<일부 Data를 선별(발췌)하는 몇 가지 방법>

Monthly User Guide(32호)-2020년 3월호

JMP Korea 신 익주 이사(ikju.shin@jmp.com)

Data 분석을 하는 경우에 있어서 전체 Data에서 일부 Data(를) 선별(발췌)해야 하는 경우가 많은데, 이번 호에서는 이러한 선별(발췌) 방법 중 **Row 단위로 선별(발췌)할 수 있는** 몇 가지 방법에 대해 배워 보도록 하겠습니다.

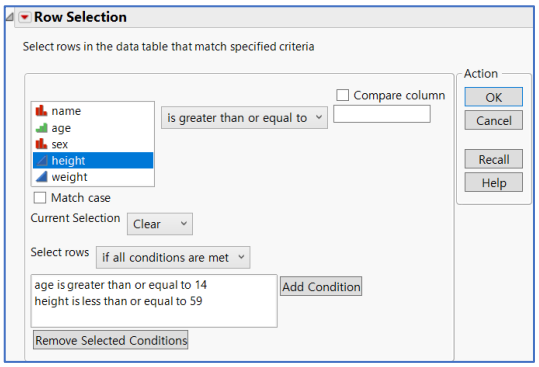
방법 1. Row / Row Selection 활용 (sample data : big class.jmp)

1. Row / Row Selection / Select Where

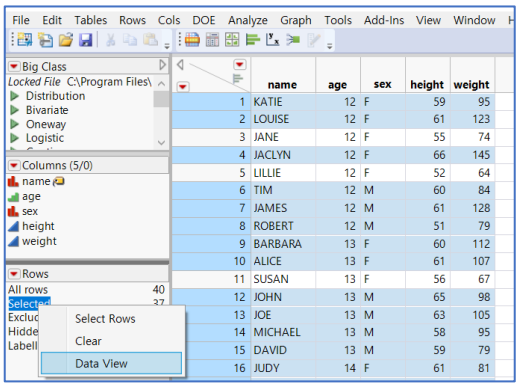
1) 엑셀의 If 함수를 사용하여 어떤 변수 기준으로 얼마보다 크거나 또는 작은 Data를 선별하는 것과 비슷한 기능

2) 각 변수별로 특정 값 기준 이상,이하,미만,초과 및 결측치(Missing Value) 여부 등을 기준으로 선별(합집합 및 교집합 개념 사용)할 수 있는 기능입니다

3) 만약, age 변수 14 이상, height 변수 59 이하 데이터를 선별하고자 한다면, 아래와 같이 입력 및 선택하면 됩니다

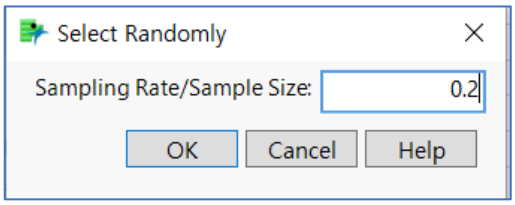


4) 그러면, 다음과 같이 해당되는 Data가 표시됩니다. 해당 Data만을 선별하여 별도의 Data Table을 만들고자 한다면, 좌측 하단 Row Panel의 Selected 위에서 우측 마우스 클릭 후 Data View를 클릭하면 됩니다



2. Row / Row Selection / Select Randomly

1) 특별한 선별 기준없이 해당 비율 또는 해당 개수 만큼 랜덤으로 선별하는 기능입니다



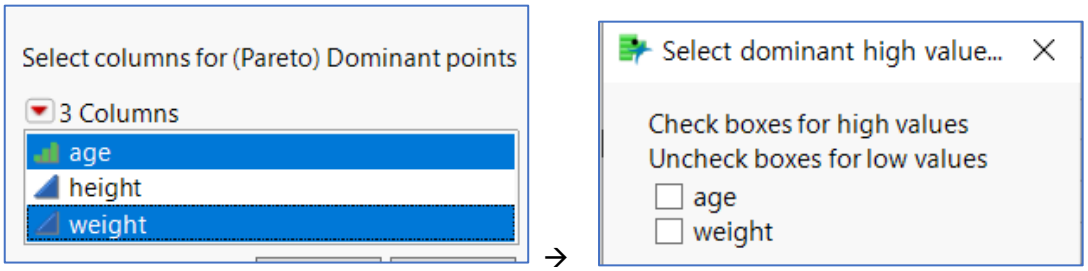
2) 위의 빈 칸에 1보다 작은 값을 입력하면 그 비율(Sampling rate) 만큼 선별하고,
1 이상의 정수 값을 입력하면 그 개수(Sampling size) 만큼 선별됩니다.

3) 해당 기능을 사용하기 전에 Data Table에 별도의 표시가 없으면 모든 변수(Column)에 대해 선별 여부가 별도의 Color로 표시되는 반면, 특정한 변수(Column)를 선택하고 위 기능을 사용하면 해당 변수에 대해서만 별도의 Color로 표시됩니다.

