データ操作の質を確保する手段としての統計的方法

東京理科大学　安井　清一

jmp

1. はじめに【背景】

DX, IoT, Industry4.0,
ビッグデータ・・・

- 様々な種類のデータ
- 大量のデータ

【元データ】

条件文は if, and, or.
所望の変換がおこなれているか心配

✓ 条件抽出
✓ 変数変換
✓ カテゴリ化

解析対象となる
データセット

【解析用データセット】
Analytical DataSet
(ADS)

JAPAN 2020 DISCOVERY SUMMIT ONLINE

jmp

1. はじめに【Motivating Example】

介護施設におけるIT支援システムと介護の質に関する研究

行動検知センサーのログを分析する。

約 3,000 件/日 × 約 1 年 = 約 100 万件
変数は 30 種類

ある時間帯のみを抽出，部屋No.からフロア階数を求める，・・・

JAPAN 2020 DISCOVERY SUMMIT ONLINE

jmp

1. はじめに【Motivating Example】

火災報告を用いた火災現象・消火活動の統計モデリング

約 50,000件/年 × 約 4 年 = 約 20 万件
変数は 約 200 種類
カテゴリ変数が多く，分類が細かい。

対象となる建築物種別・用途を抽出，分類を統合，平方根・対数変換，・・・

JAPAN 2020 DISCOVERY SUMMIT ONLINE

jmp

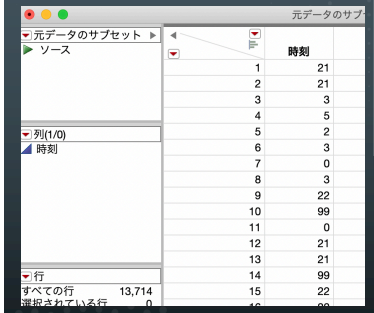
1. はじめに【問題提起】

どのようにして，『所望の変換が行われているか』を確かめるか？

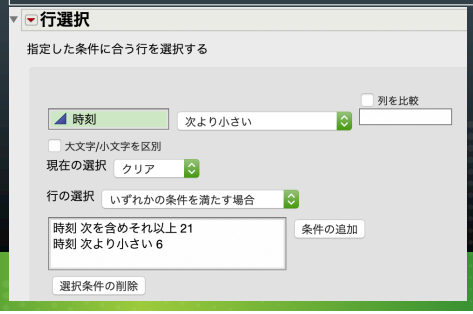
スモールデータでも，目視は危ない！

統計解析を利用して，チェックできる。

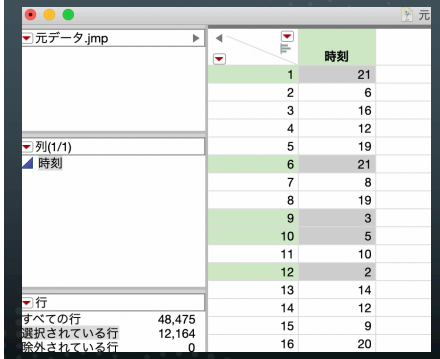
2. 例：21時から6時までを選択する。



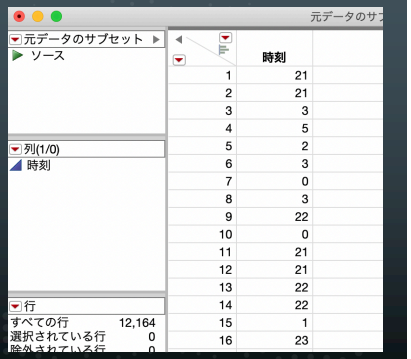
1. 行選択で
   - 「時刻　次を含めそれ以上 21」
   - 「時刻　次より小さい 6」
   - を「いずれかの条件を満たす場合」



