Auffinden von gleichen oder ähnlichen Probandennamen – eine Anwendung der Abfragesprache SQL in SAS

Ralf Minkenberg

Institut für Biometrie, Universitätsklinikum der RWTH Aachen
Telefon: 0241 / 8089890
eMail: rminkenberg@mi.rwth-aachen.de

Abstract

Bei epidemiologischen Studien, bei denen eine größere Anzahl an Probanden über einen längeren Zeitraum beobachtet wird, tritt häufig das Problem auf, gleiche oder ähnliche Probanden, die von verschiedenen Stellen (Ärzten, Krankenhäusern, Ämtern etc.) erfasst worden sind, zu erkennen. Gerade wenn von unterschiedlichsten Stellen die Daten geliefert werden, kann dies nur über Merkmale wie Name, Vorname, Geburtsdatum o.ä. gelingen. Aufgrund unterschiedlicher Schreibweisen verschiedener Namen und Flüchtigkeitsfehlern bei der Eingabe vor Ort müssen jedoch nicht nur exakt gleiche, sondern auch ähnlich aussehende Namen o.ä. gefunden werden. Es wird ein leicht zu verwendendes SAS-Makro vorgestellt, dass sich dieser Aufgabe des Auffindens gleicher, aber auch zum Teil ähnlicher Probanden widmet. Bei Vergleichen wie hier ist es sinnvoll, auch in SAS auf die Datenbanksprache SQL zurückzugreifen, die bei relationalen Datenbanken standardmäßig verwendet wird. SAS stellt dieses Werkzeug mit der Prozedur proc sql in SAS/BASE zur Verfügung. Das vorgestellte Makro verwendet innerhalb SQL auch den LIKE-Befehl zum Vergleich ähnlicher Namen. Dieser Befehl steht so nicht im üblichen DATA-Step zur Verfügung, weshalb auch dies für die Verwendung von SQL spricht. Die Vorgehensweise des Makros in seiner einfachsten Form wird erläutert.

Einleitung

Gerade bei über einen längeren Zeitraum angelegten epidemiologischen Studien, z.B. zur Bewertung der Krebsmortalität, werden die interessierenden Daten häufig von verschiedenen Institutionen (Ärzte, Krankenhäuser, Ämter etc.) erfasst und dann zentral ausgewertet. Hierbei ergibt sich vor der eigentlichen statistischen Auswertung das Problem, Probanden, die in den verschiedenen Eingabelisten mehrfach auftauchen, herauszufiltern. Wenn zu Studienbeginn die in die Studie einzuschließenden Probanden noch nicht bekannt sind (z.B. wenn nur in einem bestimmten Zeitraum diagnostizierte Krebserkrankungen oder -todesfälle erfasst werden sollen), kann bei diesem Filtern nicht auf eine eindeutige Identifikationsnummer zurückgegriffen werden. Somit muß über möglichst eindeutige demographische Merkmale, wie z.B. Name, Vorname, Geburtsdatum, Geschlecht o.ä. versucht werden, gleiche Probanden zu erkennen. Aufgrund unterschiedlicher Schreibweisen von Namen oder Flüchtigkeitsfehlern bei der Datenerfassung vor Ort hat es sich als notwendig herausgestellt, auch ähnliche Einträge, die höchstwahrscheinlich den gleichen Probanden betreffen, nach Möglichkeit herauszufinden. Im folgenden wird ein SAS-Makro vorgestellt, daß sich dieser Aufgabe widmet und somit versucht, gleiche oder nach bestimmten Kriterien definierte ähnliche Einträge aufzulisten.

Ziel des Makros ist es, aus einer Vielzahl von Beobachtungen zunächst diejenigen auszugeben, die in vorher angegebenen Merkmalen exakt übereinstimmen. Als zweiter Schritt werden dann die Beobachtungen ausgegeben, die in wenigen (normalerweise höchstens ein oder zwei) Merkmalen nicht, in den anderen aber exakt übereinstimmen. In einem dritten Schritt schließlich werden noch Beobachtungen ausgegeben, die sich nach bestimmten vorher festgelegten Kriterien in allen (oder den meisten) Merkmalen “sehr ähnlich” sind, ohne daß einzelne Merkmale exakt übereinstimmen. Ob zwei Einträge von einem gleichen Probanden stammen oder nicht, kann endgültig jeweils nur von dem Anwender beurteilt werden.

