<두 개의 범주형 변수를 그룹화하여 분석하기>

Monthly User Guide(36호)-2020년 7월호

JMP Korea 신 익주 이사(ikju.shin@jmp.com)

분석을 하다 보면 **두 개의 범주형(Category) 변수를 묶어서 하나의 변수처럼 분석**하고 싶은 경우가 많은 데, 이에 대해 살펴보겠습니다. 비슷비슷하고 약간씩 다른 상황이 몇 가지 있을 수 있으므로 여러가지 방법으로 살펴보겠습니다.

JMP 안에 있는 Sample Data 를 활용하겠습니다

Help / Sample Data Library / big class.jmp

여기서 age 와 sex 는 범주형 변수이고 height 와 weight 는 연속형 변수입니다.

	age	sex	height	weight
1	12	F	59	95
2	12	F	61	123
3	12	F	55	74
4	12	F	66	145
5	12	F	52	64

1. 먼저 성별 / 나이별로 키와 몸무게의 평균/표준편차/Min/Max 등을 구하고자 한다면 Analyze / Tabulate를 이용하여 구할 수 있을 것입니다.

		height				weight			
sex	age	Mean	Std Dev	Min	Max	Mean	Std Dev	Min	Max
F	12	58.6	5.413	52	66	100.2	33.73	64	145
	13	59.0	2.646	56	61	95.3	24.66	67	112
	14	62.6	1.517	61	65	96.6	25.64	81	142
	15	63.0	1.414	62	64	102.0	14.14	92	112
	16	62.5	3.536	60	65	113.5	2.121	112	115
	17	62.0	.	62	62	116.0	.	116	116
M	12	57.3	5.508	51	61	97.0	26.96	79	128
	13	61.3	3.304	58	65	94.3	11	79	105
	14	65.3	2.289	63	69	103.9	10.68	92	119
	15	65.2	1.924	62	67	110.8	9.985	104	128
	16	68.0	.	68	68	128.0	.	128	128
	17	69.0	1.414	68	70	153.0	26.87	134	172

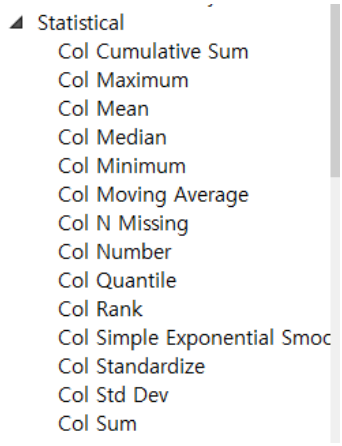
2. 여기서 성별은 2 가지 Level, 나이는 6 가지 Level인데, 두 범주형 변수 간의 교호작용 등을 고려하지 않고, **두 범주를 묶어서, 즉 12가지 Level을 가진 하나의 범주처럼 만들어 가설 검정(유의차 검정)을 하고자 한다면** 두 범주를 하나의 범주로 만들지 않고도 **Analyze / Fit Y by X 에서 Local Data Filter 등을 이용**할 수도 있습니다.

3. JMP 의 Formula 기능을 이용할 수도 있겠습니다.

Column 단위 Formula(:통계량을 구하고자 하는 변수, :범주형 구분변수, :범주형 구분변수) 이런 식으로 Formula 를 활용할 수 있습니다. 예를 들어 age 별 & sex 별 weight 의 최대 값을 구하고 싶다면, 새로운 Column 에서 Formula 를 **Col Maximum(:weight, :age, :sex)** 로 구성하면, 아래와 같이 해당 조건(12 가지)별 Max 값이 계산됩니다.

	age	sex	height	weight	Column 7
1	12	F	59	95	145
2	12	F	61	123	145
3	12	F	55	74	145
4	12	F	66	145	145
5	12	F	52	64	145
6	12	M	60	84	128
7	12	M	61	128	128
8	12	M	51	79	128
9	13	F	60	112	112
10	13	F	61	107	112
11	13	F	56	67	112

이와 관련된 Formula 는 **Statistical 아래에 Col ~~** 형태로 있습니다.

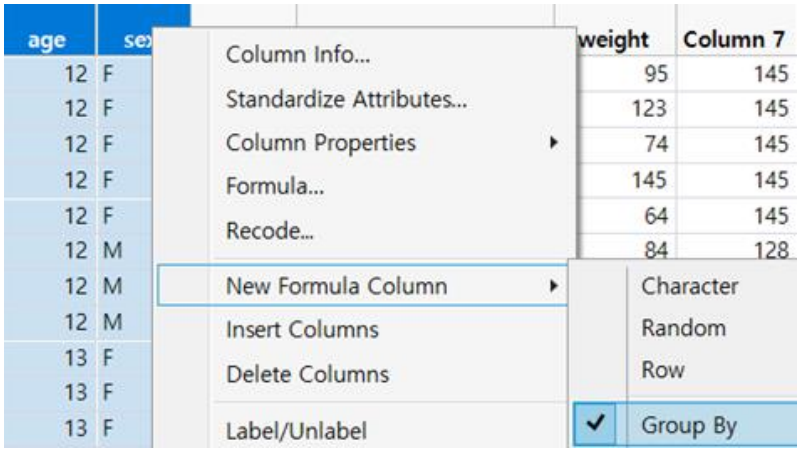


4. **New Formula Column** 기능을 활용하는 방법도 있습니다.

