Monthly User Guide from JMP Korea

제 30호 (2020년 1월) : JMP 15 Version 새로운 기능 소개 (2)

Explore Patterns

* 본 Guide 의 내용과 관련한 문의는 ikju.Shin@jmp.com 으로 연락 바랍니다

** Monthly Guide 전체 내용(지난 호 포함)은 아래 Site에서 확인 가능합니다
(https://www.jmp.com/ko_kr/newsletters.html)

JMP 15 Version 새로운 기능

지난 2019년 11월 호에 이어 JMP 15 Version 새로운 기능 소개 시리즈의 두 번째로
'Explore Patterns(패턴 탐색)' 기능에 대해 소개하고자 합니다.

2019년 11월 소개 내용

JMP 15 Version의 새로운 기능*

2020년 1월 소개 내용

Explore Patterns

* 보다 세부적인 내용은

1. JMP Korea Homepage(https://www.jmp.com/ko_kr/software/new-release/new-in-jmp-and-jmp-pro.html) 및

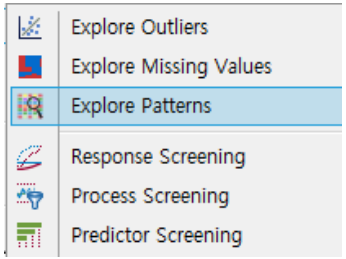
2. JMP에서 Help / New Features 를 클릭하면 됩니다

Explore Patterns(패턴 탐색)

1. Explore Patterns

1) Data에서 예기치 않는 어떤 Pattern을 Screening하는 기능
2) 주로 Data 조작(Tampering), 장비 이상 등으로 인한 어떠한 Pattern을 인식하는 기능
3) 특히, 아래와 같은 몇 가지 유형의 Data 조작 결과를 탐지해내는 기능을 가지고 있음
   -결측치(Missing Data)를 난수(Random Data)로 교체
   -다른 행 또는 열의 값을 복사하여 붙여넣기
   -선형 함수(Linear Formula)를 사용하여 값을 채워 넣기
   -Spec에 맞추어 Data 짤라내기(Spec out된 data를 삭제)

2. Menu : Analyze / Screening / Explore Patterns

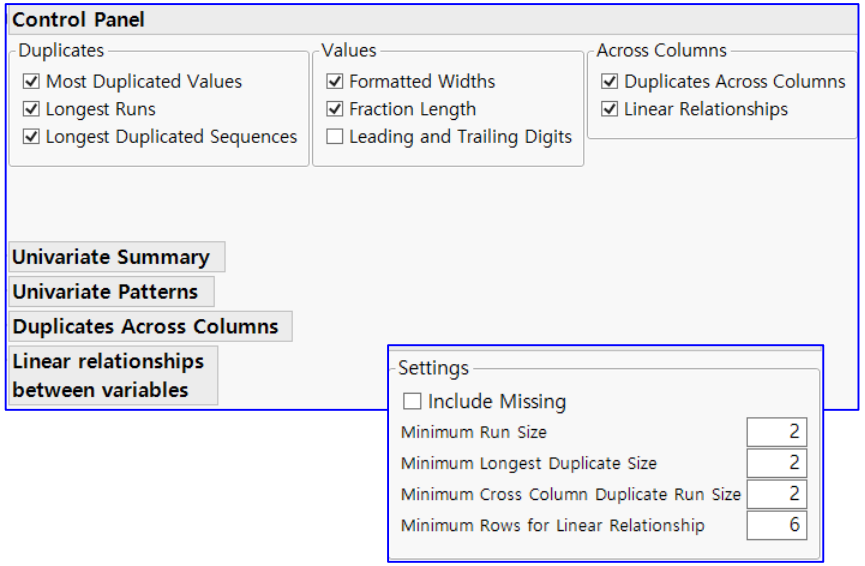


3. 활용 예제

1) Help / Sample Data Library / Nicardipine Lab Patterns.jmp
2) Lab Result 로 그룹핑된 27개 Column에 대한 Pattern 조사

4. 초기 화면

1) Analyze / Screening / Explore Patterns 에 들어가서
2) 해당 27개 Column을 선택 후 OK 클릭
3) Control Panel에서 아래와 같이 Option 선택
   -Duplicates
   -Values
   -Across Columns



Explore Patterns(패턴 탐색)

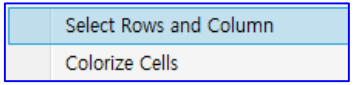
5. Univariate Summary : Longest Runs

1) 'Creatinine'를 예를 들어 설명하면, 2265번째 row부터 일곱 번 (2271번째) row까지 0.03536 이라는 숫자가 연속해서 존재한다는 뜻

