Ein SAS-Macro zur Berechnung verallgemeinerter Kendallscher Übereinstimmungsmaße

Matthias Frisch und Erich Schumacher,
Institut für Angewandte Mathematik und Statistik,
Universität Hohenheim, 70593 Stuttgart.
E-mail: frisch@uni-hohenheim.de, schumach@uni-hohenheim.de

Zusammenfassung

In der tierischen Verhaltensforschung und in anderen anwendungsbezogenen Forschungsgebieten können Sachverhalte durch Permutationen mit unbelegten Plätzen beschrieben werden. Es wird ein SAS-Macro vorgestellt, das statistische Tests für folgende Fragestellungen zur Verfügung stellt: Vergleich einer beliebigen Permutation mit unbelegten Plätzen mit einer festen Ausgangspermutation, Vergleich mehrerer beliebiger Permutationen mit unbelegten Plätzen mit einer festen Kontrollpermutation, Vergleich zweier beliebiger Permutationen mit unbelegten Plätzen.

Einleitung

In der tierischen Verhaltensforschung treten Fragestellungen auf wie z.B.: „Hängt die Reihenfolge, in der junge Ferkel an den Zitzen der Muttersau saugen von der Geburtsreihenfolge ab“? Wenn man die Geburtsreihenfolge und die Saugreihenfolge durch Permutationen beschreibt ergibt sich folgendes Bild:

Geburtsreihenfolge P_1:{1,2,3,4,5,6,7,8,9,0,0,0,0,0,0,0,0}

Saugreihenfolge P_2:{1,0,0,5,3,2,0,0,9,0,8,0,4,0,7,0,6}

Charakteristisch sind die unbelegten Stellen in beiden Permutationen. Um die oben gestellte Frage zu beantworten, ist ein Verfahren nötig, das eine beliebig zustandegekommene Permutation mit einer festen Ausgangspermutation vergleicht. Die zugehörige Nullhypothese kann wie folgt formuliert werden H_0: „Die Permutation P_2 kommt unabhängig von der Permutation P_1 zustande“. Erweitert man die Fragestellung derart, daß man die Abhängigkeit mehrerer Saugvorgänge mit der Geburtsreihenfolge untersucht, so stellt sich die Frage nach dem Vergleich von einer festen Ausgangspermutaion mit unbelegten Plätzen mit *k* beliebig

zustande gekommenen Permutationen. Grafisch läßt sich das folgendermaßen veranschaulichen:

Geburt	P_0:	{1,2,3,4,5,6,7,8,9,0,0,0,0,0,0,0,0}
Saugreihenfolge 1	P_1:	{1,0,0,5,3,2,0,0,9,0,8,0,4,0,7,0,6}
Saugreihenfolge 2	P_2:	{1,0,5,0,0,2,3,0,9,0,8,4,7,0,0,0,6}
	:	
	:	
Saugreihenfolge k	P_k:	{1,0,5,2,3,0,0,9,0,0,8,0,4,7,6,0,0}

Die zugehörige Nullhypothese lautet dann H_0: „Die Permutationen P_1 ... P_k kommen unabhängig von der Permutation P_0 zustande“.

Theoretische Grundlagen

Um einen Test auf Übereinstimmung der Permutationen zu entwickeln, wird ein Maß für die Unterschiedlichkeit von Permutationen benötigt. Aus der Anzahl der Inversionen *l*, die benötigt wird um zwei Permutationen ineinander überzuführen, kann ein solches Maß berechnet werden. So werden zum Beispiel $l = 3$ Inversionen benötigt um die Permutation {1,2,3,0,0} in die Permutation {1,0,2,0,3} zu überführen. Die maximale Distanz die eine Permutation gleichen Typs von der Permutation {1,2,3,0,0} haben kann, ist $l_{\max} = 9$. Dies ist die Permutation {0,0,3,2,1}. Durch eine lineare Transformation von *l* kann eine Normierung erreicht werden:

$$\tau = 1 - \frac{2l}{l_{\max}}.$$

Tau ist hier eine Verallgemeinerung von Kendalls Tau [2], die auf Permutationen mit unbelegten Plätzen angewandt werden kann, es liegt im Bereich $\tau \in [-1,1]$. Für den Vergleich mehrerer Permutationen mit einer Kontrollpermutation kann eine Verallgemeinerung des Kendall-Haysschen $\bar{\tau}$ [2] verwendet werden:

$$\bar{\tau} = 1 - \frac{l}{\mathrm{E}(L)} \text{ mit } l = \sum_k l_{0k}$$

Die Distanz *l* ist die Summe der Distanzen l_{0k} zwischen der Kontrollpermutation P_0 und den zu vergleichenden Permutationen P_k . Der Erwartungswert von *L* unter der Nullhypothese kann mit Hilfe der Verteilung von *L* unter H_0 berechnet werden.