3. Row / Row Selection / Select Dominant

1) 특정 변수 기준으로 가장 크거나 작은 값을 가진 row를 선별하는 기능입니다

2) 예를 들어, age 및 weight 변수 기준으로(하나 이상의 변수 선택 가능), 가장 작은 값(row)을 찾고자 할 경우에는 아래와 같이 실행하면



3) 다음과 같이 해당 값(row)이 Data Table에 표시된다.

	name	age	sex	height	weight
1	KATIE	12	F	59	95
2	LOUISE	12	F	61	123
3	JANE	12	F	55	74
4	JACLYN	12	F	66	145
5	LILLIE	12	F	52	64
6	TIM	12	M	60	84

4. Row / Row Selection / Select Matching Cells

1) 하나 이상의 변수 기준으로 Matching되는 Cell, 즉 같은 값을 가진 Row를 찾는 기능입니다

2) 예를 들어, sex는 F, height는 61인 Row를 찾고자 한다면, 아래와 같이 선택한 후
Row / Row Selection / Select Matching Cells 을 클릭하면

	name	age	sex	height	weight
1	KATIE	12	F	59	95
2	LOUISE	12	F	61	123
3	JANE	12	F	55	74
4	JACLYN	12	F	66	145

3) 아래와 같이 Matching되는 세 개의 Row가 선택됩니다.

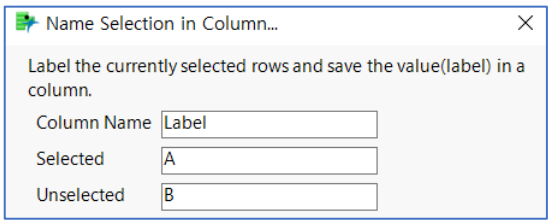
	name	age	sex	height	weigh t
1	KATIE	12	F	59	95
2	LOUISE	12	F	61	123
3	JANE	12	F	55	74
4	JACLYN	12	F	66	145
5	LILLIE	12	F	52	64
6	TIM	12	M	60	84
7	JAMES	12	M	61	128
8	ROBERT	12	M	51	79
9	BARBARA	13	F	60	112
10	ALICE	13	F	61	107
11	SUSAN	13	F	56	67
12	JOHN	13	M	65	98
13	JOE	13	M	63	105
14	MICHAEL	13	M	58	95
15	DAVID	13	M	59	79
16	JUDY	14	F	61	81
17	ELIZABETH	14	F	62	91

5. Row / Row Selection / Name Selection in Column

1) 선택한 Row와 그렇지 않은 Row를 별도로 구분(구분하는 Label을 별도의 Column에 생성)하는 기능입니다. 예를 들어, 아래와 같이 선택한 Row를 A, 그렇지 않은 Row를 B로 구분하여 Labelling 하고자 한다면,

	name	age	sex	height	weight
1	KATIE	12	F	59	95
2	LOUISE	12	F	61	123
3	JANE	12	F	55	74
4	JACLYN	12	F	66	145
5	LILLIE	12	F	52	64
6	TIM	12	M	60	84

2) 아래와 같이 입력하면 오른쪽과 같이 결과가 표시됩니다.



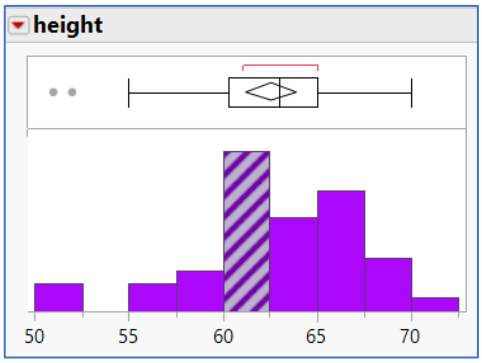
→

	name	age	sex	height	weight	Label
1	KATIE	12	F	59	95	A
2	LOUISE	12	F	61	123	A
3	JANE	12	F	55	74	A
4	JACLYN	12	F	66	145	A
5	LILLIE	12	F	52	64	A
6	TIM	12	M	60	84	B
7	JAMES	12	M	61	128	B

