Diskriminanzanalyse

Sub-Modelle der Kanonischen Analyse

Kanonische Korrelation
Diskriminanzanalyse
Korrespondenzanalyse
Optimale Skalierung

Kurt Holm

Almo Statistik-System
www.almo-statistik.de
holm@almo-statistik.de

kurt.holm@jku.at

2014

Autor des vorliegenden Dokuments: Prof. Dr. Kurt Holm, Universität Linz, Österreich

Im Text wird häufig auf das Dokument **P0** Bezug genommen. Dabei handelt es sich um das Almo-Dokument "Arbeiten mit Almo.PDF" (Dokument 0).

Weitere Almo-Dokumente

Die folgenden Dokumente können alle kostenlos von der Handbuchseite in
www.almo-statistik.de
heruntergeladen werden

```
0. Arbeiten_mit_Almo.PDF (1 MB)
1a. Eindimensionale Tabellierung.PDF (1,8 MB)
1b. Zwei- und drei-dimensionale Tabellierung.PDF (1.1 MB)
2.  Beliebig-dimensionale Tabellierung.PDF (1.7 MB)
3.  Nicht-parametrische Verfahren.PDF (0.9 MB)
4.  Kanonische Analysen.PDF (1.8 MB)
    Diskriminanzanalyse.PDF (1.8 MB)
    enthält: Kanonische Korrelation, Diskriminanzanalyse, bivariate
    Korrespondenzanalyse, optimale Skalierung
5.  Korrelation.PDF (1.4 MB)
6.  Allgemeine multiple Korrespondenzanalyse.PDF (1.5 MB)
7.  Allgemeines ordinales Rasch-Modell.PDF (0.6 MB)
7a. Wie man mit Almo ein Rasch-Modell rechnet.PDF (0.2 MB)
8.  Tests auf Mittelwertsdifferenz,  t-Test.PDF (1,6 MB)
9.  Logitanalyse.pdf  (1,2MB)  enthält Logit- und Probitanalyse
9b. Bootstrap bei Logit- und Probitanalyse.pdf
10. Koeffizienten der Logitanalyse.PDF  (0,06 MB)
11. Daten-Fusion.PDF  (1,1 MB)
12. Daten-Imputation.PDF  (1,3 MB)
13. ALM Allgemeines Lineares Modell.PDF  (2.3 MB)
13a. ALM Allgemeines Lineares Modell II.PDF  (2.7 MB)
13b. Bootstrap bei Allgemeinem Linearen Modell III.PDF
14. Ereignisanalyse: Sterbetafel-Methode,Kaplan-Meier,Cox-Regression(1,5MB)
15. Faktorenanalyse.PDF (1,6 MB)
15a. Bootstrap bei Faktorenanalyse.PDF
16. Konfirmatorische Faktorenanalyse.PDF (0,3 MB)
17. Clusteranalyse.PDF (3 MB)
18. Pisa 2012 Almo-Daten und Analyse-Programme.PDF (17 KB)
19. Guttman- und Mokken-Skalierung.PFD (0.8 MB)
20. Latent Structure Analysis.PDF (1 MB)
21. Statistische Algorithmen in C (80 KB)
22. Conjoint-Analyse (PDF 0,8 MB)
23. Ausreisser entdecken (PDF 170 KB)
24. Statistische Datenanalyse Teil I, Data Mining I
25. Statistische Datenanalyse Teil II, Data Mining II
26. Statistische Datenanalyse Teil III, Arbeiten mit Almo-Datenanalyse
27. Mehrfachantworten. Tabellierung von Fragen mit Mehrfachantworten
28. Metrische multidimensionale Skalierung (MDS)  (0,4 MB)
29. Metrisches multidimensionales Unfolding (MDU) (0,6 MB)
30. Nicht-metrische multidimensionale Skalierung (MDS)  (0,4 MB)
31. Pfadanalyse.PDF  (0,7 MB)
32. Datei-Operationen mit Almo  (1,1 MB)
33. Wählerstromanalyse und Wahlhochrechnung (1,6 MB)
34. Soziometrie. Auswertung soziometrischer Daten (0,5 MB)
35. Konfidenzintervall und p-Wert beim Bootstrap-Verfahren (200 KB)
```

Inhaltsverzeichnis

Der nachfolgende Abschnitt P29 ist dem Almo-Handbuch „Teil4 Fortgeschrittene Verfahren“ entnommen. Er wurde in einigen Teilen überarbeitet.

P29 Kanonische Korrelation
Kanonische Diskriminanzanalyse
Bivariate Korrespondenzanalyse,
Optimale Skalierung

Bei der kanonischen Korrelation werden 2 Variablengruppen miteinander korreliert:

x_1, x_2, x_3	seien die Variablen der 1. Variablengruppe
y_1, y_2	seien die Variablen der 2. Variablengruppen.

Es werden nun folgende Linearkombinationen gebildet:

(0a) $X = \alpha_1 \cdot x_1 + \alpha_2 \cdot x_2 + \alpha_3 \cdot x_3$
(0b) $Y = \beta_1 \cdot y_1 + \beta_2 \cdot y_2$

Die Koeffizienten α_1 , α_2 , α_3 bzw. β_1 , β_2 nennen wir kanonische Gewichtszahlen (oder kanonische Koeffizienten). Die beiden gewichteten Summen X und Y nennen wir kanonische Faktorwertvariable.

Das Prinzip der kanonischen Korrelation ist nun folgendes: Die kanonischen Gewichtszahlen α_i und β_j werden so gewählt, daß die beiden kanonischen Faktorwertvariablen X und Y maximal miteinander korrelieren. Wir nennen diese Korrelation die kanonische Korrelation k.

Würde die Variablengruppe Y nur aus einer Variablen bestehen, dann wären die Gewichtszahlen für die Variablen der Gruppe X, die uns der Kalkül der kanonischen Korrelation ausgeben würde, identisch mit den Regressionskoeffizienten der multiplen Regressionsanalyse. Der kanonische Korrelationskoeffizient K selbst wäre dann identisch mit dem multiplen Korrelationskoeffizienten R. Wir erkennen, daß das Verfahren der kanonischen Korrelation eine Verallgemeinerung der multiplen Regressionsanalyse für den Fall ist, daß auch die abhängige Variable aus einer Menge von Variablen besteht.

Nun besteht die Möglichkeit einen 2. Satz von kanonischen Gewichtszahlen zu bestimmen, der zum 1. orthogonal ist. Diese 2. kanonischen Faktorwertvariablen korrelieren miteinander maximal - mit den kanonischen Faktorwertvariablen der 1. Lösung jedoch mit 0. Wir sprechen hier von einem 1. kanonischen Faktor und einem 2. kanonischen Faktor, der zum 1. orthogonal ist.

Insgesamt lassen sich soviele kanonische Faktoren extrahieren, wie die kleinere der beiden Gruppen Variable umfaßt, in unserem Beispiel sind dies 2. Die Zahl der Variablen in den beiden Gruppen ist in der kanonischen Korrelation und in unserem Almo-Programm nicht beschränkt.

Diskriminanzanalyse: Wenn wir die Variablengruppe Y als abhängige Variablengruppe betrachten und X als unabhängige Variablengruppe und wenn die unabhängige Variablengruppe X aus den quantitativen Variablen x1, x2, x3, ... besteht und die abhängige Variablengruppe Y aus den 0-1 kodierten Dummies einer nominalen Variablen, dann ergibt die kanonische Korrelationsanalyse - auf diese Konstellation angewandt - die Lösung der Diskriminanzanalyse. Wir werden dies später ausführlich in P29.2 darstellen.

Bivariate Korrespondenzanalyse: Die kanonische Korrelationsanalyse kann auch auf folgende Konstellation angewendet werden: Sowohl die eine, wie auch die andere Variablengruppe besteht aus den 0-1 kodierten Dummies je einer

nominalen Variablen. Dabei entsteht dann die Lösung der sogenannten bivariaten Korrespondenzanalyse. Auch diesen Sachverhalt werden wir später ausführlicher in P29.4 darstellen.

Optimale Skalierung ("Lancaster-Skalierung"): Die unstandardisierten kanonischen Koeffizienten der 0-1 kodierten Dummies zweier nominaler Variablen, die im Rahmen der kanonischen Korrelationsanalyse ermittelt werden, können als Skalenwerte der beiden nominalen Variablen begriffen werden. Wir werden die optimale Skalierung in P29.5 ausführlicher darstellen.

Diese 3 Verfahren verwenden denselben Kalkül, den Kalkül der kanonischen Korrelation, den wir im Abschnitt P29.1.2 ausführlich vortragen werden.

P29.1. Kanonische Korrelation

Wir verwenden folgendes Beispiel:
Eine Gruppe von Variablen aus der Arbeitssituation (Überwachung durch Vorgesetzte, Monotonie der Arbeit) und eine Gruppe von Streßindikatoren (Bluthochdruck, Schlafstörung) werden miteinander korreliert.

P29.1.1 Eingabe in Maskenprogramm Prog29m1

Prog29m1.Msk
Kanonische Korrelation

Beispiel:
Eine Gruppe von Variablen aus der Arbeitssituation (Ueberwachung durch Vorgesetzte, Monotonie der Arbeit) und eine Gruppe von Stressindikatoren (Bluthochdruck, Schlafstörung) werden miteinader korreliert
Siehe Handbuch, Abschnitt P29

Was ist ein Kurzprogramm ? --> Hilfe
Bedienung --> Hilfe

1

Speicher fuer x Variable | Hilfe

Vereinbare Variable= 20 ;

2

Option: Weitere Vereinbarungen - nur wenn Almo dazu auffordert

3

Datei der Variablennamen | Hilfe

"C:\Almo7\TESTDAT\Varnamen.nam"

zeige

zeige = Namensdatei in Output zeigen
leer = nicht

4

Freie Namensfelder | Hilfe

Name 5=Ueberwachung;
Name 6=Monotonie;
Name 7=Blutdruck;
Name 8=Schlafstoerung;

erzeuge zusätzliche Namensfelder

5

Datei aus der gelesen wird | Hilfe bei Datei-Problemen

"C:\Almo7\Testdat\TESTDAT.FRE"

frei

Format der Daten | Hilfe

V1:20

der Datensatz enthält diese Variablen
Bei Format DIREKT schreiben Sie: alle_V

6

Wenn Dateiformat FIX oder Nicht-Standard-FREI | Hilfe

7

Analyse-Variable

1. Variablengruppe
unabhängige quantitative Variable

Ueberwachung,Monotonie

2. Variablengruppe
abhängige quantitative Variable

Blutdruck,Schlafstoerung

8 Option: Ein- und Ausschliessen von Untersuchungseinheiten

9 Option: Umkodierungen und Kein-Wert-Angaben

10 Option: Spezielle Kein-Wert-Behandlung

11 Option: Untersuchungseinheiten gewichten

12 verschiedene Programm-Optionen

13 Option: "Aussehen" der auszugebenden Tabelle bzw. Matrix

14 Grafik-Optionen

15 Basisstatistiken ausgeben

0

1= Basisstatistiken ausgeben
0= nicht

P29.1.1.2 Erläuterungen zu den Boxen

Box 1: Vereinbare Variable
Siehe P0.1.

Box 2: Option: Weitere Vereinbarungen - nur wenn Almo dazu auffordert
Siehe P0.2.

Box 3: Datei der Variablennamen
Siehe P0.3.

Box 4: Freie Namensfelder
Siehe P0.3.

Box 5: Datei aus der gelesen wird
Siehe P0.4.

Box 6: Wenn Dateiformat FIX oder nicht Standard-FREI
Siehe P0.4.

Box 7: Analyse-Variable

Analyse-Variable

1. Variablengruppe
unabhängige quantitative Variable

Ueberwachung,Monotonie

2. Variablengruppe
abhängige quantitative Variable

Blutdruck,Schlafstoerung

Geben Sie die unabhängigen und die abhängigen Variablen an

Box 8: Option: Ein- und Ausschliessen von Untersuchungseinheiten
Siehe P0.7.

Box 9: Kein_Wert-Angabe und Umkodierungen
Siehe P0.5.

Box 10: Option: Spezielle Kein-Wert-Behandlung

Option: Spezielle Kein-Wert-Behandlung

Besitzt eine oder mehrere Analysevariablen keinen Wert, dann verwendet Almo standardmäßig das "paarweise Ausscheiden". Der Benutzer hat die Möglichkeit eine von 7 Methoden zur Kein-Wert-Behandlung zu wählen. Dazu muß die Optionsbox geöffnet werden. Man sieht dann folgende große Box:

```
Loesche wieder diese Box

          Option:  Spezielle Kein-Wert-Behandlung
1

     0= Kein-Wert-Fälle in Analyse-Variable nicht vorhanden

     1= Paarweises Ausscheiden I    <---- ist Voreinstellung

     2= Paarweises Ausscheiden II
        a. Paarweises Ausscheiden bei quantitativen und ordinalen
           Variablen.
        b. Vollständiges Ausscheiden bei nominalen Variablen und
           deren Interaktionen, wenn auch nur eine der nominalen
           Analyse-Variablen den Wert "Kein_Wert" besitzt

     3= Vollständiges Ausscheiden
        Vollständiges Ausscheiden des gesamten Datensatzes,
        wenn auch nur eine der Analyse-Variable "Kein_Wert" ist

     4= Mittelwert-Einsetzung I              Hilfe
        Für Kein_Wert wird eingesetzt:
        a. bei quantitativen Variablen der Mittelwert
        b. bei ordinalen Variablen der Median.
           Der zum Median nächst gelegene empirische
           Skalenwert wird dann eingesetzt
        c. bei nominalen Variablen der Erwartungswert

     5= Mittelwert-Einsetzung II             Hilfe
        Für Kein_Wert wird eingesetzt:
        a. bei quantitativen Variablen der zum Mittelwert
           nächste empirisch vorkommende Wert
        b. bei ordinalen der Median (wie bei 4)
        c. bei nominalen Variablen der Erwartungswert (wie bei 4)

     6= Mittelwert-Einsetzung III            Hilfe
        Für Kein_Wert wird eingesetzt:
        a. bei quantitativen Variablen der Mittelwert
           +/- einem normalverteilten Zufallswert mit Mittelwert=0
           und Standardabweichung der Variablen
        b. bei ordinalen der Median (wie bei 4)
           Ist die Variable mit gleicher Schrittweite kodiert,
           dann wird ein Wert X errechnet, der sich ergibt aus
           Median +/- einem normalverteilten Zufallswert mit
           Mittelwert=0 und Standardabweichung in der Größe des
           halben Quartilsabstands der Variablen. Der zu X nächst
           gelegene empirische Skalenwert wird dann eingesetzt
        c. bei nominalen der wahrscheinlichste Ausprägungswert
```

```
7= Mittelwert-Einsetzung IV                    Hilfe
   a. bei quantitativen Variablen zunächst wie bei 6
      Der nächst gelegene empirische Skalenwert
      wird dann eingesetzt
   b. bei ordinalen der Median (wie bei 6)
   c. bei nominalen Variablen (wie bei 6)

      nur relevant für Allgemeines Lineares Modell (ALM) !!
1  1 = wenn abhängige Variable Kein-Wert besitzt, dann
      Datensatz aus Analyse vollständig ausschliessen
      unabhängig davon welche Kein-Wert-Behandlung
      oben im ersten Eingabefeld dieser Box gewählt wurde
   0 = gewählte Kein-Wert-Behandlung gilt auch für
      abhängige Variable

   123457
      Startwert für Zufallsgenerator                Hilfe
      fuer Kein-Wert-Behandlung 6 und 7

   1  als "gemeinsame" Fallzahl für Signifikanztest
      wird verwendet - wenn Kein-Wert-Behandlung =1 oder =2
      und wenn Kein-Wert-Fälle auftreten:
      0 = die kleinste Fallzahl, aus der die Co-Streungen zwischen
          je 2 Variablen i und k errechnet wurden
      1 = das harmonisches Mittel aus den Fallzahlen, aus denen
          die Co-Streuungen zwischen den Variablen errechnet wurden
      2 = die Zahl der Fälle, die in allen
          Analysevariablen valide Werte besitzen
      3 = die Zahl der eingelesenen Fälle
                                                    Hilfe
```

Kein-Wert-Behandlung 1: "Paarweises Ausscheiden"
Wir werden dieses Verfahren im Handbuch P45 „Data Mining“, in Abschnitt P45.12.4 sehr ausführlich darstellen. Hier wollen wir es nur kurz beschreiben.