예를 들어 age 별 & sex 별 weight 의 평균값을 구하고 싶다면

1) Age, Sex 두 Column 선택 후 우측 마우스 클릭,

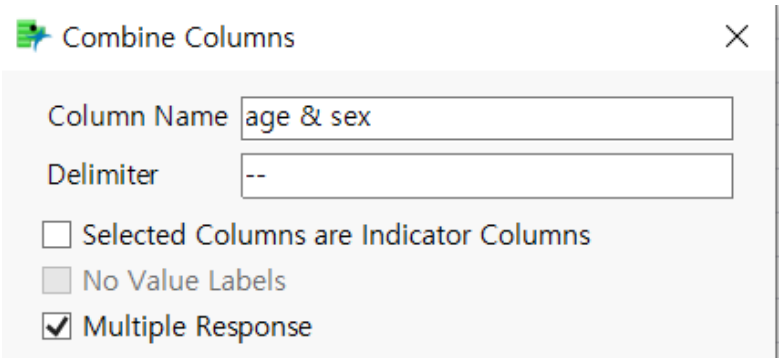
New Formula Column 에서 Group by 선택한 다음



2) weight 변수 선택 후 **New Formula Column / aggregate / mean** 을 클릭하면 다음과 같이 그 값이 표시됩니다. Formula 를 확인해 보면, 위의 3 번과 동일한 Formula 임을 알 수 있습니다.

	age	sex	height	Mean[height][age,sex]
1	12	F	59	58.6
2	12	F	61	58.6
3	12	F	55	58.6
4	12	F	66	58.6
5	12	F	52	58.6
6	12	M	60	57.333333333
7	12	M	61	57.333333333
8	12	M	51	57.333333333
9	13	F	60	59
10	13	F	61	59

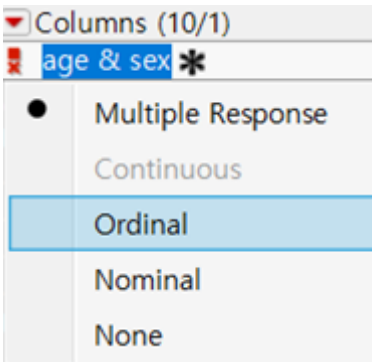
5. 이번에는 Age 변수와 Sex 변수를 하나의 변수로 만드는 방법을 알아보겠습니다.
두 변수 선택 후 **Cols / Utilities / Combine Columns** 를 클릭하여 아래와 같이 입력하면 (여기서 Delimiter 는 구분자로서 age 변수명과 sex 변수명 사이에 표시하는 구분자입니다)



아래와 같이 Age 변수와 Sex 변수가 합쳐진 하나의 변수가 만들어집니다.

	age & sex	age	sex
1	12--F	12	F
2	12--F	12	F
3	12--F	12	F
4	12--F	12	F
5	12--F	12	F
6	12--M	12	M
7	12--M	12	M

이 변수를 다른 용도로 활용하고자 한다면 Multiple Response 로 정의된 modeling Type 을 Ordinal 또는 Nominal 로 변경하여야 합니다.



6. Formula 를 활용하여 Age 변수와 Sex 변수를 하나의 변수로 만들 수도 있습니다.
약간 복잡한 Formula 입니다.
새로운 Column 에서 우측 마우스 클릭, Formula 에 들어간 다음

1) Function List 에서 **Character / Char** 선택 후 성별(sex) 변수 선택

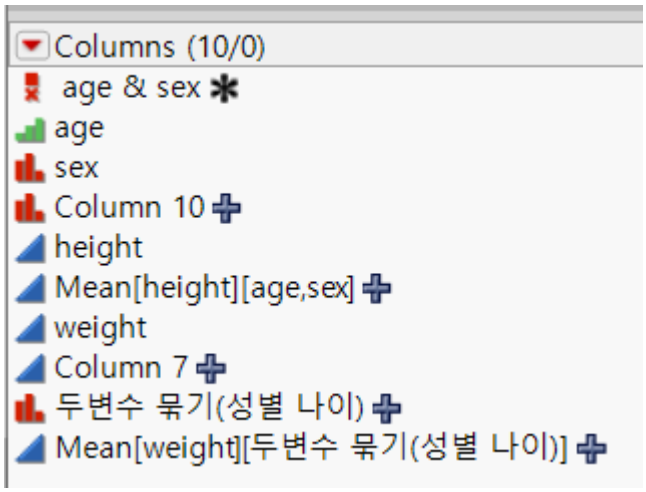
2) Function List 에서 **Character / Concat : "_"** 입력 후

3) 다시 Function List 에서 **Character / Char** 선택 후 나이(age) 변수 선택하면 아래와 같은 Formula 가 만들어지고 Data Table 에 새로운 Column 이 생성됩니다.

Char(:sex || "_" || Char(:age))

Column 10
F_12
F_12
F_12
F_12
F_12
M_12
M_12
M_12

* 별도 첨부하는 JMP File('두 명목형 변수를 하나의 변수로.jmp)의 Column Table에서 Modeling Type과 Formula를 확인해 보길 바랍니다.



---------- 끝