7-1. 고급통계

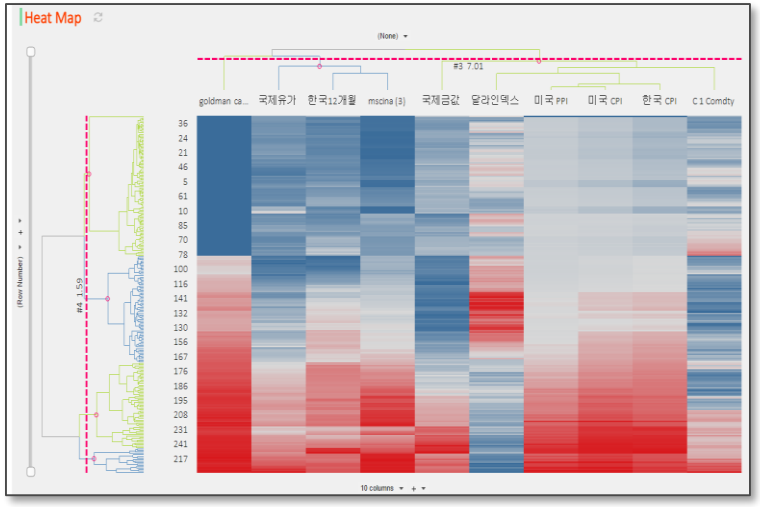
계층적 군집분석/라인 유사성/상자그래프

김 성 기

목 차

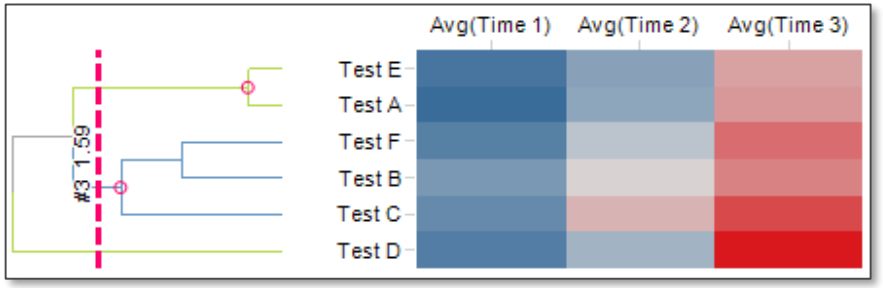
1. 계층적 군집 분석(Hierarchical Clustering)

- 처음에 **n**개의 군집으로부터 시작하여 점차 군집의 개수를 줄여나가는 방법으로 분석한다.
- 군집단계에 대한 그래프적 표현으로 계통수**(dendrogram)**를 이용하며, 유사성이 큰**(**가까운**)** 이웃을 병합하는 방법으로 군집의 단계를 보여준다.
- 계층적 군집분석 도구를 사용하는 경우 입력은 데이터 테이블이고 결과는 계통수가 있는 히트맵이다.
- 처음부터 계층적 군집분석 도구를 사용하여 시작할 수도 있고, 기존 히트맵으로부터**(**히트맵 속성의 '계통수' 이용**)** 계층적 군집분석을 사용할 수도 있다.



1. 계층적 군집 분석(Hierarchical Clustering)

- 계층적 군집분석은 '항목 간 거리' 또는 '유사성'에 따라 트리 모양 구조로 항목을 계층**(hierarchy)**으로 정리한다.
- 결과로 만들어진 계층을 그래픽으로 표시하면 계통수**(dendogram** 혹은 **phylogenetic tree)**라는 트리 구조의 그래프가 된다.
- **Spotfire**에서 계층적 군집분석 및 계통수는 히트맵 시각화와 밀접하게 연결되어 있으며, 히트맵에서 행과 컬럼을 모두 군집 분석할 수 있다.
  - 행 계통수에서는 행 사이의 거리 또는 유사성과 각 행이 속하는 노드를 군집분석의 결과로 표시한다.
  - 열 계통수에서는 변수**(**선택한 셀 값 컬럼**)** 사이의 거리 또는 유사성을 표시한다. 아래 예에는 행 계통수가 포함된 히트맵이 표시된다.

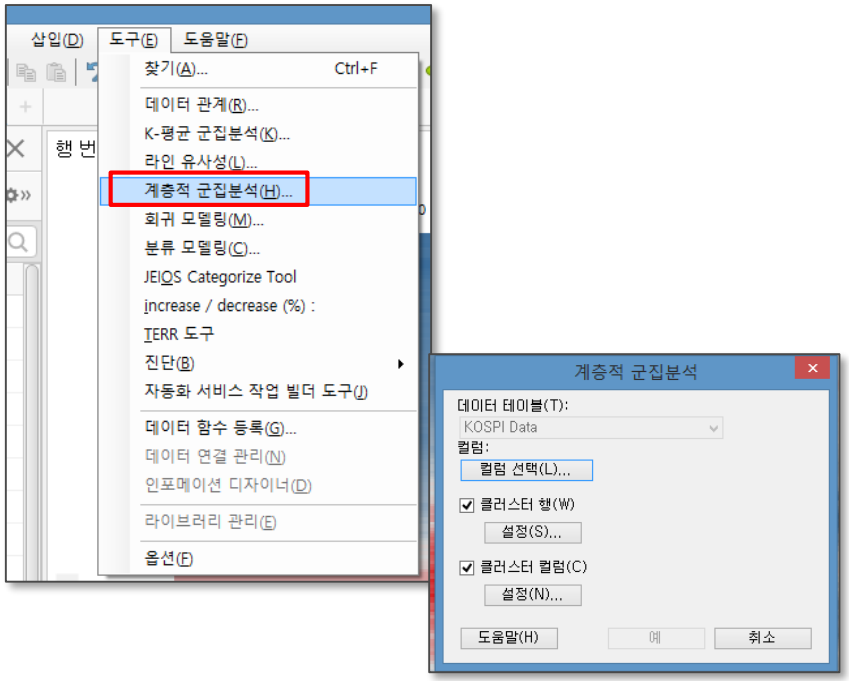


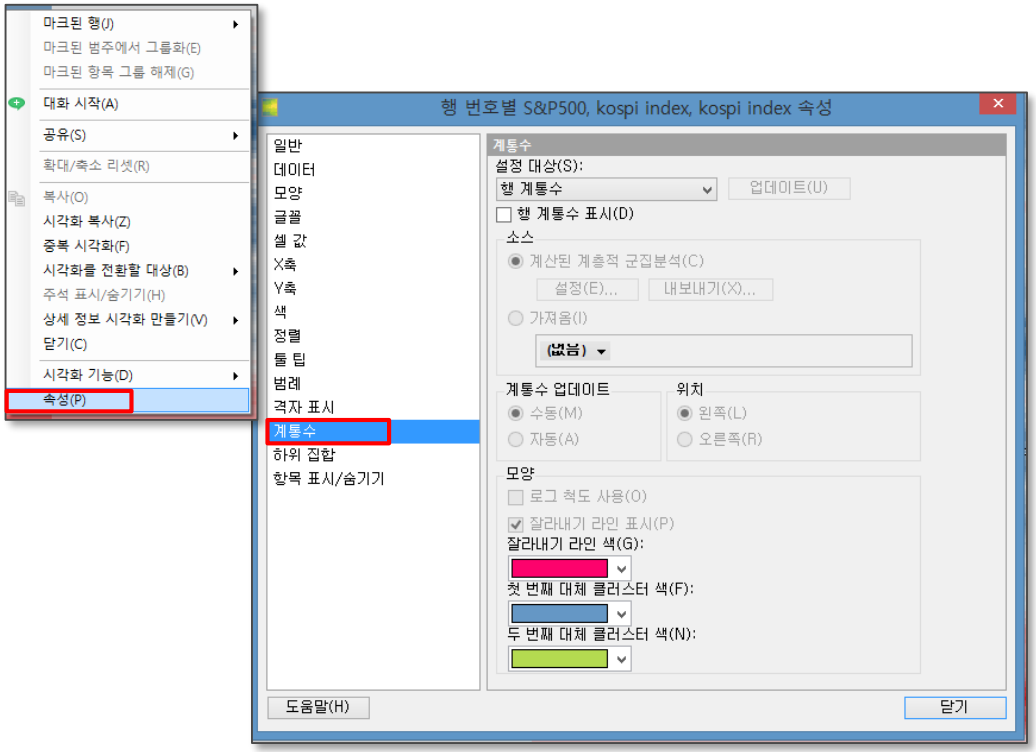
1. 계층적 군집 분석(Hierarchical Clustering)

- 계층적 군집분석은 두 가지 방법으로 수행할 수 있다.

  - 계층적 군집분석 도구 사용
    : 메인 메뉴 > 도구 > 계층적 군집분석

  - 기존 히트맵 시각화에서 계층적 군집분석
    : (히트맵 시각화 위에서) 마우스 우클릭 > 속성 > 계통수

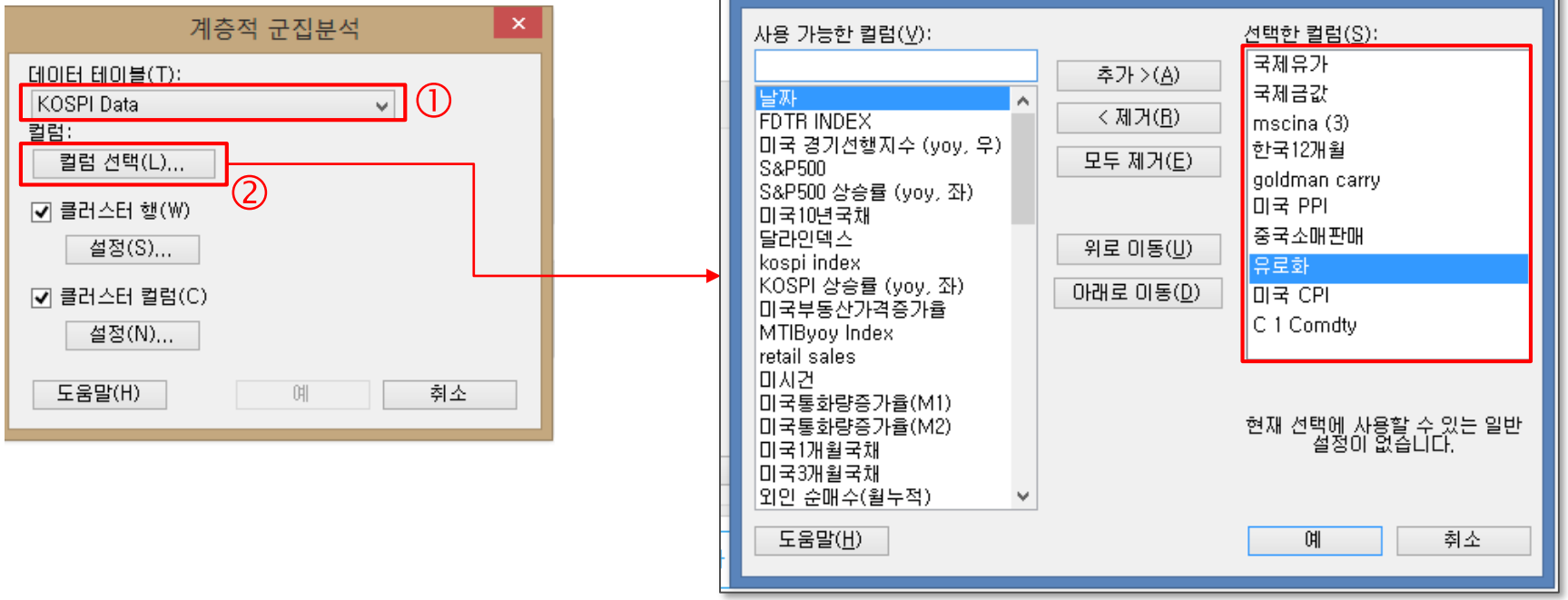




1. 계층적 군집 분석(Hierarchical Clustering)

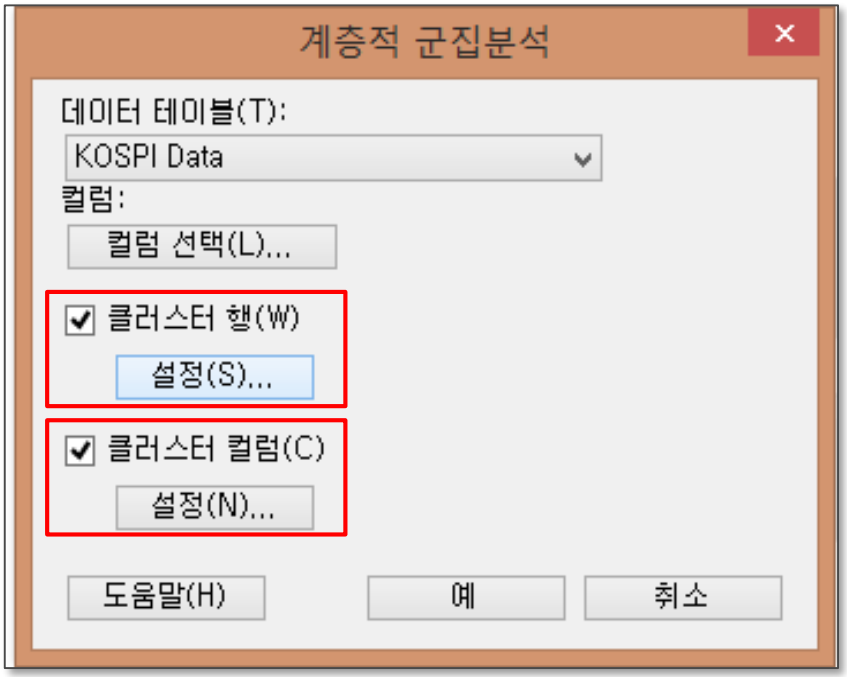
계층적 군집 분석 사용법(요약)

1. 메인 메뉴 > 도구 > 계층적 군집분석을 선택
2. 활성창에서 분석에 사용할 데이터테이블을 선택(①)하고 클러스터링에 사용한 컬럼을 선택(②)한다.



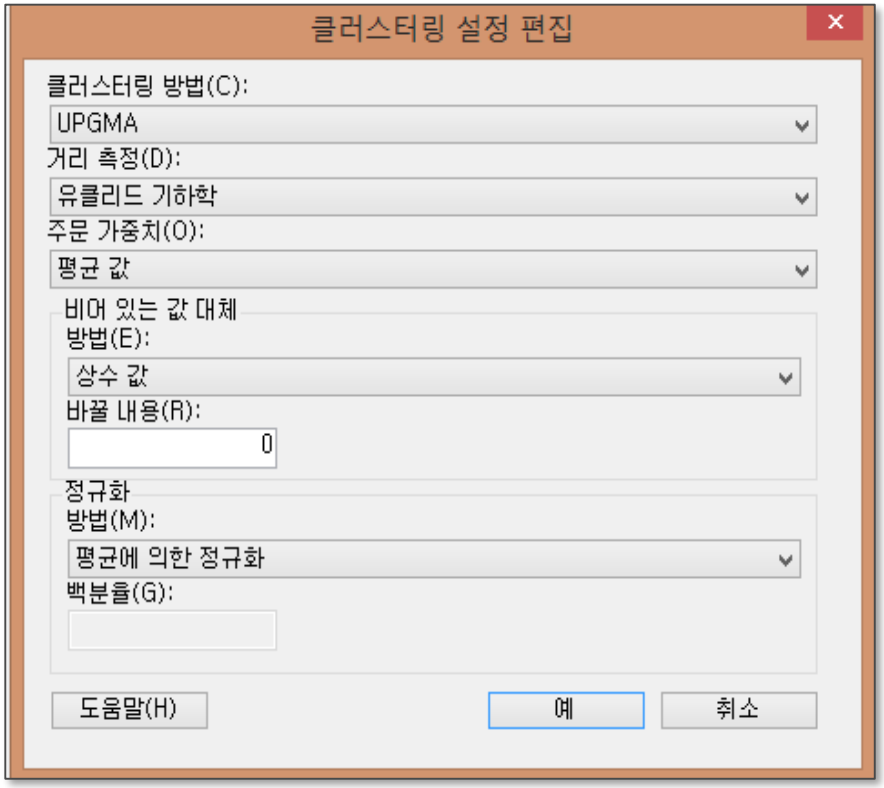
1. 계층적 군집 분석(Hierarchical Clustering)

3. 행과 열 중에서 어느 것을 클러스터링 할 것인지 유무를 체크하고, 각각 상세 옵션 설정을 위해 '설정' 버튼을 클릭한다.



1. 계층적 군집 분석(Hierarchical Clustering)

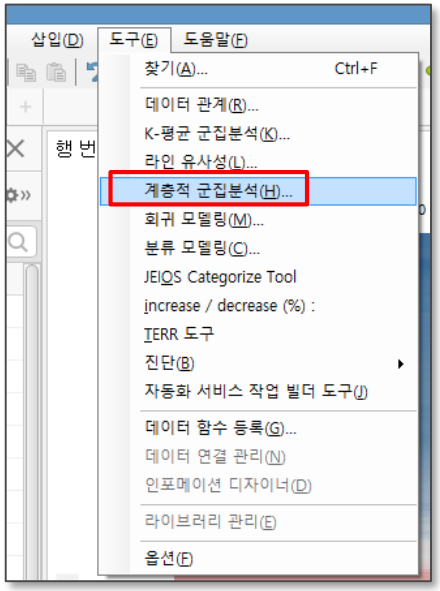
4. '설정' 활성창에서 세부 옵션을 선택한다.
   - **(1) 클러스터링 방법** : 선택된 거리 측정에서 나온 값을 이용하여 클러스터 간의 유사성 계산시 사용할 방법 선택
   - **(2) 거리 측정** : 행 또는 열의 가능한 모든 조합에 대한 관계값을 생성하기 위한 측정법 선택
   - **(3) 주문 가중치(Ordering weight)** : 클러스터링 후 덴드로그램에 표시될 순서를 결정하는 방법 선택
   - **(4) 비어있는 값 대체** : 클러스터링 수행시 사용될 데이터의 결측치 처리 방법 선택
   - **(5) 정규화(Normalization)** : 표준화 방법 선택



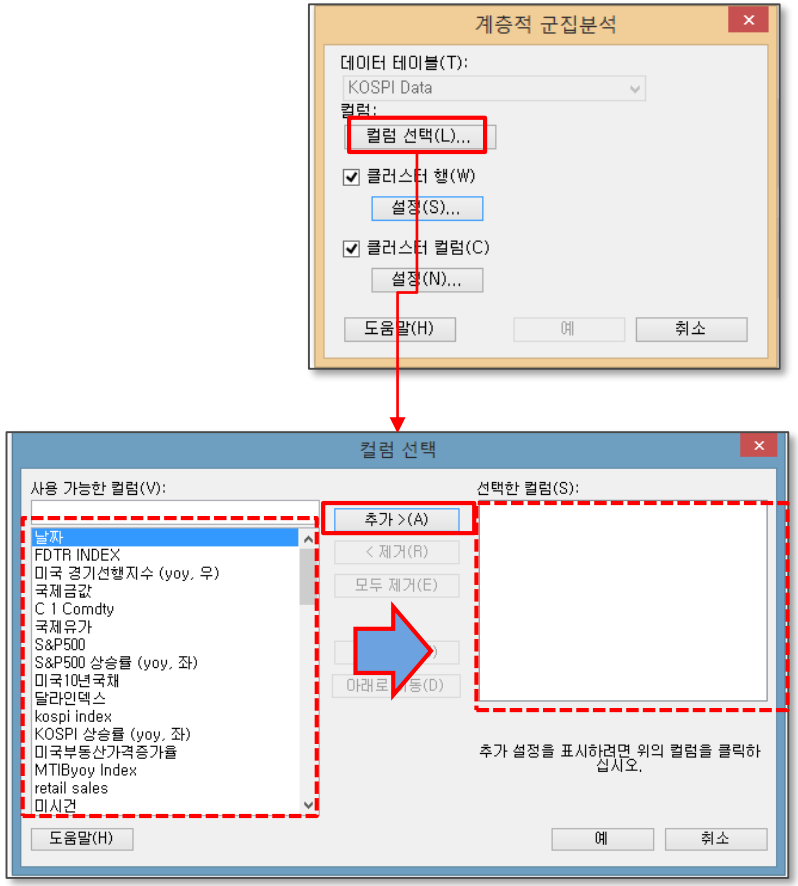
1. 계층적 군집 분석(Hierarchical Clustering)

1. 계층적 군집분석 도구를 사용할 때

1) 메인 메뉴 > 도구 > 계층적 군집분석을 선택한다.