Wird der Kalkül der kanonischen Korrelation auf die Korrelationsmatrix der Variablen angewendet, dann ist die Vorgehensweise folgende:

Jeder einzelne Korrelationskoeffizient r_{ik} für die beiden Variablen i und k wird nur aus den Untersuchungseinheiten errechnet, für die aus beiden Variablen i und k valide Werde vorhanden sind. In die Diagonale der Korrelationsmatrix wird 1.0 eingesetzt. Die Folge dieser Vorgehensweise ist, daß die verschiedenen Korrelationskoeffizienten aus verschiedenen Fallzahlen berechnet sind. Almo ermittelt standardmäßig das harmonische Mittel aus den verschiedenen Fallzahlen und verwendet dieses für Signifikanztests.

Wird der Kalkül der kanonischen Korrelation auf die Kovarianz- oder Quadratsummenmatrix angewendet, dann wird folgendermaßen verfahren:

Betrachten wir die Matrix der Abweichungsquadratsummen (kurz: Quadratsummen-matrix) zwischen den 3 Variablen V1, V2 und V3.

	V1	V2	V3
V1	SS_{11}	SS_{12}	SS_{13}
V2		SS_{22}	SS_{23}
V3			SS_{33}

Die Quadratsumme SS_{12} zwischen den Variablen V1 und V2 wird aus den Datensätzen ermittelt, die in diesen beiden Variablen valide Werte besitzen. Entsprechend wird SS_{13} und SS_{23} berechnet. Die Folge dieser Vorgehensweise ist, dass die 3 Quadratsummen auf jeweils verschiedenen n_{ij} (Zahl der

Untersuchungseinheiten) beruhen. In die Diagonale wird die Quadratsumme der Variablen selbst eingesetzt. SS_{11} ist also die Quadratsumme für die Variable V1, die sich aus den Untersuchungseinheiten ergibt, die in V1 einen validen Wert besitzen. Entsprechend wird auch SS_{22} und SS_{33} gebildet. Dann wird jede Zelle der Quadratsummenmatrix zuerst durch das zu ihr gehörende n_{ij} dividiert. Dadurch entsteht die Kovarianzmatrix. Sie ist also die „durchschnittliche" Quadratsummenmatrix. Almo ermittelt nun das harmonische Mittel n_h aus den unterschiedlichen n_{ij} des oberen Dreiecks der Matrix (ohne Diagonale). Die Kovarianzmatrix wird dann mit n_h multipliziert. Damit entsteht wieder eine Quadratsummenmatrix, diese Mal mit gleichen n_{ij}. Dieses Hochrechnen der Kovarianzmatrix zu einer neuen Quadratsummenmatrix könnte auch unterbleiben. Die Koeffizienten sind die gleichen, egal ob wir für den Kalkül die Kovarianzmatrix oder die „hochgerechnete" Quadratsummenmatrix verwenden. Dabei ist es sogar gleichgültig mit welchem n multipliziert wurde. Um die Signifikanzen ermitteln zu können, muß allerdings eine Entscheidung für ein bestimmtes n getroffen werden. Almo entscheidet sich hier für das harmonische Mittel n_h. Gelegentlich multipliziert Almo die Kovarianzmatrix mit der Zahl der eingelesenen Einheiten und verwendet aber n_h für die Signifikanztests.

Kein-Wert-Behandlung 2: „Paarweises Ausscheiden II"

a. Paarweises Ausscheiden bei ursächlichen quantitativen und ordinalen Variablen.
b. Vollständiges Ausscheiden bei ursächlichen nominalen Variablen und deren Interaktionen, wenn auch nur eine der nominalen Analyse-Variablen den Wert "Kein_Wert" besitzt

Kein-Wert-Behandlung 3: „Vollständiges Ausscheiden"

Vollständiges Ausscheiden des gesamten Datensatzes, wenn auch nur eine der ursächlichen Analyse-Variable "Kein_Wert" ist.

Kein-Wert-Behandlung 4: Mittelwert-Einsetzung I

Almo ermittelt zuerst Mittelwerte (für quantiative Variable), Median (für ordinale Variable) und den Erwartungswert (für nominale Variable).
Almo gibt diese Werte aus.

Für Kein_Wert wird eingesetzt:

b) bei quantitativen Variablen der Mittelwert

c) bei ordinalen Variablen der Median (=der mittlere Wert)
Liegt der Median nicht auf einem empirischen Wert, sondern zwischen 2 empirschen Werten, dann wird der nächst gelegene Nachbarwert als KW-Einsetzungswert verwendet.

d) bei nominalen Variablen die zum Erwartungswert nächste empirisch vorkommende Codeziffer

Die Berechnung des Erwartungswerts soll an einem Beispiel gezeigt werden. Die nominale Variable sei der Beruf mit den 3 Ausprägungen Arbeiter, Angestellte, Sonstige. Dabei wurden folgende Häufigkeiten ermittelt.

	Code	Häufigkeit	Anteil	Code*Anteil
Arbeiter	1	250	0.25	0.25
Angestellte	2	400	0.40	0.80
Sonstige	3	350	0.35	1.05

Summe				2.10

Der Erwartungswert ist 2.1
Die nächste empirisch vorkommende Codeziffer ist 2 der KW-Einsetzungswert ist also 2.

Kein-Wert-Behandlung 5: Mittelwert-Einsetzung II

Für Kein_Wert wird eingesetzt:

a. bei quantitativen Variablen der zum Mittelwert nächste empirsch vorkommende Wert
b. bei ordinalen Variablen der Median wie bei Kein-Wert-Behandlung 4
c. bei nominalen Variablen der Erwartungswert wie bei Kein-Wert-Behandlung 4

Kein-Wert-Behandlung 6: Mittelwert-Einsetzung III

Für Kein_Wert wird eingesetzt:

a. bei quantitativen Variablen der Mittelwert +/- einem normalverteilten Zufallswert mit Mittelwert=0 und Standardabweichung der Variablen
Wir könnten auch formulieren: Es wird ein normalverteilter Zufallswert mit Mittelwert und Standardabweichung der Variablen eingesetzt.
b . bei ordinalen Variablen der Median. Ist die Variable (was eher ungewöhnlich ist) mit ungleichen Schrittweiten kodiert (z.B. 1, 2, 5, 6, 23), dann wird der Median eingesetzt.
Liegt dieser zwischen zwei empirisch vorkommenden Werten, dann wird der zum Median nächst gelegene empirische Wert verwendet.

Ist die Variable mit gleicher Schrittweite kodiert, dann wird ein Wert X errechnet, der sich ergibt aus Median +/- einem normalverteilten Zufallswert mit Mittelwert=0 und Standardabweichung in der Größe des halben Quartilsabstands der Variablen. Der zu X nächst gelegene empirische Skalenwert wird dann eingesetzt.

Bei quantitativen und bei ordinalen Variablen wird also eine normalverteilte Zufallszahl mit Mittelwert=0 generiert.
Als Standardabweichung wird bei quantitativen Variablen die der jeweiligen Variablen verwendet. Bei ordinalen Variablen wird der halbe Quartilsabstand verwendet.

Betrachten wir ein Beispiel: Die quantitative Variable sei das Lebensalter. Almo errechnet für sie einen Mittelwert von 40 und eine Standardabweichung von 20. Dann wird eine normalverteilte Zufallszahl mit Mittelwert=0 und Standardabweichung=20 erzeugt. Nehmen wir an es entsteht der Zufallswert -15.25. Für den fehlenden Wert wird dann eingesetzt X = 40-15.25 = 24.75.

Bei einer ordinalen Variablen wird entsprechend verfahren. Als Standardabweichung für die Generierung der Zufallszahl wird der halbe Quartilsabstand verwendet. Der ermittelte X-Wert wird bei der ordinalen Variablen aber noch nicht als KW-Einsetzungswert verwendet. Es wird nach dem empirisch vorkommenden Wert gesucht, der am dichtesten bei X liegt. Dieser wird als KW-Einsetzungswert verwendet. So wird verhindert, daß KW-Einsetzungswerte entstehen, die empirisch nicht vorkommen.

e) Bei nominalen Variablen wird der wahrscheinlichste Ausprägungswert

eingesetzt. Die Vorgehensweise soll an einem Beispiel gezeigt werden. Die nominale Variable sei der Beruf mit den 3 Ausprägungen Arbeiter, Angestellte, Sonstige. Dabei wurden folgende Häufigkeiten ermittelt.

	Code	Häufigkeit	in %	in % kummuliert
Arbeiter	1	250	25	25
Angestellte	2	400	40	65
Sonstige	3	350	35	100

Dann wird eine gleichverteilte Zufallszahl zwischen 0 und 100 erzeugt.

Liegt sie zwischen
0 und 25, dann wird für den fehlenden Wert 1 eingesetzt
25 65 2
65 100 3

Kein-Wert-Behandlung 7: Mittelwert-Einsetzung IV

Für Kein_Wert wird eingesetzt:

a. bei quantitativen Variablen:
   Es wird zunächst ein Wert X errechnet, der sich ergibt aus dem Mittelwert +/- einem normalverteilten Zufallswert mit Mittelwert=0 und der Standardabweichung der Variablen. Dann wird der zu X nächst gelegene empirische Skalenwert für Kein_Wert eingesetzt. So wird verhindert, dass KW-Einsetzungswerte entstehen, die empirisch nicht vorkommen.
b. bei ordinalen Variablen wie bei Kein-Wert-Behandlung 6
c. bei nominalen Variablen wie bei Kein-Wert-Behandlung 6

Kein-Wert-Behandlung 4 und 5 unterscheiden sich von 6 und 7 dadurch, dass bei 6 und 7 eine Zufallsvariation dem Mittelwert bzw. Median bzw. Erwartungswert hinzugefügt wird.

Die Kein-Wert-Behandlung 4 unterscheiden sich von 5 nur dadurch dass für die quantitativen Variablen ein Mal der Mittelwert und das andere Mal der zum Mittelwert nächste empirisch vorkommende Wert als KW-Einsetzungswert verwendet wird.

Warum Zufallswert hinzufügen?

Es muß noch folgende Frage beantwortet werden: Warum wird der Mittelwert bzw. der Median bei Kein-Wert-Behandlung 6 und 7 durch einen Zufallswert überlagert?

Wird als KW-Einsetzungswert nur der Mittelwert (bzw. der Median) verwendet, dann wird die Varianz der Variablen verringert, weil für Kein-Wert immer derselbe Wert eingesetzt wird.

Werden mit den so erzeugten „vollständigen“ Daten beispielsweise Korrelationen errechnet, dann werden die Signifikanzen dieser Korrelationen überschätzt. Siehe dazu etwa R. J. A. Little/D. B. Rubin (1990, S. 381).

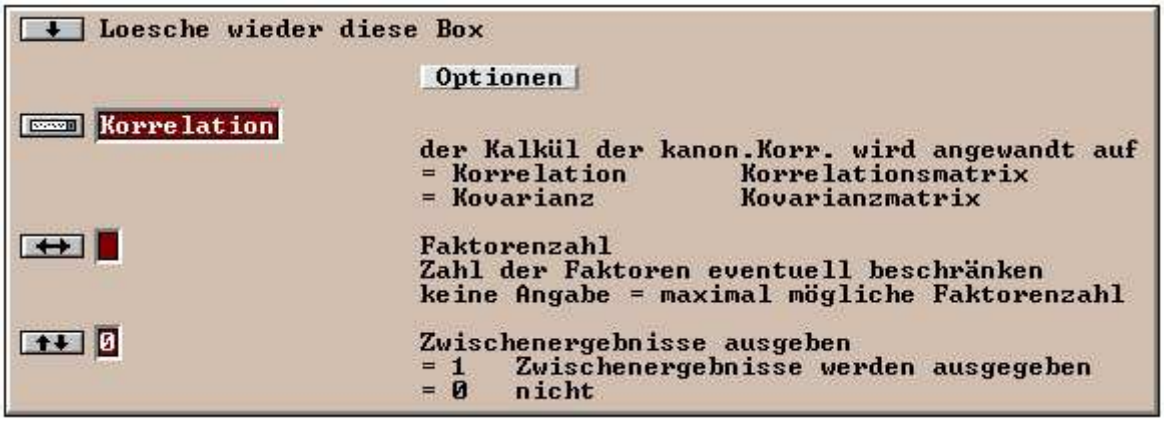
Die Überlagerung durch einen normalverteilten Zufallswert mit der Standardabweichung der Variablen bezweckt also, dass die Varianz der Variablen (fast) unverändert bleibt. Gleiches gilt auch für nominale Variable. Der Erwartungswert der Variablen ist immer derselbe. Dadurch wird die Varianz

verringert. Durch den "wahrscheinlichsten" Wert bleibt die Streuung (fast) unverändert.

Box 11: Option: Untersuchungseinheiten gewichten
Siehe P0.8.

Box 12: Verschiedene Programm-Optionen

Optionsbox geöffnet:



Eingabefeld 1:
Der Kalkül der kanonischen Korrelation kann auf die Korrelations- oder Kovarianz-Matrix angewendet werden. Normalerweise wird man die Korrelationsmatrix verwenden.

Eingabefeld 2:
Es können soviele Faktoren extrahiert werden, wie die kleinere der beiden Variablengruppen Variable umfasst. Der Benutzer kann aber diese Faktorenzahl einschränken.

Eingabefeld 2:
Es können Zwischenergebnisse angefordert werden.

Box 13: Option: "Aussehen" der auszugebenden Tabelle bzw. Matrix
Siehe P0.9.

Box 14: Grafik-Optionen
Siehe P0.10.

Box 15: Basisstatistiken ausgeben

Es können zusätzlich Basisstatitiken ausgegeben werden. Dies sind u.a.

Mittelwerte
Standardabweichungen
Zahl der diversen Werte je Variable
Zahl der fehlenden Werte je Variable

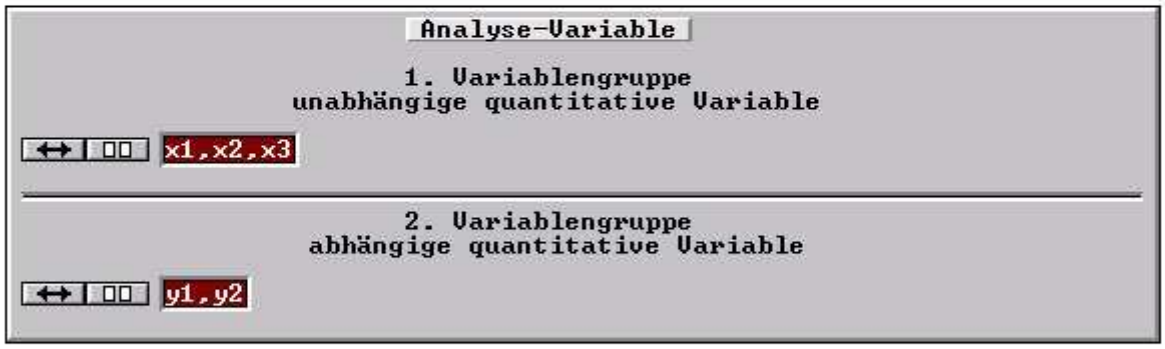
P29.1.1.5 Eingabe einer fertigen Korrelationsmatrix mit Prog29ma

Im folgenden zeigen wir ein Maskenprogramm, in dem eine kanonische Korrelation mit einer eingegebenen fertigen Korrelationsmatrix gerechnet wird.

