Übereinstimmungsmaße in der PROC FREQ – Option AGREE bei nichtsymmetrischen Tabellen – eine Macro-Lösung

Bettina Danner, Rainer Muche

Abteilung Biometrie und Medizinische Dokumentation, Universität Ulm
Telefon: 0731 / 50-26891
eMail: bettina.danner@medizin.uni-ulm.de

Abstract

In dem Beitrag werden Probleme bei der Berechnung von Übereinstimmungsmaßen (Cohens Kappa, McNemar-Test) beschrieben, die bei der Nutzung der Option AGREE der PROC FREQ (seit Version 6.11) auftreten. Angegeben werden die Lösungsmöglichkeiten und das zugehörige SAS-Macro, welches eine Berechnung auch in den beschriebenen Problemsituationen erlaubt.

Einleitung

Sowohl in Studien zum Vergleich diagnostischer Instrumente als auch in gematchten Fall-Kontroll-Studien werden Übereinstimmungsmaße und davon abgeleitete statistische Verfahren sehr häufig benutzt. Zu diesen statistischen Verfahren gehören Cohens Kappa und der McNemar-Test [1].

Seit der Version 6.11 können die beiden Verfahren mit der Option AGREE in der Prozedur PROC FREQ mit SAS berechnet werden [2]. Dabei müssen die abhängigen Beobachtungen (z.B. in der Diagnostik zwei Messungen an ein und demselben Patienten) in zwei Variablen gespeichert sein.

1. Problem: nicht-quadratische Häufigkeitstabelle

PROC FREQ unterdrückt nichtbesetzte Zeilen und Spalten in der Häufigkeitstabelle. Sie lassen sich nicht mit irgendeiner Option als Null-besetzt darstellen. In dieser dann nicht-symmetrischen Tabelle werden keine Übereinstimmungsmaße mehr berechnet.

2. Problem: inhaltlich nicht-quadratische Häufigkeitstabelle

Sollte die Tabelle zwar von der Form, aber durch nichtbesetzte Merkmalsausprägungen nicht inhaltlich symmetrisch sein, so berechnet die Prozedur die Kenngrößen, ohne zu prüfen, ob die Ausprägungen übereinstimmen. Man bekommt falsche, inhaltlich nicht interpretierbare Ergebnisse ohne irgendeine Warnung.

3. Problem: BY-Variablen

Bei der Verwendung einer BY-Variablen werden nur die quadratischen Tafeln berechnet, unabhängig davon, ob diese inhaltlich gleich sind oder nicht. Damit stellt dieses Problem eine Kombination der unter 1. und 2. genannten Situationen dar.

Lösungsstrategie

Nachfolgend wird die allgemeine Lösungsstrategie aufgezeigt und an einem Beispiel dargestellt. Da diese Strategie sehr zeitaufwendig sein kann, wurde die Macro-Lösung programmiert.

Folgende Schritte sind durchzuführen bzw. laufen im Macro ab:

1. Auslesen der Kreuztabelle in eine Out-Datei, falls in der Ausgangsdatei Einzelbeobachtungen und keine Häufigkeiten enthalten sind
2. Einlesen mindestens einer Zelle für die fehlende Zeile oder Spalte (für jede Ausprägung der BY-Variablen) mittels CARDS, wobei ein sehr kleines Gewicht, z. B $1/10^{20}$ (1E-20), für die Zellhäufigkeit vergeben wird
3. Zusammensetzen der Ausgangsdatei / der Out-Datei und der unter 2. erstellten Datei mit dem SET-Befehl
4. erneuter Aufruf des Tests für die neue Datei unter Verwendung des WEIGHT-Statements für die Zellhäufigkeiten. Die Tabelle ist nunmehr quadratisch und der Test wird gerechnet. Die kleinen Zellhäufigkeiten haben keinen Einfluß auf das Ergebnis des Tests, allerdings sind die Spalten- und Zeilenprozente unter Umständen nicht mehr korrekt.