2) 군집 분석에 사용할 중요한 컬럼들만 선택하여
'사용 가능한 컬럼' 칸에서
'선택한 컬럼' 칸으로 '추가' 한다.

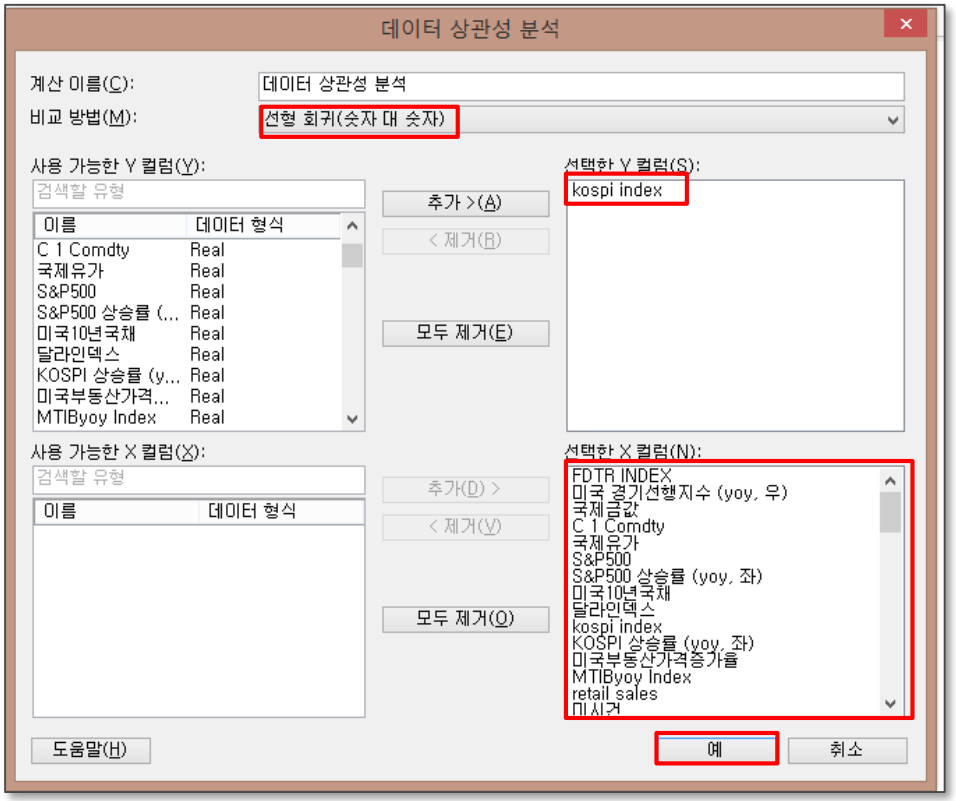


KOSPI Data.xls

1. 계층적 군집 분석(Hierarchical Clustering)

2-1) 군집 분석에 사용할 중요한 컬럼들을 선택할 때, 미리 중요한 **Key** 컬럼들을 알고 있다면 이들을 사용하면 되지만, 그렇지 못할때는 사용자가 해석하고자 하는 종속변수**(Y**값**)**를 감안하여 **Spotfire**에서 제공하는 '데이터 상관성 분석**(Data Relationship)**' 을 수행하는 것이 도움이 될 수 있다.

- 이 예제에서는 **Spotfire**의 데이터 상관성 분석 방법들 중에서 '선형 회귀' 방법을 통해서 **KOSPI** 지수**(**종속변수**; Y**컬럼**)**와 상관성이 높은 중요한 컬럼들을 찾아낸다.



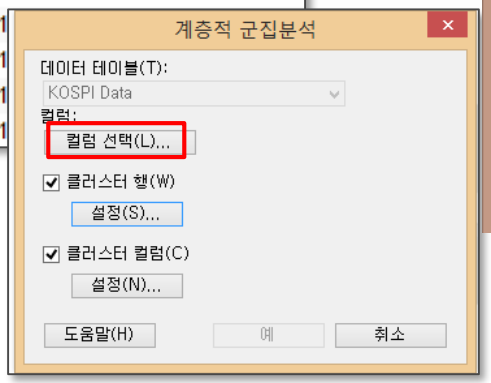
KOSPI Data.xls

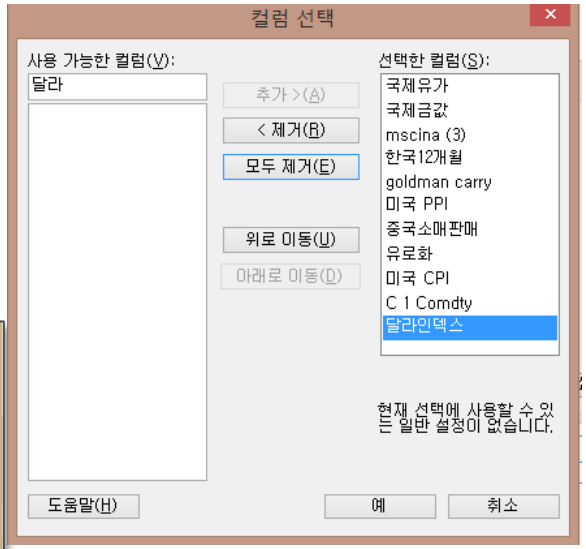
1. 계층적 군집 분석(Hierarchical Clustering)

2-2) 수행 결과 **p-value** 가 가장 작은 **X**컬럼**(**국제 유가**)**부터 차례로 검토하여 본다**.** 이러한 **X**컬럼에 있는 낮은 **p-value** 값들이 계층적 군집분석의 ‘컬럼 선택’ 에 참조될 수 있다**.**

데이터 상관성 분석 (선형 회귀)

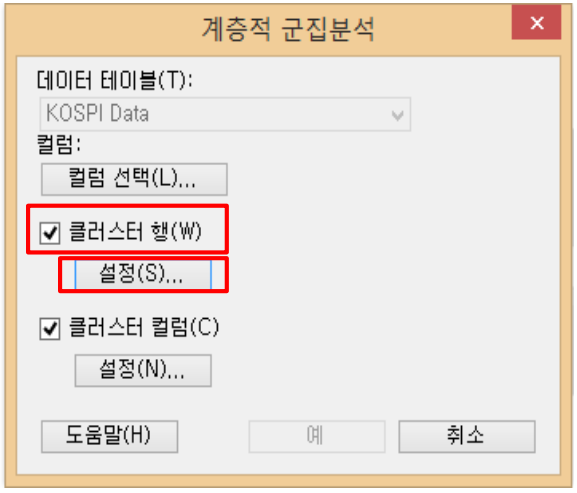
Y (numerical)	X (numerical)	p-value	FStat	RSq
kospi index	국제유가	1.27E-071	667.01	0.73
kospi index	국제금값	5.76E-069	622.38	0.72
kospi index	mscina (3)	4.57E-050	362.31	0.60
kospi index	한국12개월	2.43E-046	321.16	0.57
kospi index	goldman carry	6.98E-046	437.04	0.75
kospi index	미국 PPI	3.52E-040	259.19	0.52
kospi index	중국소매판매	2.47E-036	224.36	0.48
kospi index	유로화	7.15E-036	296.48	0.69
kospi index	미국 CPI	1.15E-031	184.67	0.43
kospi index	C 1 Comdty	1.87E-029	167.22	0.41
kospi index	달라인덱스	2.56E-027	1	
kospi index	한국 CPI	3.83E-027	1	
kospi index	달러/유로 환율	1.09E-026	1	
kospi index	한국 PPI	7.84E-024	1	



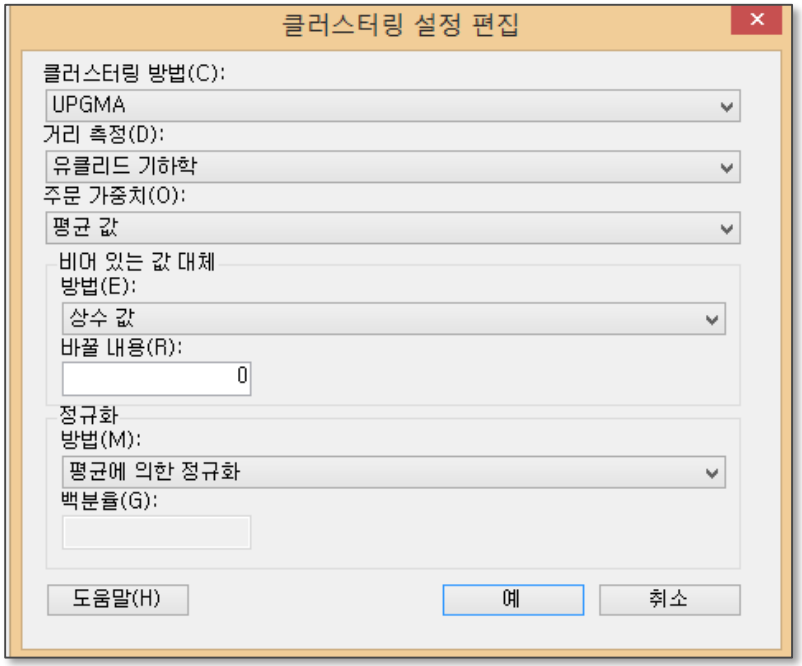


1. 계층적 군집 분석(Hierarchical Clustering)

3) 만일 데이터 의 **행 혹은 컬럼**들에 대하여 군집 분석을 하고자 한다면 ‘클러스터 행’이나 ‘클러스터 컬럼’을 체크하고 설정’ 버튼을 누른다.

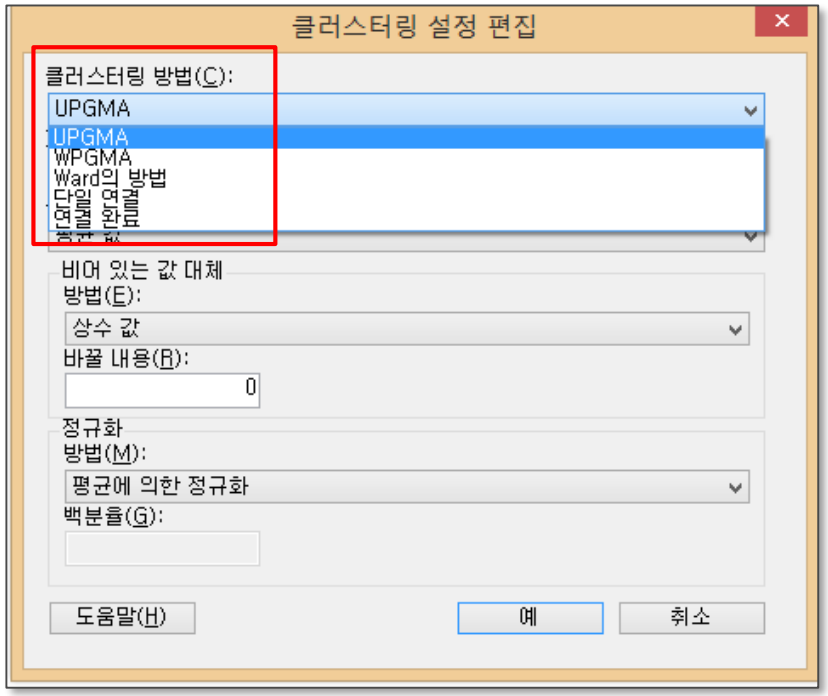


4) 클러스터링에 필요한 설정 편집 창이 나타난다.

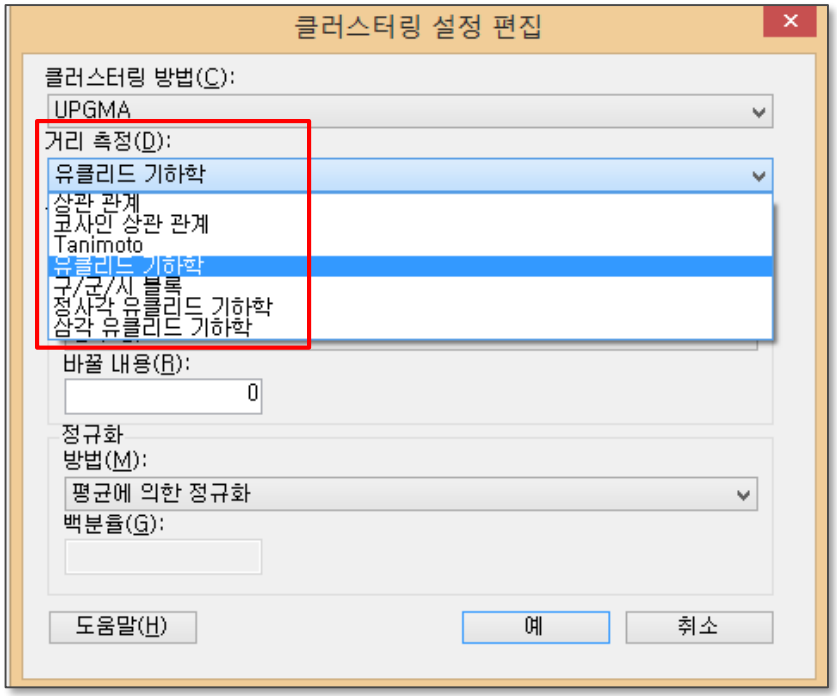


1. 계층적 군집 분석(Hierarchical Clustering)

5) 클러스터링 방법**(Clustering method)**을 선택한다. **Default** 옵션은 **'UPGMA'** 이다.

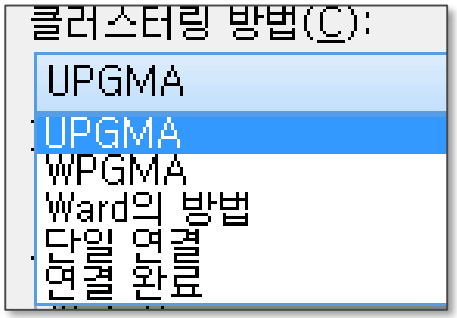


6) 거리 측정**(Distance measure)** 방법을 선택한다. **Default** 옵션은 **'유클리드 기하학(Euclidean)'** 이다.



1. 계층적 군집 분석 – 클러스터링 방법 개요

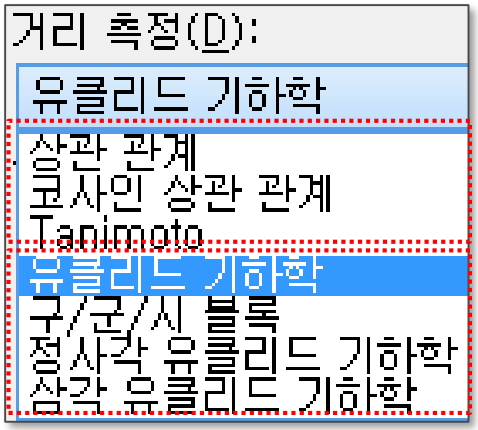
- 계층적 군집분석은 선택한 거리 측정을 사용하여 두 행 또는 컬럼의 모든 가능한 조합 간의 거리를 계산함으로써 시작된다.
- 그런 다음, 계산된 거리는 군집분석 중 행 또는 컬럼으로 형성되는 모든 클러스터 간 거리를 얻는 데 사용된다. 다음에 나오는 군집분석 방법 중 하나를 선택할 수 있다.



UPGMA	Unweighted Pair-Group Method with Arithmetic mean
WPGMA	Weighted Pair-Group Method with Arithmetic mean
Ward의 방법	군집내 제곱합 증분과 군집간 제곱합을 고려하여, 제곱의 증분 합계를 계산한다 .
단일 연결 (Single Linkage)	최단 연결법(nearest neighbor)이라고도 함. 두 군집간의 거리를 최단 거리로 하여 군집간 거리로 정의한다. 즉, 클러스터 간의 거리는 클러스터에서 가장 가까운 행(또는 컬럼) 간의 거리와 동일하다.
연결 완료 (Complete Linkage)	최장 연결법(farthest neighbor)이라고도 함. 두 군집간의 거리를 최장 거리로 하여 군집간 거리로 정의한다. 즉, 두 클러스터 간의 거리는 클러스터에서 가장 멀리 떨어진 행(또는 컬럼) 간의 거리와 동일하다.

1. 계층적 군집 분석 – 거리 측정 방법 개요

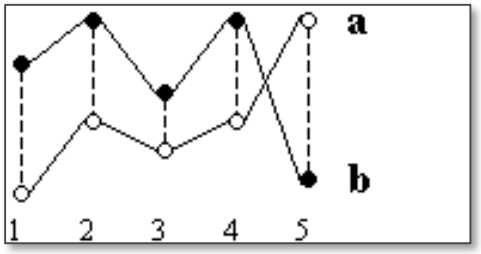
- 행 또는 컬럼 간 거리 또는 유사성을 계산하는 데 다음 방법들을 사용할 수 있다.
- 군집분석 계산의 결과는 클러스터링된 행 또는 컬럼 간 유사성 또는 거리로 제공된다.
  - **유클리드 기하학, 구/군/시 블록, 정사각 유클리드 기하학, 삼각 유클리드 기하학**
    : 행 또는 컬럼 간 거리를 제공
  - **상관 관계, 코사인 상관 관계 및 Tanimoto**
    : 행 또는 컬럼 간의 유사성으로 제공



1. 계층적 군집 분석 – 거리 측정 방법

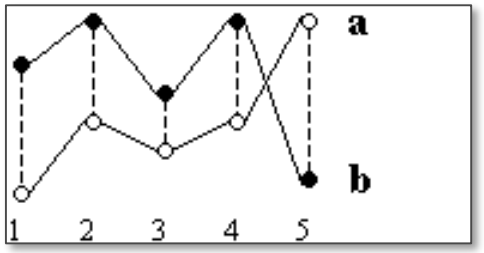
유클리드 기하학(Euclidean)
: 행 또는 컬럼 간 거리를 제공

- 유클리드 거리는 항상 **0**보다 크거나 같아야 한다. 동일한 지점에 대해서는 측정치가 **0**이 되고 유사성이 거의 없는 지점에 대해서는 측정치가 높게 나타난다.
- 아래 그림은 ***a*** 및 ***b*** 지점의 예를 보여 준다. 각 지점은 **5**개의 값으로 설명된다. 그림에서 점선은 거리 **(a_1-b_1), (a_2-b_2), (a_3-b_3), (a_4-b_4)** 및 **(a_5-b_5)**이다.