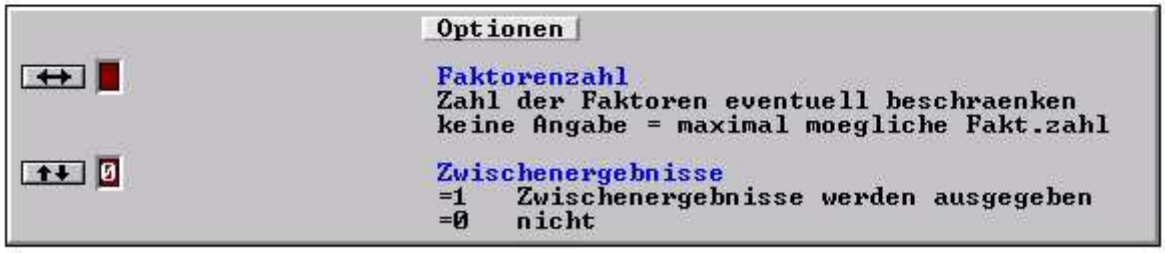
1

2

3



4



5

6

7

```
Schreiben der Matrixwerte

Schreiben Sie hier dahinter das untere
Dreieck der Matrix inklusive Diagonale

BEACHTE: Vor der Matrix steht folgendes:
      *              =Stern
      100            =Zahl der Fälle
      *              =Stern
Wenn die Zahl der Fälle, auf denen die
Korrelationsmatrix beruht nicht bekannt
ist, dann schreiben Sie auch einen Stern

Hinter der Dreiecksmatrix müssen 2 durch
ein Blank getrennte Sterne stehen

Schalten Sie dazu die Schreibsperre aus

 ↓  Schreibsperre      <--- EIN : rot
                            AUS : grau
```

```
*
100
*

1.0
0.4   1.0
0.5   0.6   1.0
0.3   0.2   0.4   1.0
0.2   0.3   0.3   0.4   1.0

* *
```

P29.1.2 Ergebnisse und Kalkül der kanonischen Korrelation

Wie werden im folgenden die Ausgabe aus dem Maskenprogramm Prog29ma (bzw. dem identischen „selbst geschriebenen“ Almo-Programm Prog29e) erläutern. Dabei werden wir auch gleich den Kalkül der kanonischen Korrelation vortragen.

Wir gehen von folgender Streuungsmatrix aus

		Variablengruppe I	Variablengruppe II
Variablengruppe	I	$\mathbf{R}_{11}$	$\mathbf{R}_{12}$
	II	$\mathbf{R}_{21}$	$\mathbf{R}_{22}$

Die Streuungsmatrix kann die Kovarianzmatrix oder die Korrelationsmatrix sein.

Selbstverständlich ist $R_{21} = R'_{12}$ bzw. $R_{12} = R'_{21}$

1. Der Hotellingsche Kalkül

Es wird das Matrizenprodukt

(1) $\mathbf{M_1 = R_{11}^{-1} \cdot R_{12} \cdot R_{22}^{-1} \cdot R_{21}}$

gebildet.

Die Eigenwerte und Eigenvektoren dieser Matrix werden ermittelt. Die Wurzel aus dem Eigenwert entspricht dem gesuchten kanonischen Korrelationskoeffizienten K.

Die Eigenvektoren entsprechen den kanonischen Gewichtszahlen der Variablengruppe I je kanonischem Faktor.

Wir können auch das Matrixprodukt

(2) $\mathbf{M_2 = R_{22}^{-1} \cdot R_{21} \cdot R_{11}^{-1} \cdot R_{12}}$

verwenden. Die Eigenwerte von $\mathbf{M_2}$ stimmen mit denen von $\mathbf{M_1}$ überein. Die Eigenvektoren von $\mathbf{M_2}$ entsprechen den kanonischen Gewichtszahlen der Variablengruppe II je Faktor. Die 1. Eigenvektoren von $\mathbf{M_1}$ und $\mathbf{M_2}$ (d.h. die Gewichtszahlen von Variablengruppe I und II für den 1. kanonischen Faktor) sind in folgender Weise miteinander verbunden

(3a) $\mathbf{G_1 = R_{11}^{-1} \cdot R_{12} \cdot G_2} \cdot \frac{1}{\sqrt{E_1}}$

(3b) $\mathbf{G_2 = R_{22}^{-1} \cdot R_{21} \cdot G_1} \cdot \frac{1}{\sqrt{E_1}}$

$\mathbf{G_1}$ = 1. Eigenvektor von M1
$\mathbf{G_2}$ = 1. Eigenvektor von M2
E_1 = 1. Eigenwert

Eine ausführliche Darstellung des Hotellingschen Kalküls ist enthalten in Morrison (1967, S.213-219).

2. Kalkül nach Paul Horst (der in Almo verwendet wird)

Es wird das Matrixprodukt

(4) $\mathbf{M_1 = T_{11}^{'-1} \cdot R_{12} \cdot T_{22}^{'-1} \cdot T_{22}^{-1} \cdot R_{21} \cdot T_{11}^{'-1}}$

oder das Matrixprodukt

(5) $\mathbf{M_2 = T_{22}^{-1} \cdot R_{21} \cdot T_{11}^{'-1} \cdot T_{11}^{-1} \cdot R_{12} \cdot T_{22}^{'-1}}$

$\mathbf{T_{11}, T_{22}}$ das ist die untere Dreiecksmatrix (Cholesky-Matrix) von $\mathbf{R_{11}}$ bzw. $\mathbf{R_{22}}$

$\mathbf{T_{22}, T'_{22}}$ das ist die jeweils obere Dreiecksmatrix (d.h. die Transponierte zu $\mathbf{T_{11}}$ bzw. $\mathbf{T_{22}}$

$\mathbf{T_{11}^{-1}, T_{22}^{-1}, T'^{-1}_{22}}$ das sind die Inversen der jeweiligen Dreiecksmatrizen.

Die positiven Eigenwerte und die dazu gehörenden Eigenvektoren von **M1** und **M2** werden ermittelt. Die Zahl dieser Eigenwerte ist gleich der Variablenzahl der kleineren Variablengruppe. Die Eigenwerte von **M1** und **M2** sind gleich. Die Wurzel aus dem Eigenwert ist die kanonische Korrelation.

(6) $k_i = \sqrt{E_i}$

E_i = i-ter Eigenwert
k_i = i-te kanonische Korrelation

Die Matrizen $\mathbf{G_1}$ und $\mathbf{G_2}$ der (unstandardisierten) kanonischen Gewichtszahlen entstehen aus

(7a) $\mathbf{G_1 = T'^{-1}_{11} \cdot v_1}$

(7b) $\mathbf{G_2 = T'^{-1}_{22} \cdot v_2}$

$\mathbf{v_1}$= die m_1 * p Matrix der Eigenvektoren aus M_1
$\mathbf{v_2}$= die m_2 * p Matrix der Eigenvektoren aus M_2

m_1= Zahl der Variablen der 1. Variablengruppe
m_2= Zahl der Variablen der 2. Variablengruppe
p= Zahl der kanonischen Faktoren (= der positiven Eigenwerte)

Die Vorzeichen der Spalte von G2 müssen überprüft werden. Wir bestimmen die kanonische Korrelation ein 2. Mal nach folgender Formel

(7c) $\mathbf{K} = \mathbf{G'_2} \cdot \mathbf{R_{21}} \cdot \mathbf{G_1}$

K= dies ist die p*p Diagonalmatrix der kanonischen Korrelation. In ihrem Diagonalglied ii steht die kanonische Korrelation k.

(7d) Ist das i-te Diagonalglied von **K** negativ, dann drehen wir in Spalte i von $\mathbf{G_2}$ das Vorzeichen um. Dadurch erreichen wir, daß k_i, nach 7c berechnet, wieder positiv wird. Alternativ wäre es auch möglich, die Spalte i von $\mathbf{G_1}$ in ihrem Vorzeichen umzudrehen.

3. In Almo wird der Kalkül nach Horst verwendet. Betrachten wir das Beispiel, das wir im Maskenprogramm Prog29ma bzw. im identischen „selbst geschriebenen“ Almo-Programm Prog29e gerechnet haben.

Die Ergebnisse dieses Almo-Programms sollen im folgenden besprochen werden. Der Benutzer wird zuerst feststellen, daß Almo die 1. Variablengruppe auch als "unabhängige" und die 2. Variablengruppe auch als "abhängige" bezeichnet. Diese Einteilung ist für die kanonische Korrelationsanalyse bedeutungslos. Für die (später dargestellte) Diskriminanz-Analyse ist sie jedoch wichtig.

Die Korrelationsmatrix war folgende

	Variablengruppe 1			Variablengruppe 2	
	x1	x2	x3	y1	y2
x1	1.0000	0.4000	0.5000	0.3000	0.2000
x2	0.4000	1.0000	0.6000	0.2000	0.3000
x3	0.5000	0.6000	1.0000	0.4000	0.3000
y1	0.3000	0.2000	0.4000	1.0000	0.4000
y2	0.2000	0.3000	0.3000	0.4000	1.0000

Almo liefert zuerst die Ergebnisse für die 1. Variablengruppe und dann für die 2. Der besseren Übersicht halber werden wir die Ergebnisse zusammenfassen.

4. Zuerst werden die Inversen der Cholesky-Matrizen $\mathbf{T_{11}^{-1}}$ und $\mathbf{T_{22}^{-1}}$ ermittelt (die wir von Almo erhalten, wenn die Option "Zwischergeb=1;" gesetzt wird)

   Inverse der oberen Dreiecksmatrix $\mathbf{T_{11}}$ der Cholesky-Matrix

Spalte 1	Spalte 2	Spalte 3
1.0000	-0.4364	-0.41370
0	1.0911	-0.63660
0	0	1.3369

   Inverse der unteren Dreiecksmatrix $\mathbf{T_{22}}$ der Cholesky-Matrix

Spalte 1	Spalte 2
1.0000	0
-0.4364	1.0911

5. Die Matrizen $\mathbf{M_1}$ und $\mathbf{M_2}$ sind

Matrix $\mathbf{M_1}$

0.0976	0.0457	0.0863
0.0457	0.0577	0.0281
0.0863	0.0281	0.0804

Matrix $\mathbf{M_2}$

0.1778	0.0500
0.0500	0.0579

6. Die Zahl der Eigenwerte größer .0 in $\mathbf{M_1}$ und $\mathbf{M_2}$ ist gleich die Zahl der Variablen in der kleineren Variablengruppe, also 2. Aus den Matrizen $\mathbf{M_1}$ und $\mathbf{M_2}$ extrahiert Almo 2 Eigenwerte.

1.Eigenwert: 0.19602 1. kanonische Korrelation $\sqrt{0.19602} = 0.44274$

2.Eigenwert: 0.03980 2. kanonische Korrelation $\sqrt{0.03980} = 0.19950$

Das vollständige Ergebnis ist folgendes:

Faktor	Kanonische Korrelation	Eigenwert	Wilks' Lambda	Chi-Quadrat	df	Signifikanz (1-p)*100
1	0.44274	0.19602	0.77199	24.84375	6	99.93769 %
2	0.19950	0.03980	0.96020	3.89891	2	85.96303 %
Summe (=Pillais Spur)		0.23582				

```
Koeffizienten fuer Gesamtmodell
-------------------------------
multiple Korrelation                    0.343377
  beruhend auf Pillais Spur
  siehe Handbuch P 29.1.2,(10b),(10c)
F-Wert                                  4.277387
Freiheitsgrade Nenner =    6
               Zaehler=  192
Signifikanz: p                          0.000694
Signifikanz: (1-p)*100                 99.930554 %
Teststaerke von F                       0.979128
-----------------------------------------------
```

********** **Erläuterung:**

Für die beiden (orthogonalen) Faktoren wird auch die aus Wilks Lambda abgeleitete Signifikanz mitgeteilt. Wir werden im nachfolgenden Punkt 8 darauf eingehen. Die beiden Variablengruppen korrelieren mit einem multiplen Korrelationskoeffizienten von 0.343377 miteinander. Er ist mit (1-p)*100 = 99.930554 % signifikant.

Die zu den beiden Eigenwerten gehörenden Eigenvektoren werden mitgeteilt, wenn der Benutzer Zwischenergebnisse anfordert.

Eigenvektoren v_1

	Faktor 1	Faktor 2
x1	0.7042	0.1011
x2	0.3576	-0.9055
x3	0.6133	0.4119

Eigenvektoren v_2

	Faktor 1	Faktor 2
y1	0.9401	-0.3408
y2	0.3408	0.9401

7. Die Matrizen der (unstandardisierten) kanonischen Gewichtszahlen sind dann (gemäß 7a und 7b)

G_1 Kanonische

G_2 Kanonische

```
Gewichtszahlen fuer                 Gewichtszahlen fuer 2.
1.(unabhaengige)                    (abhaengige)
Variablengruppe                     Variablengruppe
(unstandardisiert)                  (unstandardisiert)
     Faktor 1   Faktor 2                 Faktor 1   Faktor 2
 x1   0.2943     0.3259              y1   0.7914     0.7511
 x2  -0.0002    -1.2503              y2   0.3718    -1.0258
 x3   0.8199     0.5506
```

Die Vorzeichen in Spalte 2 von G2 sind entsprechend der Vorschrift 7d umgedreht worden.

8. Die Signifikanz der kanonischen Korrelation wird über das Wilks'sche Lambda durch den Bartlett-Test ermittelt.
   Wilks Lambda und der Chi-Quadrat-Wert nach Bartlett für die i-te kanonische Korrelation sind

(8) $W_i = (1-E_i).(1-E_{i+1}) \; .... \; .(1-E_p)$

(9) $\text{Chi-Quadrat}_i = (n-1-0.5(m_1+m_2+1) \; \ln(W)$

(10) $df_i = (m_1-i+1)(m_2-i+1)$

W_i = Wilks Lambda für die i-ten kanonische Korrelation
Chi-Quadrat_i = Chi-Quadrat für i-ten kanonische Korrelation
df_i = Freiheitsgrade
Ei = i-ter Eigenwert aus M_1
E_p = letzter (kleinster) positiver Eigenwert aus M_1
p = Zahl der positiven Eigenwerte aus M1
n = Zahl der Untersuchungseinheiten
m_1, m_2 = Zahl der Variablen der 1. bzw. der 2. Variablengruppe
i = i-te kanonische Korrelation, die getestet werden soll.

Für die 1. kanonische Korrelation entsteht

$W_1 = (1-E_1)(1-E_2) = (1-0.196)(1-0.0398) = 0.772$
$\text{Chi-Quadrat}_1 = -(100-1-0.5(3+2+1)).\ln(0.772) = 24.84$
$df_1 = (m_1-1+1)(m_2-1+1) = 6$

Signifikanz p = 0.000623
(1-p).100 = 99.9377

Für die 2. kanonische Korrelation entsteht

$W_2 = (1-E_2) = (1-0.0392) = 0.9602$
$\text{Chi-Quadrat}_2 = 3.8989$
$df_2 = (m_1-2+1)(m_2-2+1) = 2$

Signifikanz p = 0.1404
(1-p).100 = 85.96

Pillais Spur und Cramers V

Die Summe der Eigenwerte aus der kanonischen Korrelation ist identisch mit Pillais Spur aus dem Allgemeinen Linearen Modell. Siehe dazu Handbuch zu P20, Abschnitt P20.9.4.1, Punkt 12. Aus Pillais Spur kann nun ein mulitpler Korrelationskoeffizient errechnet werden.