Beispiel:

Quadratische, aber inhaltlich nicht symmetrische Tafel

TABLE OF GRUPPE_A BY GRUPPE_B

GRUPPE_A / Frequency / Percent / Row Pct / Col Pct	GRUPPE_B 2	GRUPPE_B 3	Total
1	21	14	35
	31.82	21.21	53.03
	60.00	40.00	
	52.50	53.85	
2	19	12	31
	28.79	18.18	46.97
	61.29	38.71	
	47.50	46.15	
Total	40	26	66
	60.61	39.39	100.00

Bei Gruppe_A fehlt Merkmalsausprägung 3
Bei Gruppe_B fehlt Merkmalsausprägung 1
Die Tabelle ist **quadratisch**, aber inhaltlich **nicht symmetrisch**, dennoch wird die Statistik erstellt:

STATISTICS FOR TABLE OF GRUPPE_A BY GRUPPE_B
McNemar's Test

Statistic = 0.758 DF = 1 Prob = 0.384

Simple Kappa Coefficient

Kappa = -0.013 ASE = 0.122
95% Confidence Bounds -0.251 0.225

Sample Size = 66

Quadratische und inhaltlich symmetrische Tafel (nach Macro)

TABLE OF GRUPPE_A BY GRUPPE_B

GRUPPE_A / Frequency / Percent / Row Pct / Col Pct	GRUPPE_B 1	GRUPPE_B 2	GRUPPE_B 3	Total
1	**1E-20**	21	14	35
	0.00	31.82	21.21	53.03
	0.00	60.00	40.00	
	100.00	52.50	53.85	
2	0	19	12	31
	0.00	28.79	18.18	46.97
	0.00	61.29	38.71	
	0.00	47.50	46.15	
3	0	0	**1E-20**	**1E-20**
	0.00	0.00	0.00	0.00
	0.00	0.00	**100.00**	
	0.00	0.00	0.00	
Total	**1E-20**	40	26	66
	0.00	60.61	39.39	100.00

STATISTICS FOR TABLE OF GRUPPE_A BY GRUPPE_B
Test of Symmetry

Statistic = 47.000 DF = 3 Prob = **0.001**

Kappa Coefficients

Statistic	Value	ASE	95% Confidence Bounds	
Simple Kappa	0.004	0.042	-0.078	0.087
Weighted Kappa	-0.000	0.000	-0.000	-0.000

Sample Size = 66

Das SAS-Macro: AGREE_NS.MAC

%MACRO agree_ns (datname,varlist1,varlist2,n,wt,wherevar,m,byvar,j);	**PROC FREQ/AGREE (McNemar Test/Bowker Test auf Symmetrie, Kappa-Koeffizient) - wenn Tafel nicht quadratisch und symmetrisch ist**
OPTIONS NONUMBER NOCENTER NODATE PS=63 LS=88; %Macro agree_ns (datname,varlist1,varlist2,n,wt,wherevar,m,byvar,j);	**datname** Dateiname **varlist1** Variablenliste 1 (z.B. Fälle) - gleiche Länge wie varlist2 **varlist2** ariablenliste 2 (z.B. Kontrollen) - gleiche Länge wie varlist1 **n** Anzahl der zu testenden Variablen von varlist1 bzw varlist2 **wt** WEIGHT-Variable - kann auf missing gesetzt werden **wherevar** WHERE-Variable - kann auf missing gesetzt werden **m** enthält Operationszeichen und Wert für WHERE-Variable, (z.B. m= ^=3) **byvar** BY-Variable - enthält Anzahl der Ausprägungen der BY-Variablen, bzw. der ersten **j** Ausprägungen der aufsteigend sortierten BY-Variablen, für die Berechnung durch geführt werden soll
TITLE2; TITLE3;	TITLE1 - kommt aus ausrufendem SAS-Programm