구/군/시 블록 거리(City Block)
: 행 또는 컬럼 간 거리를 제공

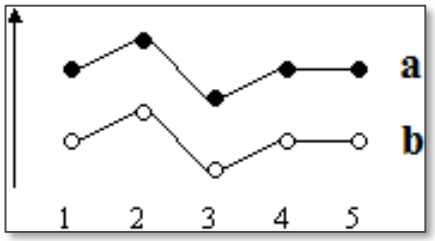
- *구/군/시 블록 거리*는 항상 **0**보다 크거나 같아야 한다.
- 대부분의 경우, 이 거리 측정은 *유클리드 거리* 와 유사한 결과를 제공한다. 그러나 *구/군/시 블록 거리*의 경우, 거리가 제곱으로 계산되지 않으므로 단일 치수로 표시되는 큰 차이의 영향은 적다.



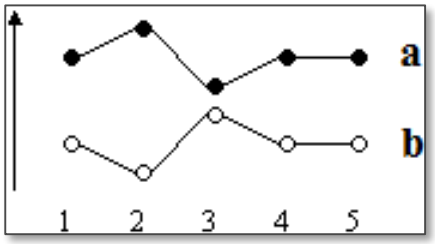
1. 계층적 군집 분석 – 거리 측정 방법

상관 관계(Corelation relation)
: 행 또는 컬럼 간의 유사성으로 제공

- 이 상관 관계는 **'Pearson**의 상관 관계' 또는 **'Pearson**의 **r'** 이라고도 한다. 범위는 **+1 ~ -1**이고, **+1**이 가장 높은 상관 관계이다. 정반대인 지점은 상관 관계가 **-1**이다.



a 및 ***b***가 동일하면 최고의 상관 관계이다.



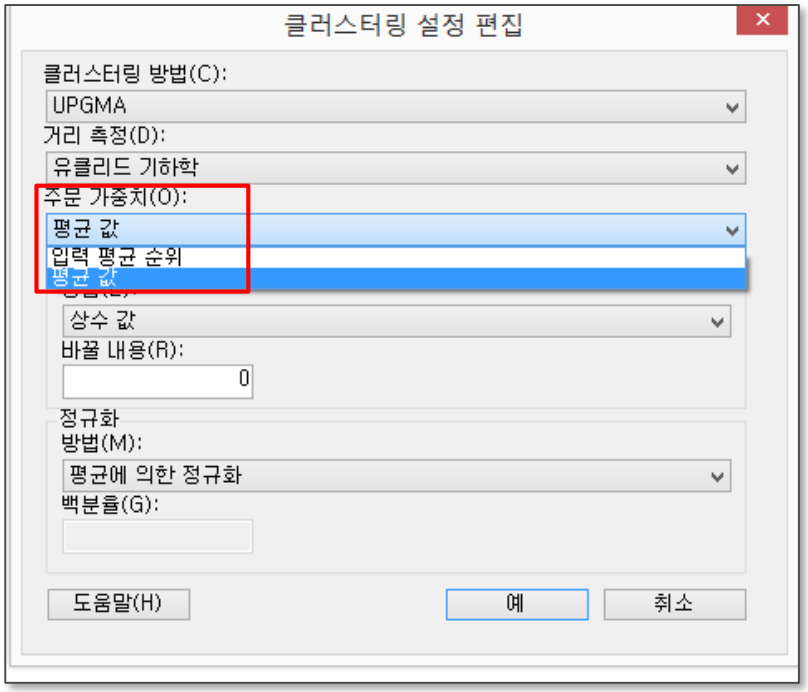
a 및 ***b***가 완벽하게 미러링되면 최고의 음수 상관 관계를 나타낸다.

코사인 상관 관계 (Cosine correlation)
: 행 또는 컬럼 간의 유사성으로 제공

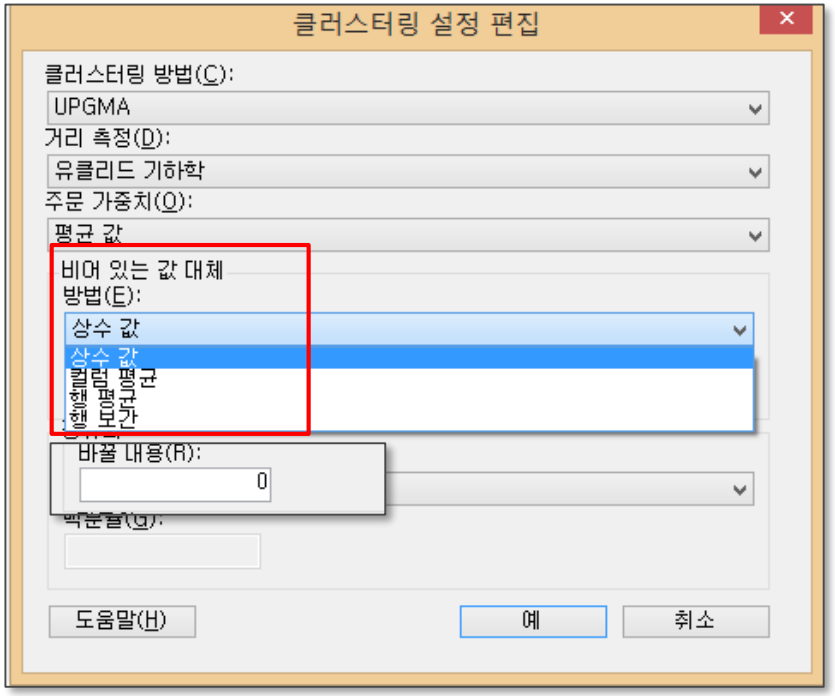
- 코사인 상관 관계의 범위는 **+1 ~ -1**이고, **+1**이 가장 높은 상관 관계다.
- 정반대인 지점은 상관 관계가 **-1**이다.
- 코사인 상관 관계 및 상관 관계의 차이점은 상관 관계에서 평균값을 뺀다는 점이다.

1. 계층적 군집 분석(Hierarchical Clustering)

7) 주문 가중치**(Ordering weight)** 방법을 선택한다. **Default** 옵션은 ‘평균값’이다.



8) 비어있는 값 대체**(Empty Value replacement)** 방법을 선택한다. **Default** 옵션은 ‘상수값 **(Constant value)**’이다.



1. 계층적 군집 분석 – 주문 가중치 (ordering weight) 방법

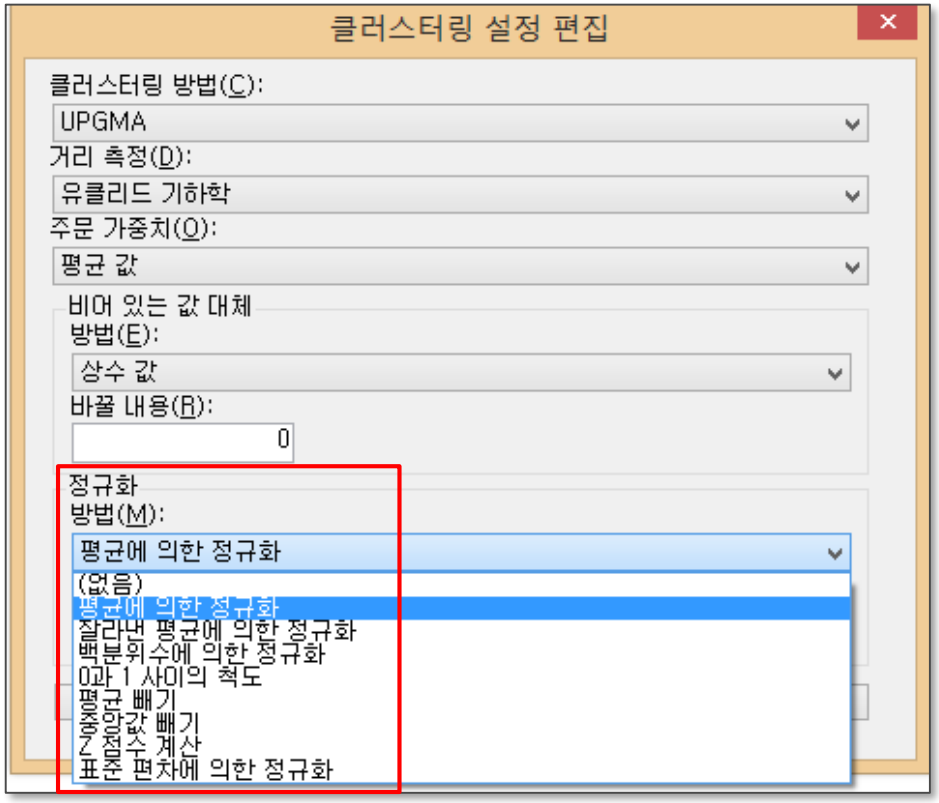
- 주문 가중치는 행 계통수에서 표시되는 세로 순서를 관리한다.
- 컬럼 계통수의 경우 컬럼의 가로 순서를 관리한다.
- 한 클러스터(항상 정확히 두 개 하위 클러스터가 있음)에 속한 두 개의 하위 클러스터가 가중되고 가중치가 낮은 클러스터가 다른 클러스터 왼쪽 위에 배치된다. 가중치는 다음 중 한 가지이다.
  - 행(또는 컬럼)의 입력 평균 순위. 이는 **Spotfire**로 행(또는 컬럼)을 가져오는 순서이다.
  - 행(또는 컬럼)의 평균 값.

*** 계층적 군집 분석 알고리즘**

Spotfire에서 계층적 군집분석에 사용되는 알고리즘은 계층적 응집 방법이다. 행 클러스터링의 경우, 클러스터 분석은 개별 클러스터에 배치된 각 행에서 시작된다. 그런 다음, 두 행의 가능한 모든 조합 간 거리가 선택한 '거리 측정' 을 사용하여 계산된다. 가장 유사한 두 개의 클러스터가 그룹화되어 새 클러스터를 형성한다. 이후 단계에서는 새 클러스터와 모든 나머지 클러스터 사이의 거리가 선택된 '군집분석 방법' 을 사용하여 다시 계산된다. 따라서 단계를 반복할 때마다 클러스터 수가 하나씩 감소된다. 결과적으로 모든 행이 하나의 대형 클러스터에 그룹화된다. 계통수**(dendogram)**의 행 순서는 선택한 '주문 가중치' 로 정의된다. 클러스터 분석은 컬럼 클러스터링과 같은 방법으로 수행된다.

1. 계층적 군집 분석(Hierarchical Clustering)

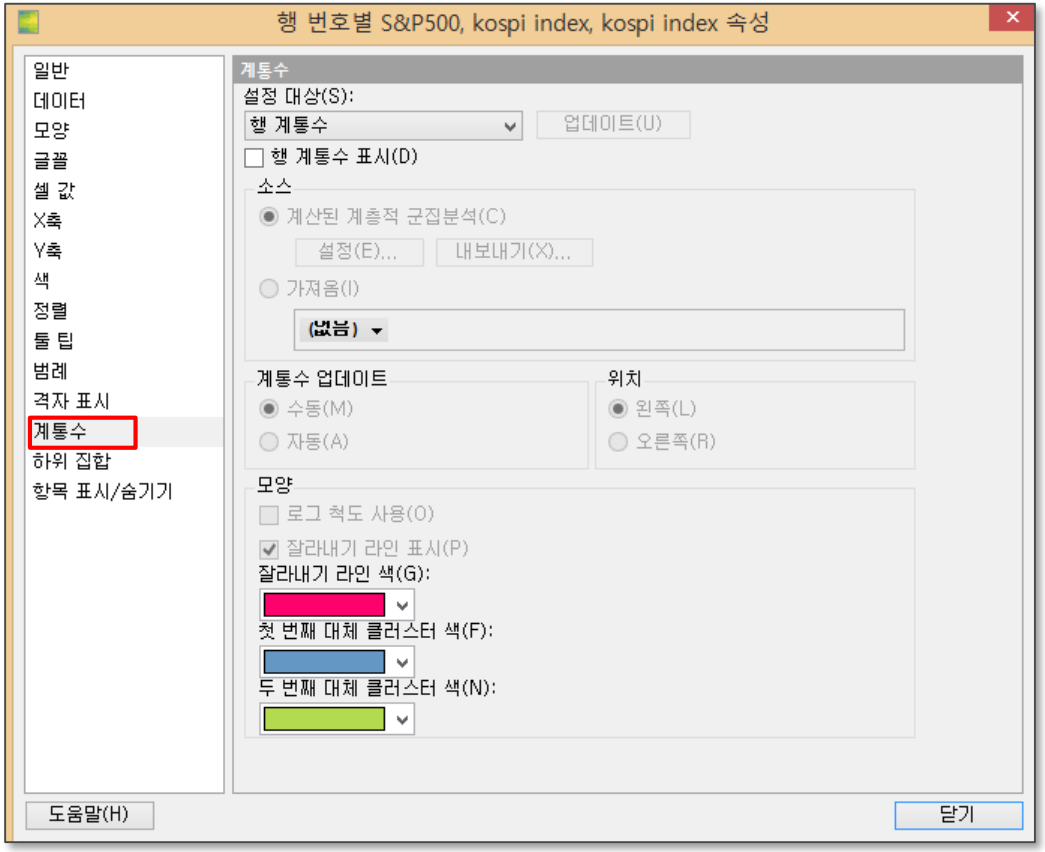
9) 정규화 방법**(Normalization)**을 선택한다.
Default 옵션은 ‘평균에 의한 정규화**(Normalize by mean)’** 이다.



1. 계층적 군집 분석(Hierarchical Clustering)

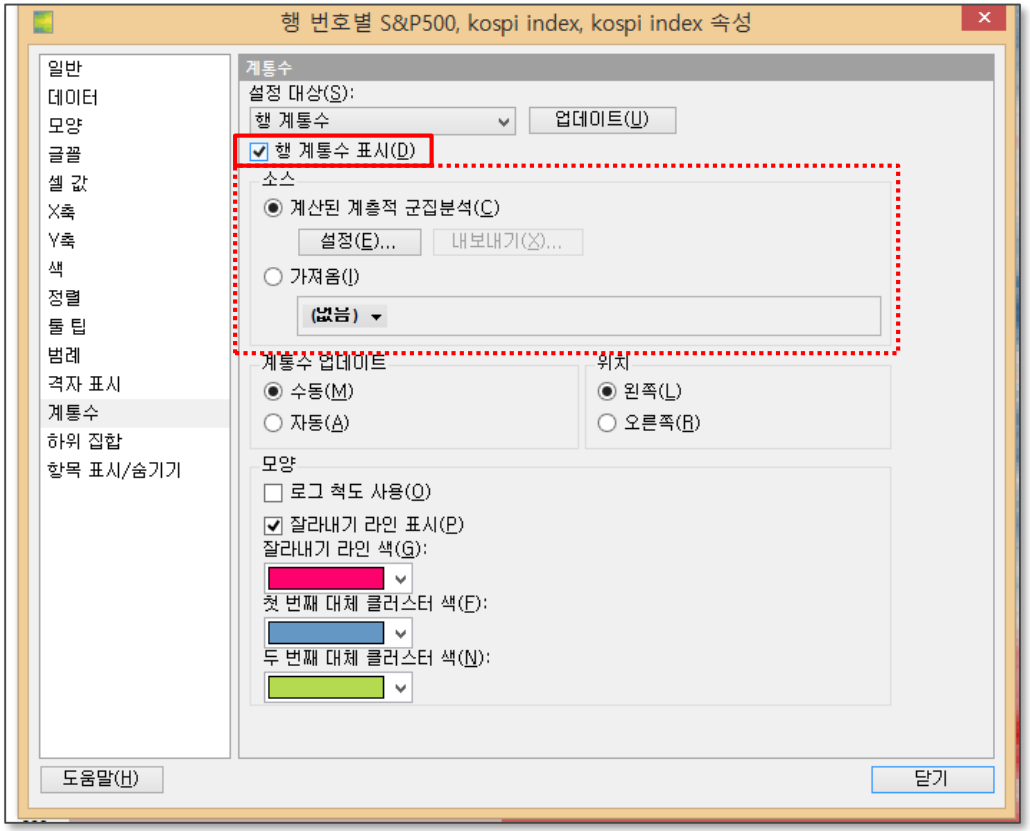
2. 작성해 놓은 히트맵으로부터 속성을 사용할 때

1) 히트맵에서 마우스 우클릭 > 속성 > 계통수를 선택한다.



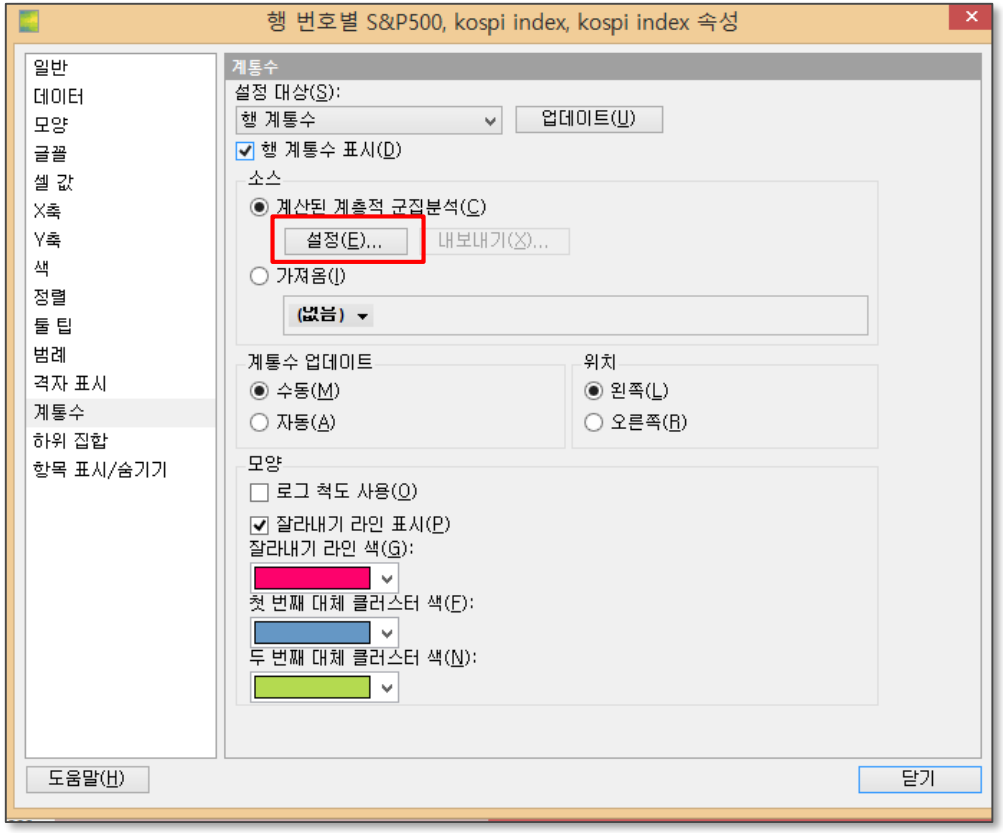
1. 계층적 군집 분석(Hierarchical Clustering)

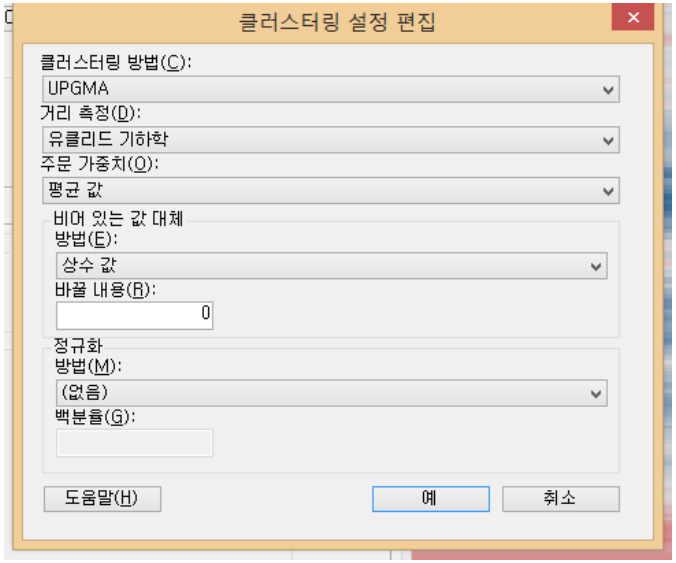
2) ‘행 계통수 표시’의 체크박스를 선택하면 그 아래의 ‘소스’ 칸이 활성화되면서 설정이 가능해진다.