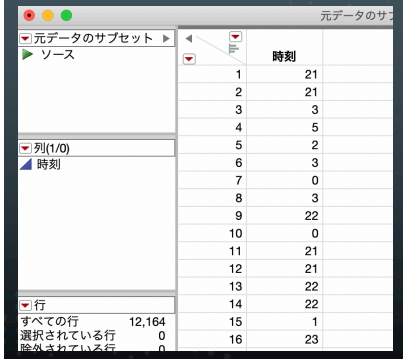
2. 例：21時から6時までを選択する。



サブセット

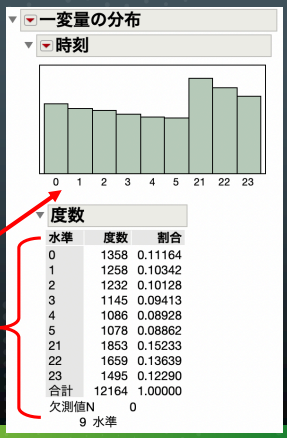


2. 例：21時から6時までを選択する。

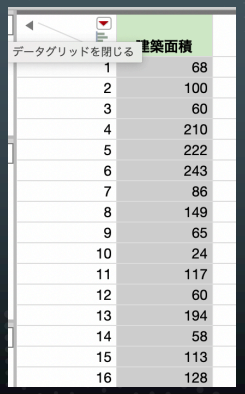


- 順序尺度に変える。
- 1変量の分布。

21 ~ 23，0 ~ 5 なのでOK



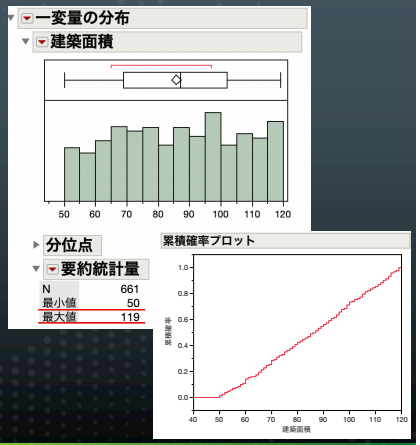
3. 例：連続量の切り出し



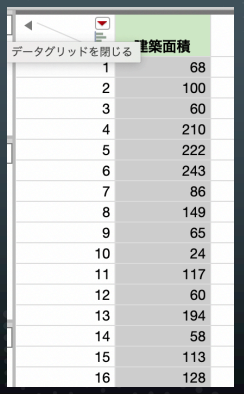
	建築面積
1	68
2	100
3	60
4	210
5	222
6	243
7	86
8	149
9	65
10	24
11	117
12	60
13	194
14	58
15	113
16	128

50以上，120未満

- 行選択（orルール）
- サブセット



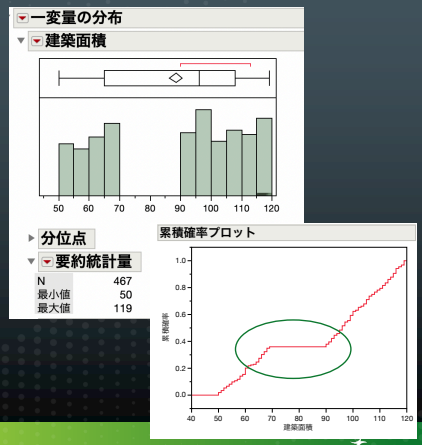
4. 例：連続量の切り出し（歯抜けの発見）



	建築面積
1	68
2	100
3	60
4	210
5	222
6	243
7	86
8	149
9	65
10	24
11	117
12	60
13	194
14	58
15	113
16	128

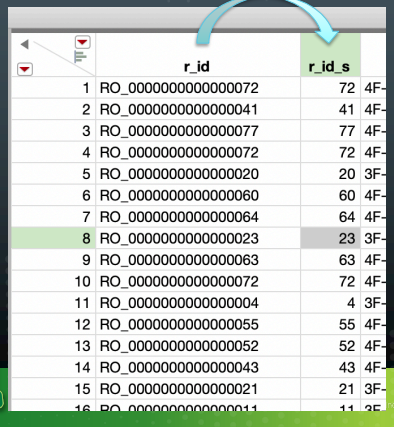
50以上，120未満

- 行選択（orルール）
- サブセット



5. 例：余計な文字を消去した変数を作る。

RO_・・・を消す『right(“r_id”,2)』（1対1の変換）

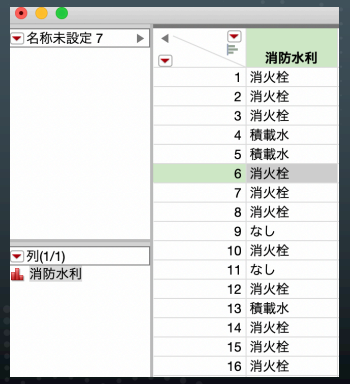


	r_id	r_id_s	
1	RO_0000000000000072	72	4F-
2	RO_0000000000000041	41	4F-
3	RO_0000000000000077	77	4F-
4	RO_0000000000000072	72	4F-
5	RO_0000000000000020	20	3F-
6	RO_0000000000000060	60	4F-
7	RO_0000000000000064	64	4F-
8	RO_0000000000000023	23	3F-
9	RO_0000000000000063	63	4F-
10	RO_0000000000000072	72	4F-
11	RO_0000000000000004	4	3F-
12	RO_0000000000000055	55	4F-
13	RO_0000000000000052	52	4F-
14	RO_0000000000000043	43	4F-
15	RO_0000000000000021	21	3F-

