複雑なデータを解析しやすくするためのSASによるデータ加工

加工プログラム例
―少数ステップでの処理―

持田製薬　髙田康行

本日の内容

- 本日示すプログラムの目的
- 素直な方法
  - TRANSPOSEプロシジャ
  - LENGTH関数とSUBSTR関数
  - RETAINステートメント
  - first.by変数
- 簡潔な方法
  - ARRAYステートメント
  - VNAME関数

本日示すプログラムの目的

タイプ1：1症例1オブザベーション

ID	TC1	HDL1	TC2	HDL2	TC3	HDL3
1	212	50	224	64	204	73
2	206	58	208	63	212	58
3	221	47	236	70	242	38
4	251	60	239	37	236	42

加工

タイプ4：症例ごと項目ごと時点ごとの繰り返し

ID	LBCD	LB	VISIT	VAL
1	1	TC	1	212
1	1	TC	2	224
1	1	TC	3	204
1	2	HDL	1	50
1	2	HDL	2	73
1	2	HDL	3	73
2	1	TC	1	206
2	1	TC	2	208
2	1	TC	3	212
2	2	HDL	1	58
2	2	HDL	2	58
2	2	HDL	3	58
3	1	TC	1	221
3	1	TC	2	236

- 測定値が横並び
- 検査項目名と時点が併合された変数名

- 測定値が縦並び
- 検査項目名 (LB) と時点(VISIT)が別々

素直な方法

横に並んでいるものを
縦の並びに転置

何も考えないで転置すると･･･

```
Proc TRANSPOSE data=TMP1
               out =TMP2 ;
run ;
```

被験者？

不要

OBS	_NAME_	COL1	COL2	COL3	COL4
1	ID	1	2	3	4
2	TC1	212	206	221	251
3	HDL1	50	58	47	60
4	TC2	224	208	236	239
5	HDL2	64	63	70	37
6	TC3	204	212	242	?
7	HDL3	73	58	38	?

並びはこれでいい？

イメージと異なる！

被験者ごとに転置すると
目的のイメージに近いが・・・

```
Proc TRANPOSE data=TMP1
              out =TMP3 ;
  by ID ;
run ;
```

OBS	ID	TC1	HDL1	TC2	HDL2	TC3	HDL3
1	1	212	50	224	64	204	73
2	2	206	58	208	63	212	58
3	3	221	47	236	70	242	38
4	4	251	60	239	37	236	42

OBS	ID	_NAME_	COL1
1	1	TC1	212
2	1	HDL1	50
3	1	TC2	224
4	1	HDL2	64
5	1	TC3	204
6	1	HDL3	73
7	2	TC1	206
8	2	HDL1	58
9	2	TC2	208
10	2	HDL2	63
11	2	TC3	212
12	2	HDL3	58
13	3	TC1	221
14	3	HDL1	47
15	3	TC2	236
16	3	HDL2	70
17	3	TC3	242
18	3	HDL3	38
19	4	TC1	251
20	4	HDL1	60
21	4	TC2	239
22	4	HDL2	37
23	4	TC3	236
24	4	HDL3	42

問題1：検査項目名と時点の情報が併合されていると使いにくい

問題1：検査項目名と時点の情報を分離

- Ifステートメントを用いて全てを条件付け

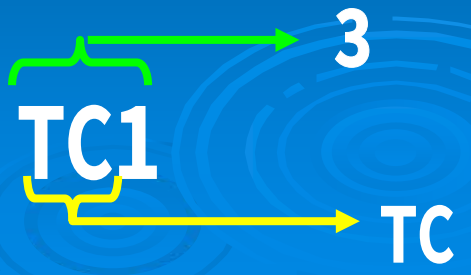
```
data TMP4 ;
  set TMP3 ;
  if _NAME_ = "TC1" then do ;
    LB = "TC " ;
    VISIT = 1 ;
  end ;
  if _NAME_ = "TC2" then do ;
    LB = "TC " ;
    VISIT = 2 ;
  end ;
  ....
run ;
```