1. 계층적 군집 분석(Hierarchical Clustering)

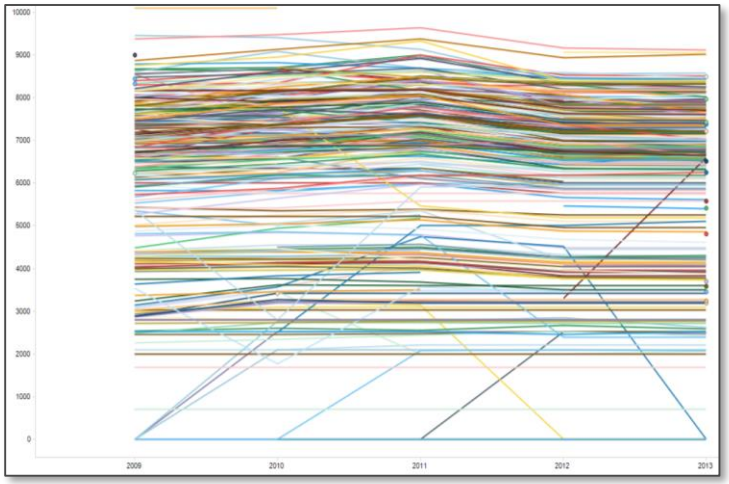
3) ‘계산된 계층적 군집 분석’ 아래 ‘설정’ 버튼을 누르면 ‘클러스터링 설정 편집’ 창이 나타난다. 이 부분의 사용 방법은 이전의 설명들과 동일하다.

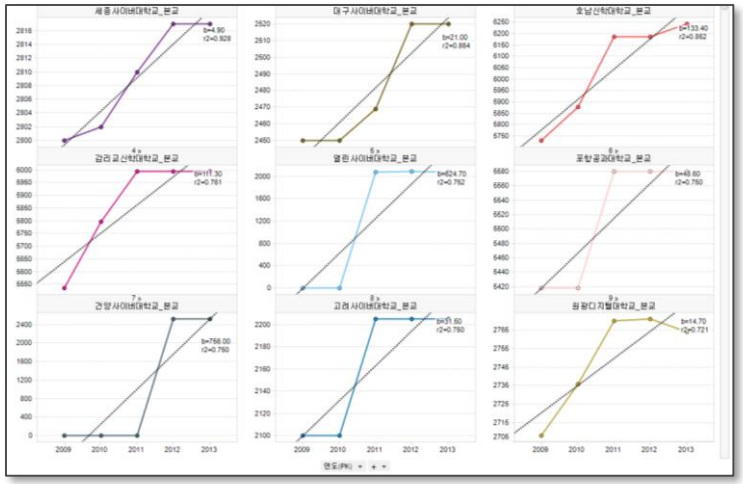




2. 라인 유사성(Line Similarity)

- 라인 형태의 차트에서 여러 라인들간의 유사성을 분석하여 특정한 추세나 유사한 라인을 우선순위로 서열을 생성하여 주는 기능이다
- **Spotfire**에서는 ‘선 그래프’와 ‘평형좌표 그래프’에서 사용할 수 있다.
- 라인 유사성을 실행하면 **Output**이 시각화 형태로 자동 생성되지 않는다. 대신에 이용한 데이터 테이블컬럼에 ‘라인 유사성’ 과 ‘라인 유사성**(rank)**’라는 **2**개의 컬럼들이 새로 생성된다.
- 따라서 **Spotfire**의 선 그래프에서, 결과로 도출된 **Line Similarity (rank)**라는 **Column**을 ‘격자 표시’ 기능을 활용하면 어느 선**(**항목**)**이 가장 원하는 추세나 유사한 라인인지 쉽게 이해 할 수 있다.

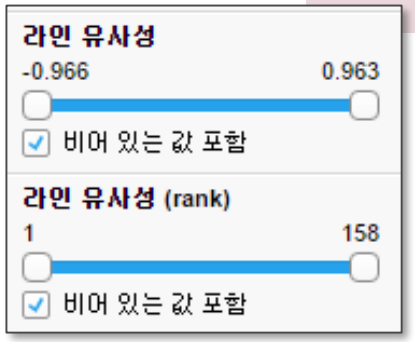


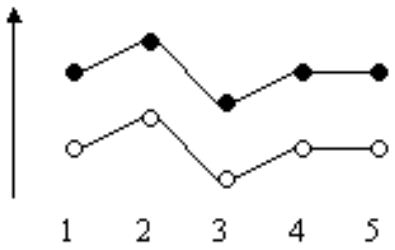


2. 라인 유사성(Line Similarity)

라인 유사성(Similarity) 계산 :

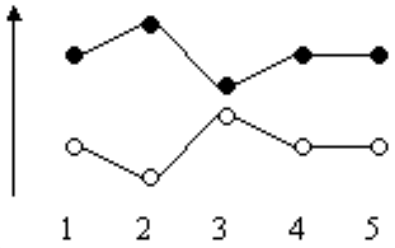
- 결과로 라인유사성 상관 계수 (r)를 산출
- -1 ~ +1 사이의 범위





동일한 형태를 갖는 프로파일은 최대 상관계수를 갖는다.

$r = +1$

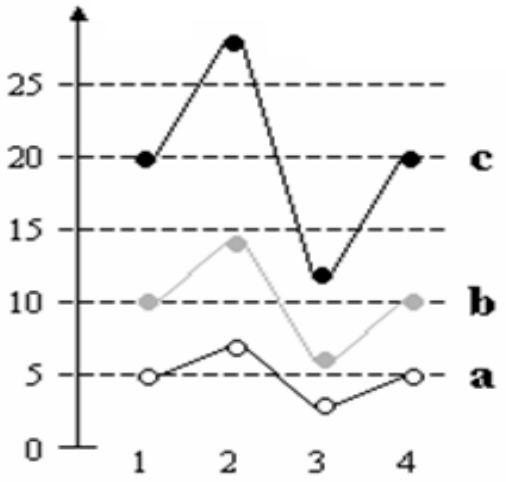


완전히 대칭인 프로파일은 최대의 음의 상관계수를 갖는다.

$r = -1$

a, b, c *모두*

$r = +1$



2. 라인 유사성(Line Similarity)

라인 유사성은 몇 가지 제약 조건이 있다. 계산의 기준이 되는 적절한 선 그래프를 만들지 않았다면 라인 유사성 도구를 사용할 수 없다.

- 여러 개의 **Y**축 척도를 사용할 수 없다.
- 연속적이고 저장함**(binned)**에 저장된 **X**축을 사용할 수 없다.
- 선 그래프에서 비교하고자 하는 모든 선에서 **X**축의 시작점 및 끝점이 같아야 한다.

2. 라인 유사성(Line Similarity)

- **Spotfire**의 라인 유사성은 2가지 목적으로 사용할 수 있다.

1. **특정한 추세에 가장 유사한 라인 형태 찾기**

   예) 12개월동안 매출이 가장 심한 증가 추세를 보이는 대리점은?

   과거 5년간 등록금이 가장 가파르게 오르고 있는 대학은?

   우리회사의 전체 제품들 중에서 가장 매출 하락세가 심한 영업 품목은?

2. **특정 라인과 가장 유사한(동일한) 라인 찾기**

   예) 우리 회사에서 가장 영업 실적이 좋은 영업사원과 동일한 영업 실적 패턴을 보이는 사원들은?

   예) 여러 대학 중에서 과거 5년간 우리 대학의 취업률과 가장 동일한(유사한) 취업률 패턴을 보이고 있는 대학들은?

2. 라인 유사성(Line Similarity)

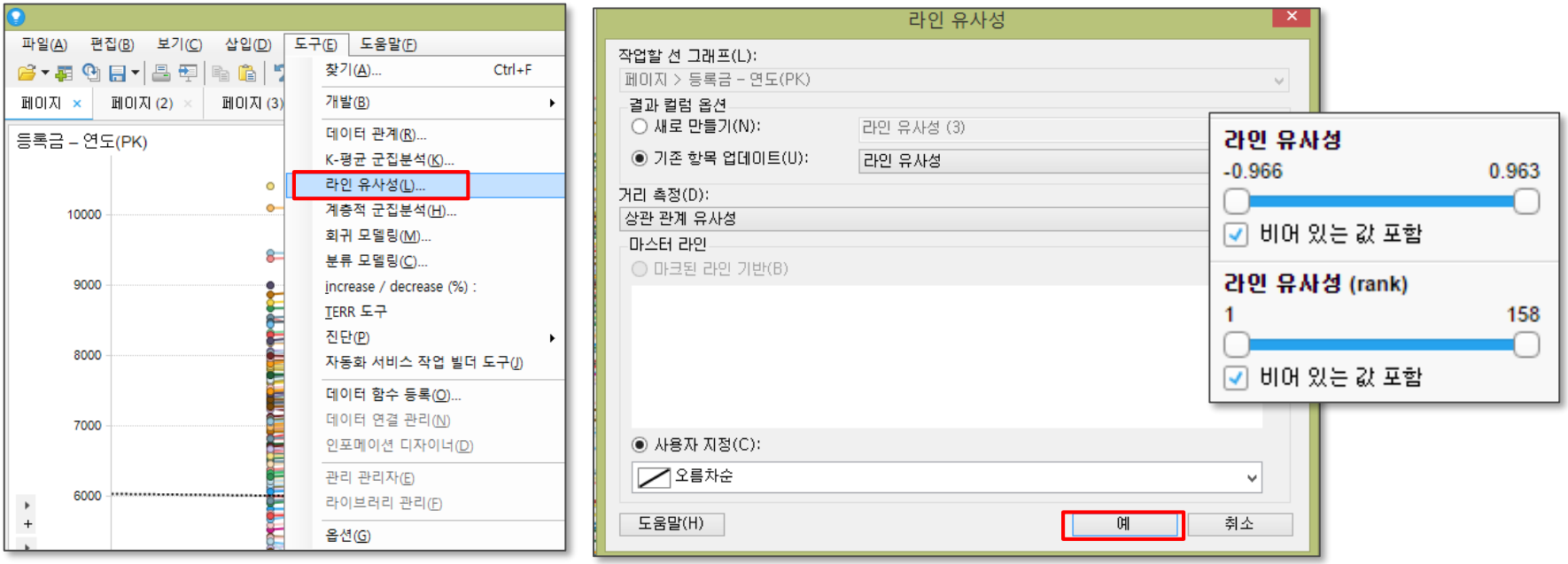
1. **특정한 추세에 가장 유사한 라인 형태 찾기 실행 방법**

   1) 먼저 선 그래프 시각화를 생성하고 필요한 설정(X축, Y축, 색, 선 지정 등)을 한다.

   2) 메인 메뉴에서 '도구' > '라인 유사성' 을 선택하면 설정 창이 표시된다.

   3) 필요한 설정을 변경한 후 '예'를 누른다.

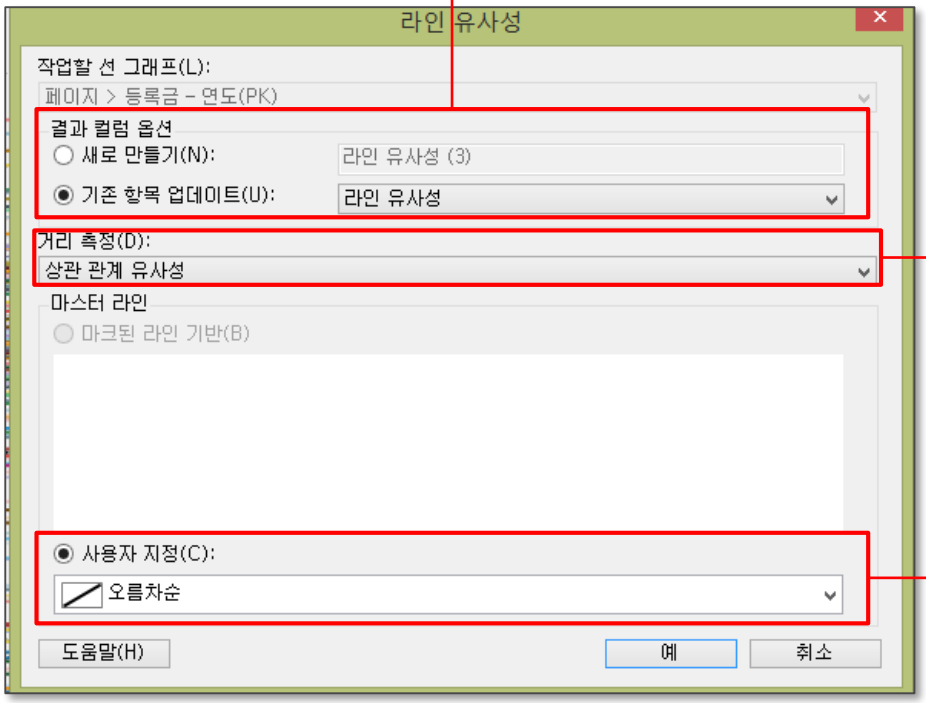
      실행 결과 : 사용자가 지정한 패턴과 가장 유사한 패턴을 갖는 라인부터 1위로 순위가 정해지는 컬럼(rank)과 필터가 생성된다.

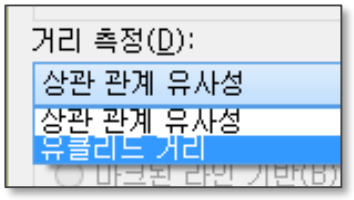


2. 라인 유사성(Line Similarity)

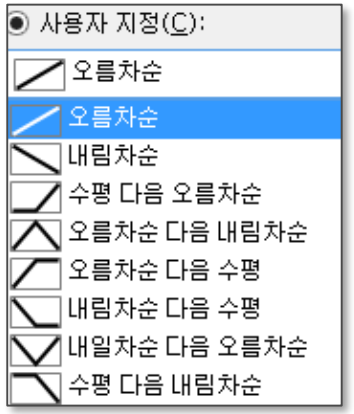
라인 유사성을 실행 후 생성되는 **2**개의 컬럼들을, 기존에 생성한 컬럼들 이외에 새로 추가할 것인지 아니면 기존 결과물에 **update** 해서 덮어 쓸 것인지 결정하는 항목

유사성 계산의 기반으로 사용할 거리 측정 방법을 선택한다. **(**알고리즘 상세 내용은 이전 **15page** 참조**)**





Spotfire에서 미리 정의해 놓은 패턴으로서, 이중에서 사용자가 원하는 패턴을 선택한다.



대학정보공시-주요지표(통합)-대학.txt

2. 라인 유사성(Line Similarity) - 실습

- 문제 : 학생수가 **1,100**명 이하인 대학들 중에서 **2009~2013**년 5년간 등록금이 가장 가파르게 오르고 있는 대학 Top 20은?

  * 힌트 :

    - line chart(색 지정 기준 : 학교명, Y축에 다중 척도로 표시)
    - 라인 유사성
    - 격자 표시
    - Line & Curve(추세선/직선 맞춤)

2. 라인 유사성(Line Similarity)

2. 특정 라인(master line)과 가장 유사한(동일한) 라인 찾기 실행 방법

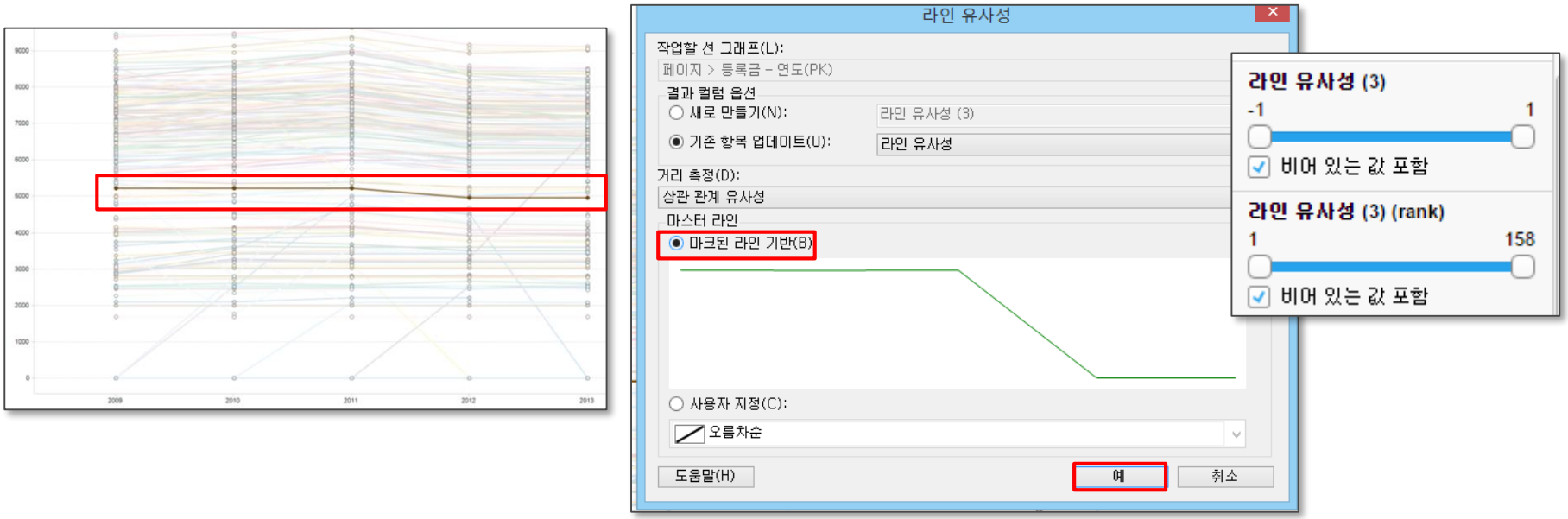
1) 먼저 선 그래프 시각화를 생성하고 필요한 설정(X축, Y축, 색, 선 지정 등)을 한다.

2) 선 그래프에서 사용자가 대상으로 삼고 싶은 한 개의 선(라인)만 선택한다.*

3) 메인 메뉴에서 '도구' > '라인 유사성' 을 선택하면 설정 창이 표시된다.