- 名義尺度 vs 名義尺度の2変量の関係

R2乗（U） = 1.000

6. 例：項目をチェックし，分類を統合する。



	消防水利
1	消火栓
2	消火栓
3	消火栓
4	積載水
5	積載水
6	消火栓
7	消火栓
8	消火栓
9	なし
10	消火栓
11	なし
12	消火栓
13	積載水
14	消火栓
15	消火栓
16	消火栓

（多対1の変換）

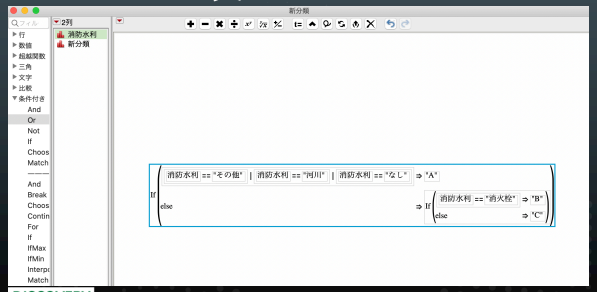
項目の種類をチェックする。

1変量の分布

6. 例：項目をチェックし，分類を統合する。

分類を統合する。

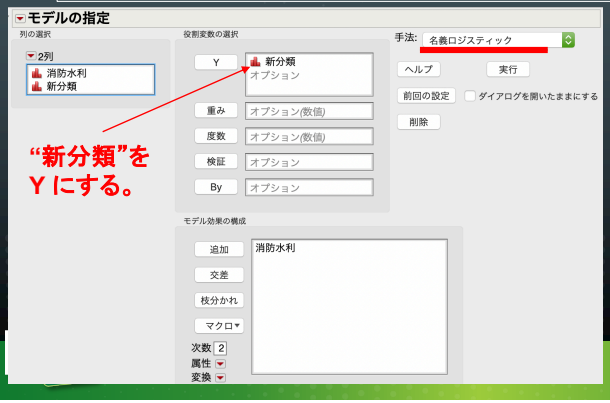
if 文



	消防水利	新分類
1	消火栓	B
2	消火栓	B
3	消火栓	B
4	積載水	C
5	積載水	C
6	消火栓	B
7	消火栓	B
8	消火栓	B
9	なし	A
10	消火栓	B
11	なし	A
12	消火栓	B
13	積載水	C
14	消火栓	B
15	消火栓	B
16	消火栓	B

6. 例：項目をチェックし，分類を統合する。

まずは，名義尺度 vs 名義尺度 の『**モデルのあてはめ**』で1対1対応になっているかチェックする。



“新分類”を Y にする。

（多クラス）ロジステック回帰で，新分類を元の分類で予測（判別）する。寄与率 = 1，誤判別率 = 0 であれば，まずは大丈夫。

6. 例：項目をチェックし，分類を統合する。

名義尺度 vs 名義尺度 の『**モデルのあてはめ**』でチェックする。

名義ロジスティックのあてはめ 新分類

効果の要約

勾配で収束しました，19回の反復

反復履歴

モデル全体の検定

モデル	(-1)*対数尤度	自由度	カイ2乗	p値(Prob>ChiSq)
差	1111.9079	8	2223.816	<.0001*
完全	1.8276e-6			
縮小	1111.9079			

R2乗(U)	1.0000
AICc	20.1582
BIC	72.4566
オブザベーション(または重みの合計)	1402

あてはめの詳細

指標	学習	定義
エントロピーR2乗	1.0000	1-Loglike(model)/Loglike(0)
一般化R2乗	1.0000	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
平均 -Log p	0.0000	∑ -Log(p[j])/n
RASE	0.0000	√ ∑(y[j]-p[j])²/n
平均 絶対偏差	0.0000	∑ |y[j]-p[j]|/n
誤分類率	0.0000	∑ (p[j]≠pMax)/n
N	1402	n

R2乗（U） = 1.000

混同行列

学習

実測値	予測値 度数		
新分類	A	B	C
A	113	0	0
B	0	972	0
C	0	0	317

誤判別率0%

6. 例：項目をチェックし，分類を統合する。

続いて，『2変数の関係』で変換前後を**分割表**でチェックする。

消防水利

新分類	度数 全体% 列% 行%	その他	なし	河川等	消火栓	積載水	合計
A		4	109	0	0	0	113
		0.29	7.77	0.00	0.00	0.00	8.06
		100.00	100.00	0.00	0.00	0.00	
		3.54	96.46	0.00	0.00	0.00	
B		0	0	0	972	0	972
		0.00	0.00	0.00	69.33	0.00	69.33
		0.00	0.00	0.00	100.00	0.00	
		0.00	0.00	0.00	100.00	0.00	
C		0	0	55	0	262	317
		0.00	0.00	3.92	0.00	18.69	22.61
		0.00	0.00	100.00	0.00	100.00	
		0.00	0.00	17.35	0.00	82.65	
合計		4	109	55	972	262	1402
		0.29	7.77	3.92	69.33	18.69	