Zur Berechnung der Verteilung der Distanzen L unter H_0 wird folgende Schreibweise eingeführt: Eine Permutation mit fehlenden Werten läßt sich durch den Ausdruck

$[\, n\, ,\, r\, ,\, \underline{d}\,]$ mit

n: Anzahl der Elemente der Permutation,

r: Anzahl der nicht belegten Plätze,

$\underline{d}$: Vektor der Anzahl der Elemente zwischen den unbelegten Stellen,

beschreiben. So wird z. B. die Permutation {0,1,0,2,3,0,4} durch [7,3,(1,2)] charakterisiert. Die Anzahl der Permutationen, die von einer festen Ausgangspermutation die Distanz l besitzen, wird beschrieben durch $[\, n\, ,\, r\, ,\, \underline{d}\, ,\, l\,]$. So ist [7,3,(1,2),0] = 1 oder [7,3,(1,2),6] = 15, was sich durch kombinatorische Überlegungen leicht zeigen läßt. In dieser Notation ist die Verteilung von L unter H_0 :

$$(\, l\, ,\, [\, n\, ,\, r\, ,\, \underline{d}\, ,\, l\,]\, /\, \Sigma_k\, [\, n\, ,\, r\, ,\, \underline{d}\, ,\, k\,]\,)$$

Die Berechnung der Verteilung der Distanzmaße unter H_0 erfolgt über Rekursionsformeln. Definiert man $s = \sum_{\nu} d_{\nu}$, dann lassen sich zwei Typen von Distanzen unterscheiden:

Typ 1: $s<n\text{-}r$

Hierunter fallen z.B. Distanzen zu Permutationen wie {0,1,0,2,3,0,4}, die durch [7,3,(1,2)] beschrieben wird. Durch vollständige Induktion läßt sich zeigen, daß

$$[\, n+1,\, r,\, \underline{d},\, l\,] = \sum_{j=0}^{n} [\, n,\, r,\, \underline{d},\, l-j\,].$$

Typ 2: $s=n\text{-}r$

Hierunter fallen z.B. Distanzen zu Permutationen wie {0,0,1,0,2,3,0}, die durch [7,4,(0,1,2)] beschrieben wird. Hier läßt sich zeigen, daß

$$[\, n+1,\, r,\, \underline{d},\, l\,] = [\, n,\, r-1,\, (d_2,\, d_3,\, ...,\, d_k),\, l\,] + \sum_{i=1}^{k} \sum_{j=s_{i-1}}^{s_i+i-1} [n,\, \mathrm{r},\, (d_1,\, ...,\, d_{i-1},\, d_i-1,\, d_{i+1},\, ...,\, d_k)\,]$$

mit $s_i = \sum_{j=1}^{i} d_j$ und $s_0 = 0$.

Die Anwendung der Rekursionsformeln veranschaulicht folgendes Rechenbeispiel:

[7,3,(1,2),2] = [6,3,(1,2),0] + [6,3,(1,2),1] + [6,3,(1,2),2] = ... = [6,3,(1,2),0] + [5,3,(0,2),0] + [4,2,(2),0] + [4,2,(2),0] + [3,2,(1),0] + [3,2,(1),0] + [3,2,(1),0] + [1,2,(1),0] + [2,1,(),0] +

[1,2,(1),0] + [2,2,(0),0] + [2,1,(),1] + [2,1,(),0] + [2,1,(),1] + [2,1,(),0] + [2,1,(),1] + [2,1,(),0] = 18.

Durch die Faltung von Verteilungen von Distanzen können Tests zum Vergleich einer Kontrollpermutation mit mehreren Permutationen entwickelt werden. Berechnet man l als Summe der Distanzen l_{0j} einer festen Kontrollpermutation P_0 zu j beliebigen Permutationen P_j, so ist L unter H_0 eine Faltung der Verteilungen L_{0j}.

Ein Test zum Vergleich von zwei beliebig zustande gekommenen Permutationen mit unbelegten Plätzen, kann mit Hilfe der Mischung der Verteilung der Distanzen aller möglichen Ausgangspermutationen P_1 zur Permutation P_2 konstruiert werden.

SAS-Macro

Im Macro `khtau.mak` sind drei Tests mit unterschiedlichen Optionen implementiert.