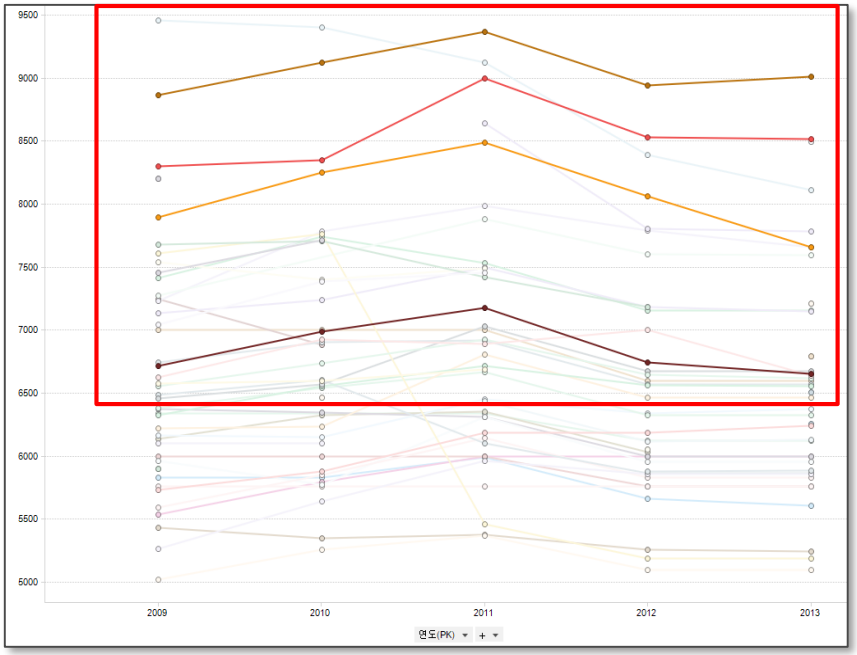
4) '마스터 라인'에서 '마크된 라인 기반'의 체크박스를 선택하고 '예'를 누른다.

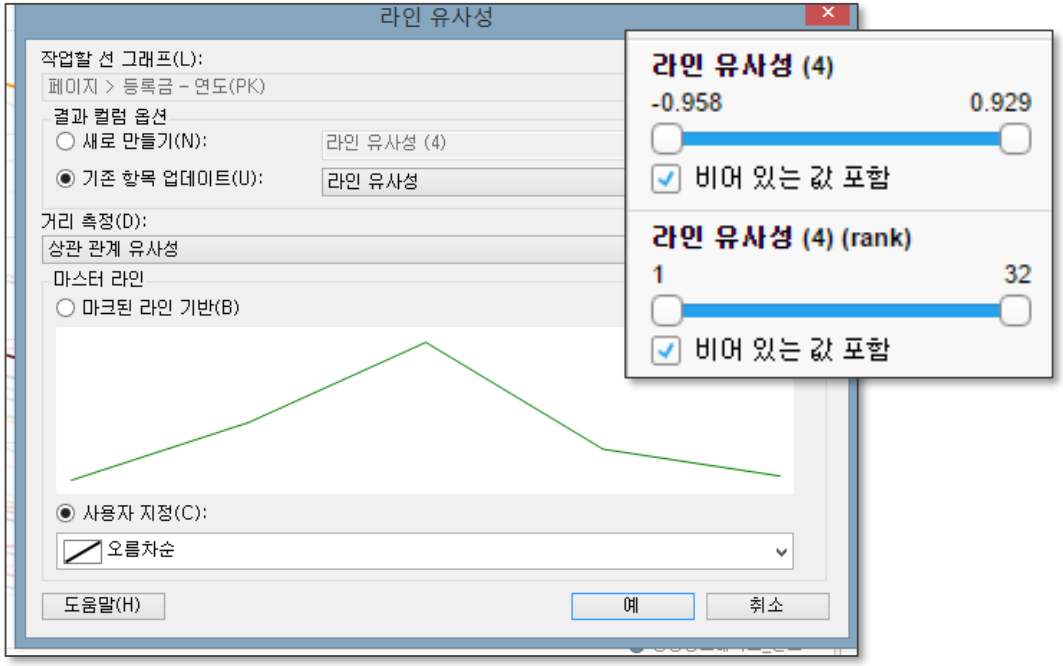
실행 결과 : 마크된 라인과 가장 유사한 패턴을 갖는 라인부터 1위로 순위가 정해지는 컬럼(rank)과 필터가 생성된다.



2. 라인 유사성(Line Similarity)

* '2)번 단계'에서 선 그래프에서 사용자가 대상(master line)으로 하고싶은 선(라인)이 여러 개일 경우에, 원하는 선들을 모두 선택하여 이용할 수 있다.(ctrl 사용)
  이 경우에는 선택한 선들의 평균이 마스터 라인 기반으로 지정된다.(아래 그림 참조)





대학정보공시-주요지표(통합)-대학.txt

2. 라인 유사성(Line Similarity) - 실습

- 문제 :

1. 전체 대학들 중에서 2009~2013년 5년간 '강원대학교_본교'와 등록금의 추세가 가장 유사한 대학 Top 20은?

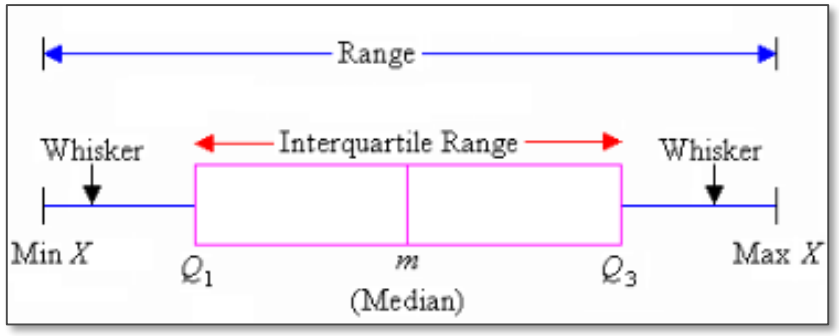
* 힌트 :

- line chart(색 지정 기준 : 학교명, Y축에 다중 척도로 표시)
- 라인 유사성
- 격자 표시
- Line & Curve(추세선/직선 맞춤)

2. 전체 대학들 중에서 2009~2013년 5년간 '강원대학교_본교'와 등록금의 추세가 가장 반대인 대학 Top 5는?

3. 상자 그래프(Box Plot)

- 상자 그래프의 정확한 명칭은 skeletal box-and-whisker plot이다.



출처 :
http://techntalk.tistory.com/entry/%EB%B0%95%EC%8A%A4%ED%94%8C%EB%A1%AF-Box-Plot%EA%B3%BC-%EC%A0%95%EA%B7%9C%EB%B6%84%ED%8F%ACnormal-distribution%EC%9D%98-%EA%B4%80%EA%B3%84-%EB%B0%95%EC%8A%A4%ED%94%8C%EB%A1%AF-%EA%B7%B8%EB%A6%AC%EB%8A%94-%EB%B2%95

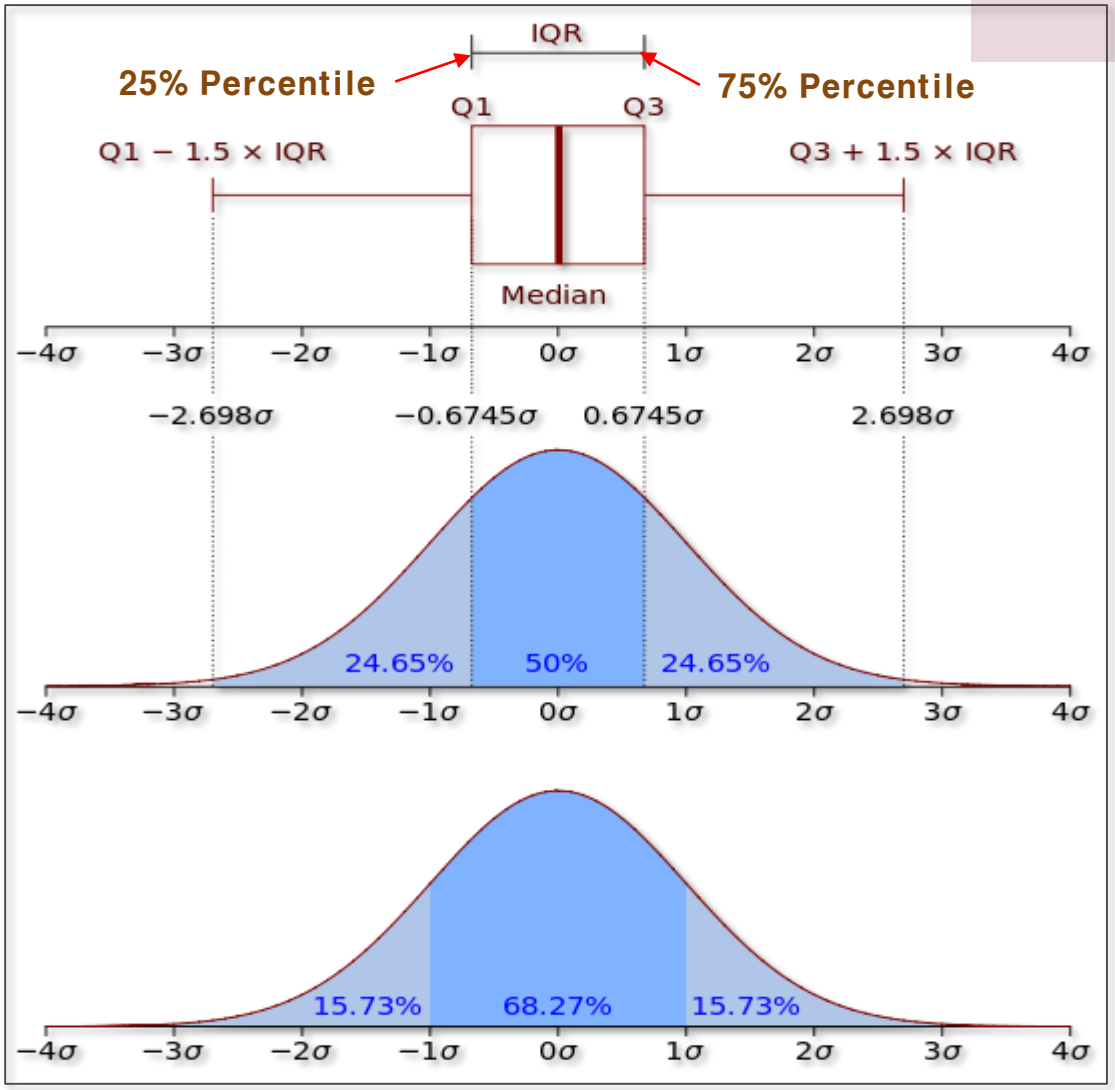
- 상자 그래프는 중앙값(mean), 평균(median), 사분위수(**quartiles**), 범위, 이상치 등과 같은 주요 통계 측정치를 시각화하는 그래픽 도구이다.
- 단일 상자 그래프를 사용하여 모든 데이터를 표시할 수 있다.
- 상자 그래프는 기본 통계 분포를 가정 하지 않고 통계 모집단의 샘플에 변형을 표시한다. 상자의 다른 부분 사이의 간격은 데이터의 분산 (확산) 및 왜도 **(skewness)** 의 정도를 나타내며 이상치를 나타낸다.

출처 : **https://en.wikipedia.org/wiki/Box_plot**

3. 상자 그래프(Box Plot)

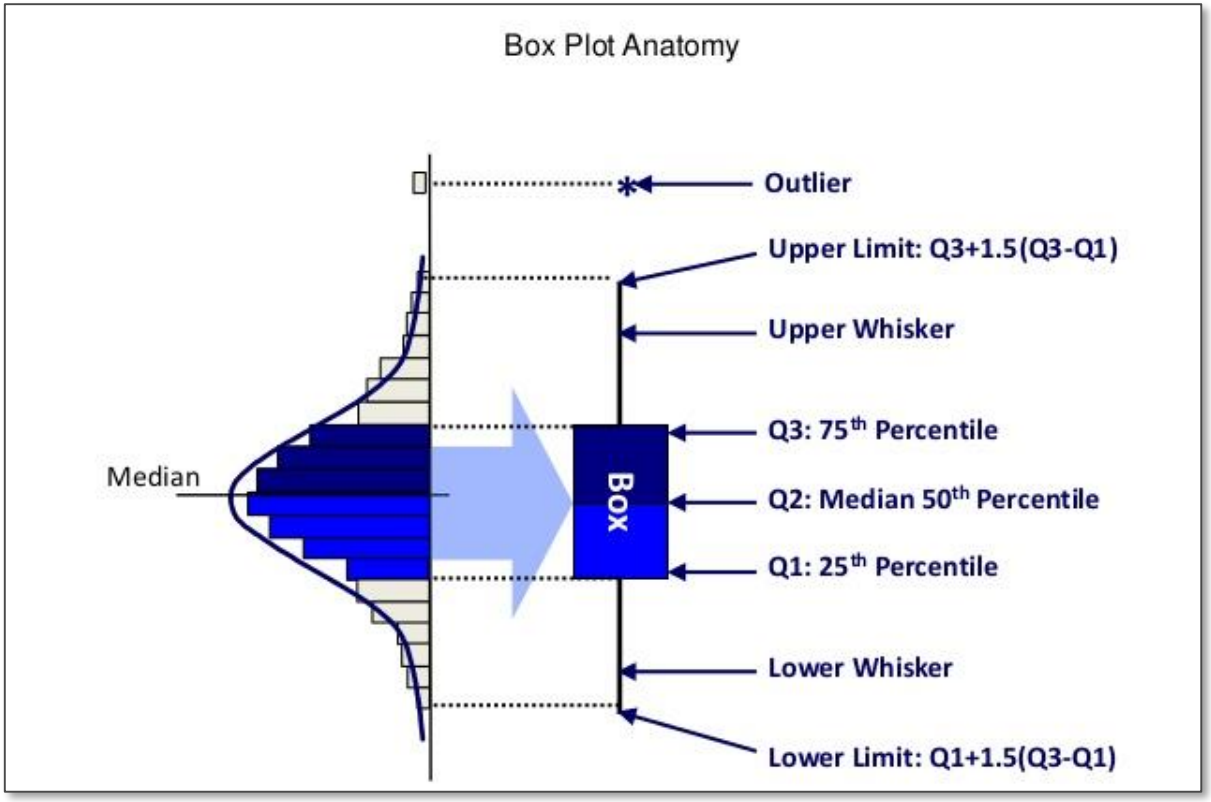
- 우측 그림은 데이터가 정규 분포를 따를때이며, 각 용어와 의미들을 설명하고 있다.

* σ (시그마) : 모집단의 표준편차



출처 :
https://en.wikipedia.org/wiki/Box_plot

3. 상자 그래프(Box Plot)



출처 :**Application of Lean Six Sigma In Food Processing Process Improvement , Nov 12, 2014**

Aizad Ahmad, MBA , Manager, Strategic Review (Lean Six Sigma), International Banking

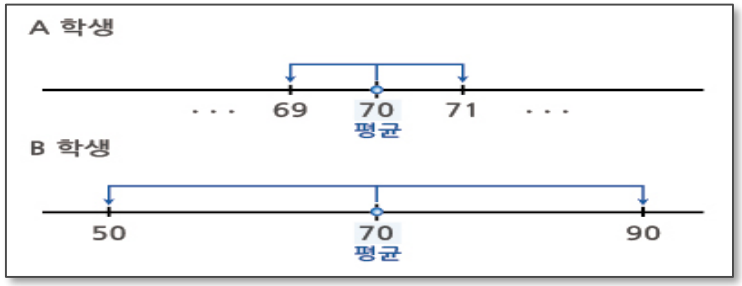
3. 상자 그래프(Box Plot) - 표준편차

표준편차 (Standard deviation)와 평균(Average) 및 분산(variance)

표준편차 : 자료의 값이 얼마나 흩어져 분포되어 있는지 나타내는 산포도 값의 한 종류.
일반적으로 모집단의 표준편차는 σ (시그마)로 표시한다.

과목	국어	영어	수학	평균
점수	69	70	71	70

과목	국어	영어	수학	평균
점수	50	70	90	70



A학생은 모든 과목이 평균 **70**점에 아주 가까이 분포하지만 **B**학생은 국어,수학 성적이 평균과 **20**점이나 떨어져있다. 이 경우 **A**보다 **B**의 표준편차가 더 크다.

표준편차를 구하려면 먼저 각 자료값과 평균의 차이를 구하는데, 이를 '편차'라 한다. 편차는 (자료값)-(평균)이다. 편차를 구하여 그 평균값을 표준편차라 하면 편리하겠지만, 편차의 합은 항상 **0**이기 때문에 불가능하다. 왜냐하면 평균 자체가 모든 자료값들의 평균값이기 때문이다.

과목	국어	영어	수학	평균
점수	50	70	90	70
편차	-20	0	20	0

출처 : **http://terms.naver.com/entry.nhn?docId=1158826&cid=40942&categoryId=32212**

3. 상자 그래프(Box Plot) - 표준편차

편차의 합이 **0**이 되는 문제는 편차의 값 중 음수가 발생하기 때문인데, 편차는 음수이든 양수이든 자료가 평균으로부터 얼마나 차이가 나는지 그 절대값을 알고자 구하는 값이므로 편차가 음수가 되지 않도록 제곱하여 모두 양수가 되게 한다. 그리고 편차를 제곱한 값들의 평균을 내면 자료값들이 평균으로부터 어느 정도 떨어져 있는지를 알 수 있다. 그러나 아래 **B**학생의 경우를 보면 편차를 제곱하는 바람에 자료값의 분산도가 **266.67**로 너무 커진 것을 확인할 수 있다.

과목	국어	영어	수학	평균
점수	50	70	90	70
편차	-20	0	20	0
$(\text{편차})^2$	400	0	400	266.67

분산

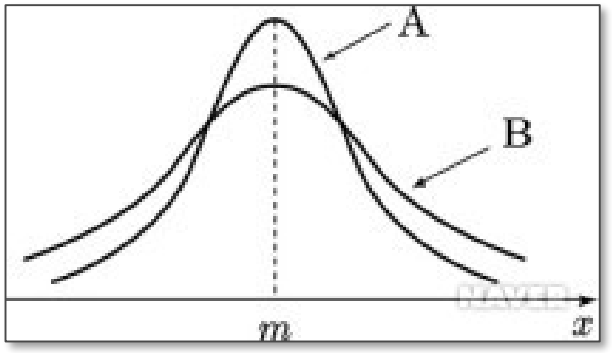
제곱해서 과도하게 부풀려진 값을 다시 원래의 차이값에 비슷하게 맞춰주기 위해서는 제곱근을 적용한다. 학생 **B**의 경우에는 $\sqrt{266.67} = 16.33$ 이 **표준편차**가 된다.

제곱근을 적용하기 직전, **(편차)²**의 평균을 '**분산**'이라 하고, **표준편차**는 $\sqrt{(\text{분산})}$ 이 된다. 다시 말해, 표준편차는 편차 제곱의 평균으로 구할 수 있다. 표준편차가 **0**일 때는 자료값이 모두 같은 값을 가지고, 표준편차가 클수록 자료값 중에 평균에서 떨어진 값이 많이 존재한다.

