Tipps & Tricks

Wo ist Mister X?
oder: Welche Variablen sind (wie) belegt?

Heribert Ramroth
Institute of Public Health
INF 324
69120 Heidelberg
Heribert.Ramroth@uni-heidelberg.de

Zusammenfassung

Stellen Sie sich vor, Sie haben eine Anzahl von n (Indikator-)Variablen. Sie möchten gerne namentlich wissen, welche Variablen davon je Datenzeile ein bestimmtes Muster enthalten (z.B. das Zeichen X oder die Zahl 1 als Indikatorwert, bzw. die Zeichenkette XXX für nicht-lesbare Werte). Unübersichtlich wird es dann, wenn es sich um 30, 50 oder 100 Variablen handelt. Ziel der hier vorgestellten Lösung ist es, eine zusammenfassende Variable zu erzeugen, welche alle Variablen namentlich auflistet, die dieses Muster enthalten.

Schlüsselwörter: Indikatorvariablen, Mustererkennung, namentliche Auflistung, Makro %SucheMrX

1 Ausgangssituation

Gegeben seien n Variablen, in denen nach einem bestimmten Muster gesucht werden soll. Die Variablen seien zum Beispiel mit dem Zeichen X, der Zeichenkette XXX, oder einem Indikatorwert 1 belegt. Auf dem Bildschirm ist die Darstellungsmöglichkeit der Variablen begrenzt: es passt nur eine bestimmte kleine Anzahl von Variablen auf die Bildschirmbreite. Wenn viele Variablen durchsucht werden sollen, aber nur wenige das Suchmuster enthalten, bietet sich die Möglichkeit, die Variablennamen zu extrahieren und an eine Zielvariable zu übergeben. Die Zielvariable soll also nur die verketteten Namen der Variablen enthalten, die dem Suchmuster entsprechen.

1.1 Vereinfachtes Beispiel

Gegeben seien drei Text-Variablen a, b und c. Diese seien entweder mit dem Wert X belegt, oder leer. Die Zielvariable WhereX sollte analog zu Abbildung 1 aussehen:

Obs	a	b	C
1	X		
2			X
3		X	X
4	X		X
5	X	X	
6			
7		X	

Suchmuster → X

WhereX
a
c
b, c
a, c
a, b
b

Abbildung 1: Wertetabelle der Variablen a, b, c und der Zielvariablen WhereX

2 Lösungsweg

Der Lösungsweg ist als Makro skizziert. Die Identifikation der in Frage kommenden Variablen ist über die zugrunde liegende SAS-Tabelle realisiert. Diese könnte gegebenenfalls mit einer *where*-Bedingung auf Text-Variablen eingeschränkt werden.

Skizze des Lösungsweges:

1. Identifizierung der Variablennamen mittels der von PROC CONTENTS per ODS bereitgestellten Ausgabe-Tabelle *variables*.
2. Erzeugung einer Variablenliste als Makrovariable *VarList* mit allen in Frage kommenden Variablennamen.
3. Extraktion der *n* Einzelnamen aus der Makrovariablen *VarList* und Übergabe an *n* Makro-Variablen $DVar_i$, i=1, …, *n*, für alle Datensatz-Variablen Var_i, i=1, …, *n*, die das Suchmuster enthalten (=Namensidentifikation).
4. Iterativ: Übergabe der Variablennamen $DVar_i$ an Makrovariablen $NVar_i$, i=1, …, *n*.
5. Anwendung der Funktion CATX zur Zusammenfassung aller nicht leeren Einträge der Variablen $NVar_i$, i=1, …, *.n.*

```
%macro SucheMrX (DatIn, DatOut);

/* Schritt 1
   Identifizierung der Variablennamen mit PROC CONTENTS */

ods output variables= ContentsVars;
proc contents data=&DatIn. Position;
run;
```

Dem ODS ungeübten SAS-Benutzer sei eine kurze Einführung [1] empfohlen.