DATA zzzz;
%DO k=1 %TO &n %BY 1;
call symput ("a&k",scan(&varlist1,&k," "));
call symput ("b&k",scan(&varlist2,&k," "));
%END;
RUN;
	Liest alle Variablen aus varlist1 und varlist2 ein
	gezielter Zugriff auf alle Variablen aus varlist1 über &&a&i
	gezielter Zugriff auf alle Variablen aus varlist2 über &&b&i

%IF &byvar^=. %THEN %DO; PROC SORT DATA=&datname OUT=file; BY &byvar; RUN;	Sortiert Datei &datname nach &byvar, wenn diese Variable gesetzt wurde
DATA file; SET file; BY &byvar; IF FIRST.&byvar THEN FIRST=1;	Markiert jede neue Ausprägung von &byvar mit 1 in SAS-interner Variable First
DATA file; SET file; IF FIRST=1 THEN x +1; RUN;	Zählt die Anzahl der Ausprägungen von &byvar - ggf. Umwandlung von char in num
PROC FREQ DATA=file; %IF &wt^=. %THEN %DO; TABLE &byvar*x/MISSING; WEIGHT &wt; %END; %ELSE %DO; TABLE &byvar*x/MISSING; %END; TITLE2 "Tabelle mit ursprünglicher (in &byvar) und neuer"; TITLE3 "(in x) Codierung der BY-Variablen für &datname"; RUN;	Gegenüberstellung der ursprünglichen (in byvar gespeicherten) und neuen Klassierung (in x gespeicherten) Ausprägung der BY-Variablen
%DO i=1 %TO &j %BY 1; DATA d_&i; SET file; IF x=&i THEN OUTPUT d_&i; RUN; %END;	Teilt ursprüngliche Datei &datname in &j Einzeldateien, eine für jede Ausprägung von &byvar - es entstehen die Dateien d_1 bis d_&j
PROC DATASETS; DELETE file; RUN; QUIT; %END;	Löschen der Hilfsdatei file aus PROC SORT

```
%ELSE %DO;
 %LET j=%EVAL (1);
 DATA d_1;
     SET &datname;
 RUN;
%END;

%DO  k=1 %TO &n %BY 1;
    %DO i=1 %TO &j %BY 1;

    PROC FREQ DATA=d_&i;
        %IF &wherevar^=. %THEN %DO;
           WHERE &wherevar&m;
           %IF &wt^=. %THEN %DO;
              TABLE &&a&k*&&b&k/ OUT=xxx&k&i NOROW
                                    NOCOL;
              WEIGHT &wt;
           %END;
           %ELSE %DO;
             TABLE &&a&k*&&b&k/ OUT=xxx&k&i NOROW
                                    NOCOL;
           %END;
           TITLE2 "PROC FREQ (AGREE) für Datei &datname,
                    mit WHERE = &wherevar mit &m";
           TITLE3 "mit Gruppierung nach BY = &byvar mit neu
                    codierter Ausprägung &i";
        %END;
        %ELSE %DO;
           %IF &wt^=. %THEN %DO;
              TABLE &&a&k*&&b&k/ OUT=xxx&k&i NOROW
                                    NOCOL;
              WEIGHT &wt;
           %END;
           %ELSE %DO;
             TABLE &&a&k*&&b&k/ OUT=xxx&k&i NOROW
                                    NOCOL;
           %END;
           TITLE2 "PROC FREQ (AGREE) für Datei &datname,
                    ohne WHERE";
           TITLE3 "mit Gruppierung nach BY = &byvar mit neu
                    codierter Ausprägung &i";
        %END;
    RUN;

    DATA yyy&k&i;
        SET xxx&k&i;
        IF &&a&k^=&&b&k THEN count_n=1e-20;
        IF &&a&k>&&b&k  THEN f_=&&b&k;
        IF &&a&k<&&b&k  THEN f_=&&b&k;
        IF &&a&k>&&b&k  THEN k_=&&a&k;
        IF &&a&k<&&b&k  THEN k_=&&a&k;
    RUN;

    DATA yy1&k&i yy2&k&i;
        SET yyy&k&i;
        IF &&a&k^=f_ THEN OUTPUT yy1&k&i;
        IF &&b&k^=k_ THEN OUTPUT yy2&k&i;
    RUN;
    DATA yy1&k&i;
        SET yy1&k&i;
        KEEP  f_  &&b&k  count_n;
    DATA yy1&k&i;
        SET yy1&k&i;
        RENAME count_n=count  f_=&&a&k;
    DATA yy2&k&i;
        SET yy2&k&i;
        KEEP  k_  &&a&k  count_n;
```