출처 : **http://terms.naver.com/entry.nhn?docId=1158826&cid=40942&categoryId=32212**

3. 상자 그래프(Box Plot) - 표준편차

평균과 분산의 쌍두마차 : 분산은 편차를 다 더하는 경우 항상 **'0'** 이 됨을 고려하여 편차의 제곱을 더하여 이들의 평균을 낸 것이다. 평균은 자료의 핵심 요소를 파악하게 해 주지만 분포 정도를 나타내 주지는 못한다.
예컨대 수학 시험을 **3**회 실시한 점수가 **80, 60, 40**인 **B**학생과 **70, 60, 50**인 **A**학생을 비교해 보자. 전자와 후자 모두 평균이 **60**점으로 동일하다.



3회 시험을 실시한 점수를 일일이 명시하지 않고 평균만을 기재한다면 학생의 수학 성적이 항상 평균 수준에서 안정을 보이고 있는지, 혹은 종잡을 수 없이 큰 등락을 보이고 있는지 확인할 길이 없다.

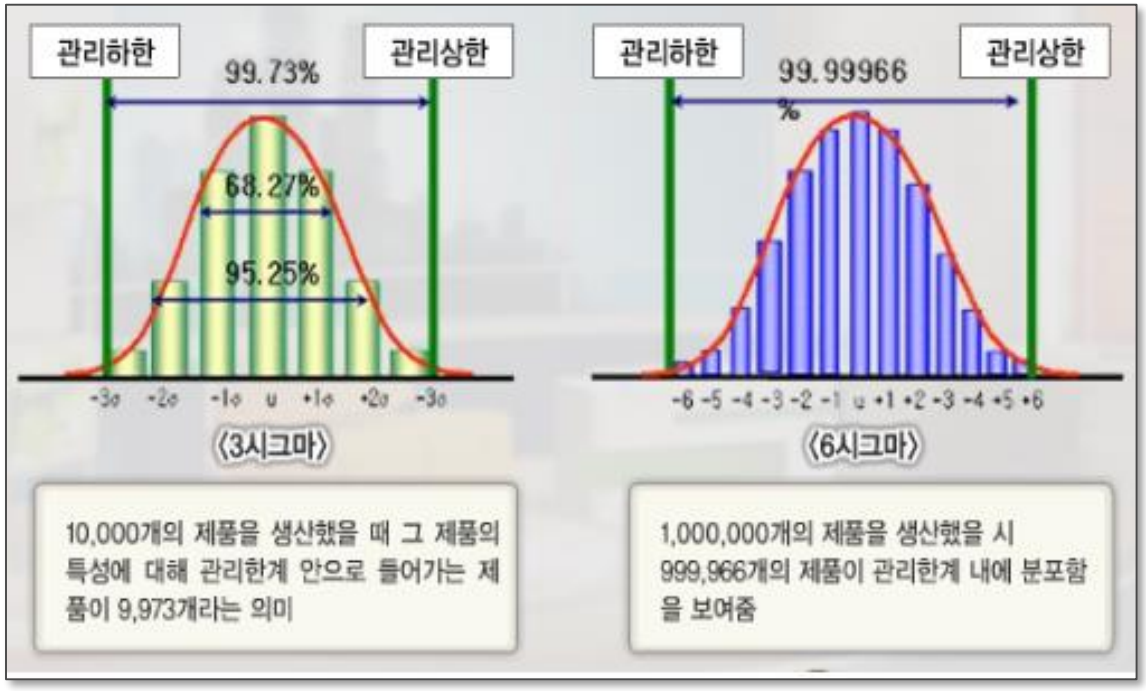
따라서 등락의 안정성 즉 점수의 평균에의 수렴정도를 나타내주는 지표가 필요했고 산포도 중 가장 많이 사용하고 있는 것이 **분산**과 **표준편차**인 것이다. 수학 점수에 있어 가장 바람직한 것은 높은 평균 점수와 등락 없는 분포이다. 경영학의 재무관리파트에서는 평균을 '수익'에 분산을 '위험'에 대비시킨다. 그러므로 분산이 작을수록 분포는 안정성을 지닌다. 그래프 상으로는 평균 주위에 밀집되어 있는 ***A***의 분산이 ***B***의 분산보다 작다.

출처 : **http://terms.naver.com/entry.nhn?docId=2073806&cid=47324&categoryId=47324**
산포도와 표준편차 (통합논술 개념어 사전, 2007. 12. 15., 청서출판)

3. 상자 그래프(Box Plot) – 6시그마

- 6 시그마

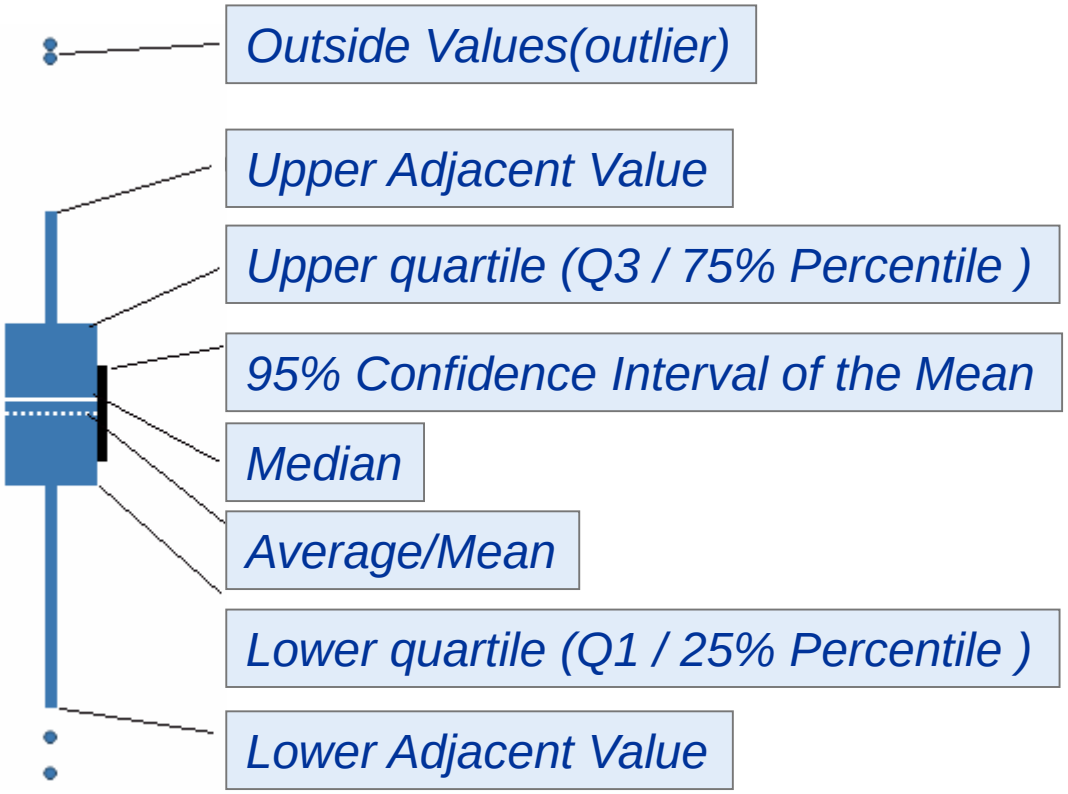
① 사전적 의미 : 백만 번 가운데 **3.4**회의 불량이 발생하는 수준을 의미**(3.4/1,000,000)**
② 통계학적 의미 : 표준편차(산포, 변동)을 의미 -> 즉 **SPEC** 대비 ±**6**σ의 상태

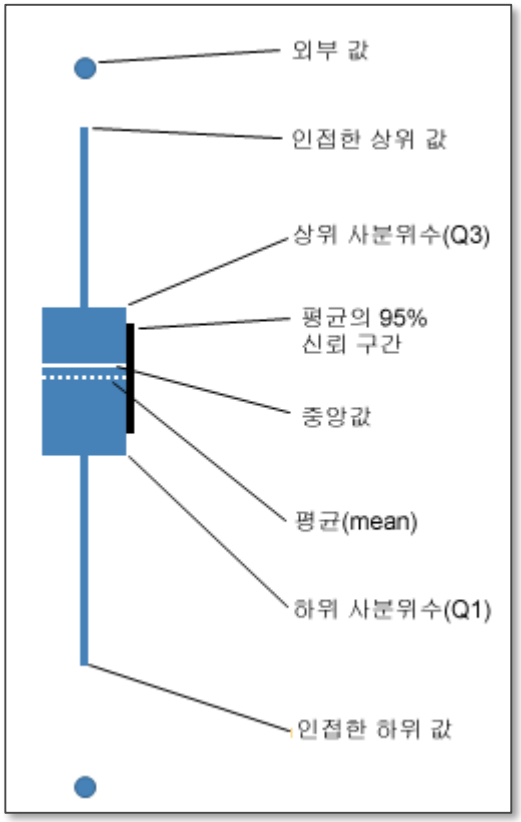


출처 : **http://kiyoo.tistory.com/213#sthash.0c5w6n0a.dpuf**

3. 상자 그래프(Box Plot)

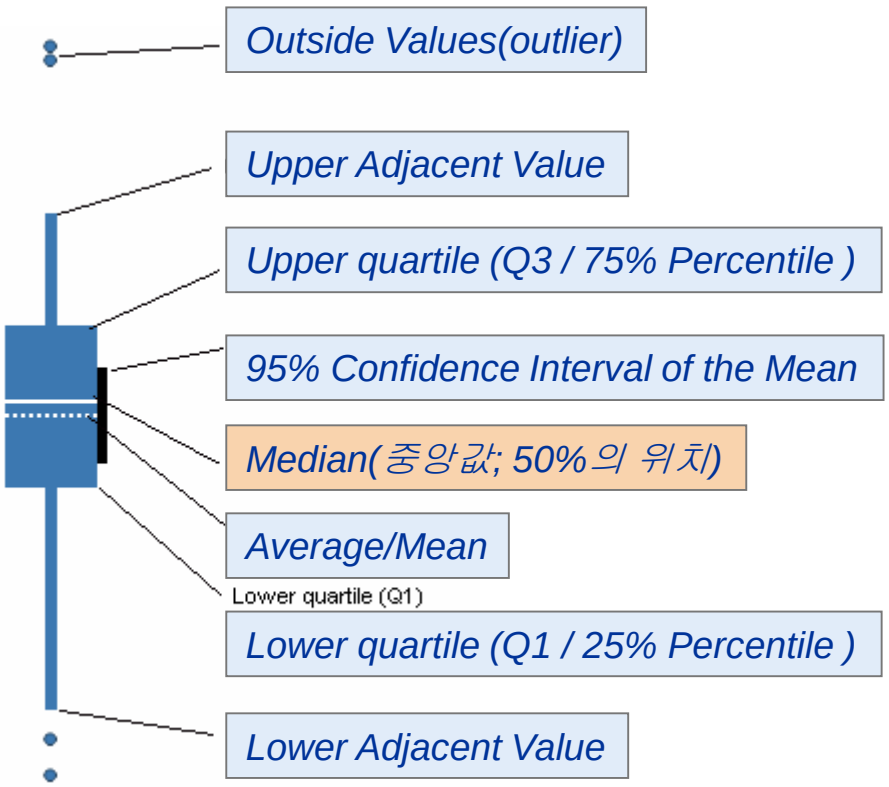
Spotfire 의 상자 그래프에 표시되는 정보들의 의미 :





3. 상자 그래프(Box Plot)

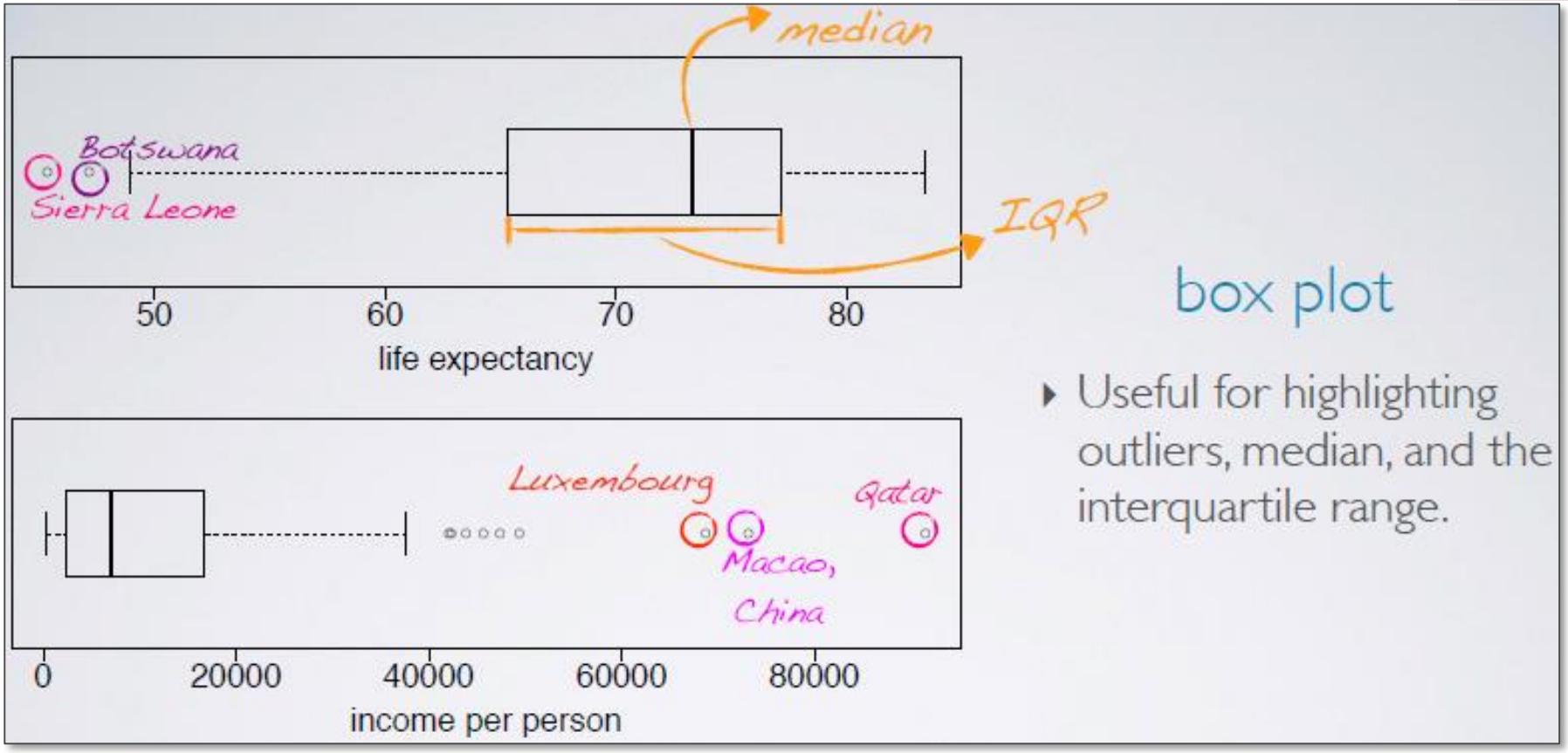
Spotfire 의 상자 그래프에 표시되는 정보들의 의미 :



- **중앙값(median): 50%**의 위치.
  중앙 값은 짝수일 경우 **2**개가 될 수도 있는데, 그 둘의 평균이 중앙값이 된다. 홀수일 경우, 중앙값은 **1**개가 된다.
- **박스(Box): Q3-Q1**
  IQR(Interquartile Range; 사분범위**)**이라고 함. 분포의 양끝 **1/4**을 제외한 범위
- **수염 (whiskers):** 박스의 각 모서리 **(Q1, Q3)**로 부터 **IQR**의 **1.5**배 내에 있는 가장 멀리 떨어진 데이터 점까지 이어져 있는 것
- **왜도(Skewness):** 분포의 비대칭의 정도, 즉 분포가 기울어진 방향과 그 기울어진 정도를 나타내는 척도
- **이상치(Outlier):** 수염**(whiskers)**보다 바깥쪽에 존재하는 데이터

출처 : **http://goodtogreate.tistory.com/216**

3. 상자 그래프(Box Plot)



출처 : **http://researchhubs.com/post/ai/data-analysis-and-statistical-inference/visualizing-numerical-data.html**

3. 상자 그래프(Box Plot)

Spotfire 에서 제공되는 각종 집계 방법(aggregation measures)들 :

- 합계**(Sum)**
- 평균**(Avg)**
- 카운트**(Count)**
- 고유한 수 **(Unique Count)**
- 최소값**(Min)**
- 최대값**(Max)**
- 중앙값**(Median)**
- 표준 편차**(Standard Deviation ; StdDev)**
- 표준 오차**(Standard Error ; StdErr)**
- 분산**(Variance ; Var)**
- **95%** 신뢰 구간의 하위 끝점**(L95)**
- **95%** 신뢰 구간의 상위 끝점**(U95)**
- 제**1** 사분위수**(First Quartile ; Q1)**
- 제**3** 사분위수**(Third Quartile ; Q3)**
- 인접한 하위 값**(Lower Adjacent Value;LAV)**
- 인접한 상위 값**(Upper Adjacent Value; UAV)**
- 큰 정수개수**(CountBig)**
- 고유한 연결 **(Unique Concatenate)**
- 연결**(Concatenate)**
- 처음**(First)**
- 기하학적 평균**(Geometric Mean)**
- 사분위수 범위**(Interquartile Range ; IQR)**
- 마지막**(Last)**
- 하위 내부 펜스**(Lower Inner fence; LIF)**
- 하위 외부 펜스**(Lower Outer fence; LOF)**
- 평균 편차 **(Mean Deviation)**
- 중앙값 절대편차**(Mean Absolute Deviation;MAD)**
- 가장 공통적 **(Most Common)**
- 이상값 개수**(Ourlier Count;Outliers)**
- **10**번째 백분위수**(10th Percentile;P10)**
- **90**번째 백분위수**(90th Percentile; P90)**
- 이상값 비율**(Outlier Percentage;PctOutliers)**
- 제곱**(Product)**
- 범위**(Range)**
- 상위 내부 펜스**(Upper Inner fence ;UIF)**
- 상위 외부 펜스**(Upper Outer fence ;UOF)**

3. 상자 그래프(Box Plot)

상자 그래프에서 다음 3가지를 해석(파악)할 수 있다.