全ての組み合わせを
記載するのは大変

LENGTH関数とSUBSTR関数

- LENGTH関数
  - LENNGTH(変数)
  - 文字変数に使うと、内容の文字数が分かる
- SUBSTR関数
  - SUBSTR(文字変数,x,y)
  - 文字列から、x文字めからy文字分を抽出

- LENGTH(_NAME_)
- SUBSTR(_NAME_,1,2)



関数を使うと

```
data TMP5 ; set TMP3 ;
   L      = LENGTH(_NAME_) ;
   LB     = SUBSTR(_NAME_,1,L-1) ; **項目名は最後以外;
   VISIT = SUBSTR(_NAME_,L,L) ;    **時点は最後の1文字;
run ;
```

ID	_NAME_	COL1	L	LB	VISIT
1	TC1	212	3	TC	1
1	HDL1	50	4	HDL	1
1	TC2	224	3	TC	2
1	HDL2	64	4	HDL	2
1	TC3	204	3	TC	3
1	HDL3	73	4	HDL	3
2	TC1	206	3	TC	1
2	HDL1	58	4	HDL	1
2	TC2	208	3	TC	2
2	HDL2	63	4	HDL	2
2	TC3	212	3	TC	3
2	HDL3	58	4	HDL	3

問題2：
数字に見えるが
実は文字型

問題3：
表示したい順は
TC→HDL
ソートすると
HDL→TC

問題2：変数の文字型を数値型に

```
data TMP6 ; set TMP3 ;
   L      = LENGTH(_NAME_) ;
   LB     = SUBSTR(_NAME_,1,L-1) ;
   VISIT = SUBSTR(_NAME_,L,L) + 0 ;
run ;
```

数値計算式を
入れることで
数値とみなされる

#	変数	タイプ	長さ	ラベル
1	ID	数値	8	
2	_NAME_	文字	8	前の変数名
3	COL1	数値	8	
4	L	数値	8	
5	LB	文字	8	
6	VISIT	文字	8	

#	変数	タイプ	長さ	ラベル
1	ID	数値	8	
2	_NAME_	文字	8	前の変数名
3	COL1	数値	8	
4	L	数値	8	
5	LB	文字	8	
6	VISIT	数値	8	

問題3：表示順を示す変数を作成

- Ifステートメントを用いて全てを条件付け

```
data TMP7 ; set TMP3 ;
   L      = length(_name_) ;
   VISIT = SUBSTR(_NAME_,L,L) + 0 ;
   LB     = SUBSTR(_NAME_,1,L-1) ;
   if LB = "TC"  then LBCD = 1 ;
   if LB = "HDL" then LBCD = 2 ;
run ;
```

OBS	ID	COL1	_NAME_	VISIT	LB	LBCD
1	1	212	TC1	1	TC	1
2	1	50	HDL1	1	HDL	2
3	1	224	TC2	2	TC	1
4	1	64	HDL2	2	HDL	2
5	1	204	TC3	3	TC	1
6	1	73	HDL3	3	HDL	2
7	2	206	TC1	1	TC	1
8	2	58	HDL1	1	HDL	2
9	2	208	TC2	2	TC	1
10	2	63	HDL2	2	HDL	2

たくさんの変数では大変

RETAINステートメント

- RETAINステートメント
  - RETAIN 変数名 [初期値]
  - 次のオブザベーションを読むとき、前の値を保持

```
data SMP1 ;
  RETAIN SUM 0 ;
  do i = 1 to 3 ;
    RET   = SUM ;      *RETAINで保持された計算前の値;
    SUM   = SUM  + i ;*計算後、保持される値;
    SUMX  = SUMX + i ;*RETAINなしの場合;
    OUTPUT ;
  end ;
run ;
```

OBS	i	RET	SUM	SUMX
1	1	0	1	.
2	2	1	3	.
3	3	3	6	.