Die Abfragesprache SQL innerhalb SAS

Im Bereich relationaler Datenbanken hat sich von Anfang an als Abfragesprache SQL (*Structured Query Language*) durchgesetzt. Mit Hilfe weniger relativ einfacher Befehle ist es mit Hilfe von SQL möglich, beliebig komplexe Abfragen zu erstellen, die sich auf Tabellen einer relationalen Datenbank beziehen.
In SAS ist zwar das Konzept einer relationalen Datenbank in den SAS-Datensätzen nicht verwirklicht worden, trotzdem stehen die standardisierten SQL-Befehle auch innerhalb SAS zur Verfügung. Somit können auch hier entsprechende Abfragen erstellt werden, die sich dann auf entsprechende SAS-Datensätze beziehen. Die Vorteile von SQL gegenüber einem üblichen DATA-Step liegen zunächst in der einfacheren Möglichkeit, gewünschte Teilmengen aus einem oder mehreren Datensätzen zu generieren. Weiterhin stehen in SQL einige Befehle und Verknüpfungen zur Verfügung, die so im DATA-Step keine Verwendung finden. Da SQL-Befehle innerhalb eines SAS-Programms an beliebiger Stelle mittels der hierfür vorgesehenen Prozedur PROC SQL eingefügt werden können, können alle weiteren SAS-Prozeduren o.ä. weiterhin verwendet werden. Die Prozedur PROC SQL ist im SAS/BASE-Modul enthalten. Die allgemeine Syntax dieser Prozedur läßt sich wie folgt zusammenfassen, wobei hier nur auf den SELECT-Befehl zur Erstellung von Abfragen eingegangen wird, die dann im Output-Fenster ausgegeben werden:

PROC SQL *Optionen*;
SELECT <DISTINCT> *object-items* <INTO *host-variable(s)*> FROM *from-list* <WHERE *where-expression*> <GROUP BY *group-by-item(s)*> <HAVING *having-expression*> <ORDER BY *order-by-item(s)*>;
QUIT;

Der SELECT-Befehl kann also aus folgenden Teilen bestehen:

SELECT-clause	legt fest, welche Merkmale oder Ausdrücke ausgegeben werden sollen.
INTO-clause	legt fest, in welche Makro-Variablen abgespeichert werden soll.
FROM-clause	legt fest, aus welchen Tabellen (Datensätzen) die Abfrage erstellt werden soll, evtl. auch aus welchen Kombinationen von Tabellen.
WHERE-clause	legt fest, welche einschränkenden Bedingungen für die Abfrage erfüllt sein sollen.
GROUP BY-clause	legt fest, in welche Untergruppen die Abfrage getrennt werden soll.
HAVING-clause	legt fest, welche mit Kenngrößen zusammenhängende Einschränkungen erfüllt sein sollen.
ORDER BY-clause	legt fest, wie die Abfrage sortiert werden soll.

Zur genauen Syntax der einzelnen Teile des SELECT-Befehls muß auf entsprechende Literatur verwiesen werden [1].

Auffinden gleicher und ähnlicher Beobachtungen

Das im folgende näher beschriebene SAS-Makro zum Auffinden gleicher und ähnlicher Beobachtungen geht davon aus, daß zwei Datensätze mit gleichen Merkmalsnamen vorliegen, die daraufhin untersucht werden sollen, ob gleiche oder ähnliche Beobachtungen in diesen vorhanden sind. Es werden nun aus diesen beiden Dateien (“Datei1, Datei2”) zunächst diejenigen Beobachtungen ausgegeben, die in einer vorgegebenen Anzahl (im Beispiel drei – “Var1, Var2, Var3”) von Merkmalen exakt übereinstimmen. Dies geschieht innerhalb der Prozedur SQL folgendermaßen:

```
PROC SQL;
      SELECT * FROM Datei1, Datei2 WHERE Datei1.Var1=Datei2.Var1 AND
      Datei1.Var2=Datei2.Var2 AND Datei1.Var3=Datei2.Var3;
QUIT;
```