- Variance(분산 ; 변수의 흩어진 정도)
  - 만약 사분범위가 길면 보다 흩어진 분포이고, 짧으면 밀집된 분포임을 알 수 있다.
  - 자료의 극단적인 값들에 의한 영향을 덜 받는 장점이 있다.
- Skewness(비대칭도/왜도 : 평균값에 관한 비대칭의 방향과 그 정도)
  - Whisker(수염)의 길이를 비교하여 비대칭 여부를 판단할 수 있다.
- Outliers(이상치 : 다른 변수값과 다른 유형을 보이는 변수값)
  - 수염(whiskers)보다 바깥쪽에 존재하는 데이터

3. 상자 그래프(Box Plot) - 왜도

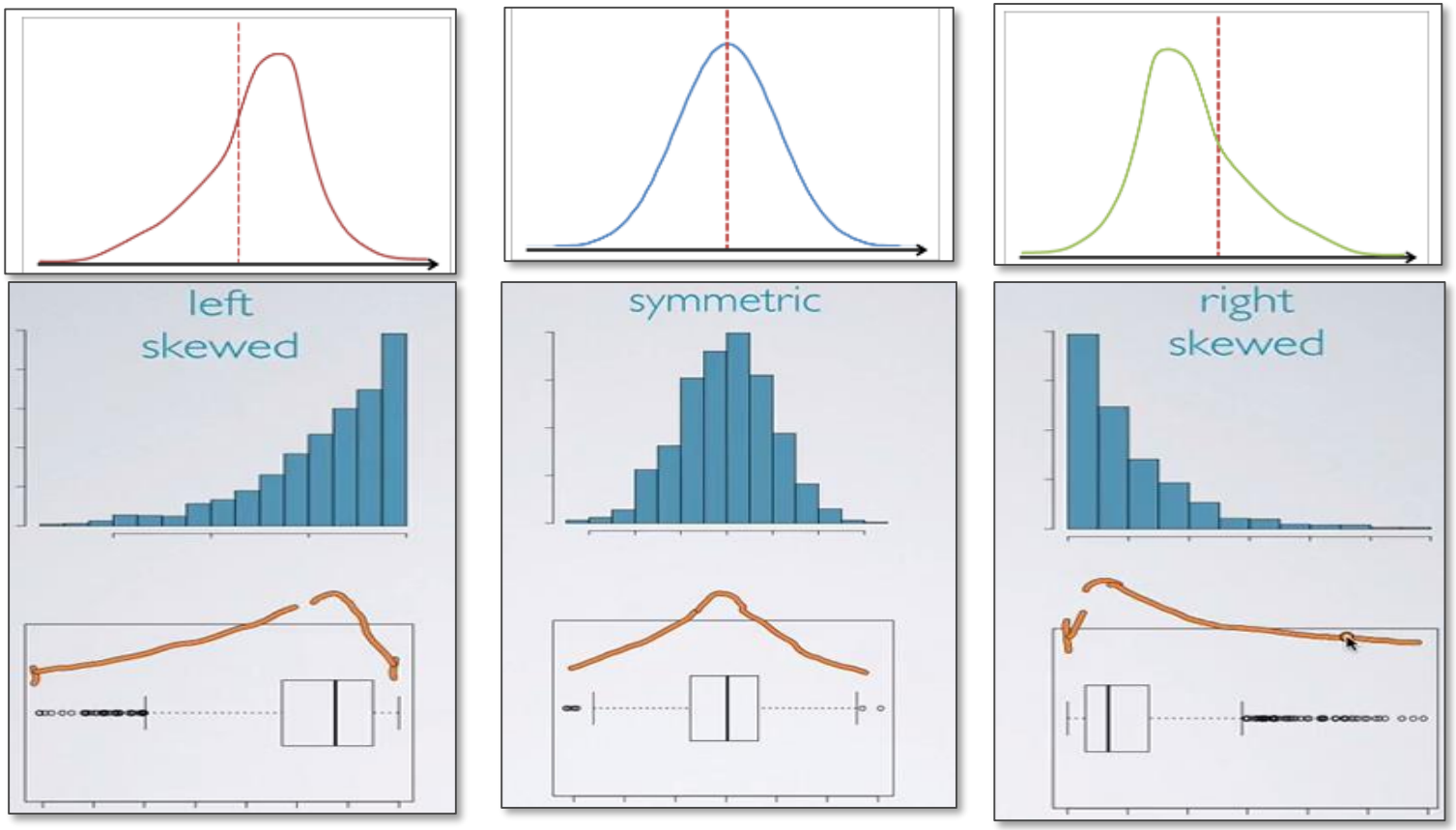
Skewness(왜도)

: 분포의 비대칭의 정도, 즉 분포가 기울어진 방향과 그 기울어진 정도를 나타내는 척도이다.

왜도 < **0**,좌측으로 긴 꼬리 **(부적 비대칭,negative Skew]**

왜도 = **0**,좌우 대칭 분포

왜도 > **0**,우측으로 긴 꼬리**(**정적 비대칭,**positive Skew)**



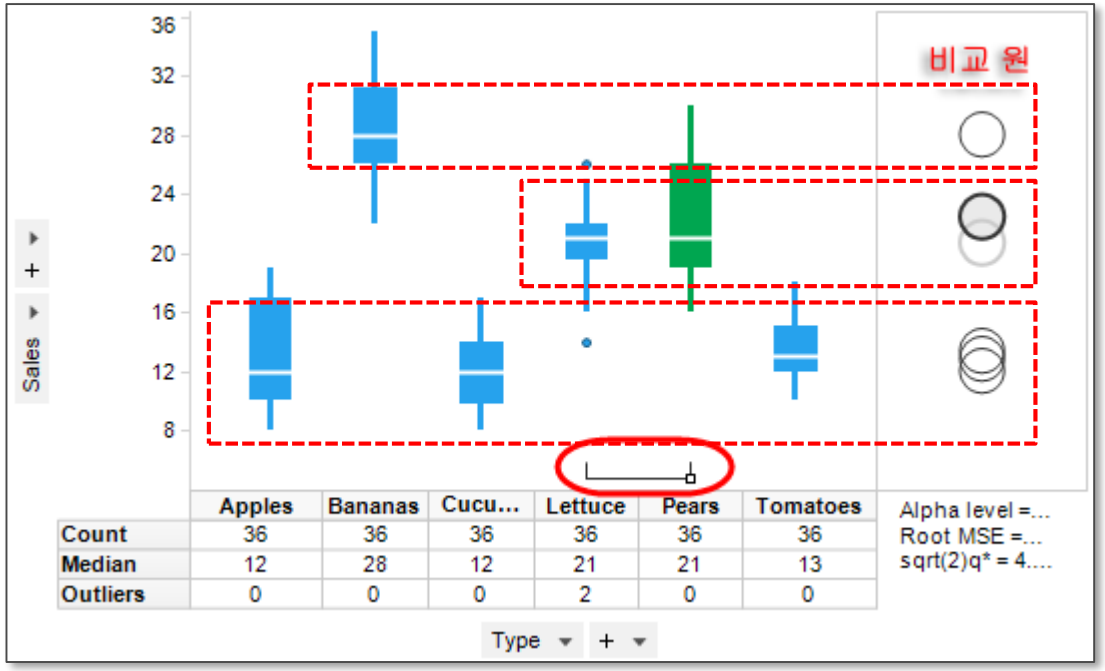
출처 : **http://blog.naver.com/chochila/40144022678**

출처 : **http://researchhubs.com/post/ai/data-analysis-and-statistical-inference/visualizing-numerical-data.html**

3. 상자 그래프(Box Plot) – 비교원

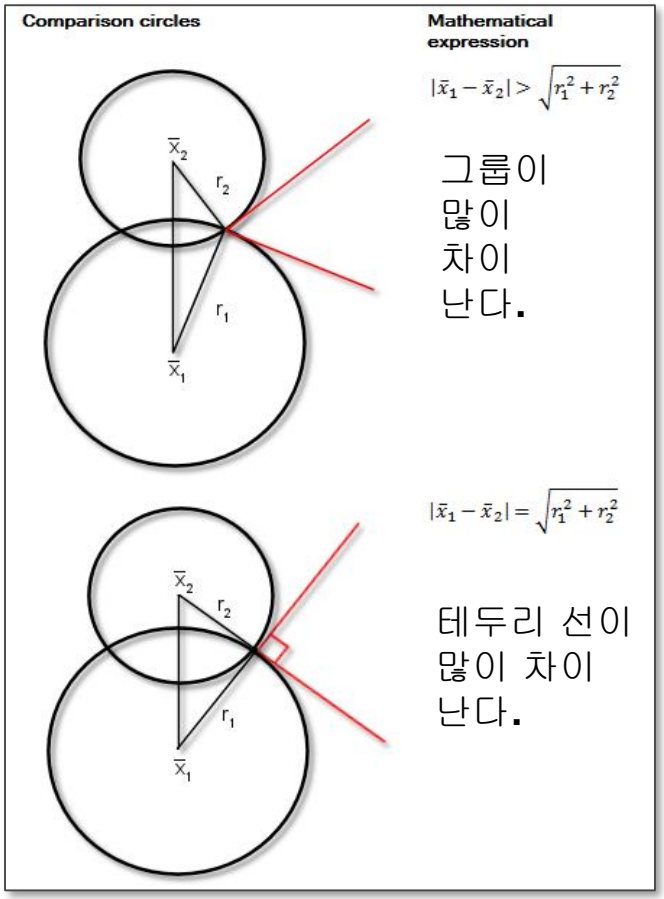
비교원(Comparison Circles)

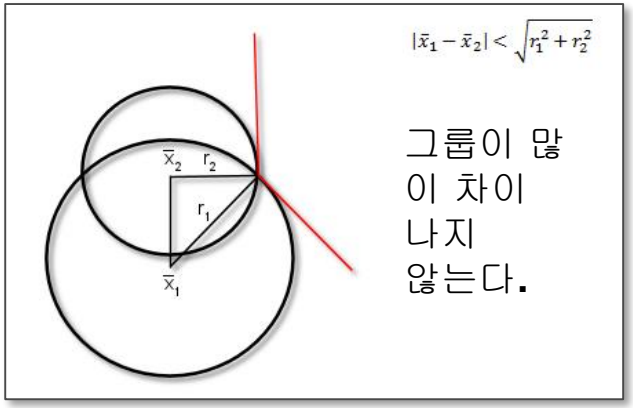
- 비교 원을 그려서 다양한 범주에 대한 평균 값(상자 그래프의 상자)이 서로 많은 차이가 나는지 여부를 표시할 수 있다. 일치하는 상자에 대한 평균값을 중심으로 원이 그려진다.
- 두 그룹의 원이 겹치지 않으면 일반적으로 두 그룹의 평균이 많이 차이 나는 것이다. 원이 크게 겹치는 경우 평균이 많이 차이 나지 않는 것이다.
- 관계 표시기의 사각형은 마크된 상자를 나타내고 관계 표시기의 라인은 마크된 상자와 많이 차이 나지 않는 상자로 확장된다.
- 상자 그래프를 강조 표시하거나 반대로 비교 원을 강조 표시하여 상자 그래프를 강조 표시할 수도 있다. 강조 표시된(마크된) 비교 원은 이중 라인 테두리로 그려진다.



3. 상자 그래프(Box Plot) – 비교원

- 비교 원에는 **Tukey-Kramer** 방법이 계산에 사용된다.
- 각 그룹(각 상자 그래프)에서는 원의 중심이 그룹 평균 값과 일치하는 원을 가져온다.
- 원이 크게 겹치는 경우 평균이 많이 차이 나지 않는 것이다.
- 일반적으로 원의 크기가 크면 분포가 안정적(밀집됨)이고, 크기가 작으면 분포가 넓은 것을 알 수 있다.



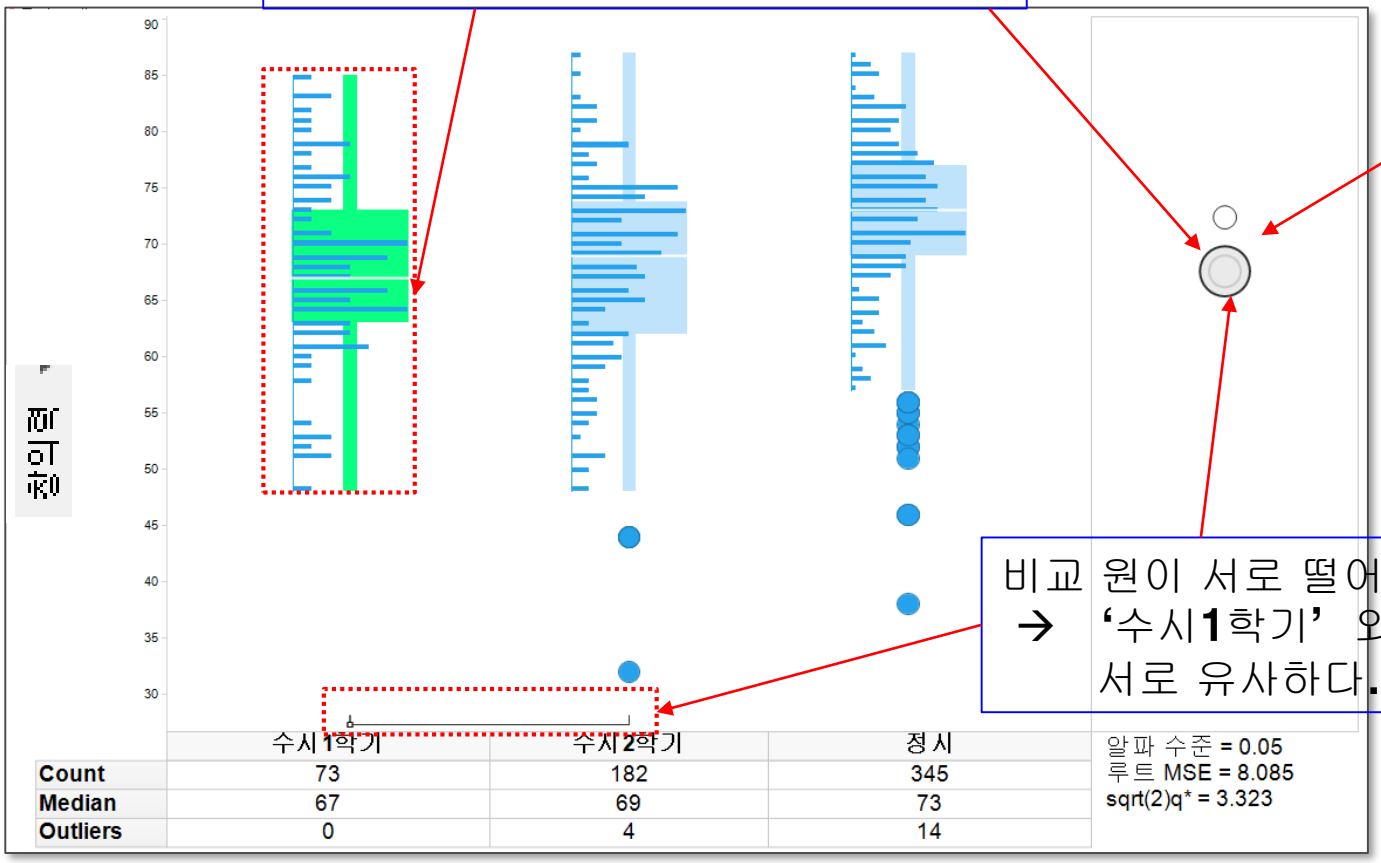


3. 상자 그래프(Box Plot) – 비교원

최대값과 최소값의 차이가 가장 적다.
→ 비교 원의 반지름이 가장 크다.

'정시' 의 비교 원이 다른 무리와 많이 떨어져 있다.
→ '수시**1**학기' 와 '수시**2**학기' 보다 '창의력' 이 많이 우수하다.

비교 원이 서로 떨어져 있지 않다.
→ '수시**1**학기' 와 '수시**2**학기' 가 서로 유사하다.

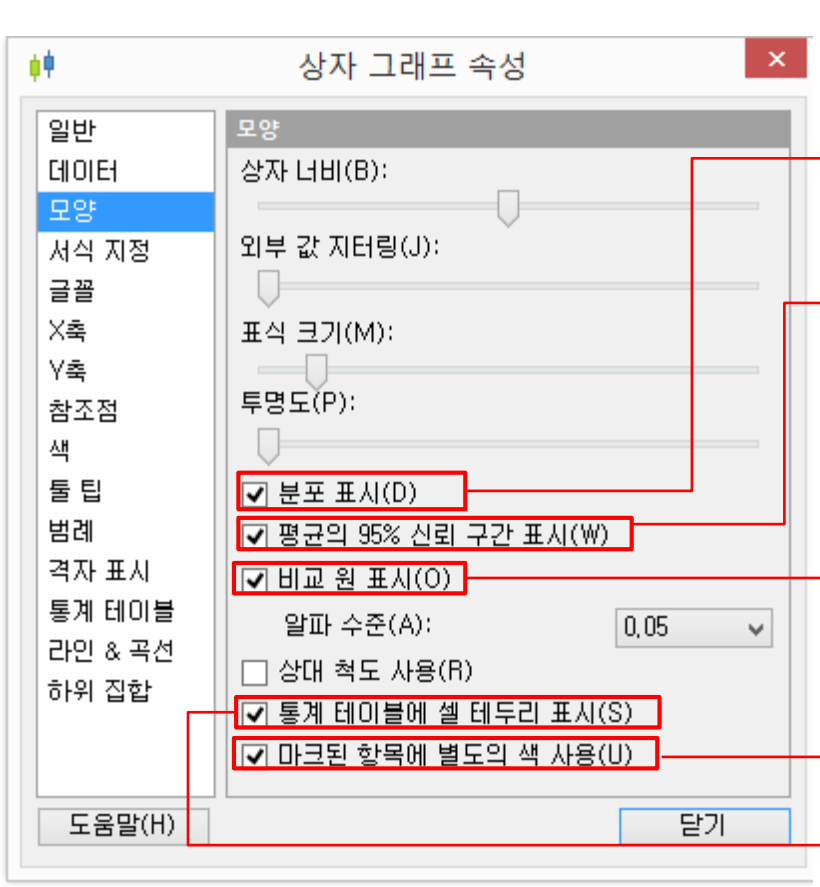


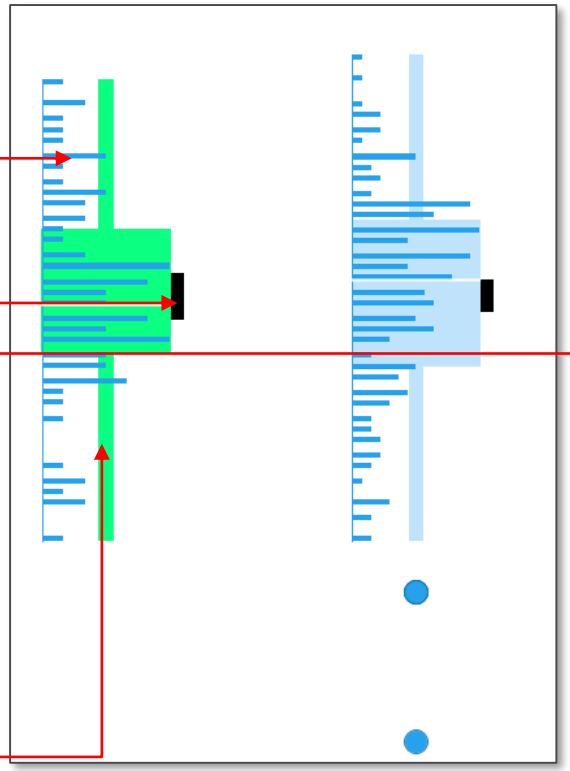
	수시1학기	수시2학기	정시
Count	73	182	345
Median	67	69	73
Outliers	0	4	14

