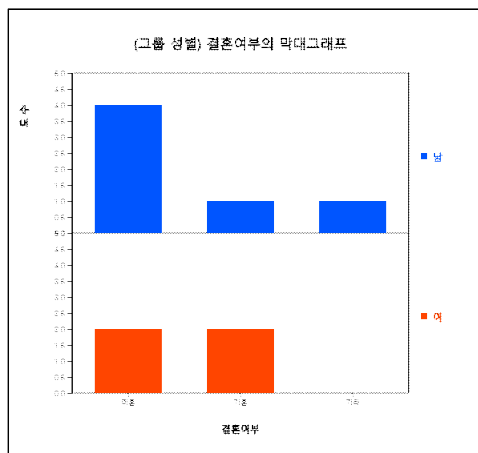
성별	결혼여부
1	1
2	2
1	1
2	1
1	2
1	1
1	1
2	2
1	3
2	1

- 『eStat』에 성별과 결혼여부 데이터를 입력한다([그림 4.11]). ‘변량편집’을 이용하여 변량명 ‘성별’을 입력하고 변량값 1과 2에 대한 변량값명 ‘남자’, ‘여자’를 입력한다. 같은 방법으로 변량명 ‘결혼여부’를 입력하고 변량값 1, 2, 3에 대한 변량값명 ‘미혼’, ‘기혼’, ‘기타’를 입력한다. 이와 같이 변량값명에 대한 편집을 한 데이터는 JSON 형식으로 저장(아이콘  클릭)한다. 다시 불러올 때도 JSON 형식으로 불러오는 아이콘  을 클릭하여야 한다.

성별	결혼여부	V3	V4
1	1	1	
2	2	2	
3	1	1	
4	2	1	
5	1	2	
6	1	1	
7	1	1	
8	2	2	
9	1	3	
10	2	1	

[그림 4.11] 성별 결혼여부 데이터 입력

- 마우스로 첫째 변량(‘분석변량’) ‘결혼여부’와, 둘째 변량(‘by 그룹’) ‘성별’의 변량명을 차례로 클릭하면 기본적으로 선택되어있는 [그림 4.12]와 같은 성별 결혼여부의 막대그래프가 나타난다. 이때 도수분포표 아이콘  을 클릭하면 성별 결혼여부의 교차표가 결과저장창에 표시된다([그림 4.13]). 교차표에서는 행변량이 그룹변량이 되고 열변량이 분석변량이 된다. 이 교차표를 이용하여 성별 결혼여부에 대한 막대그래프가 그려진 것이다.



[그림 4.12] 성별 결혼여부의 막대그래프

교차표	올변량		(결혼여부)		
	미혼	기혼	기타	합계	
성별					
남	4 66.7%	1 16.7%	1 16.7%	6 100%	
여	2 50.0%	2 50.0%	0 0.0%	4 100%	
합계	6 60.0%	3 30.0%	1 10.0%	10 100%	
	결측수	0			
독립성검정					
카이제곱값	1.667	자유도	2	p-값	0.4346