(10b) $R_P = \sqrt{\frac{\sum E}{p}}$

R_P = „Pillais Korrelation“
ΣE = Summe der Eigenwerte
p = Zahl der kanonischen Faktoren

Bestehen die beiden Variablengruppen aus den Dummies zweier nominal-polytomer Variablen (das ist der Fall der bivariaten Korrespondenzanalyse, siehe P29.4), dann ist „Pillais Korrelation“ identisch mit Cramers V, wie wir es mit Prog10 aus einer 2-dimensionalen Tabellenanalyse erhalten. Siehe dazu auch P20.9.5.1.

9. Aus Pillais Spur können wir nun die Signifikanz des Gesamtmodells ermitteln. Die Formeln dafür sind in P20.9.4.1, Punkt 12 angegeben. Almo liefert folgende Ausgabe:

```
---------------------------------------------------------------------------

Faktor  Kanonische  Eigenwert   Wilks' Lambda  Chi-Quadrat  df  Signifikanz
        Korrelation                                            (1-p)*100

 1       0.44274     0.19602       0.77199      24.84375     6   99.93769 %
 2       0.19950     0.03980       0.96020       3.89891     2   85.96303 %
---------------------------------------------------------------------------
Summe                0.23582
(=Pillais Spur)

Koeffizienten fuer Gesamtmodell
-------------------------------
multiple Korrelation                    0.343377
  beruhend auf Pillais Spur
  siehe Handbuch P 29.1.2,(10b),(10c)
F-Wert                                  4.277387
Freiheitsgrade Nenner =    6
               Zaehler=  192
Signifikanz: p                          0.000694
Signifikanz: (1-p)*100                 99.930554 %
Teststaerke von F                       0.979128
-----------------------------------------------
```

Wenn wir die 1. Variablengruppe als unabhängige Variable betrachten, die die 2. Variablengruppe erklären, dann interessiert natürlich auch die Signifikanz der kanonischen Gewichtszahlen der einzelnen Variablen der 1. Gruppe hinsichtlich der 2. Variablengruppe. Diese Signifikanzprüfung wird im Rahmen unseres Programms zur kanonischen Korrelation nicht vorgenommen. Sie kann jedoch mit Programm 20 vorgenommen werden. Zu diesem Zweck braucht der Benutzer im „selbst geschriebenen“ Almo-Programm Prog29e nur die Programm-Nr. von 29 auf 20 zu ändern.

Almo rechnet in diesem Fall eine multivariate Regressionsanalyse und gibt dabei u.a. für unser Beispiel aus

	Wilks lambda	F-Wert	df1	df2	Signifikanz (1-p)*100
x1	0.9816	0.8905	2	95	58.33%
x2	0-9610	1.9288	2	95	85.10%
x3	0.9101	4.6903	2	95	98.86%

Wir sehen, daß nur x3 eine signifikante Wirkung besitzt. Beachte, daß hier die Trennung in orthogonale kanonische Faktoren keine Rolle spielt.

10. Die Matrizen der standardisierten kanonischen Gewichtszahlen **C1** und für **C2**

die beiden Variablengruppen ergeben sich aus

(11a) $\mathbf{C_1} = \mathbf{G_1}' \cdot \mathbf{D_1}$
(11b) $\mathbf{C_2} = \mathbf{G_2}' \cdot \mathbf{D_2}$

$\mathbf{G_1}, \mathbf{G_2}$ = siehe bei 7a, 7b
$\mathbf{D_1}$ = Diagonalmatrix der Wurzel aus den Diagonalgliedern von $\mathbf{R_{11}}$
$\mathbf{D_2}$ = Diagonalmatrix der Wurzel aus den Diagonalgliedern von $\mathbf{R_{22}}$

Wird von der Korrelationsmatrix (wie in unserem Beispiel) ausgegangen, dann sind $\mathbf{D_1}$ und $\mathbf{D_2}$ Einheitsmatrizen, so daß $\mathbf{C_1}$ gleich $\mathbf{G_1}$ und $\mathbf{C_2}$ gleich $\mathbf{G_2}$ sind.

Der Unterschied zwischen unstandardisierten und standardisierten kanonischen Koeffizienten ist vergleichbar dem zwischen unstandardisierten und standardisierten Regressionskoeffizienten in der Regressionsanalyse. Die standardisierten kanonischen Koeffizienten erzeugen - eingesetzt in die Gleichung 0a bzw. 0b - kanonische Faktorwerte mit Mittelwert 0 und Standardabweichung 1.0.

Die kanonischen Koeffizienten können nur als standardisierte miteinander verglichen werden, sofern die Variablen einer Gruppe in verschiedenen Meßeinheiten gemessen wurden.

11. Als "kanonische Strukturkoeffizienten" $\mathbf{S_1}$ bzw. $\mathbf{S_2}$ bezeichnet man die Korrelationen zwischen den kanonischen Faktorwertvariablen $\mathbf{X_i}$ bzw. $\mathbf{Y_i}$ (des i-ten kanon Faktors) mit den Originalvariablen x1, x2, ... bzw. y1, y2, ... (siehe Gleichung 0a und 0b).

    (12a) $\mathbf{S_1} = \mathbf{R_{11}} \cdot \mathbf{G_1}$
    (12b) $\mathbf{S_2} = \mathbf{R_{22}} \cdot \mathbf{G_2}$

    $\mathbf{G_1}, \mathbf{G_2}$ = siehe 7a, 7b
    $\mathbf{R_{11}}, \mathbf{R_{22}}$ = Korrelationsmatrix der 1. bzw. 2. Variablengruppe

    Beachte: Wurde die kanonische Korrelationsanalyse nicht auf die Korrelationsmatrix, sondern z.B. auf die Kovarianzmatrix angewendet, dann muß zuvor für Gleichung 12a bzw. 12b die Korrelationsmatrix $\mathbf{R_{11}}$ bzw. $\mathbf{R_{22}}$ gebildet werden.

Für unser Beispiel erhalten wir

S_1
Kanonische Strukturkoeffizienten der 1. (unabhaengigen) Variablengruppe (=Korrelation der Variablen mit den kanonischen Faktoren)

	Faktor 1	Faktor 2
x1	0.7042	0.1011
x2	-0.6094	-0.7895
x3	0.9669	0.0365

S_2
Kanonische Strukturkoeffizienten der 2.(unabhaengigen) Variablengruppe (=Korrelation der Variablen mit den kanonischen Faktoren)

	Faktor 1	Faktor 2
y1	0.9401	0.3408
y2	0.6884	-0.7253

Der 1. Koeffizient 0.7042 bedeutet, daß die Originalvariable x1 mit der kanonischen Faktorwertvariable X1 (für den 1. kanonischen Faktor) mit 0.7042 korreliert.

12. Erklärte Varianz und Redundanzanalyse.

Almo liefert folgende Ausgabe:

```
-----------------------------------------------------------------------------------------------------
-------------------
Prozent erklaerte (standardisierte) Varianz in der 1.(unabhaengigen)
Variablengruppe
durch eigene kanonische Faktoren erklaert
 Faktor 1  60.0805 %
 Faktor 2  21.1641 %

durch kanonische Faktoren der anderen Variablengruppe erklaert
 Faktor 1  11.7767 %
 Faktor 2  0.8423 %
-----------------------------------------------------------------------------------------------------
-------------------
```

Entsprechende Werte werden auch für die 2. Variablengruppe ausgegeben.

Für die 1. Variablengruppe gilt

(13a) $SS_i = \sum s^2 / m_1$

(13b) $Rd_i = SS_i * E;$

SS_i = der in der Variablengruppe 1 durch den eigenen i-ten kanonischen Faktor erklärten Varianzanteil

Rd_i = der in der 1. Variablengruppe durch den i-ten kanonischen Faktor der 2. Variablengruppe erklärten Varianzanteil. Rd_i wird auch bezeichnet als "Redundanz der Variablengruppe 1 - gegeben Variablengruppe 2 - bezogen auf die kanonische Beziehung i"

$\sum s^2$ = Summe der quadrierten Strukturkoeffizienten (gemäß 12a) des i-ten kanonischen Faktors, d.h. Summe der quadrierten Koeffizienten in der i-ten Spalte von S_1.

m_1 = Zahl der Variablen in der Variablengruppe 1.

E_i = i-ter Eigenwert, bzw. i-te quadrierte kanonische Korrelation.

Unsere Ausführungen gelten "spiegelbildlich" auch für erklärte Varianz und Redundanz der 2. Variablengruppe.

P29.1.3 Die kanonischen Faktorwerte

P29.1.3.1 Kalkül

Die kanonischen Faktorwertvariablen werden gemäß folgender Gleichung berechnet:

$$X_i = \alpha_{1i} \cdot x_1 + \alpha_{2i} \cdot x_2 + \ldots + \alpha_{m1i} \cdot x_{m1}$$

$$Y_i = \beta_{1i} \cdot y_1 + \beta_{2i} \cdot y_2 + \ldots \beta_{m2i} \cdot y_{m2}$$

X_i = Faktorwertvariable für den kanonischen Faktor i (1.Variablengruppe)

Y_i = Faktorwertvariable für den kanonischen Faktor i (2. Variablengruppe)

$x_1, x_2, \ldots, x_{m1}$ = Variable der 1. Variablengruppe

$y_1, y_2, \ldots y_{m2}$ = Variable der 2. Variablengruppe

m_1 = Zahl der Variablen der 1. Gruppe

m_2 = Zahl der Variablen der 2. Gruppe

α_{1i} = unstandardisierte kanonische Gewichtszahl für die Variable x_1 aus der 1. Variablengruppe hinsichtlich des kanonischen Faktors i

β_{1i} = entsprechend α_{1i} für Variable y_1.

Wird die Kovarianzmatrix analysiert, dann sind x_1, ... xm_1, y_1, ... ym_2 als Abweichungen von ihrem jeweiligen Mittelwert gemessen.

Wird die Korrelationsmatrix analysiert, dann sind x_1, ... x_{m1}, y_1, ... y_{m2} standardisiert. In diesem Fall sind dann auch die unstandardisierten und die standardisierten kanonischen Gewichtszahlen gleich.

P29.1.3.2 Eingabe in Almo-Syntax-Programm

Es ist sehr ungewöhnlich, dass im Rahmen einer kanonischen Korrelationsanalyse Faktorwertvariable je Untersuchungsobjekt ermittelt werden. Wir haben deswegen für diesen Fall auch keine Programm-Maske entwickelt. Es soll der Rechengang aber trotzdem vorgeführt werden.
Wir wollen die Eingabe in Almo an einem Datenbeispiel aus Hartung/Elpelt (1989, S.178) vorführen. Das nachfolgende „selbst geschriebene“ Almo-Syntax-Programm ist als Beispielprogramm unter dem Namen „Hartu178.Alm“ in Almo enthalten. Sie finden das Programm im Menü „Almo/Liste aller Almo-Programme“.

Im 1. Anfang-Ende-Block des nachfolgenden Almo-Programms wird die Kovarianzmatrix für die beiden Variablengruppen gebildet und dann dem Kalkül der kanonischen Korrelation unterworfen.

Im 2. Anfang-Ende-Block werden dann die kanonischen Faktorwertvariable gebildet.

Im 3. Anfang-Ende-Block werden dann die kanonischen Faktorwertvariable mit Programm 19 interkorreliert.

```
#

        ┌──────────────────────────────────────────────────────────┐
        │                      Hartu178.Alm                        │
        │                  Kanonische Korrelation                  │
        │                                                          │
        │ Bei 15 Frauen wird die kanonische Korrelation zwischen   │
        │                                                          │
        │     1. Variablengruppe: V1 Hämoglobingehalt im Blut      │
        │                         V2 Oberfläche der Erythrozyten   │
        │        und                                               │
        │     2. Variablengruppe: V3 Blutdruck                     │
        │                         V4 Alter                         │
        │ ermittelt                                                │
        │                                                          │
        │ Beispiel aus Hartung/Elpelt: Multivariate Statistik      │
        │                              1989, S. 178                │
        └──────────────────────────────────────────────────────────┘
                                                                    #

VEREINBARE
  Variable=[20];             # Speicher fuer 20 Variable vereinbaren        #
                             # Anfang des eigentlichen Almo-Programms       #
ANFANG

Name1=HbGehalt;              # Den Variablen werden Namen gegeben           #
Name2=Oberflaeche;
Name3=Blutdruck;
Name4=Alter;

PROGRAMM = 29;               # Kanonische Korrelation hat die Programm-Nr.29#

  u_quantitative_V = V1,2;   # erste  (unabhaengige) Variablengruppe        #
  a_quantitative_V = V3,4;   # zweite (abhaengige)   Variablengruppe        #

  Matrix      = Kovarianz;
  Kov_Nenner  = -1;          # die Kovarianzmatrix wird (bei Hartung/Elpelt)#
```

```
                              # mit n-1 dividiert                           #

  SA_Nenner   = -1;           # die Standardabwg. wird (bei Hartung/Elpelt)  #
                              # mit n-1 dividiert                           #

  Faktoren    = ;             # Zahl der Faktoren eventuell beschraenken     #

  Zwischergeb = 0;            # 1= Zwischenergebnisse ausgeben  0= nicht     #
ENDE_PROGRAMM_PARAMETER       # Ende des Blocks der Programmparameter        #


Lese V1:4;                    # Lese Datensatz hinter dem Wort ENDE          #

Schreibe V1:4                 # Schreibe die einzelnen Datensätze            #
  in Zwischendatei            # für die nachfolgende Berechnung der          #
  Format frei;                # kanonischen Faktorwert-Variable in eine      #
                              # Zwischendatei                                #

Gehe_in_Programm              # Gehe mit eingelesenen Daten in Programm      #
Gehezu Lese                   # Zurueck und naechsten Datensatz lesen        #

                              # BEACHTE:                                     #
                              # Die Matrix der (unstandardisierten und nicht #
                              # normalisierten) kanonischen Gewichtszahlen   #
                              # wird von Almo in die Datei 21 gespeichert.   #
                              # Dies ist eine interne Zwischendatei          #
ENDE
13.6  92 123 36
15.4 103 137 57
17.2 104 139 61
12.7  95 127 42
13.9  87 125 46
14.5  95 120 31
17.6 108 132 49
15.2 105 118 27
13.8  84 125 35
15.0 102 140 58
14.7  97 142 63
15.5  96 126 44
13.9  93 131 47
14.2  95 118 32
15.3 102 112 25
*
#------ Beachte: Der Stern hinter den Daten ist obligatorisch            ---#
#------                    2. ANFANG-ENDE-Block                       ------#
#------ kanonische Faktorwert-Berechnung durch Programm 27            -- ---#
ANFANG

PROGRAMM=27;                  # kanon.Faktorwert-Berechnung                  #

                              # BEACHTE:                                     #
                              # Die Matrix der (unstandardisierten und nicht #
                              # normalisierten) kanonischen Gewichtszahlen   #
                              # wird von Almo aus der internen Datei 21 ge-  #
                              # lesen, in die sie in obigem 1.Programm-Block #
                              # geschrieben wurde                            #

  u_quantitative_V = V1,2;    # erste (unabhaengige) Variablengruppe         #
  a_quantitative_V = V3,4;    # zweite(abhaengige)   Variablengruppe         #

  u_Faktorwert_Variable = V5,6; # kanon.Faktorwert-Variable für unabhängige #
                                # quantitative Variable                     #

  a_Faktorwert_Variable = V7,8; # kanon.Faktorwert-Variable für abhängige   #
                                # quantitative Variable                     #

  Matrix   = Kovarianz;

  Option 1 = 50;              # Maximal 50% der unabhängige bzw. abhängigen  #
```

```
                              #quantitativen Variablen, also 1 von den       #
                              # 2 Variablen (V1,2 bzw. V3,4) duerfen Kein_   #
                              # Wert besitzen. Fuer sie wird der Mittelwert  #
                              # eingesetzt. Sonst werden die Faktorwert-     #
                              # Variable V5,6 bzw. V7,8 auf Kein_Wert gesetzt#

                              # Berechne und zeige:                          #
  Zeige=kanon_Faktorwert;     # kanonische Faktorwert-Variable               #

  Zwischergeb = 0;            # 1=   einige Zwischenergebnisse ausgeben      #
                              # 0=   nicht ausgeben                          #

ENDE_PROGRAMM_PARAMETER

Lese V1:4                     # Lese die im 1. Anfang-Ende-Block zwischen-   #
  aus Zwischendatei           # gespeicherten Daten                          #
  Format frei
  leerzu Ende;

Gehe_in_Programm              # gehe mit eingelesenen Daten in Programm 27   #

Schreibe V1:8                 # Schreibe den um die Faktorwert-Variable V5:8 #
  in Datei 2                  # verlaengerten Datensatz in neue Datei        #
  "C:\Almo6\Progs\KanFakw.fre"
  Format frei;

Gehezu Lese                   # zurueck und naechsten Datensatz verarbeiten  #

ENDE


#----- 3. ANFANG-ENDE-Block: kanonische Faktorwert-Variable korrelieren ----#
ANFANG

Name 5=kanFak11;              # den kanonischen Faktorwert-Variablen werden  #
Name 6=kanFak12;              # Namen gegeben                                #

Name 7=kanFak21;
Name 8=kanFak22;


PROGRAMM=19;                  # Korrelationsprogramm                         #
  quantitative_V = V1,2,      # = die unabh. quantitaiven Variablen          #
                   3,4,       # = die abh.   quantitaiven Variablen          #
                   5:8;       # = die kanonischen Faktorwert-Variablen       #
ENDE_PROGRAMM_PARAMETER


Lese V1:8                     # Lese Datensatz, der in 2.Anfang-Ende-Block   #
  aus Datei 1                 # gespeichert wurde. V1:4 sind die unabhäng.   #
  "C:\Almo6\Progs\KanFakw.fre"    # und abh. quantit. Variablen. V6 bis V8  #
  Format frei                     # sind die kanonischen Faktorwert-        #
  leerzu Ende;                    # Variablen.                              #

Gehe_in_Programm              # gehe mit eingelesenen Daten in Programm 19   #

Gehezu Lese                   # zurueck und naechsten Datensatz verarbeiten  #

ENDE
```