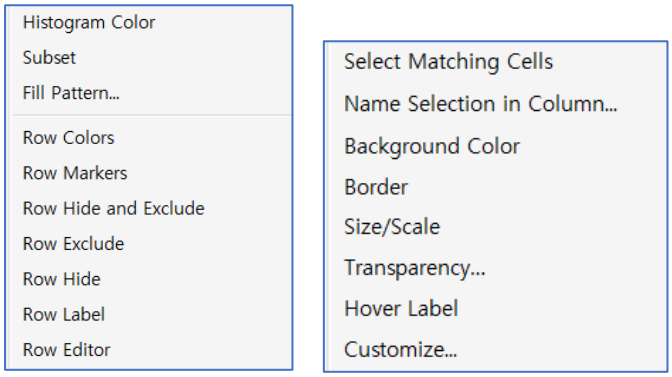
방법 2. Graph 활용

1. Histogram 활용

1) Analyze / Distribution에서 특정한 변수에 대해 Histogram을 그린 다음, 아래와 같이 특정 영역을 선택하고



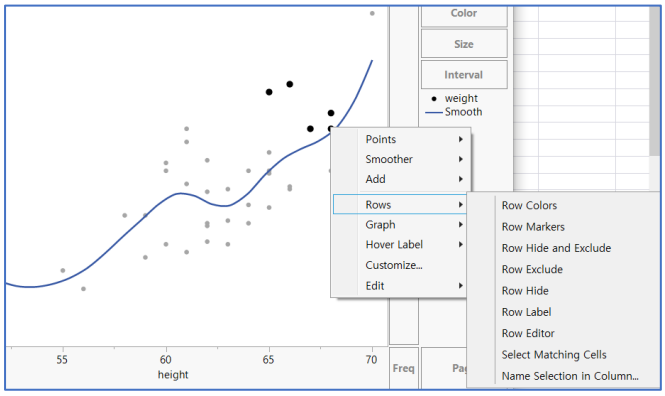
2) 우측 마우스를 클릭하면 Subset, Select Matching Cells 및 Name Selection in Column 기능 등을 활용할 수 있습니다.



2. 다른 Graph 활용

뿐만 아니라 다른 Graph 에서도 거의 동일한 방법으로 Data 선별할 수 있습니다.

Graph Builder 또는 다른 Graph 메뉴에서 Graph를 그린 다음, Graph의 특정한 영역을 선택한 후 우측 마우스를 클릭하면, 아래와 같이 Select Matching Cells 및 Name Selection in Column 기능 등을 활용할 수 있습니다(하위 Menu는 Graph의 종류 및 해당 Graph를 그린 JMP Menu에 따라 약간씩은 다를 수 있음)



방법 3. Tables / Subset 활용

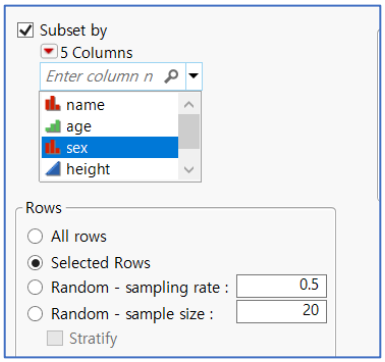
1. 샘플링(Random Sampling, 층화 랜덤 샘플링 등)을 할 때 많이 활용된다.

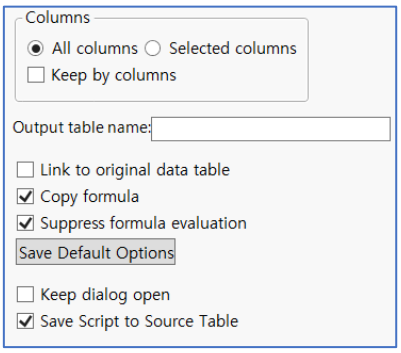
2. Menu에 대한 설명

1) Subset by : 층별 변수별로 Subset하는 기능이다. 여기서 'sex'를 선택하면 남녀별로 분리된 두 개의 Data Table이 생성된다.

2) Rows / Selected Rows : Data Table에서 선택한 Row만 Subset(Row Panel의 'Selected' 위에서 우측 마우스 클릭 후 Data View를 선택한 결과와 같다)

3) Rows / Stratify : 층화(Stratified) 랜덤 샘플링

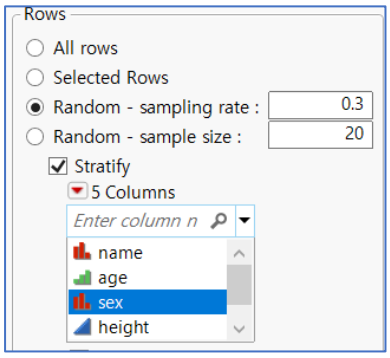




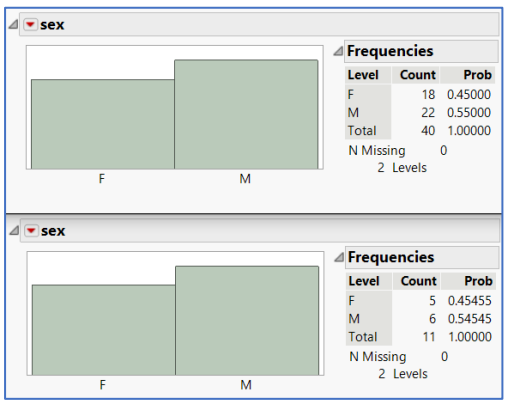
3. 층화(Stratified) 랜덤 샘플링

(층별) 변수별로 동일한 비율만큼 샘플링을 하는 것을 층화 랜덤 샘플링이라고 한다. 이러한 부분 Data를 만들고자 할 경우에는

1) 아래와 같이 'Stratify' 클릭 후 층별(Stratification) 변수를 선택한 뒤 OK 클릭하면 된다.



2) Analyze / Distribution에서 Raw data와 Subset data에 대해 확인해 보면 동일 비율만큼 부분 data가 만들어 졌음을 확인할 수 있다.



끝.