Longest Runs

Column	Starting Row	Count	Value	Rarity
Creatinine	2265	7	0.03536	18.9
Hematocrit	3048	4	28.999999	13.1
Prothrombin Time	2931	6	12	10.1
BUN	2267	5	2.142	8.5
LDH	3456	3	602	7.1
Activated PTT	1460	3	30.299999	6.2
Platelet	521	3	150	5.8

2) 그 결과를 확인하고자 한다면 위의 'Creatinine' 변수 명 위에서 우측 마우스 클릭, Select Rows and Columns 선택하면



3) 아래 처럼 Data Table에서 확인 가능

	um	CO2	Chloride	Creatine Kinase	Creatinine	Er
2262	1.5	104.352	108	•	0.0884	4
2263	00...	•	112	•	0.05304	3
2264	•	•	•	•	0.0442	
2265	•	•	•	•	0.03536	
2266	•	•	•	•	0.03536	
2267	•	•	•	•	0.03536	
2268	•	•	•	•	0.03536	
2269	•	•	•	•	0.03536	
2270	•	•	•	•	0.03536	
2271	•	•	•	•	0.03536	
2272	•	•	•	•	•	

4) Rarity의 의미

-If P (x) is the probability of observing phenomenon x, according to some probability model, then Rarity = **-log2(P (x))**

-간단히 말하면, 동전 던지기를 하여 앞면(또는 뒷면)이 해당 횟수 만큼 나올 확률을 말함. Creatinine 변수의 Rarity 18.9의 의미는 연속해서 18.9 번 만큼 던졌을 때 동전 앞면(또는 뒷면)이 나올 확률

6. Univariate Summary : Longest Duplicate Runs

1) 아래 그림과 같이 몇 개의 연속된 열이 같은 값을 가지는 경우를 확인하는 기능이다.

	Col1
1	0.7675205788
2	0.0030591879
3	0.8760613939
4	0.2370808288
5	0.0556239646
6	0.5823953294
7	0.2075038038
8	0.6209054205
9	0.5104548738
10	0.2346275104
11	0.2772667364
12	0.1994321193
13	0.0030591879
14	0.8760613939
15	0.2370808288
16	0.0556239646
17	0.5823953294
18	0.1879393037

Explore Patterns(패턴 탐색)

2) 아래와 같이 'Activated PTT' 변수 명 위에서
우측 마우스 클릭, Select Rows and Columns 선택

Longest Duplicate Sequences

Column	Starting Row I	Starting Row J	Count	Rarity
Activated PTT	2816	3034	4	18.2
Prothrombin Time	2942	2966	6	14.9
Hematocrit	3051	3057	3	14.6
Leukocytes	547	1275	3	11.1
Creatinine	1464	1681	8	11.0
Leukocytes	975	1485	3	10.8
ALT	840	2425	3	10.5
Sodium	2339	2976	6	9.8

3) Data Table에서 Tables / Subset, selected rows 선택한 뒤
OK 클릭
: 다른 Column을 확인하여 해당 Data에 대한 세부적인 내용 확인

	STUDYID	SITEID	USUBJID	AGE	SEX	Activated PTT	ALT	ALP	A
1	NICSA...	40	401005	48	M	25.700001	192	179	
2	NICSA...	40	401006	48	F	29.799999	11	108	
3	NICSA...	40	401006	48	F	26.9	11	68	
4	NICSA...	40	401006	48	F	28.299999	36	82	
5	NICSA...	44	441011	37	M	25.700001	•	214	
6	NICSA...	44	441011	37	M	29.799999	•	150	
7	NICSA...	44	441012	65	F	26.9	•	251	
8	NICSA...	44	441012	65	F	28.299999	•	231	

7. Univariate Summary : Fractional Length >= 15

1) 단순한 합 또는 비율이거나 인위적으로 만든 값인지를 확인하는 데 활용된다.
2) Fraction Length(분수 길이, 또는 continued fraction)
-연속적으로 나누어진 항의 순서를 의미
(number as a sequence of continually divided terms)
3) 기준 값으로 15를 많이 활용하며, 15 이상일 경우 난수(random number) 사용 또는 특정한 함수 사용으로 인위적으로 만든 Data일 가능성을 의심한다.
4) 아래의 결과는 'Urate' 데이터의 약 22%가 Fractional Length가 15 이상이라는 의미이다

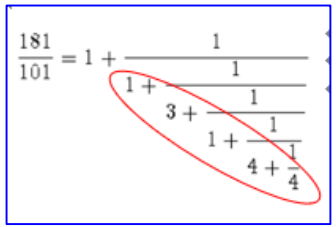
Fraction Length>=15

Column	Rate
Urate	0.2212
Creatinine	0.1764
Phosphate	0.1754
pH	0.0691
Erythrocytes	0.0684
BUN	0.0652
Glucose	0.0503
Calcium	0.0409
CO2	0.0192
PCO2	0.0151
Partial Pressure Oxygen	0.0092
Bilirubin	0.0067

5) Fractional Length에 대한 세부 사항은 아래 링크 참조
https://en.wikipedia.org/wiki/Continued_fraction

Explore Patterns(패턴 탐색)

6) Fractional Length에 대해 이미지로서 간략히 설명하면,
특정한 분수는 아래와 같이 표현될 수 있으며
이럴 경우 continued fraction은 보통 {1; 1, 3, 1, 4, 4}로 나타내고
fraction length 는 5 로 표현된다.