Von den vielen Ergebnissen, die dieses Programm ausgibt, sollen hier folgende ausgewählt werden:

1. Im 1. Block werden für die beiden Variablengruppen folgende kanonische Korrelationen errechnet:
   1. kanonische Korrelation: 0.3715

2. kanonische Korrelation: 0.1519

2. Im 2. Block werden die kanonischen Faktorwerte ausgegeben. Almo liefert folgenden Output

Als Beispiel wird die kanonische Faktorwert-Berechnung für den 1. Datensatz gezeigt.

Wurde die Korrelationsmatrix analysiert, dann wird jede einzelne Variable standardisiert und mit dem kanonischen Faktorwert-Koeffizienten multipliziert.

Die Formel ist folgende:

```
(Variablenwert - Mittelwert) * kanonFakwKoeff / Standabwg
```

Wurde die Kovarianzmatrix analysiert, dann wird in obiger Formel Standabwg = 1 gesetzt.

```
V1                              (13.6 - 14.8333) * 1.01805 / 1
V2                            + (92 - 97.2) * -0.0732857 / 1
kanonische Faktorwert-Variable V5 = -0.874475


V1                              (13.6 - 14.8333) * -0.620049 / 1
V2                            + (92 - 97.2) * 0.217486 / 1
kanonische Faktorwert-Variable V6 = -0.366221


V3                            + (123 - 127.667) * -0.214266 / 1
V4                            + (36 - 43.5333) * 0.223848 / 1
kanonische Faktorwert-Variable V7 = -0.686335


V3                            + (123 - 127.667) * 0.44893 / 1
V4                            + (36 - 43.5333) * -0.282173 / 1
kanonische Faktorwert-Variable V8 = 0.0305376
```

Datensatz	kanonische Faktorwert-Variable V5	V6	V7	V8
1	-0.874	-0.366	-0.686	0.031
2	0.152	0.910	1.015	0.390
3	1.911	0.011	1.482	0.159
4	-2.011	0.844	-0.200	0.133
5	-0.203	-1.640	1.124	-1.893
6	-0.178	-0.272	-1.163	0.095
7	2.025	0.633	0.295	0.403
8	-0.198	1.469	-1.630	0.325
9	-0.085	-2.230	-1.339	1.211
10	-0.182	0.941	0.596	1.455
11	-0.121	0.039	1.287	0.942
12	0.767	-0.674	0.462	-0.880
13	-0.642	-0.335	0.062	0.518
14	-0.484	-0.086	-0.510	-1.085
15	0.123	0.755	-0.792	-1.804

3. Im 3. Block werden u.a. die Interkorrelationen zwischen den Faktorwertvariablen ausgegeben.

	kanFak11	kanFak12	kanFak21	kanFak22
kanFak11	1.0000			
kanFak12	0.0000	1.0000		

kanFak21	0.3715	0.0000	1.0000	
kanFak22	0.0000	0.1519	0.0000	1.0000

kanFak11 = Faktorwertvariable des 1. Faktors aus Variablengruppe 1
kanFak12 = Faktorwertvariable des 2. Faktors aus Variablengruppe 1
kanFak21 = Faktorwertvariable des 1. Faktors aus Variablengruppe 2
kanFak22 = Faktorwertvariable des 2. Faktors aus Variablengruppe 2

Die 1. kanonische Korrelation zwischen den beiden Variablengruppen war 0.3715 und die 2. war 0.1519. Siehe oben Punkt 1. Wir sehen, daß die beiden Faktorenwertvariable kannFak11 mit kanFak21 und kanFak12 mit kanFak22 mit diesen Werten korrelieren.

P29.1.4 Kanonische Korrelation und Regressionsanalyse

Wir wollen für die Korrelationsmatrix in P29.1.2, Satz 3 nochmals eine kanonische Korrelation rechnen - wobei wir diese Mal die 2. Variablengruppe als nur aus y_1 bestehend betrachten. Die Korrelationsmatrix ist also folgende

	Variablengruppe 1			Variablengruppe 2
	x1	x2	x3	y1
x1	1.0000	0.4000	0.5000	0.3000
x2	0.4000	1.0000	0.6000	0.2000
x3	0.5000	0.6000	1.0000	0.4000
y1	0.3000	0.2000	0.4000	1.0000

Wir erhalten folgende Ergebnisse (gekürzt):

Faktor	Kanonische Korrelation	Eigenwert	Wilks' Lambda	Chi-Quadrat	df	Signifikanz (1-p)*100
1	0.42175	0.17787	0.82213	18.90045	3	99.94571 %
Summe (=Pillais Spur)		0.17787				

Kanonische Gewichtszahlen fuer 1.(unabhaengige) Variablengruppe
(unstandardisiert, nicht normalisiert)

x1	V1	0.3430
x2	V2	-0.2018
x3	V3	0.8980

Prozent erklaerte (standardisierte) Varianz
in der 2.(abhaengigen) Variablengruppe

durch kanonische Faktoren der anderen Variablengruppe erklaert
Faktor 1 17.7872 %

Nun rechnen wir mit Programm 20 eine Regressionsanalyse mit y_1 als abhängiger und x_1, x_2, x_3 als unabhängiger Variablen. Dazu verwenden wir die Programm-Maske Prog20mn. Der Benutzer findet sie durch Klick auf den Knopf "Verfahren/ Regressonsanalyse". Sehr viel kürzer ist das folgende „selbst geschriebene“ Syntax-

Programm:

```
Vereinbare
Variable = 10;
Anfang
Programm = 20;
Uquantitative_V = V1:3;
Aquantitative_V = V4;
Ende_Programmparameter;
Lese Matrix aus Eingabe;
GEHE_IN_PROGRAMM
Ende
*
100
1.0
0.4   1.0
0.5   0.6   1.0
0.3   0.2   0.4   1.0
*
*
```

Zur Eingabe einer Matrix siehe Handbuch, Teil 2, Abschnitt 43.1.1 Almo liefert folgende Ergebnisse (gekürzt)

(a) Multipler Korrelationskoeff.: 0.421749
Er stimmt mit der kanonischen Korrelation aus der kanonischen Korrelationsanalyse überein

(b) Signifikanz (1-p) ·100= 99.949072%
Sie stimmt bis zur 2. Kommastelle überein. Die kleine Differenz entstand nur aus der andersartigen approximativen Berechnung der Signifikanz aus F- und Chi-Quadrat-Wert.

(c) Regressionskoeffizienten

```
V1        0.1447
V2       -0.0851
V3        0.3787
```

Werden die kanonischen Gewichtszahlen mit der kanonischen Korrelation multipliziert, dann stimmen sie mit den Regressionskoeffizienten überein.

(d) Erklärte Streuung: 0.177872
Stimmt voll überein mit dem Ergebnis aus der kanonischen Korrelationsanalyse.

Hinweis: Wenn Sie eine kanonische Korrelationsanalyse mit einer abhängigen Variablen rechnen, dann können Sie die einzelne Variable auch als 1. Variablengruppe angeben und die unabhängigen Variablen als 2. Variablengruppe. Tun Sie das, wenn für die einzelne Variable eine negative kanonische Gewichtszahl von -1.0 angegeben wird.

P29.2 Diskriminanzanalyse und Klassifikation

Bei der "Diskriminanzanalyse als kanonische Korrelationsanalyse" besteht die 1. Variablengruppe aus den unabhängigen quantitativen Variablen und die 2. Variablengruppe aus den 0-1 kodierten Dummies der abhängigen nominalen Variablen (siehe Handbuch zu P20, Abschnitt P20.3). Daß die klassische Fisher'sche Diskriminanzanalyse und die kanonische Korrelationsanalyse äquivalent sind, wurde schon 1953 von Tatsuoka nachgewiesen (siehe dazu Tatsuoka, 1971, S.

177ff). Im Almo berechnen wir alle Koeffizienten der kanonischen Korrelationsanalyse - so wie im vorausgegangenen Abschnitt P29.1.2 dargestellt.

P29.2.1 Eingabe in Programm-Maske Prog29m3

```
Prog29m3.Msk
Kanonische Diskriminanzanalyse

als kanonische Korrelation zwischen den unabhängigen
quantitativen Variablen und den Dummies der abhängigen
nominalen Variablen

Beispiel:
Es soll die Wahl des Studiums der Physik,Soziologie,Rechts-
wiss. erklärt werden durch die Schulnoten in Mathe,Deutsch,
Englisch,Geschichte

Was ist ein Kurzprogramm ? --> [Hilfe]
Bedienung --> [Hilfe]
```

1
```
Speicher fuer x Variable [Hilfe]
Vereinbare Variable= 20 ;
```

2
```
[↓] Option: Weitere Vereinbarungen - nur wenn Almo dazu auffordert
```

3
```
Datei der Variablennamen [Hilfe]
[↔][💾] "C:\Almo7\Testdat\Varnamen.nam"
[↔][↓] zeige        zeige = Namensdatei in Output zeigen
                    leer  = nicht
```

4
```
Freie Namensfelder [Hilfe]
                                  abhängige nominale Variable
[↔] Name3=Studium:Physik,Soziologie,Recht;

[↔] Name5=MatheNote;              unabh. quantitative Variable
[↔] Name6=DeutschNote;
[↔] Name7=EnglischNote;
[↔] Name8=GeschichteNote;

[...] erzeuge zusätzliche Namensfelder
```

5
```
Datei aus der gelesen wird [Hilfe]
                                  bei Datei-Problemen
[💾] "C:\Almo7\Testdat\TESTDAT.FRE"
[▭] frei            Format der Daten [Hilfe]
[↔][▭] V1:20        der Datensatz enthält diese Variablen
                    Bei Format DIREKT schreiben Sie: alle_V
```

6
```
[↓] Wenn Dateiformat FIX oder Nicht-Standard-FREI [Hilfe]
```

7
```
Analyse-Variable

unabhängige quantitative Variable

[↔][□□] MatheNote,DeutschNote,EnglischNote,GeschichteNote

abhängige nominale Variable
(nur eine erlaubt) [Hilfe]

[↔][□□] Studium
```

8
```
[↓] Option: Ein- und Ausschliessen von Untersuchungseinheiten
```

9 Option: Umkodierungen und Kein-Wert-Angaben

10 Option: Spezielle Kein-Wert-Behandlung

11 Option: Untersuchungseinheiten gewichten

12 verschiedene Programm-Optionen

13 Option: Diskriminanzkoeffizienten speichern

14 Option: "Aussehen" der auszugebenden Tabelle bzw. Matrix

15 Grafik-Optionen

16 Basisstatistiken ausgeben

0

1= Basisstatistiken ausgeben
0= nicht

P29.2.2 Erläuterungen zu den Boxen

Box 1: Vereinbare Variable
Siehe Dokument Nr. 0 "Arbeiten mit Almo", Abschnitt P0.1.

Box 2: Option: Weitere Vereinbarungen - nur wenn Almo dazu auffordert
Siehe P0.2.

Box 3: Datei der Variablennamen
Siehe P0.3.

Box 4: Freie Namensfelder
Siehe P0.3.

Box 5: Datei aus der gelesen wird
Siehe P0.4.

Box 6: Wenn Dateiformat FIX oder nicht Standard-FREI
Siehe P0.4.

Box 7: Analyse-Variable

Analyse-Variable

unabhängige quantitative Variable

MatheNote,DeutschNote,EnglischNote,GeschichteNote

abhängige nominale Variable
(nur eine erlaubt)

Hilfe

Studium

Eingabefeld 1: Geben Sie die unabhängigen Variablen an. Sie müssen quantitativ sein

Eingabefeld 2: Geben Sie die abhängige Variablen an. Sie muß nominal sein. Es ist nur eine erlaubt.

Box 8: Option: Ein- und Ausschliessen von Untersuchungseinheiten
Siehe P0.7.