[그림 4.13] 성별 결혼여부에 대한 교차표

4.3 측도를 이용한 데이터 요약

중심위치의 측도

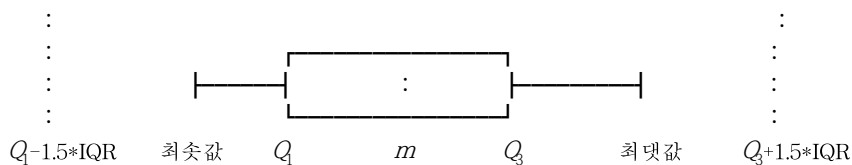
- 중심위치의 측도(measure of central tendency)에는 평균, 중앙값, 최빈값 등이 있는데 이 중 가장 많이 사용되는 것이 **평균(mean)**이다.(이것을 산술평균이라고도 한다.) 평균은 데이터를 대표하는 일종의 무게중심으로 볼 수 있다.
- 주어진 데이터가 모집단일 때의 평균을 **모평균(population mean)**이라 하고 보통 μ (그리스 문자로 '뮤'라고 읽음)로 표시한다. 또한, 주어진 데이터가 표본일 때의 평균을 **표본평균(sample mean)**이라 하고 \bar{x} '엑스 바아'라고 읽음)로 표시한다. 평균은 어느 한 데이터값이 아주 크거나 작은 극단점의 영향을 많이 받는다. 하지만 표본평균은 모평균을 예측하기 위한 좋은 성질을 가지고 있어서 데이터 분석에 자주 사용된다.
- **중앙값(median)**은 데이터를 크기 순서로 나열할 때 중앙에 놓이는 값으로 데이터가 표본일 경우 m , 모집단일 경우 M 으로 표시한다. 즉, 데이터의 수를 n 이라 할 때, n 이 홀수이면 $(n+1)/2$ 번째의 값을 중앙값으로, n 이 짝수이면 $n/2$ 번째와 $(n/2 + 1)$ 번째 데이터값의 평균을 중앙값으로 정의한다. 중앙값은 극단점이 있는 경우에도 민감하지 않아 극단점이 있는 경우에는 평균보다 중심위치의 측도로 더 자주 쓰인다.
- **최빈값(mode)**은 데이터 중 가장 빈도가 많은 값이다. 하지만 연속형 데이터일 경우 거의 많은 데이터값들이 한번만 나타나기 때문에 단순히 빈도수가 많은 값을 최빈값으로 정하는 것은 불합리하다. 이런 경우 연속형 데이터를 몇 개의 계급구간으로 나누어서 각 구간에 대한 도수분포표로 정리한 후 가장 도수가 높은 구간의 중간값을 최빈값으로 정하기도 한다.

산포도의 측도

- 어느 체조시합에서 '갑'선수의 경기에 대한 네 심판의 채점이 3, 4, 6, 7점이었다. 또 '을'선수의 경기에 대한 채점은 2, 4, 6, 8점이었다. 두 선수 모두 평균은 5점이지만 '을'은 '갑'에 비해 점수의 편차가 크다는 것을 쉽게 알 수 있다. 데이터가 흩어진 정도를 수치로 측정하는 것을 **산포도의 측도(measure of dispersion)**라

한다. 많이 쓰이는 산포도의 측도는 분산 또는 표준편차이고, 그밖에 범위, 사분위수범위 등이 있다.


- **분산**(variance)이란 각 데이터값과 평균과의 거리를 제곱하여 합을 구한 후 이를 데이터의 수로 나눈 것이다. 따라서 데이터가 평균을 중심으로 많이 흩어져 있으면 분산이 커지고, 데이터가 평균주위에 몰려 있으면 분산이 작게 된다. 모집단의 분산을 **모분산**(population variance)이라 부르며 σ^2 (시그마 제곱)으로 표시하고, 표본의 분산을 **표본분산**(sample variance)이라 부르며 s^2 로 표시한다.
- 표본분산을 계산할 때 표본의 수 n 대신 $n-1$ 을 사용하는 데 그 이유는 6장에서 설명한다.
- **표준편차**(standard deviation)는 분산의 제곱근으로 정의한다. 모집단의 표준편차를 모표준편차라고 부르며 σ 로 표시하고, 표본의 표준편차를 표본표준편차라고 부르며 s 로 표시한다. 분산은 제곱거리의 평균이어서 현실적인 해석이 쉽지 않으나 표준편차는 분산의 제곱근이어서 각 값과 평균과의 평균거리의 측도로 해석이 가능하다.
- **범위**(range)는 데이터의 최댓값에서 최솟값을 뺀 차이를 나타낸다. 범위는 계산하기가 간편하나 극단점이 있을 경우 올바른 산포의 측도가 되지 못한다.
- 범위의 단점을 보완한 것이 사분위수범위인데 이것을 알기 위해서 먼저 백분위수를 살펴보자. **p% 백분위수**(percentile)는 데이터를 작은 것부터 큰 것까지 순서대로 늘어놓았을 때 대략 p%번째 데이터를 뜻한다.
- 백분위수 중 25% 백분위수를 **일사분위수**(1st quartile, Q_1 으로 표시), 50% 백분위수를 **이사분위수**(2nd quartile, Q_2 로 표시) 또는 **중앙값**(m 으로 표시), 75% 백분위수를 **삼사분위수**(3rd quartile, Q_3 로 표시)라고 부른다. **사분위수범위**(interquartile range, IQR로 표시)는 삼사분위수에서 일사분위수를 뺀 값 즉, $Q_3 - Q_1$ 이다.
- **상자그래프**(box-whisker plot)는 이러한 사분위수에 관한 데이터의 정보를 그래프로 나타낸 것으로 최근에 많이 사용되기 시작한 데이터정리 방법이다. 상자그래프는 먼저 일사분위수(Q_1)와 삼사분위수(Q_3)를 네모상자로 연결한 다음 중앙값(m)을 상자 안에 표시한다. $Q_1 - 1.5 \cdot \text{IQR}$ 이내인 값 중에서 최솟값과 $Q_3 + 1.5 \cdot \text{IQR}$ 의 이내의 최댓값을 상자와 선으로 연결한다([그림 4.14] 참조). 상자그래프를 이용하면 데이터분포의 대칭성, 데이터의 중심위치, 산포의 정도 등을 잘 알아볼 수 있다. 극단점이 있을 경우 $Q_1 - 1.5 \cdot \text{IQR}$ 과 $Q_3 + 1.5 \cdot \text{IQR}$ 의 선을 넘는 데이터는 극단점으로 간주하기도 한다. 통계패키지에서는 상자그래프의 좌측선을 $\max(\text{최솟값}, Q_1 - 1.5 \cdot \text{IQR})$, 우측선을 $\min(\text{최댓값}, Q_3 + 1.5 \cdot \text{IQR})$ 으로 표시하기도 한다.