3) Formatted Widths와 Fraction Length 확인 가능

Formatted Widths

Width	Overall Count	Decimal Count
0	0	889
1	0	606
2	888	0
3	1	0
4	606	0
5	0	0
6	0	635
7	0	0
8	0	0
9	635	0

Fraction Length

Length	Count
0	889
1	255
2	239
3	112
4	125
5	172
6	28
7	173
8	137

'25' 와 같이 자릿수(Width)가 2 이고
소수점(Decimal) 이하의 수가 0 인
Data가 888 개 있다는 뜻

8. Univariate Pattern

1) 변수 별로 확인 가능, 여기서는 'Activated PTT' 를 선택
2) Missing Row의 개수, 특정 값이 몇 번 나오는 지, Longest Runs
등을 확인할 수 있으며

Number of rows	Number Missing	Number of Unique Values
3463	1333	222

Most Duplicated Values

Value	Count
23	95
24	90
25	83
27	82
22	74
26	66
28	65
29	57
21	50
30	47

Longest Runs

Starting Row	Count	Value	Rarity
1460	3	30.299999	6.2
3435	3	29.5	4.6
1032	3	19	3.1
1400	3	28	0.5
1456	3	28	0.5
469	3	26	0.5
1051	3	22	0.3
1056	3	22	0.3
758	3	27	0.2
960	3	24	0.1
933	3	23	0.1

Longest Duplicate Sequences

Starting Row I	Starting Row J	Count	Rarity	First few values
2816	3034	4	18.2	25.7 29.8 26.9
1050	1055	4	9.4	20 22 22
1528	2386	4	9.1	27 31 24
773	2587	4	8.4	26 27 26

Explore Patterns(패턴 탐색)

9. Duplicates Across Columns

: 둘 이상의 변수에서 동일한 값이 있는 지를 확인할 수 있다

Duplicates Across Columns

With runs at least 2 Colorize

from Row	to Row	Number Rows	Number Equal	Duplicate Sets
603	604	2	2	Erythrocytes, Potassium
1593	1594	2	2	Platelet, Sodium
2052	2053	2	2	Creatine Kinase, LDH
2516	2517	2	2	Activated PTT, Hematocrit
2572	2573	2	2	Activated PTT, Hematocrit
2580	2581	2	2	Activated PTT, Hematocrit
3378	3379	2	2	ALT, AST

	Creatine Kinase	Creatinine	Erythrocytes	Potassium
599	•	0.10608	3.5999999	3.8
600	•	0.10608	3.4000001	3.7
601	•	0.15912	3	3.5999999
602	•	0.12376	2.5999999	4
603	•	0.07072	4	4
604	•	0.05304	3.4000001	3.4000001
605	•	0.16796	3.3	4.5

10. Linear relationships between variables

1) 변수 간에 k 개의 연속된 row에 대해 선형 관계 여부를 확인할 수 있는 기능이다
2) 결과가 아래와 같다면 그 의미는
   -2030번 row ~ 2035번 row 에서
   -Creatinine 변수와 Chloride 변수 간에 다음과 같은 함수식을 가지는 선형 관계가 있다는 뜻이다.
   Creatinine = -0.81328 + 0.00884*Chloride

Exact relation for at least 6 Rows. Colorize

Y	X	Rows	Count	Constant	Slope
Creatinine	BUN	2267-2271 2273	6	-0.0177	0.02476
Creatinine	Chloride	2030-2035	6	-0.8133	0.00884

	rubin	BUN	Calcium	CO2	Chloride	Creatine Kinase	Creatinine
2027	•	15.708	•	134....	111	•	0.09724
2028	•	10.353	•	121....	107	•	0.07956
2029	•	•	•	•	•	•	•
2030	0066...	2.856	2.125	104....	101	86	0.07956
2031	0029...	5.712	2.2750001	143....	100	27	0.07072
2032	0066...	3.927	2.200000...	113....	99	15	0.06188
2033	0166...	5.355	2.299999...	117....	100	108	0.07072
2034	0266...	3.927	2.2750001	121....	98	255	0.05304
2035	•	1.428	•	121....	98	•	0.05304