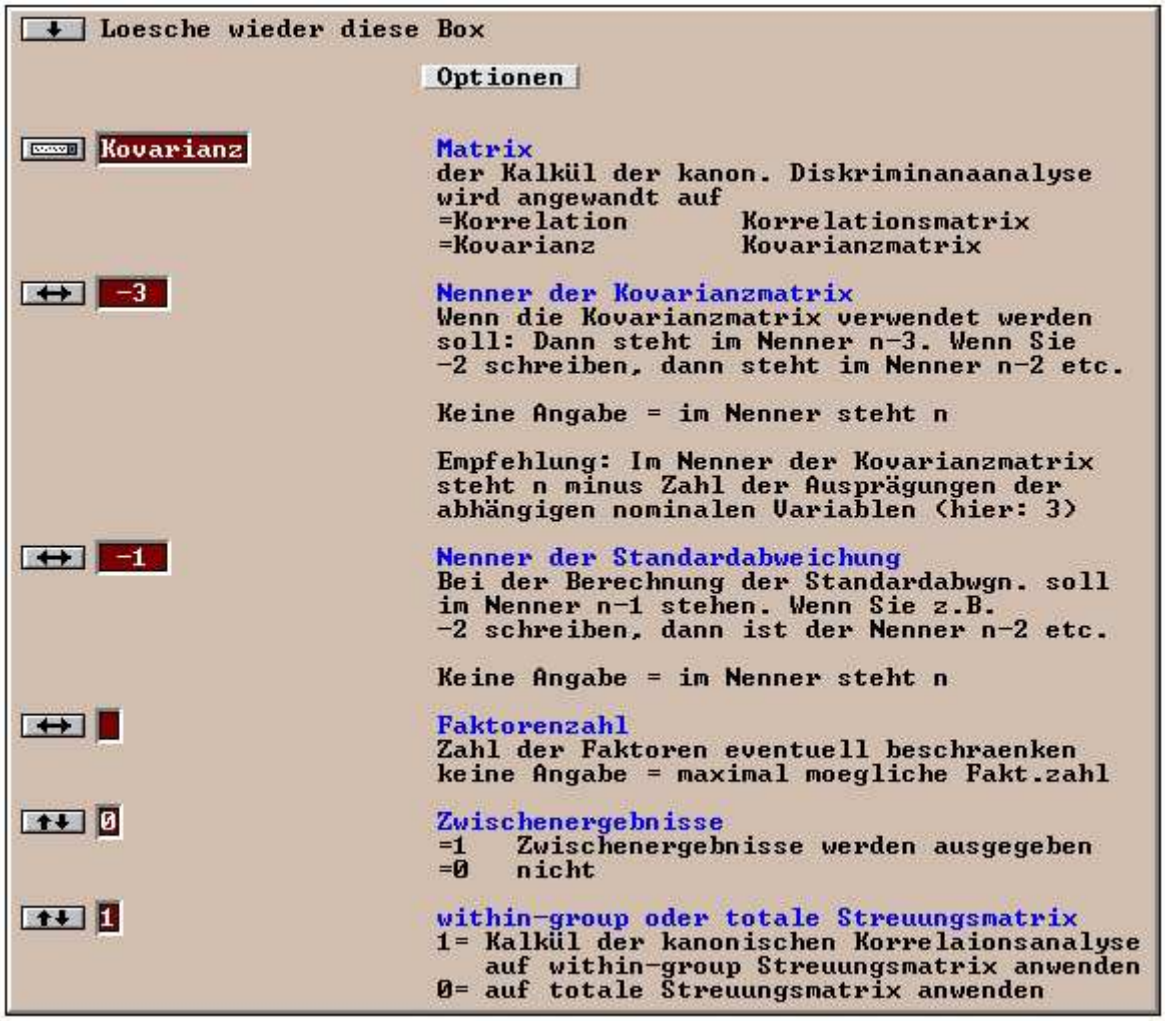
Box 9: Kein_Wert-Angabe und Umkodierungen
Siehe P0.5.

Box 10: Option: Spezielle Kein-Wert-Behandlung
Besitzt eine oder mehrere Analysevariablen keinen Wert, dann verwendet Almo standardmäßig das "paarweise Ausscheiden". Der Benutzer hat die Möglichkeit eine von 7 Methoden zur Kein-Wert-Behandlung zu wählen. Dazu muß die Optionsbox geöffnet werden. Wir haben diese Optionsbox bei der kanonischen Korrelation in Abschnitt P29.1.1.2 bereits dargestellt und erläutert.

Box 11: Option: Untersuchungseinheiten gewichten
Siehe P0.8.

Box 12: Verschiedene Programm-Optionen

Optionsbox geöffnet:



Eingabefeld 1: Der Kalkül der kanonischen Diskriminanzanalyse kann auf die Korrelations- oder Kovarianz-Matrix angewendet werden. Normalerweise wird man die Kovarianzmatrix verwenden.

Eingabefeld 2: Nenner der Kovarianzmatrix
Nur sinnvoll, wenn die Kovarianzmatrix verwendet werdens soll. Wenn Sie beispielsweise -3 eintragen, dann steht bei der Berechnung der Kovarianzmatrix im Nenner n-3. Wenn Sie -2 schreiben, dann steht im Nenner n-2 etc. Keine Angabe: Im Nenner steht n oder es wurde die Korrelationsmatrix verwendet. Empfehlung: Im Nenner der Kovarianzmatrix steht n minus Zahl der Ausprägungen der abhängigen nominalen Variablen. In unserem Beispiel steht deswegen -3 im Eingabefeld

Eingabefeld 3: Nenner der Standardabweichung
Wenn Sie beispielsweise -1 eintragen, dann steht bei der Berechnung der Standardabweichung im Nenner n-1. Wenn Sie -2 schreiben, dann steht im Nenner n-2 etc. Keine Angabe: Im Nenner steht n.

Eingabefeld 4: Es können soviele Faktoren extrahiert werden, wie die kleinere der beiden Variablengruppen Variable umfasst. Der Benutzer kann aber diese Faktorenzahl einschränken.

Eingabefeld 5: Es können Zwischenergebnisse angefordert werden.

Eingabefeld 6: within-group oder totale Streuungsmatrix
1= Der Kalkül der kanonischen Korrelaionsanalyse wird auf die within-group Streuungsmatrix angewendet.
0= Er wird auf die totale Streuungsmatrix angewendet.
Empfehlung: 1

Eingabefeld 7: kanonische Gewichtszahlen
1= kanonische Gewichtszahlen nur für unabhängige Variablengruppe ausgeben
0= auch für abhängige Variablengruppe ausgeben
Empfehlung: 1. Die Eingabe "0" ist bei Diskriminanzanalyse nicht sinnvoll.

Box 13: Diskriminanzkoeffizienten speichern

Optionsbox geöffnet:

Die Matrix der errechneten kanonischen Diskriminanzkoeffizienten kann gespeichert werden. Sie kann dann im Klassifikationsprogramm Prog 27 eingelesen werden. Siehe dazu Abschnitt P27.1.1, Erläuterung zur Box 7. Geben Sie den vollen Pfad- und Dateinamen an.

Box 14: Option: "Aussehen" der auszugebenden Tabelle bzw. Matrix
Siehe P0.9.

Box 15: Grafik-Optionen
Siehe P0.10.

Box 16: Basisstatistiken ausgeben

Basisstatistiken ausgeben

0

1= Basisstatistiken ausgeben
0= nicht

Es können zusätzlich Basisstatitiken ausgegeben werden. Dies sind u.a.

Mittelwerte
Standardabweichungen
Zahl der diversen Werte je Variable
Zahl der fehlenden Werte je Variable

P29.2.4 Ausgabe

Wir wollen die Ergebnisse aus der Diskriminanzanalyse an einem sehr einfachen Beispiel darstellen und erläutern. Dazu verwenden wir Daten von Tatsuoka (1971, S.180). Die zuvor abgebildete Programm-Maske Prog29m3 wird entsprechend ausgefüllt. Die so ausgefüllte Maske ist zu finden unter dem Menü "Almo/Liste aller Almo-Programme/Tatsuoka2.Alm". Das Programm liegt auch als Syntax-Programm vor. Zu finden unter "Almo/Liste aller Almo-Programme/Tatsuok1.Alm".

Vom Maskenprogramm Tatsuoka2.Alm zeigen wir hier nur die relevanten Dialogboxen:

Datei aus der gelesen wird

Hilfe
bei Datei-Problemen

"C:\Almo15\TESTDAT\Tatsuoka.fre"

frei

Format der Daten

Hilfe

V1:3

der Datensatz enthält diese Variablen
Bei Format DIREKT schreiben Sie: alle_V

Analyse-Variable

unabhängige quantitative Variable

V1,2

abhängige nominale Variable
(nur eine erlaubt)

Hilfe

V3

Die Daten "Tatsuoka.fre" sind folgende (gekürzt):

```
V1 V2 V3
--- --- --

20  6 1
21 10 1
15 12 1
15  8 1
11 11 1
24 17 1
18 13 1
14  4 1
17 12 2
11 11 2
15 14 2
20 16 2
14 16 2
 .  . .
 .  . .
```

V1 und V2 sind die unabhängigen quantitativen Variablen. V3 ist die eine abhängige nominale Variable.

Die Box "verschieden Programm-Optionen" wird geöffnet und so ausgefüllt:

```
X  Loesche wieder diese Box (dann Voreinstellungen wieder gueltig)

                          Optionen

Kovarianz          Matrix
                   der Kalkül der kanon. Diskriminanaanalyse
                   wird angewandt auf
                   =Korrelation  Korrelationsmatrix
                   =Kovarianz    Kovarianzmatrix (Voreinstellung)

-3                 Nenner der Kovarianzmatrix
                   Wird beispielsweise in das Eingabefeld  -3
                   geschrieben, dann steht im Nenner      n-3

                   Eingabefeld leer = im Nenner steht
                   n minus Zahl der Ausprägungen der abhängigen
                   nominalen Variablen (empfohlen).
                   Dies ist auch die Voreinstellung,
                   wenn die Optionsbox nicht geöffnet wird

-1                 Nenner der Standardabweichung
                   Bei der Berechnung der Standardabwgn. soll
                   im Nenner n-1 stehen. Wenn Sie z.B.
                   -2 schreiben, dann ist der Nenner n-2 etc.

                   Eingabefeld leer = im Nenner steht n
                   Dies ist auch die Voreinstellung,
                   wenn die Optionsbox nicht geöffnet wird

1                  Zwischenergebnisse
                   =1  Zwischenergebnisse werden ausgegeben
                   =0  nicht (Voreinstellung)

1                  within-group oder totale Streuungsmatrix
                   =1  Kalkül der kanonischen Diskriminanzanalyse
                       auf within-group Streuungsmatrix anwenden
                       (Voreinstellung)
                   =0  auf totale Streuungsmatrix anwenden
```

Gemäß der Anweisung "Matrix=Kovarianz;" ermittelt Almo aus den eingelesenen Daten die Kovarianzmatrix. Da wir `Kov_Nenner=-3;` gesetzt haben, wurde bei der Berechnung der Kovarianzmatrix im Nenner n-3 verwendet (da die abhängige nominale Variable 3 Ausprägungen besitzt). Es mag übrigens durchaus sinnvoll sein, die Kovarianzmatrix für die kanonische Korrelationsanalyse mit n im Nenner zu rechnen.

Almo liefert folgende Kovarianzmatrix

Kovarianz-Matrix
(Varianz/Kovarianz ist mit n-3 dividiert)

	V1	V2	V3-1	V3-2	V3-3
V1	28.474074	6.170370	1.200000	0.481481	-1.681481
V2	6.170370	17.476543	-0.279012	1.123457	-0.844444
V3-1	1.200000	-0.279012	0.217284	-0.098765	-0.118519
V3-2	0.481481	1.123457	-0.098765	0.246914	-0.148148
V3-3	-1.681481	-0.844444	-0.118519	-0.148148	0.266667

Da das in der Optionsbox verlangt wurde, rechnet Almo die "within-groups"-Streuungsmatrix $\mathbf{R^*_{11}}$ der Submatrix $\mathbf{R_{11}}$ der unabhängigen Variablen gemäß folgender Formel:

$$\mathbf{R^*_{11} = R_{11} - R_{12} \,.\, R_{22}^{-1} \,.\, R_{21}}$$

zur Bezeichnung der Submatrizen siehe Graphik in P29.1.2.

$\mathbf{R_{22}^{-1}}$ ist die Inverse von $\mathbf{R_{22}}$
$\mathbf{R^*_{11}}$ ist

	V1	V2
V1	16.63	2.65
V2	2.65	12.20

Der Benutzer muß entscheiden, ob der Kalkül der kanonischen Diskriminanzanalyse auf die "within-groups"-Matrix R^*_{11} oder die Gesamtstreuungsmatrix R_{11} angewendet werden soll. In der statistischen Literatur wird die within-groups-Matrix präferiert.

Almo liefert nun folgende weiteren Ergebnisse und Zwischenergebnisse (die in der Optionsbox angefordert wurden).

Faktor	Kanonische Korrelation	Eigenwert	Wilks'Lambda	Chi-Quadrat	df	Signifikanz (1-p)*100
1	0.65454	0.42842	0.42622	22.59897	4	99.96589 %
2	0.50428	0.25430	0.74570	7.77599	1	99.47306 %

Die Eigenwerte, die bei der klassischen Fisher'schen Diskriminanzanalyse ermittelt werden, sind nicht mit den oben angegebenen Eigenwerten aus der kanonischen Korrelationsanalyse identisch. Sie lassen sich jedoch gemäß folgender Formel leicht aus ihnen ableiten (Tatsuoka, 1971, S.179)

$$D_i = \frac{E_i}{1 - E_i}$$

D_i = i-ter Eigenwert aus klassischer Fisher'scher Diskriminanzanalyse
E_i = i-ter Eigenwert aus kanonischer Diskriminanzanalyse

Almo gibt für D folgende Werte aus:

0.7495 0.3410

Die unstandardisierten kanonischen Gewichtszahlen müssen mit einer Konstanten je Faktor multipliziert werden, um die unstandardisierten Diskriminanzkoeffizienten der klassischen Fisher'schen Diskriminanzanalyse zu erhalten. Diese Konstanten ergeben sich gemäß der Formel

$$g_i = \sqrt{\frac{D_i}{E_i}}$$

g_i = Multiplikationskonstante für Faktor i
D_i, E_i = siehe oben

Almo liefert folgende Werte für g_i

1.3227 1.15803

Mit diesen Zahlenwerten sind die im folgenden von Almo ausgegebenen verschie-

denen kanonischen Gewichtszahlen bereits multipliziert.

Die unstandardisierten Diskriminanzkoeffizienten sind dann folgende:

```
Kanonische Gewichtszahlen fuer 1.(unabhaengige) Variablengruppe
(unstandardisiert, nicht normalisiert)
```

	Faktor 1	Faktor 2
V1	0.219936	-0.117989
V2	0.087659	0.277849

Exkurs: Vergleich mit anderen Statistikprogrammen:
Das Problem der Vorzeichen-Umkehr

Wenn Sie dieselben Daten mit einem anderen Statistikprogramm, z.B. mit SAS (Prozedur Candisc) rechnen, dann kann eventuell der Fall auftreten, daß im Vergleich zu den Almo-Ergebnissen, in einer Spalte der obigen Matrix das Vorzeichen umgedreht ist. Es ist prinzipiell möglich Spalte i in obiger Matrix umzudrehen. Wir können beispielsweise die Vorzeichen in Spalte 2 (=Faktor 2) umdrehen. Wir würden dann erhalten

	Faktor 1	Faktor 2
V1	0.219...	0.117...
V2	0.087...	-0.277...

Wir könnten auch noch zusätzlich die 1. Spalte umdrehen.
Das Vorzeichen muß aber dann auch in Spalte i in den folgenden Matrizen umgedreht werden:

(1) Gruppen-Zentroide
(2) Standardisierte kanonische Gewichtszahlen
(3) Kanonische Strukturkoeffizienten

Der Almo-Benutzer muß diese Vorzeichen-Umkehr von Hand vornehmen. Die Möglichkeit sie über eine Option vorzunehmen existiert nicht. Die Vorzeichen-Umkehr wirkt sich in folgender Weise aus: Das Vorzeichen der kanonischen Faktorwert-Variable (der Diskriminanzwerte) wird dadurch umgedreht. Auf die Bestimmung der Wahrscheinlichkeiten der Gruppenzugehörigkeiten in der Klassifikation hat das jedoch keine Auswirkungen. Siehe dazu die nachfolgenden Abschnitte P29.2.6.