Die in Frage kommenden Variablen des Datensatzes werden an eine Makrovariable als Liste weitergegeben, um n einzelne Makro-Variablen zu erzeugen (Schritte 2 und 3).

```
/* Schritt 2  */
  proc sql noprint;
     select distinct variable into :VarList separated by " "
        from ContentsVars
        order by pos;
  quit;

/*   Ergebnis: VarList = "a b c"   */

/* Schritt 3 */
  %LET AnzVar=0;
    %DO %WHILE (%SCAN (&VarList, &AnzVar.+1, %STR( ))^=);
        %LET AnzVar = %EVAL (&AnzVar.+1);
        %LET DVar&AnzVar. = %SCAN(&VarList, &AnzVar., %STR( ));
    %END;

/*   Zerlegt VarList in n-Einzelvariablen (Beispiel aus SAS-Help)
     Ergebnis: DVar1 = a  /  DVar2 = b  /  DVar3 = c / AnzVar= 3
*/

/* Kombination von Datastep-Syntax und Macro-Syntax
   (Schritte 4 und 5)  */
  data &DatOut.;
     set &DatIn.;

  /* Schritt 4
     Übergabe relevanter Variablennamen (d.h. Variablen mit
     Suchmuster) als Wert einer neuen Makro-Variablen */

     %DO i=1 %TO &AnzVar.;
       IF %Upcase(&&&DVar&i.)="X" THEN NVar&i="&&&DVar&i.";
     %END;

  /* Ergebnis:   NDVar1 = "a"  /  NVar2 = "b"  / NVar3 = "c"      */

  /* Schritt 5
     Zusammenfassen der nicht-leeren Einzelwerte */

     WhereX=catx(", ", of NVar1 - NVar&AnzVar.);

run;

%mend SucheMrX;
```

3 Bemerkungen

Die Belegung mit dem Zeichen X, der Zeichenkette XXX oder der Zahl 1 lässt sich im Makro durch eine zusätzliche Makrovariable *value* realisieren. Der Aufruf wäre wie folgt:

```
%SucheMrX (DatIn, value, DatOut);
```

Die hier gelöste Datensituation ergab sich in einem realen Projekt mit afrikanischen Projektpartnern. Dabei sollte die wahrscheinlichste Todesursache aufgrund eines Interviews mit den nächsten Angehörigen der/des Verstorbenen bestimmt werden, da in vielen Entwicklungsländern kein Totenschein ausgestellt werden kann. Die Datentabelle enthielt nur Indikatoren, die den Wert y (für„yes“) enthielten oder leer waren.
Das Problem ergab sich in einer zweiten Datensituation, in der gescannte Fragebögen für die Dateneingabe nicht lesbar waren. Die Originalfragebogen lagen in Burkina Faso/Afrika und wurden ursprünglich eingescannt, um unnötige Transportkosten zu vermeiden. Anstelle nicht-lesbarer Daten wurde der Wert X eingegeben. Mit dem obigen Makro wurden die Variablennamen mit Suchmuster X extrahiert und dem Projektpartner vor Ort per E-Mail zur Klärung geschickt. [2]

Scottland Yard [3] hatte eine *ähnliche* Fragestellung: Mr. X taucht regelmäßig in London auf. Seine Standorte (=Straßennamen) sollen lokalisiert werden. Ähnlichkeiten mit lebenden Autoren sind rein zufällig (Abb. 2).

Alle 3 Problemsituationen, insbesondere die letzte, ergaben einen dringenden Handlungsbedarf. Alle konnten mit diesem Makro gelöst werden.

Abbildung 2:

Literatur

[1] Ramroth, H.: SAS/ODS (Output Delivery System) - eine Einführung. KSFE 2010 - Proceedings der 14. Konferenz der SAS-Anwender in Forschung und Entwicklung. Shaker Verlag; S. 227-235.

[2] Ramroth, H., Lorenz, E., Rankin, J.C., Fottrell, E., Yé, M., Neuhann, F., Ssennono, M., Sié, A., Byass, P., Becher, H.: Cause of death distribution with InterVA and physician coding in a rural area of Burkina Faso. Trop Med Int Health. 2012 Jul;17(7), S. 904-13.

[3] Scotland Yard. Auf der Jagd nach Mr. X. Ravensburger Spieleverlag.