first.by変数とlast.by変数

- first.by変数
  - Byステートメントで指定済みグループの最初のオブザベーションのとき1、それ以外は0
- last.by変数
  - first.by変数と同様

by ID ; の場合

ID	first.by変数	last.by変数
1	1	0
1	0	0
1	0	0
1	0	0
1	0	0
1	0	1
2	1	0
2	0	0
2	0	0

結果に残らない
変数として作成される

RETAINステートメントとfirst.by変数を利用

```
Proc SORT data=TMP6 ;
  by ID VISIT ;
run ;
data TMP8 ; set TMP6 ;
  by ID VISIT ;
  RETAIN LBCD ;
  if first.VISIT then LBCD = 0 ;
  LBCD = LBCD + 1 ;
run ;
```

first.VISIT

ID	COL1	VISIT	LB	LBCD
1	212	1	TC	1
1	50	1	HDL	2
1	224	2	TC	1
1	64	2	HDL	2
1	204	3	TC	1
1	73	3	HDL	2
2	206	1	TC	1
2	58	1	HDL	2
2	208	2	TC	1
2	63	2	HDL	2
2	212	3	TC	1
2	58	3	HDL	2
3	221	1	TC	1
3	47	1	HDL	2
3	236	2	TC	1
3	70	2	HDL	2
3	242	3	TC	1
3	38	3	HDL	2
4	251	1	TC	1
4	60	1	HDL	2
4	239	2	TC	1
4	37	2	HDL	2
4	236	3	TC	1
4	42	3	HDL	2

RETAIN

最後に順番を整えて･･･

```
Proc SORT data=TMP8 ;
  by ID LBCD VISIT ; *被験者ごと検査項目ごと時点ごと;
run ;
```

タイプ4のデータ

ID	LBCD	LB	VISIT	COL1
1	1	TC	1	212
1	1	TC	2	224
1	1	TC	3	204
1	2	HDL	1	50
1	2	HDL	2	64
1	2	HDL	3	73
2	1	TC	1	206
2	1	TC	2	208
2	1	TC	3	212
2	2	HDL	1	58
2	2	HDL	2	63
2	2	HDL	3	58

簡潔な方法

一つのデータステップで

ARRAYステートメント

- ARRAYステートメント
  - 配列が定義可能
  - ARRAY 配列名 {要素数} [配列要素] ;
    - ARRAY aaa{3} TC1 TC2 TC3 ;
  - ARRAYステートメントの内容イメージ

TC1	TC2	TC3

➡

aaa{1}	aaa{2}	aaa{3}

配列 aaa と定義する

TC1	TC2	TC3
HDL1	HDL2	HDL3

➡

bbb{1,1}	bbb{1,2}	bbb{1,3}
bbb{2,1}	bbb{2,2}	bbb{2,3}

配列 bbb と定義する

ARRAYステートメントの使い方 (1)

- TCのみ実行

```
data ARRY1 ;
  set TMP1 ;
  by ID ;
  ARRAY aaa{3} TC1 TC2 TC3 ;
  do VISIT = 1 to 3 ;
    VAL = aaa{VISIT} ;
    output ;
  end ;
run ;
```

aaa{1}

aaa{2}

aaa{3}

OBS	ID	TC1	TC2	TC3	VISIT	VAL
1	1	212	224	204	1	212
2	1	212	224	204	2	224
3	1	212	224	204	3	204
4	2	206	208	212	1	206
5	2	206	208	212	2	208
6	2	206	208	212	3	212
7	3	221	236	242	1	221
8	3	221	236	242	2	236
9	3	221	236	242	3	242
10	4	251	239	236	1	251
11	4	251	239	236	2	239
12	4	251	239	236	3	236

ARRAYステートメントの使い方 (2)

- TCとHDLを同時に実行

```
data ARRY2 ;
  set TMP1 ;
  by ID ;
  array bbb{2, 3} TC1   TC2   TC3
                  HDL1  HDL2  HDL3 ;
  do LBCD = 1 to 2 ;
    do VISIT = 1 to 3 ;
      VAL = bbb{ LBCD, VISIT} ;
      output ;
    end ;
  end ;
run ;
```