Im Verlauf des Kalküls hat Almo auch die Mittelwerte der quantitativen unabhängigen Variablen je Ausprägung der abhängigen Variablen ermittelt.

```
Mittelwerte fuer 1.(unabhaengige) Variablengruppe
je Auspraegung der abhaengigen nominalen Variablen
```

	V1	V2
V3-1	17.2500	10.1250
V3-2	14.5000	14.1000
V3-3	9.4167	9.1667

Der Wert 17.25 im linken oberen Eck bedeutet z.B., daß alle Untersuchungseinheiten mit der Ausprägung 1 in der nominalen Variablen V3 in der quantitativen Vari-

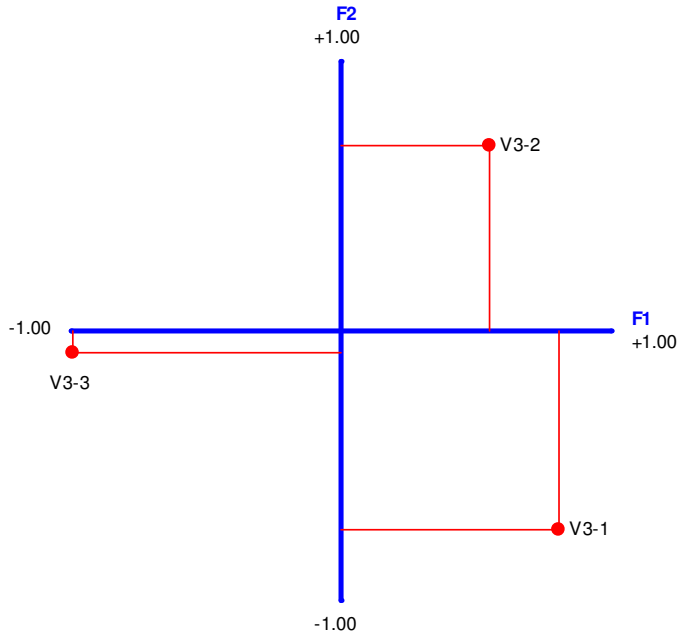
ablen V1 einen Mittelwert von 17.25 besitzen.

Nun ist es auch möglich, im orthogonalen Koordinatensystem der kanonischen (Diskriminanz-) Faktoren den Mittelpunkt der Gruppe 1, 2 und 3 (also der Untersuchungseinheiten in Ausprägung 1, 2 und 3) einzutragen. Wir sprechen hier von den "Gruppen-Zentroiden im kanonischen Raum". Almo liefert folgende Gruppen-Zentroide und folgende Grafik.

```
"Gruppen-Zentroide"
Lage der Auspraegungen der abhaengigen nominalen Variablen
im (durch die orthogonalen Faktoren aufgespannten) kanonischen Raum
```

	Faktor 1	Faktor 2
V3-1	0.8082	-0.7395
V3-2	0.5518	0.6894
V3-3	-0.9986	-0.0815

Gruppen-Zentroide
Diskriminanzanalyse



Die Gruppen-Zentroide werden gemäß folgender Gleichung bestimmt (siehe auch Cooley/Lohnes, 1971, S.250):

$$\mathbf{Z} = (\mathbf{M} - \mathbf{M_G})\ \mathbf{B}$$

M = $m_2 * m_1$ - Matrix der Ausprägungsmittelwerte (siehe oben)
m_1 = Zahl der unabhängigen quantitativen Variablen
m_2 = Zahl der Ausprägungen der abh. nominalen Variablen (=Zahl der Gruppen)
B = m_1*f-Matrix der unstandardisierten Diskriminanzkoeffizienten (kanonische Gewichtszahlen) - siehe oben
$\mathbf{M_G}$ = $m_2 * m_1$ - Matrix des Gesamtmittelwertes der unabhängigen quantitativen Variablen. In der 1. Spalte steht über alle Zeilen hinweg der Gesamtmittelwert von V1, in der 2. Spalte steht über alle Zeilen hinweg der Gesamtmittelwert von V2.
Z = $m_2 * m_1$ - Matrix der Gruppen-Zentroide.

Almo berechnet dann noch die standardisierten Diskriminanzkoeffizienten und die Strukturkoeffizienten.

Standardisierte kanonische Gewichtszahlen der 1. (unabhaengigen) Variablengruppe (=kanon.Gew.zahl * Wurzel aus Diagonalglied der Streuungsmatrix)
BEACHTE: Errechnet aus "within-groups"-Matrix

```
Standardisierte kanonische Gewichtszahlen
der 1.(unabhaengigen) Variablengruppe
(=kanon.Gew.zahl * Wurzel aus Diagonalglied der Streuungsmatrix)
BEACHTE: Errechnet aus "within-groups"-Matrix
```

	Faktor 1	Faktor 2
V1	0.8968	-0.4811
V2	0.3062	0.9705

```
Kanonische Strukturkoeffizienten
der 1.(unabhaengigen) Variablengruppe
(=Korrelation der Variablen mit kanonischen Faktoren)
```

	Faktor 1	Faktor 2
V1	0.9537	-0.3009
V2	0.4727	0.8812

Signifikanz der Diskriminanzkoeffizienten

Natürlich will man wissen, ob die unabhängigen quantitativen Variablen eine signifikante diskriminierende Wirkung besitzen, bzw. wie groß ihre Signifikanz ist. Zu diesem Zwecke muß ein Allgemeines Lineares Modell (ALM) gerechnet werden. Zu diesem Zweck braucht im vorausgegangenem „selbst geschriebenem“ Almo-Programm „Tatsuok1.Alm“ nur die Programm-Nr. auf 20 geändert werden.

Almo liefert unter anderem folgendes Ergebnis.

Streuungsquelle	generalisierte Streuung	Wilks Lambda	Korrel Koeff.	F-Wert	df	Signifikanz p	(1-p)100
Gesamtstreuung	0.0439						
Fehlerstreuung	0.0187				52		
alle unabh. Var. zusammen	0.0252	0.4262	0.5843	6.9124	4	0.0003	99.9691
V1	0.0119	0.6105	0.6241	8.2944	2	0.0020	99.7994
V2	0.0069	0.7299	0.5197	4.8098	2	0.0164	98.3557

V1 besitzt also einen F-Wert von 8.2944 (mit df1=2, df2=26) und einer Signifikanz (1-p).100 von 99,8 %. Für V2 wurde ein F=4.8098 und eine Signifikanz von 98.36 % ermittelt (zusätzlich gibt Almo noch die partielle Korrelation für V1 und V2 aus).

Vergleich zu SPSS:

Wird in der DISCRIMINANT-Prozedur von SPSS die Anweisung "Method=Wilks" verwendet (und über entsprechendes Setzen von TOLERANCE, FIN, FONT die Aufnahme aller unabhängigen quantitativen Variablen erzwungen), dann wird eine schrittweise Aufnahme der unabhängigen quantitativen Variablen in das Modell angefordert. Für den letzten Schritt - in unserem Beispiel der 2. Schritt - wird eine Tabelle ausgegeben mit der Überschrift "Variables in the Analysis after Step..." (in unserem Beispiel: Step 2). Die dort unter "F to remove" angegebenen F-Werte

entsprechen den F-Werten aus der obigen Almo-Ausgabe - nicht jedoch die Wilks'schen Lambda-Werte; diese haben eine andere Bedeutung. Das Wilks'sche Lambda und der zu ihm äquivalente F-Wert, die in SPSS nach dem letzten Schritt ausgegeben werden, sind identisch mit dem Wilks'schen Lambda und dem F-Wert, die von Almo für das Gesamt-Modell ausgegeben werden. In SPSS ist die Ausgabe überschrieben mit "At step2, V2 was included in the analysis."

Das Almo-Programm 20 liefert hier folgende Ausgabe:

```
generalisierte Gesamtstreuung           0.043896
================================================

Koeffizienten fuer Gesamt-Modell

Durch alle unabh. Variable
erklaerte generalisierte Streuung       0.025186
generalisierte Fehlerstreuung           0.018709
------------------------------------------------
Wilks Lambda                            0.426224
F-Wert f. erklaerte Streuung            6.912447
Freiheitsgrade Nenner =    4
               Zaehler=   52
Signifikanz: p                          0.000309
Signifikanz: (1-p)*100                 99.969127 %
Teststaerke von F                       0.989929
------------------------------------------------
Pillais Spur                            0.682725
F-Wert f. erklaerte Streuung            6.996862
Freiheitsgrade Nenner =    4
               Zaehler=   54
Signifikanz: p                          0.000274
Signifikanz: (1-p)*100                 99.972555 %
Teststaerke von F                       0.990932
------------------------------------------------
multiple Korrelation (aus Pillais Spur) 0.584262
quadriert                               0.341363
```

Vergleich zu SAS:

Bei SAS wird in der Funktion STEPDISC nach der Aufnahme der letzten Variable eine entsprechende mit "Statistics for Removal" überschriebene Tabelle ausgegeben. Wie in Almo werden auch hier die (quadrierten) partiellen Korrelationen für V1 und V2 ausgegeben.
Zu beachten ist, daß bei dieser Betrachtung die Trennung in 2 kanonische Faktoren nicht berücksichtigt wird. Es ist also z.B. nicht möglich zu sagen: V1 besitzt in der 1. kanonischen Diskriminanzfunktion einen F-Wert von x1 mit einer Signifikanz von y1 und in der 2. kanonischen Diskriminanzfunktion einen F-Wert von x2 und eine Signifikanz von y2.

P29.2.6 Diskriminanzwerte und Klassifikation

Diskriminanzwerte sind - im Rahmen der kanonischen Korrelationsanalyse - kanonische Faktorwerte. Unsere Ausführungen von P29.1.3 gelten hier also uneingeschränkt - wobei es allerdings in der klassischen Diskriminanzanalyse nicht üblich ist, daß die kanonischen Faktorwerte auch für die Gruppe der Dummies der abhängigen nominalen Variablen ermittelt werden.

Von "Klassifikation" sprechen wir, wenn es darum geht, die Gruppenzugehörigkeit der Untersuchungseinheiten aus ihren Werten in den unabhängigen quantitativen Variablen zu erklären bzw. zu prognostizieren.

Betrachten wir ein Beispiel: Die Präferenz für die politischen Parteien A, B, C soll durch die unabhängigen quantitativen Variablen Einkommen, Kinderzahl, Bildungsniveau erklärt werden. Wir haben also eine abhängige nominale Variable und mehrere unabhängige quantitative Variable. Von "Klassifikation im engeren Sinne" sprechen wir, wenn folgende Konstellation gegeben ist: In einer vorausgehenden Diskriminanzanalyse wurde die Wirkung (die kanonischen Diskriminanzkoeffizienten) der unabhängigen Variablen hinsichtlich der abhängigen nominalen Variablen ermittelt. Jetzt geht es darum, für einige Individuen (deren Parteipräferenz nicht bekannt ist) auf Grund der Kenntnis ihre Werte in den unabhängigen Variablen ihre Parteipräferenz zu prognostizieren.
Von "Klassifikation im weiteren Sinne" sprechen wir, wenn die Parteipräferenz bekannt ist und wir nun überprüfen wollen, ob die ermittelten unstandardisierten Diskriminanzkoeffizienten (kanonische Gewichtszahlen) diese Parteipräferenz richtig "prognostizieren".

P27.2.7 Eingabe in Maskenprogramm Prog29mb

```
                        Prog29mb.Msk
             Kanonische Diskriminanzanalyse
                            mit
            Ermittlung der Diskriminanz-Werte
            Reproduzieren der Gruppenzugehörigkeit

        optional:
        Speichern der Diskriminanz-Werte
        Streudiagramm der Objekte im kanonischen Raum

 Das Programm rechnet in 4 Schritten

 1. Zuerst wird eine kanonische Diskriminanzanalyse
    gerechnet. Sie ermittelt die Diskriminanzkoeffizienten
    Almo verwendet dafür (intern) das Programm zur
    kanonischen Diskriminanzanalyse Prog 29
 2. Dann werden aus diesen die Diskriminanzwerte für jede
    Untersuchungseinheit ermittelt und ausgegegeben.
    Almo verwendet dafür (intern) das Klassifikations-
    programm Prog 27
 3. Optional: Die Diskriminanzwert-Variable werden an die
    Datensätze als zusätzliche Variable angefügt und in
    ein neue Datei geschrieben
 4. Optional: Abschliessend wird ein Streudiagramm der
    Objekte mit den Diskriminanzfaktoren als Koordinaten-
    achsen gezeichnet. Wurden 3 oder mehr Diskriminanz-
    faktoren gefunden, so werden nur die ersten 3 verwendet

Beispiel:
Für das Beispiel werden die bekannten Irisdaten von Fisher
verwendet. R.A.Fisher: The use of multiple measurements in
taxonomic problems, in Annals of Eugenics, 1936

Für 3 Liliensorten werden verschiedene Merkmale wie
Blütenlänge, Blattbreite etc. verwendet

Die Diskriminanzkoeffizienten aus dem 1. Schritt werden
dann verwendet, um im 2. Schritt die Diskriminanzwerte
für jedes Objekt zu errechnen und die Gruppenzugehörigkeit
(=die Lilienart) zu reproduzieren. Dann wird noch ein
Streudiagramm der Objekte im kanonischen Raum gezeichnet

Handbuch Teil 4 Fortgeschrittene Verfahren, Abschnitt P29.2
```

Programm-Bedienung ---> [Hilfe]

Speicher fuer x Variable | Hilfe

Vereinbare Variable= 20

Option: Weitere Vereinbarungen - nur wenn Almo dazu auffordert

Variablennamen

Datei der Variablennamen | Hilfe

zeige = Namensdatei in Output zeigen
leer = nicht zeigen

Freie Namensfelder | Hilfe

Leere alle Eingabefelder dieser Sub-Box

Name1=Bluetenlaenge;
Name2=Bluetenbreite;
Name3=Blattlaenge;
Name4=Blattbreite;
Name5=Lilie:SETOSA,VERSICOLOR,VIRGINICA;

erzeuge zusätzliche Namensfelder

Variablennamen in Datei speichern | Hilfe
Eingabefeld leer = nicht speichern

Datei aus der gelesen wird | Hilfe bei Datei-Problemen

"C:\Almo15\Testdat\Irisdata.fre"

frei

Format der Daten | Hilfe

V1:5

der Datensatz enthält diese Variablen
Bei Format DIREKT schreiben Sie: alle_V

Wenn Dateiformat FIX oder Nicht-Standard-FREI | Hilfe

Analyse-Variable

unabhängige quantitative Variable

Bluetenlaenge,Bluetenbreite,Blattlaenge,Blattbreite

abhängige nominale Variable
(nur eine erlaubt)

Hilfe

Lilie

Option: Ein- und Ausschliessen von Untersuchungseinheiten

BEACHTE: Umkodierungen wirken sich aus auf:
1. die Berechnung der kanonischen Diskriminanzkoeffizienten und
2. die Berechnung der Diskrinanzwert-Variablen (im Beispiel: V21,22)

Option: Umkodierungen und Kein-Wert-Angaben

Zulässiger Kein-Wert

50

Maximal 50% der unabhängigen quantitativen Variablen, dürfen Kein_Wert besitzen. Für sie wird der Mittelwert eingesetzt. Sonst werden die Diskriminanzwert-Variable (im Beispiel: V21,22) auf Kein_Wert gesetzt

die "Spezielle Kein-Wert-Behandlung" wirkt sich nur auf die Berechnung der kanonischen Diskriminanzkoeffizienten aus

Option: Spezielle Kein-Wert-Behandlung

Option: Ausreisser vom Typ 1 identifizieren

Hilfe

die Gewichtung wirkt sich nur auf die
Berechnung der kanonischen Diskriminanz-
koeffizienten aus

Option: Untersuchungseinheiten gewichten

Berechne und zeige

p_fuer_Gruppe — Wahrscheinlichkeit der Gruppenzugehörigkeit — Hilfe

Bayes_p_fuer_Gruppe — Bayes-Wahrscheinlichkeit der Gruppenzugehörigkeit — Hilfe

kanon_Faktorwert — kanonische Faktorwert-Variable (Diskriminanzwert-Variable)