알파 수준 = 0.05
루트 MSE = 8.085
sqrt(2)q* = 3.323

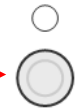
3. 상자 그래프(Box Plot) – 속성(모양) 설정

상자 그래프로 커서를 이동한 후에, 마우스 우클릭 > 속성 > 모양 에서 속성 설정





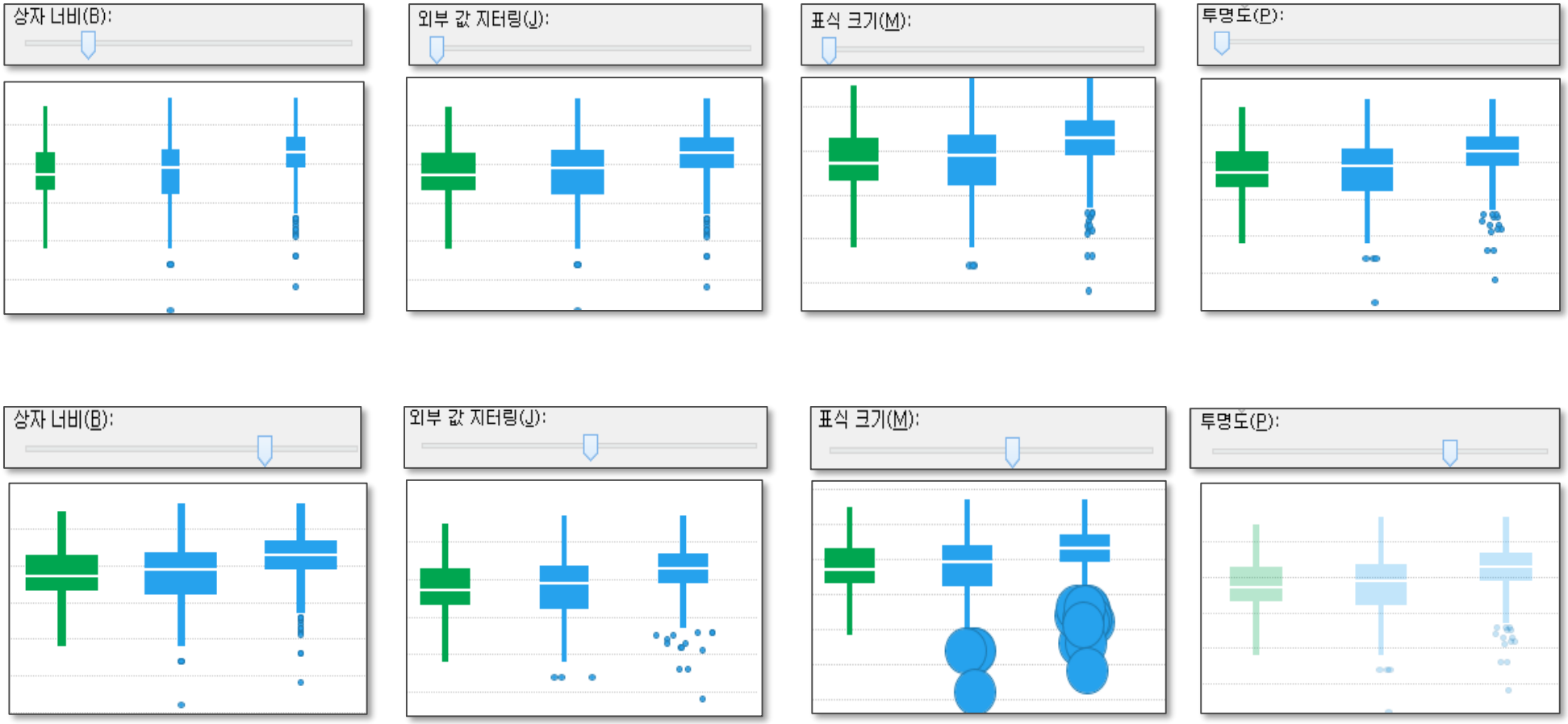
수시 1학기	수시 2학기
73	182
67	69
0	4



알파 수준 = 0.05
루트 MSE = 8.085
sqrt(2)q* = 3.323

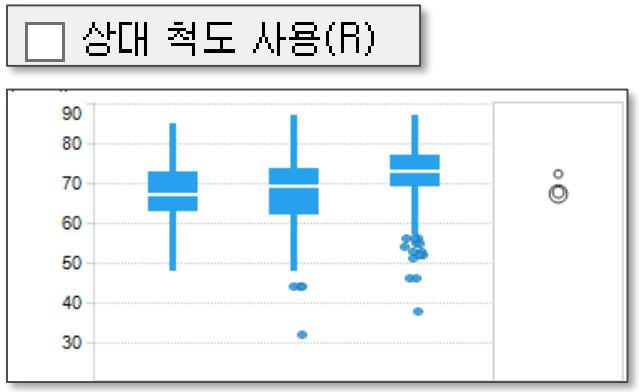
3. 상자 그래프(Box Plot) – 속성(모양) 설정

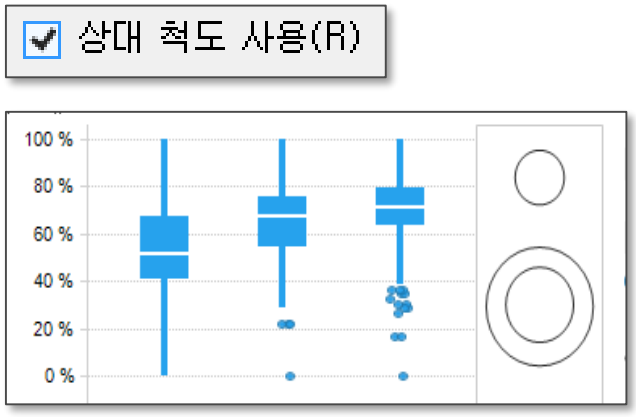
상자 그래프로 커서를 이동한 후에, 마우스 우클릭 > 속성 > 모양에서 속성 변경 결과

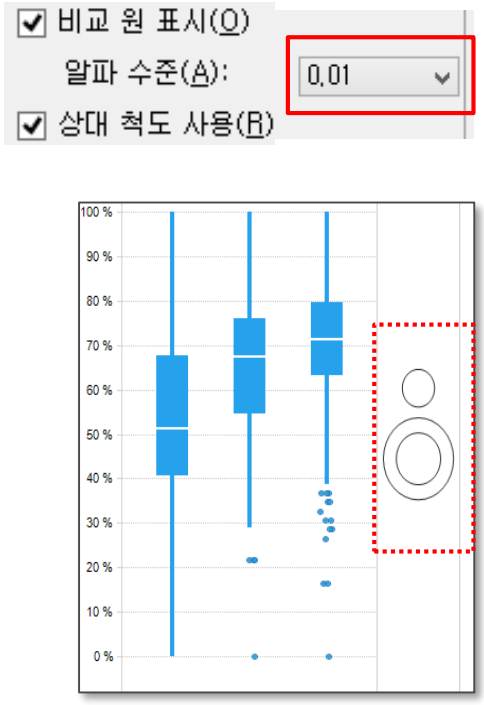


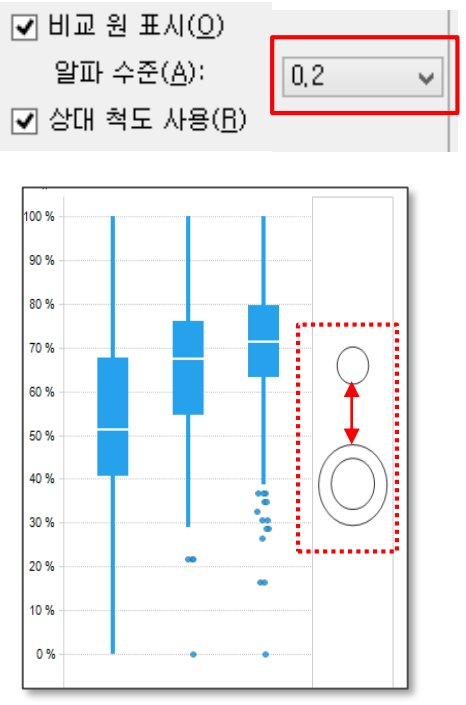
3. 상자 그래프(Box Plot) – 속성(모양) 설정

상자 그래프로 커서를 이동한 후에, 마우스 우클릭 > 속성 > 모양





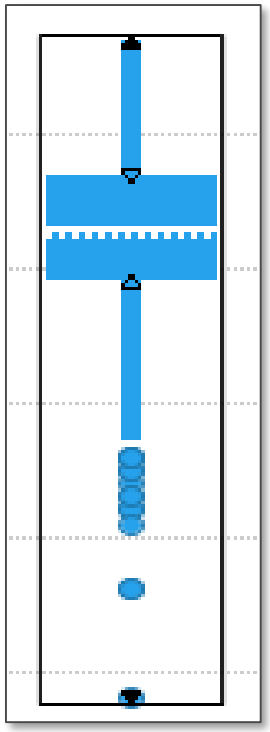


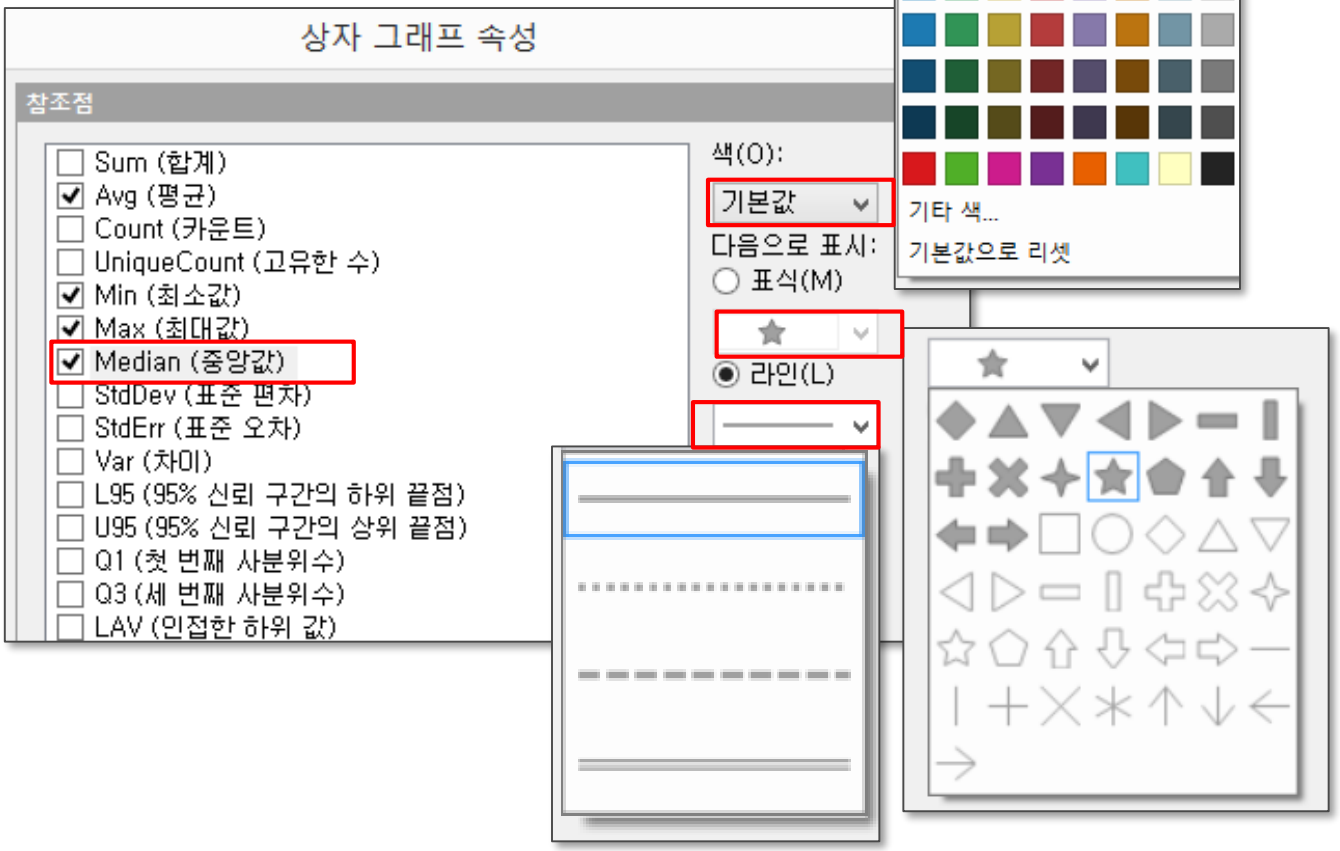


알파 수준 : 그룹간에 차이가 나는 정도**(0 < x < 1)**

3. 상자 그래프(Box Plot) – 속성(참조점) 설정

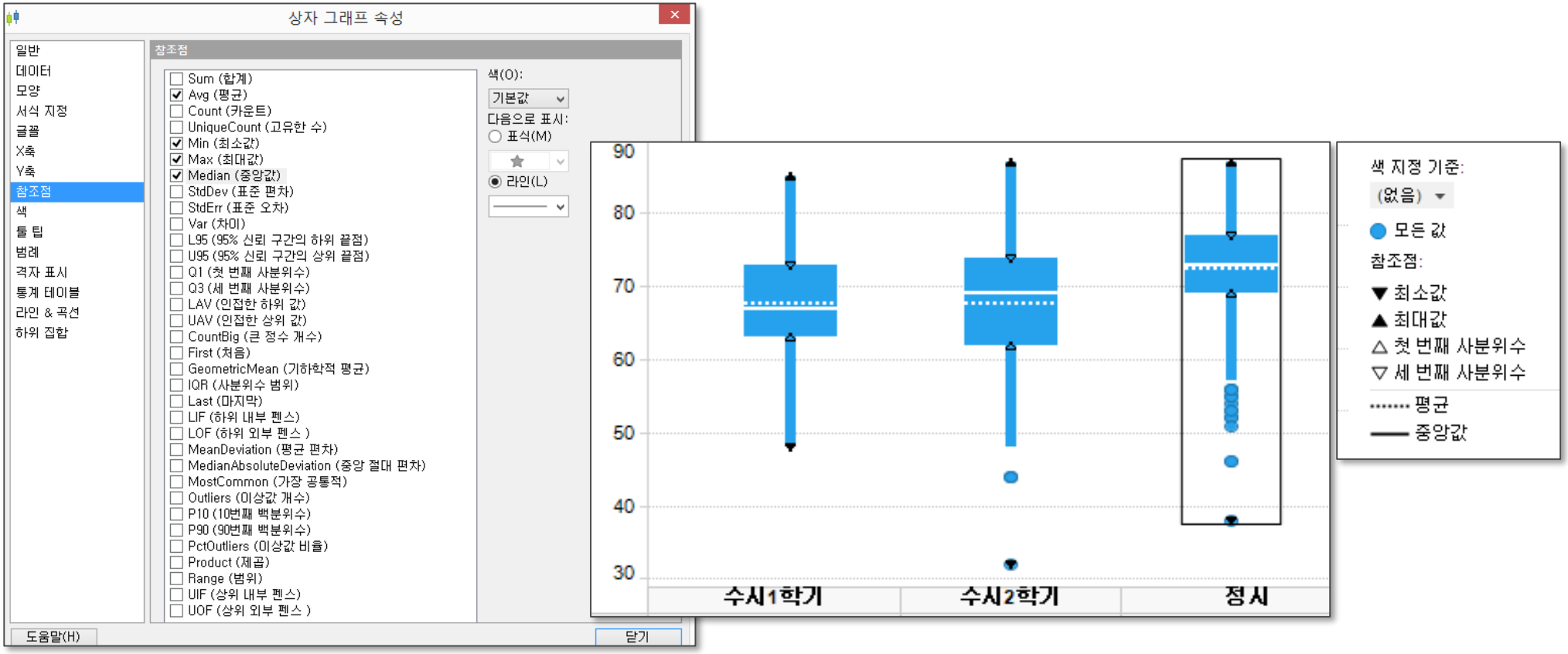
- 상자 그래프로 커서를 이동한 후에, 마우스 우클릭 > 속성 > 참조점 선택
- 참조점은 표식이나 라인, 2개 중에서 선택하여 상자 그래프 내에 표시할 수 있다.





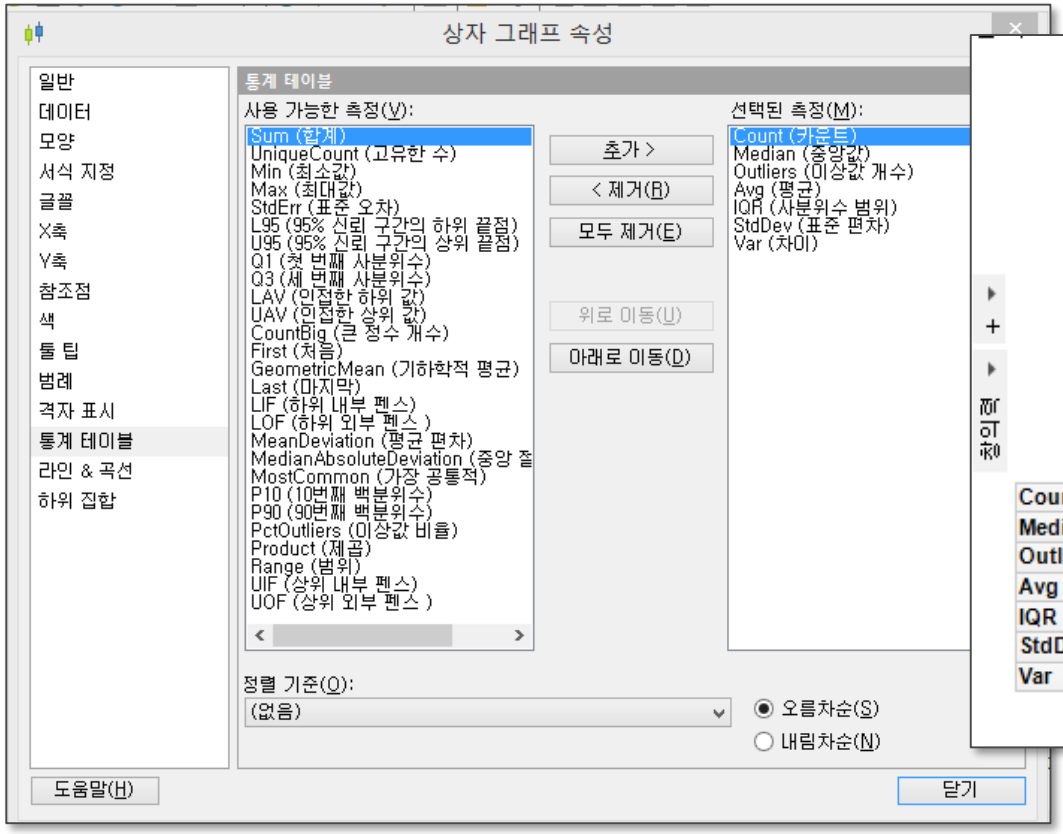
3. 상자 그래프(Box Plot) – 속성(참조점) 설정

상자 그래프로 커서를 이동한 후에, 마우스 우클릭 › 속성 › 참조점



3. 상자 그래프(Box Plot) – 속성(통계테이블) 설정

상자 그래프로 커서를 이동한 후에, 마우스 우클릭 > 속성 > 통계테이블



	수시1학기	수시2학기	정시
Count	73	182	345
Median	67	69	73
Outliers	0	4	14
Avg	67.6164	67.6484	72.4174
IQR	10	11.75	8
StdDev	8.23379	8.83598	7.62796
Var	67.7953	78.0746	58.1857

미국 프로야구 선수 데이터**.xls**

3. 상자 그래프(Box Plot) – 실습

- 문제 :
  1. 전체 포지션 중에서 연봉이 가장 낮은 포지션은?

     * 힌트 :
       - 상자 그래프
         - 연봉
         - 비교원
       - Bar chart(연봉 평균)과 대비
  2. 전체 포지션 중에서 연봉의 편차가 가장 심한 포지션은?

     - 표준 편차와 비교원의 2가지 관점에서 평가
  3. 전체 포지션 중에서 연봉에 있어서 이상치 비율이(Percent of Outliers) 가장 높은 포지션은?