Es werden so alle exakt gleichen Beobachtungen komplett (hier kann natürlich auch eine Einschränkung auf wenige Variablen geschehen) ausgegeben.
In einem nächsten Schritt sollen nun die Beobachtungen ausgegeben werden, die nur in einem Teil der vorgegebenen Variablen exakt übereinstimmen. Auch dies läßt sich ähnlich wie obige Ausgabe komplett übereinstimmender Beobachtungen in SQL lösen. Sollen aus obigem Beispiel die Beobachtungen angegeben werden, die in zwei der drei Variablen "Var1, Var2, Var3" übereinstimmen, so lautet der entsprechende Befehl z.B.:

```
PROC SQL;
      SELECT * FROM Datei1, Datei2 WHERE
         (CASE WHEN Datei1.Var1=Datei2.Var1 THEN 1 ELSE 0 END +
          CASE WHEN Datei1.Var2=Datei2.Var2 THEN 1 ELSE 0 END +
          CASE WHEN Datei1.Var3=Datei2.Var3 THEN 1 ELSE 0 END) = 2;
QUIT;
```

Die obigen beiden Abfragen werden schließlich nur noch so in ein Makro eingebettet, daß die Anzahl der übereinstimmenden Variablen und deren Namen vom Benutzer frei gewählt werden können. Somit erhält man mit relativ einfachen Mitteln eine Möglichkeit, die Beobachtungen herauszusuchen, die zumindest in einigen Merkmalen übereinstimmen. Dieser Fall tritt besonders in Zusammenhang mit unterschiedlichen Schreibweisen von Eigennamen oder bei Flüchtigkeitsfehlern während der Eingabe auf. Hier ist es häufig so, daß nur einzelne die Beobachtung kennzeichnende Merkmale sich unterscheiden, während andere gleich sind. Hier ist es für den Benutzer bei Ausgabe einer Liste solcher Beinahe-Übereinstimmungen meistens sehr leicht zu sehen, ob tatsächlich gleiche Beobachtungen vorliegen oder nicht.

Weitere Möglichkeiten mit der Funktion LIKE

Bisher können einander ähnliche Beobachtungen nur dann aufgefunden werden, wenn sie zumindest in einem Merkmal, das zur Identifizierung geeignet ist, übereinstimmen. Gerade wenn wenige solche Merkmale vorliegen ist eine Erweiterung notwendig, bei der auch nur ähnliche Ausprägungen einer Variablen unter definierten Vorgaben als potentiell den gleichen Probanden charakterisierend ausgegeben werden. Hierbei leistet die Funktion LIKE wertvolle Dienste, die nur innerhalb PROC SQL zur Verfügung steht. Eine vergleichbare Funktion für den DATA-Step in SAS gibt es wohl nicht. Mit Hilfe der LIKE-Funktion können zwei Zeichenketten miteinander verglichen werden, wobei drei Klassen von Zeichen unterschieden werden:

_	(Underscore)	steht für irgendein einzelnes Zeichen.
%	(Prozentzeichen)	steht für keines oder beliebig viele Zeichen.
alle anderen Zeichen		stehen für genau dieses Zeichen.

Es ist somit u.a. möglich, Zeichenketten herauszusuchen, die sich nur am Anfang oder Ende unterscheiden, was sich z.B. bei Doppelnamen als sinnvoll erweist. Mit der folgenden Abfrage werden sich nur am Anfang oder Ende unterscheidende Merkmale ausgegeben. (Die COMPRESS-Funktion entfernt alle Leerzeichen aus dem Funktionsargument.)

```
PROC SQL;
        SELECT * FROM Datei1, Datei2 WHERE
            Datei1.Var1 LIKE COMPRESS(Datei2.Var1 || '%') OR
            Datei2.Var1 LIKE COMPRESS(Datei1.Var1 || '%');
QUIT;
```

Je nach Fragestellung lassen sich durch die LIKE-Funktion viele andere Abfragen erstellen, die unter gewissen, interessierenden Aspekten ähnliche Merkmalsausprägungen herausfinden können. Es sollte jedoch hier genau überlegt werden, ob durch solche Abfragen wirklich noch zusätzliche, tatsächlich gleiche Beobachtungen gefunden werden können und ob der Aufwand zum Erstellen gerade komplexer Abfragen noch im Verhältnis zum erzielten Ergebnis steht.