[그림 4.14] 상자그래프

- **변이계수**(coefficient of variation)는 표준편차를 평균으로 나눈값에 100을 곱하여 %를 계산한 것으로서 단위가 다른 두 종류의 데이터를 비교할 때 이용된다.

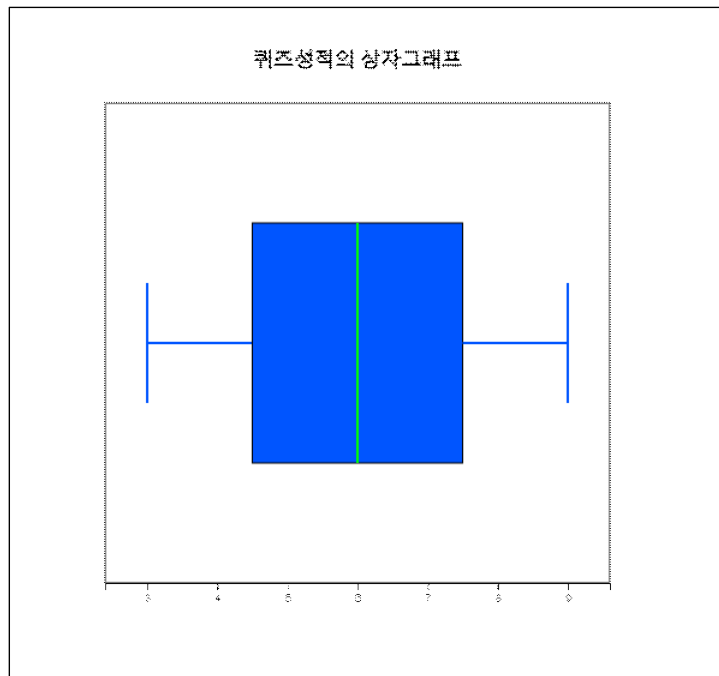
『eStat』을 이용한 측도의 계산

- 통계학 클래스의 7명 표본 학생을 대상으로 10점 만점인 퀴즈를 본 결과 5, 6, 3, 7, 9, 4, 8점 이었다. 이 데이터의 측도를 『eStat』을 이용하여 구해 보자.
- 『eStat』을 이용하여 평균과 중앙값을 구하려면 시트의 V1열에 데이터를 입력한 후 기초통계량 아이콘 을 클릭한다. 그러면 결과저장창에 [그림 4.15]와 같은 결과가 나타난다. 평균, 중앙값과 함께 표준편차, 최솟값, 최댓값, 일사분위수, 삼사분위, 범위, 사분위수범위, 변위계수를 계산하여 준다. 이때 분산, 표준편차, 변이계수는 데이터가 모집단인지(n공식) 또는 표본인지(n-1 공식)에 따라 분석자가 선택하여 사용한다.