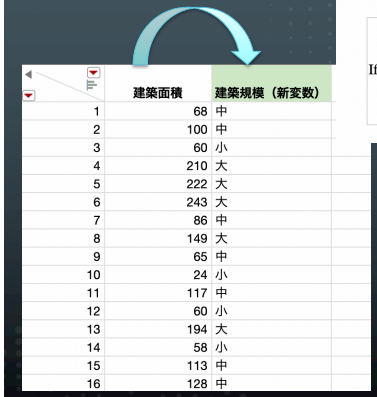
検定

N	自由度	(-1)*対数尤度	R2乗(U)
1402	8	1111.9079	0.8718

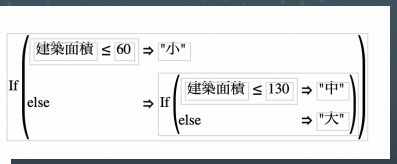
検定	カイ2乗	p値(Prob>ChiSq)
尤度比	2223.816	<.0001*
Pearson	2804.000	<.0001*

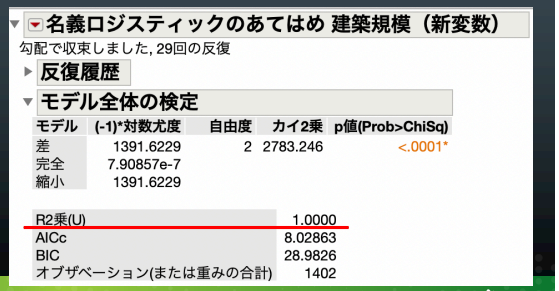
列もしくは行ごとに見て行って，0でないセルの行分類と列分類が所望のものかチェックする。

7. 例　連続量をカテゴリ(離散)化する。



	建築面積	建築規模（新変数）
1	68	中
2	100	中
3	60	小
4	210	大
5	222	大
6	243	大
7	86	中
8	149	大
9	65	中
10	24	小
11	117	中
12	60	小
13	194	大
14	58	小
15	113	中
16	128	中





名義ロジスティックのあてはめ 建築規模（新変数）

勾配で収束しました, 29回の反復

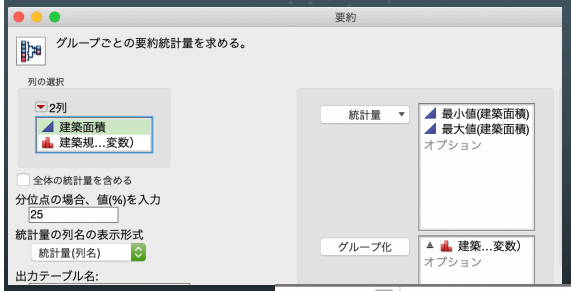
反復履歴

モデル全体の検定

モデル	(-1)*対数尤度	自由度	カイ2乗	p値(Prob>ChiSq)
差	1391.6229	2	2783.246	<.0001*
完全	7.90857e-7			
縮小	1391.6229			

R2乗(U)	1.0000
AICc	8.02863
BIC	28.9826
オブザベーション(または重みの合計)	1402

7. 例　連続量をカテゴリ(離散)化する。



続いて,『要約(最大、最小)』で所望のカテゴリ化が行われているか確認する。

	建築規模（新変数）	行数	最小値(建築面積)	最大値(建築面積)
1	小	188	0	60
2	大	569	131	4012
3	中	645	61	130

8　まとめ

- データの規模が大きくなるとカテゴリ項目数も多くなる。カテゴリ項目、数値の範囲等の確認は『1変数の分布』を利用する。
- 変換の前後は『2変量の関係』、『モデルのあてはめ』で寄与率『RO(U)』= 1.00 で「1対1」「多対1」変換になっているかをチェック。その後に、分割表で所望の変換になっているかをチェックする。

8　まとめ

データ解析

- 確率論的
- 汎化能力の確保、過学習の防止
- 寄与率 = 1.00 は NG.

データ準備

- 決定論的
- 変換においては、寄与率 = 1.00 が正しい。