Aufbau des SAS-Makros

Das zum Auffinden gleicher oder ähnlicher Beobachtungen verwendete SAS-Makro benötigt in seiner einfachsten Form als Übergabeparameter lediglich die Namen der beiden zu vergleichenden SAS-Dateien und die Liste der für die Überprüfung zu verwendenden Variablennamen, die in beiden Dateien vorhanden sein müssen. Um Fehlermeldungen zu vermeiden, stellt das Makro nun zunächst fest, ob die beiden Dateien existieren; weiterhin wird die Anzahl der zu benutzenden Variablen bestimmt (siehe auch [2]).

```
%macro find_dop(datei1= , datei2= , Vars= );

/* Teste die Existenz von &datei1 */
   %if %sysfunc(exist(&datei1)) = 0 %then %do;
      %put Datei &datei1 existiert nicht ;
      %put Makro beendet ;
      %goto ende;
   %end;

/* Teste die Existenz von &datei2 */
( . . . )

/* Bestimme Anzahl an Variablen und speichere diese in Makro-Variablen var1-varN */
   %let z = 1;
   %do %until(%scan(&vars,&z,%str( ))=%str());
      %let var&z = %scan(&vars,&z,%str( ));
      %let z = %eval(&z + 1);
   %end;
   %let zahl = %eval(&z – 1);
```

Als nächstes werden zunächst die Beobachtungen ausgegeben, die sich bezüglich aller angegebener Merkmale nicht unterscheiden. Hierbei muß nur die im vorigen Abschnitt angegebene Abfrage für das Makro verallgemeinert werden:

```
   proc sql;
      select * from &datei1, &datei2 where
      %do i=1 %to %eval(&zahl–1);    &datei1..&&var&i = &datei2..&&var&i and    %end;
      &datei1..&&var&zahl = &datei2..&&var&zahl;
```

In der hier vorgestellten einfachsten Form überprüft das Makro dann, welche Beobachtungen (ohne die exakt gleichen – diese werden durch den EXCEPT-Befehl ausgeschlossen) in allen bis auf höchstens ein Merkmal übereinstimmen:

```
    select * from &datei1, &datei2 where
    (%do i=1 %to &zahl; case when &datei1..&&var&i = &datei2..&&var&i then 1 else 0
    end +  %end;  0) >= %eval(&zahl–1) except (select * from &datei1, &datei2 where
    %do i=1 %to %eval(&zahl–1);  &datei1..&&var&i = &datei2..&&var&i and  %end;
    &datei1..&&var&zahl = &datei2..&&var&zahl);
  quit;

  %ende:

%mend;
```

Das Makro existiert in einer noch allgemeineren Form, in der einerseits festgelegt werden kann, wieviele der angegebenen Merkmale mindestens übereinstimmen müssen, damit die Beobachtungen ausgewählt werden. Andererseits ist es dort auch möglich, innerhalb eines oder mehrerer Merkmale Zeichenketten zu finden, die sich nur am Anfang bzw. Ende unterscheiden, indem die LIKE-Funktion wie zuvor beschrieben verwendet wird. Das Makro kann jederzeit angefordert werden: per eMail rminkenberg@mi.rwth-aachen.de oder unter http://www.klinikum.rwth-aachen.de/webpages/mbio/user/rminkenberg/home.html .

Literatur

[1] SAS Institute Inc.; SAS Guide to the SQL Procedure: Usage and Reference, Version 6, First Edition; Cary, NC, 1989

[2] Art Carpenter; Carpenter's Complete Guide to the SAS Macro Language; Cary, NC: SAS Institute Inc., 1998