3

2

OBS	ID	LBCD	VISIT	VAL
1	1	1	1	212
2	1	1	2	224
3	1	1	3	204
4	1	2	1	50
5	1	2	2	64
6	1	2	3	73
7	2	1	1	206
8	2	1	2	208
9	2	1	3	212
10	2	2	1	58
11	2	2	2	63
12	2	2	3	58
13	3	1	1	221
14	3	1	2	236
15	3	1	3	242
16	3	2	1	47
17	3	2	2	70
18	3	2	3	38
19	4	1	1	251
20	4	1	2	239
21	4	1	3	236
22	4	2	1	60
23	4	2	2	37
24	4	2	3	42

問題4：
検査項目名は
どこに？

VNAME関数

- VNAME関数
  - VNAME(変数)
  - 指定した変数の名前を取得

```
data SMP2 ; set TMP1 ;
  NM1 = VNAME(TC1) ;
  NM2 = VNAME(HDL1) ;
run ;
```

OBS	ID	TC1	HDL1	NM1	NM2
1	1	212	50	TC1	HDL1
2	2	206	58	TC1	HDL1
3	3	221	47	TC1	HDL1
4	4	251	60	TC1	HDL1

問題4：検査項目名の取得

```
data ARRY3 ;
  set TMP1 ;
  by ID ;
  ARRAY bbb{2, 3} TC1  TC2  TC3
                  HDL1 HDL2 HDL3 ;
  do LBCD = 1 to 2 ;
    do VISIT = 1 to 3 ;
      VAL   = bbb{ LBCD, VISIT} ;
      NM    = VNAME(bbb{ LBCD, VISIT}) ;
      LB    = SUBSTR(NM,1,LENGTH(NM)-1) ;
      output ;
    end ;
  end ;
run ;
```

VNAME関数と
ARRAYを同時利用

素直な方法の
2行分を1行で

ID	LBCD	LB	VISIT	VAL
1	1	TC	1	212
1	1	TC	2	224
1	1	TC	3	204
1	2	HDL	1	50
1	2	HDL	2	64
1	2	HDL	3	73
2	1	TC	1	206
2	1	TC	2	208

タイプ4のデータ

まとめ

- データ加工におけるプログラム例を提示
  - 素直な方法と簡潔な方法
  - 知っていると便利な機能
- ここで記載したのは個人的なノウハウ
  - SASマクロを使った手法を林さんから発表
  - SASのトレーニングやQ&Aを活用してください！
  - もっとスマートな方法があれば教えてください！

付録：本スライド中のメインプログラム

```
data TMP1 ;
   input ID  TC1   HDL1  TC2   HDL2  TC3   HDL3 ;
cards ;
1  212  50   224   64    204   73
2  206  58   208   63    212   58
3  221  47   236   70    242   38
4  251  60   239   37    236   42
;
run ;
```

付録：本スライド中のメインプログラム

```
Proc TRANPOSE data=TMP1
              out =TMP3 ;
  by ID ;
run ;
data TMP6 ; set TMP3 ;
  L     = LENGTH(_NAME_) ;
  LB    = SUBSTR(_NAME_,1,L-1) ;
  VISIT = SUBSTR(_NAME_,L,L) + 0 ;
run ;
Proc SORT data=TMP6 ;
  by ID VISIT ;
run ;
```

付録:本スライド中のメインプログラム

```
data TMP8 ; set TMP6 ;
  by ID VISIT ;
  RETAIN LBCD ;
  if first.VISIT then LBCD = 0 ;
  LBCD = LBCD + 1 ;
  rename COL1 = VAL ;
run ;
Proc SORT data=TMP8 ;
  by ID LBCD VISIT ;
run ;
```

付録:本スライド中のメインプログラム

```
data ARRY3 ;
  set TMP1 ;
  by ID ;
  ARRAY bbb{2, 3} TC1  TC2  TC3
                  HDL1 HDL2 HDL3 ;
  do LBCD = 1 to 2 ;
    do VISIT = 1 to 3 ;
      VAL   = bbb{ LBCD, VISIT} ;
      NM    = VNAME(bbb{ LBCD, VISIT}) ;
      LB    = SUBSTR(NM,1,LENGTH(NM)-1) ;
      output ;
    end ;
  end ;
run ;
```