Tests:

- **`test1`** Vergleich einer festen Ausgangspermutation mit unbelegten Stellen mit einer beliebigen Permutation mit unbelegten Stellen
- **`test2`** Vergleich einer festen Ausgangspermutation mit unbelegten Stellen mit mehreren beliebigen Permutationen mit unbelegten Stellen
- **`test3`** Vergleich von zwei beliebigen Permutationen mit unbelegten Stellen

Optionen:

- **`tau`** Verallgemeinertes Kendallsches Tau / Kendall-Hayssches Tau
- **`l`** Distanz / Multiple Distanz der zu vergleichenden Permutationen
- **`e`** Erwartungswert der Distanz unter der Nullhypothese
- **`v`** Varianz der Distanz unter der Nullhypothese
- **`d`** Verteilung der Distanz unter der Nullhypothese
- **`p`** P-values

Aufruf des Macros

Im folgenden Beispiel wird der Aufruf des Macros demonstriert. In einem Data-Step werden die Daten eingegeben. Die Permutationen stehen in Spalten nebeneinander. Die feste Aus-

gangspermutation wird mit `p0` bezeichnet, die zu Vergleichenden Permutationen mit `p1-pk`. Beim Vergleich zweier beliebiger Permutationen werden diese mit `p1` und `p2` bezeichnet. Daran anschließend wird der Quelltext des Macros eingefügt, eventuell mit Angabe des Verzeichnisses, in dem er abgespeichert ist. Beim Aufruf des Macros wird an erster Stelle der Parameterliste die Bezeichnung des Tests erwartet, dann die zu berechnenden Optionen. Der Name des Eingabe-Datasets ist notwendig, die Angaben zu `symsize` und `wrksize` sind optional.

```
data data;
 input p0 p1-p2;
 cards;
 1  2 1
 2  0 0
 3  3 0
 4  0 0
 5  5 4
;

%inc 'khtau.mak';

%khtau ('test2','tau','l',
        'e','v','p',
        daten   = data,
        symsize = 2048
        wrksize = 2048);
```

In der Ausgabe erscheint nach der Bezeichnung des durchgeführten Tests zur Kontrolle nochmals die Datenmatrix (die Permutationen sind in Zeilen dargestellt), dann die berechneten Ergebnisse:

```
test2

datenmatrix
1  2  3  4  5
2  0  3  0  5
1  0  0  0  4

tau
0.6842105

l
2

e          v
6.3333333  4.7583333

P_VALUE                    P
p_u=P(X<=x)           0.0275
```

```
p_o=P(X>=x)          0.9933333
p_z=2*min(p_u,p_o)       0.055
```

Der mutiple Korrelationskoeffizient $\bar{\tau}$ beträgt in diesem Beispiel 0,68. Die multiple Distanz *l*, die sich aus der Summe der Distanzen der Kontrollpermutation mit den zu vergleichenden Permutationen (l_{01}=1 und l_{02}=1) zusammensetzt, beträgt 2. Erwartungswert und Varianz der Verteilung der multiplen Distanz unter der Nullhypothese betragen 6,33 und 4,76.

Der in der Praxis sicher am häufigsten auftretende Fall ist der Test

> H_0: "Die Vergleichspermutationen kommen unabhängig von der Kontrollpermutation zustande."
>
> H_A: "Beim Zustandekommen der Vergleichspermutationen liegt eine positive Abhängigkeit zur Kontrollpermutation vor."

zum Niveau α. Die Entscheidungsvorschrift für diesen Test lautet: Ist **p_u** < α, dann liegt eine positive Abhängigkeit vor.

In unserem Beispiel erhält man

p_u = 0.0275 < 0,05

und somit eine signifikante positive Abhängigkeit. Die beiden anderen möglichen Testprobleme lassen sich durch Vergleich von **p_o** und **p_z** mit dem nominellen α durchführen.

Literatur

[1] Bosch, K. (1993):
Statistik Taschenbuch.
R. Oldenbourg Verlag, München Wien.

[2] Kendall, M. G. (1970):
Rank Correlation Methods.
Griffin, London.

[3] SAS-Institute (1990):
SAS-Guide to Macro Processing, Version 6 Edition.
SAS-Institute Inc, Cary, NC (USA).

[4] Schumacher, E. (1980):
Kendalls TAU, used as a coefficient of disarray between permutations with unoccupied places.
Colloquia Mathematica Societis Janos Bolyai. 32 Nonparametric Statistical

Interference, Budapest.