기초통계량	분석변량 (V1)
자료수	7
결측수	0
평균	6.000
분산 (n)	4.000
분산 (n-1)	4.667
표준편차 (n)	2.000
표준편차 (n-1)	2.160
최솟값	3.000
1사분위수	4.500
중앙값	6.000
3사분위수	7.500
최댓값	9.000
범위	6.000
사분위수범위	3.000
변위계수 (n)	33.33 %
변위계수 (n-1)	36.00 %

[그림 4.15] 『eStat』을 이용한 측도의 계산

- 상자그래프 아이콘을 클릭하면 [그림 4.16]과 같은 그래프가 보여진다.



[그림 4.16] 퀴즈성적의 상자그래프

『eStatH』를 이용한 측도의 계산

- 『eStatH』 메뉴에서 ‘상자그래프 - 기초통계량’을 선택하면 [그림 4.17]과 같은 데이터 입력 화면이 나타난다. 여기에 데이터를 입력하면 그 즉시 기초통계량이 계산되고 [실행] 버튼을 클릭하면 [그림 4.18]과 같은 점그래프와 상자그래프가 나타난다. 이 그래프의 아래부분에 있는 점들은 마우스로 움직일 수 있어 통계량의 변화를 관찰할 수 있다.

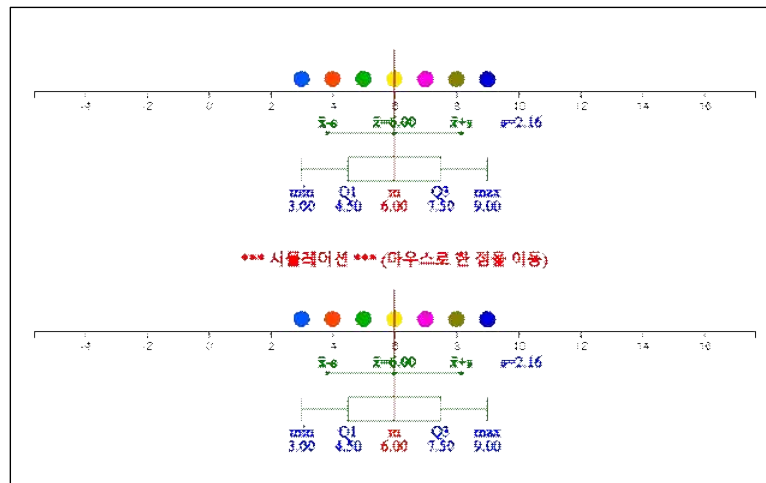
상자그래프 - 기초통계량 메뉴

[자료 입력]

[기초통계량]

자료수	n	=	<input type="text" value="7"/>	최솟값	min	=	<input type="text" value="3.00"/>
평균	μ, \bar{x}	=	<input type="text" value="6.00"/>	1사분위수	$Q1$	=	<input type="text" value="4.50"/>
모분산(n)	σ^2	=	<input type="text" value="4.00"/>	중앙값	m	=	<input type="text" value="6.00"/>
표본분산(n-1)	s^2	=	<input type="text" value="4.67"/>	3사분위수	$Q3$	=	<input type="text" value="7.50"/>
모집단 표준편차	σ	=	<input type="text" value="2.00"/>	최댓값	max	=	<input type="text" value="9.00"/>
표본 표준편차	s	=	<input type="text" value="2.16"/>	범위	$range$	=	<input type="text" value="6.00"/>
				사분위수범위	IQR	=	<input type="text" value="3.00"/>

[그림 4.17] 『eStatH』를 이용한 퀴즈성적의 기초통계량



[그림 4.18] 『eStatH』를 이용한 퀴즈성적의 상자그래프