Wenn keine &byvar gesetzt wurde, wird ursprüngliche Datei &datname nach d_1 geschrieben

Schleife über die Anzahl der Variablen in den Variablenlisten (&n)
Schleife über die Ausprägungen von &byvar

Die Kreuztabellen werden am Bildschirm angezeigt, die Teststatistik aber unterdrückt, da an dieser Stelle quadratische aber nicht-symmetrische Tafeln vorkommen können, für die eine falsche Berechnung erscheinen würde

Neue Variable count_n wird mit sehr kleinem Wert gesetzt, wenn Inhalt von Variable a und b ungleich ist
Neue Variable f_, wenn Inhalt von a und b nicht gleich ist; f_ bekommt Inhalt von b
Neue Variable k_, wenn Inhalt von a und b nicht gleich ist; k_ bekommt Inhalt von a

Trennen der Datei in zwei Teildateien, wenn

Inhalt von a ungleich dem Inhalt von f_
Inhalt von b ungleich dem Inhalt von k_

In der Datei yy1&k&i werden drei Variablen behalten

Umbenennung von zwei Variablen

In Datei yy2&k&i werden drei Vairablen behalten

DATA yy2&k&i; SET yy2&k&i; RENAME count_n=count k_=&&b&k; RUN;	Umbenennung von zwei Variablen
DATA yyy&k&i; SET yy1&k&i yy2&k&i; RUN;	Zusammenbinden der beiden Dateien yy1&k&i und yy2&k&i zu yyy&k&i
DATA test&k&i; SET xxx&k&i yyy&k&i; WHERE count^=.; RUN;	Zusammenbinden der Datei yyy&k&i mit Out-Datei xxx&k&i zu test&k&i, bei denen count-Variable nicht missing
PROC SORT DATA=test&k&i OUT=test&k&i NODUP; BY &&a&k &&b&k;	Sortieren nach Kombination von &&a&k und &&b&k
DATA test&k&i; SET test&k&i; BY &&a&k &&b&k; IF FIRST.&&b&k THEN first=1; RUN;	Markieren jeder neuen Ausprägung von &&b&k in SAS-interner Variablen First mit 1
DATA test&k&i; SET test&k&i; IF first=1 THEN OUTPUT test&k&i; RUN;	Schreiben aller Datensätze mit first=1 nach test&k&i
PROC FREQ DATA=test&k&i; TABLE &&a&k*&&b&k/AGREE NOPRINT NOCOL NOROW; WEIGHT count; TITLE2; TITLE3; RUN;	Anzeige der Teststatistik für die nunmehr symmetrische Kreuztabelle, die Darstellung der ergänzten Tabelle am Bildschirm wird unterdrückt, falls man diese sehen möchte NOPRINT-Anweisung löschen
PROC DATASETS; DELETE xxx&k&i yyy&k&i yy1&k&i yy2&k&i test&k&i aaa&k&i; RUN; QUIT;	
%END; %END;	
PROC DATASETS; %DO i=1 %TO &j %BY 1; DELETE d_&i; %END; PROC DATASETS; DELETE zzzz; RUN; QUIT; %MEND;	

Literatur

[1] I. Guggenmoos-Holzmann, K.-D. Wernecke: Medizinische Statistik. Blackwell Wissenschafts-Verlag, Berlin 1996.

[2] SAS / STAT Software. Changes and Enhancements through Release 6.11 (55356). SAS Inst. Inc, Cary, NC 27513.