Identifikationsvariable — Hilfe

Hilfe

apriori-Wahrscheinlichkeit für Bayes-Wahrscheinlichkeit
Wahrscheinlichkeit für die Ausprägungen der abhäng. Variablen vorgeben

0

= 0 Keine apriori-Wahrscheinlichkeit vorgeben
= 1 empirische Randverteilung als Wahrscheinlichkeit vorgeben
= 40,25,35
z.B. diese %-Werte als Wahrscheinlichkeiten für die Ausprägungen der abhängigen nominalen Variablen vorgeben

verschiedene Programm-Optionen

Diskriminanzwerte in eine Datei speichern

2- bzw. 3-dimensionales Streudiagramm

0

=1 Streudiagramm zeichnen
=0 nicht

Option: "Aussehen" der auszugebenden Tabelle bzw. Matrix

Grafik-Optionen

Basisstatistiken ausgeben

0

1= Basisstatistiken ausgeben
0= nicht

P29.2.8 Erläuterung zu den Boxen

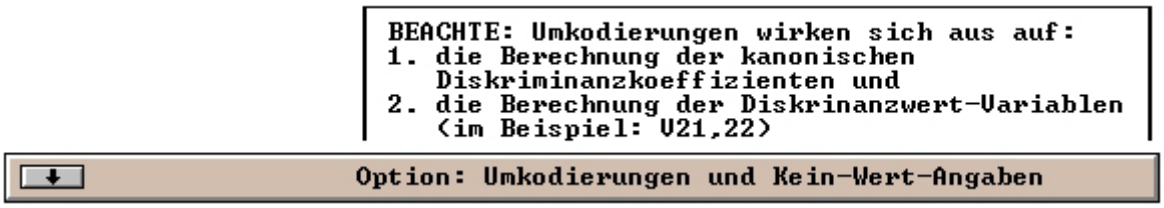
Für das in der Programm-Maske eingesetzte Beispiel werden die bekannten Lilien-Daten von Fisher verwendet (R.A.Fisher: The use of multiple measurements in taxonomic problems, in Annals of Eugenics, 1936)

Für 3 verschiedene Arten von Lilien werden verschiedene Merkmale wie Blütenlänge, Blattbreite etc. verwendet. Die Aufgabe ist es, anhand dieser Merkmale zu prognostizieren, welche der 150 untersuchten Blumen welcher der 3 Lilienarten angehören

In einem 1. Schritt werden die Diskriminanzkoeffizienten ermittelt. Aus diesen werden dann in einem 2. Schritt die Diskriminanzwerte für jedes der 150 Objekte errechnet und die Gruppenzugehörigkeit (=die Lilienart) reproduziert. Dann wird noch ein Streudiagramm der 150 Objekte im kanonischen Raum gezeichnet

Die Boxen der Programm-Maske stimmen weitgehend überein mit den in Abschnitt P29.2.2 bereits erklärten Boxen des Maskenprogramms Prog29m3. Wir wollen hier nur die neu hinzugekommenen Boxen erläutern.

Box: Umkodierungen und Kein-Wert-Angaben



Zur Art und Weise wie Variable umkodiert und ihre Kein-Wert-Codes deklariert werden, siehe Abschnitt P0.5.

Zu beachten ist folgendes:

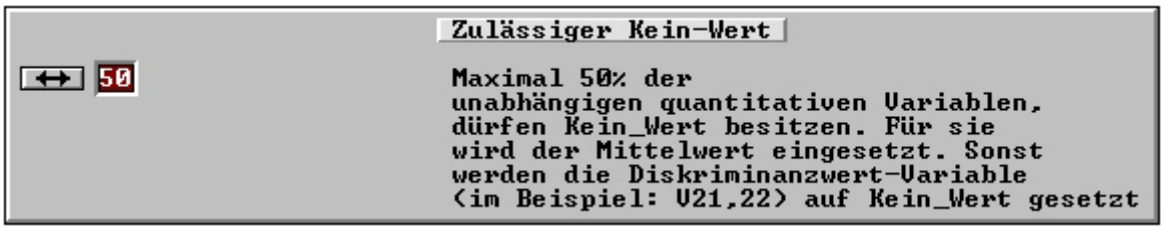
1. Es dürfen nur die unabhängigen quantitativen Variablen umkodiert werden. Die Umkodierungen wirken sich dabei nur aus auf:

   a. die Berechnung der kanonischen Diskriminanzkoeffizienten und
   b. die Berechnung der Diskriminanzwert-Variablen (im Beispiel: V21,22)

   Werden in Box 15 die Daten inklusive der Diskriminanzwert-Variablen in eine neue Datei gespeichert, dann werden die unabhängigen quantitativen Variablen mit ihren Orginal-Werten und nicht mit ihren umkodierten Werten gespeichert.

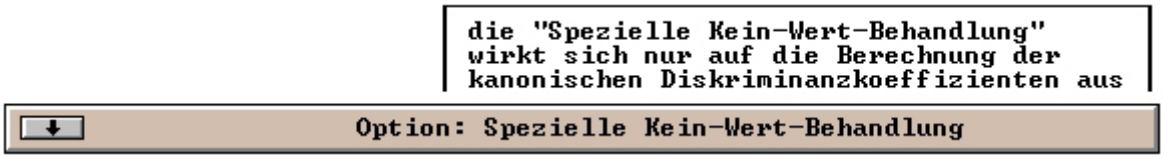
2. Die Kein-Wert-Angabe wird für alle Variable vorgenommen. Es sei denn, in der eingelesenen Datei sind die Kein-Wert-Codes schon in der Almo-internen Form enthalten.

Box: Zulässiger Kein-Wert



Im Eingabefeld wird angegeben, wieviel Prozent der unabhängigen quantitativen Variablen Kein-Wert sein dürfen - damit trotzdem die Diskriminanzwerte berechnet werden. Im Beispiel haben wir "50" eingesetzt. Das bedeutet, dass maximal 50 % der unabhängigen quantitativen Variablen, in unserem Beisiel also 2 von den 4 unabhängigen Variablen Kein_Wert besitzen dürfen. Für sie wird der Mittelwert eingesetzt.
Sonst werden die Diskriminanzwert-Variable V21,22 auf Kein_Wert gesetzt.
Wenn also bei einem Untersuchungsobjekt in unserem Beispiel nicht mehr als 2 unabhängige Variable Kein-Wert sind, dann wird für dieses Objekt trotzdem ein Diskriminanzwert berechnet.
Zu beachten ist: Diese Regelung gilt nur bei der Berechnung der Diskriminanzwerte – nicht bei der Berechnung der Diskriminanzkoeffizienten.

Box: Spezielle Kein-Wert-Behandlung



Besitzt eine oder mehrere Analysevariablen keinen Wert, dann verwendet Almo standardmäßig das "paarweise Ausscheiden". Der Benutzer hat die Möglichkeit eine von 7 Methoden zur Kein-Wert-Behandlung zu wählen. Dazu muß die Optionsbox geöffnet werden. Wir haben diese Optionsbox bei der kanonischen Korrelation in Abschnitt P29.1.1.2 bereits dargestellt und erläutert.

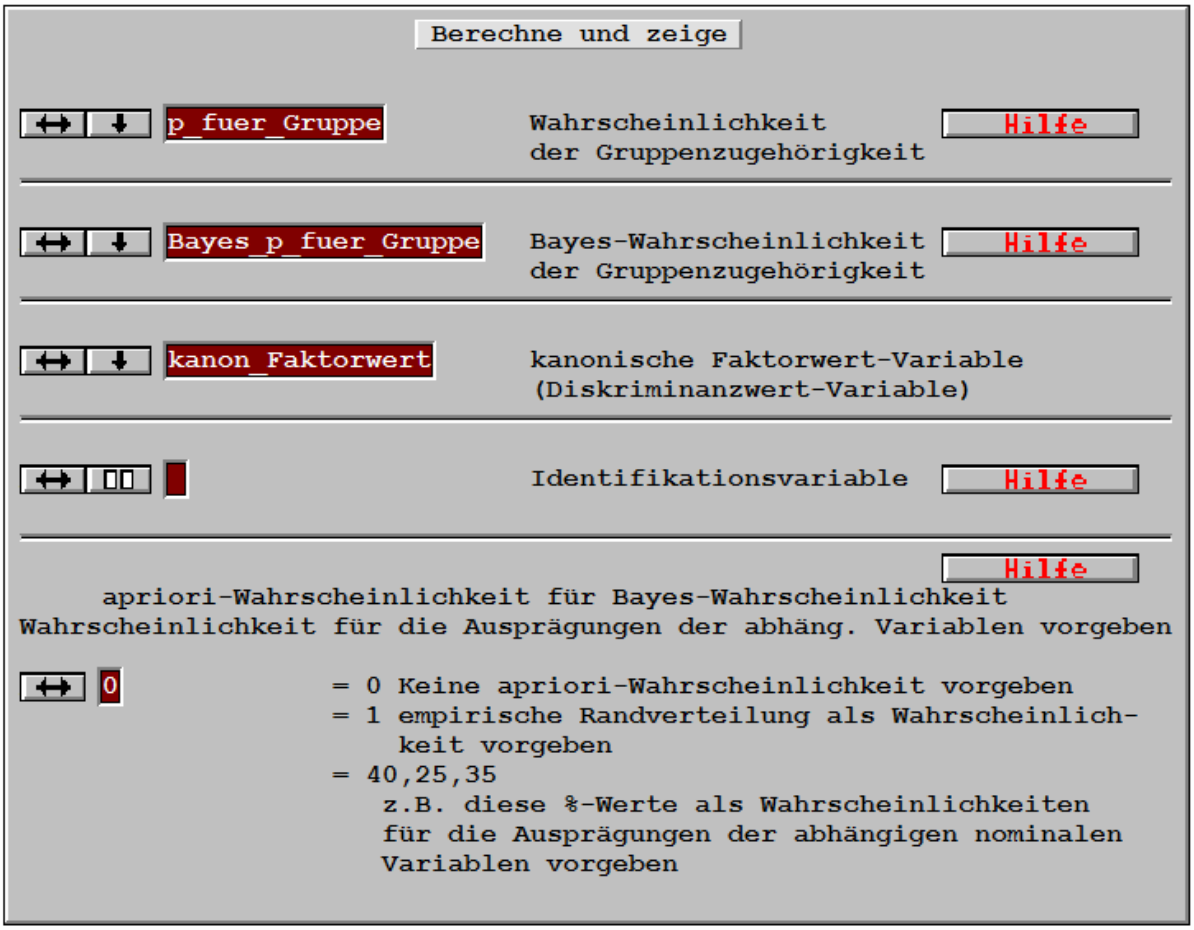
Zu beachten ist: Die Spezielle Kein-Wert-Behandlung gilt nur bei der Berechnung der Diskriminanzkoeffizienten - nicht bei der Berechnung der Diskriminanzwerte.

Empfehlung: Man sollte (man muß aber nicht) die Kein-Wert-Behandlung 4 verwenden. Dabei werden fehlende Werte bei den quantitativen unabhängigen Variablen durch ihre Mittelwerte ersetzt. Das ist sinnvoll, weil bei der Ermittlung der Diskriminanzwerte von Almo auch so verfahren wird. Siehe dazu die Erläuterungen zu Box 10.

Box: Option: Untersuchungseinheiten gewichten
Siehe dazu Dokument Nr. 0 Arbeiten mit Almo, Abschnitt P0.8.

Zu beachten ist: Gewichtet wird nur bei der Berechnung der Diskriminanzkoeffizienten - nicht bei der Berechnung der Diskriminanzwerte; also nur bei Schritt 1, nicht aber bei Schritt 2 und 3.

Box: Berechne und zeige ...



Wir werden in Abschnitt P29.2.9.1 diese Eingabefelder erläutern. Wir empfehlen, die Almo-Vorgaben zu akzeptieren. Hier wird nur das Eingabefeld "Identifikations-Variable" erläutert:

Almo liefert beispielsweise für die vom Modell prognostizierte Gruppenzugehörigkeit der Objekte folgenden Output:

```
          Die Gruppe mit maximaler Wahrscheinlichkeit ist
          mit * markiert

          | Wahrscheinlichkeit
          | der Zugehoerigkeit
          | zu Gruppe
Datensatz |   1      2      3   |
----------|-----------------------
   1      | 0.901  0.929* 0.399 |
   2      | 0.649* 0.529  0.133 |
   3      | 0.282  0.399* 0.076 |
   4      | 0.100  0.196* 0.072 |
   5      | 0.908* 0.779  0.351 |
   6      | 0.653* 0.462  0.433 |
   7      | 0.203  0.356* 0.153 |
   8      | 0.715  0.876* 0.415 |
   9      | 0.927  0.987* 0.654 |
  10      | 0.608* 0.415  0.385 |
   .      |   .      .      .   |
   .      |   .      .      .   |
   .      |   .      .      .   |
```

Wenn nun durch eine Ein- bzw. Ausschluss-Anweisung oder durch eine Ausreisser-Bereinigung die 3. Person ausgeschlossen worden wäre, dann würde das nicht ersichtlich

werden, da die Datensatz-Nummern in der 1. Spalte von Almo einfach fortlaufend nummeriert werden. Der Output würde also folgendermaßen ausschauen:

Datensatz	Wahrscheinlichkeit der Zugehoerigkeit zu Gruppe 1	2	3	
1	0.901	0.929*	0.399	
2	0.649*	0.529	0.133	
3	0.100	0.196*	0.072	<--- tatsächlich ist dies die
4	0.908*	0.779	0.351	4. Untersuchungseinheit
5	0.653*	0.462	0.433	
6	0.203	0.356*	0.153	
7	0.715	0.876*	0.415	
8	0.927	0.987*	0.654	
9	0.608*	0.415	0.385	
10	0.049	0.097	0.203*	
.	.	.	.	
.	.	.	.	
.	.	.	.	

Die Datensatznummer 3 wird jetzt an die 4. Untersuchungseinheit aus der Orginaldatei vergeben usw. Die Untersuchungseinheiten sind also nicht mehr korrekt identifizierbar. Ist in der Orginaldatei eine Variable vorhanden, die die Untersuchungseinheiten identifiziert, z.B. eine Fragebogen-Nummer dann sollte diese Variable als "Identifikationsvariable" angegeben werden. Almo liefert dann folgenden Output, der es ermöglicht, die Untersuchungseinheiten zu identifizieren:

Datensatz	Wahrscheinlichkeit der Zugehoerigkeit zu Gruppe 1	2	3	ID-Nr. V5	
1	0.901	0.929*	0.399	1	
2	0.649*	0.529	0.133	2	
3	0.100	0.196*	0.072	4	<---die Identifikations-
4	0.908*	0.779	0.351	5	variable zeigt, dass
5	0.653*	0.462	0.433	6	der 3. Datensatz die
6	0.203	0.356*	0.153	7	4.Untersuchungseinheit
7	0.715	0.876*	0.415	8	enthält
8	0.927	0.987*	0.654	9	
9	0.608*	0.415	0.385	10	
10	0.049	0.097	0.203*	11	

Hier wurde V5 als "Identifikationsvariable" angegeben.

Bleibt das Eingabefeld für die Identifikationsvariable leer, dann gibt Almo die letzte Spalte nicht aus.

Box: Optionen
Siehe dazu die Erläuterungen zu Box 12 in Abschnitt P29.2.2.

Box: Diskriminanzwerte in eine Datei speichern

Optionsbox geöffnet:

```
X  Loesche wieder diese Box (dann Voreinstellungen wieder gueltig)

                 Diskriminanzwerte in eine Datei speichern

"C:\Almo15\Progs\Dskmwert"

                          Geben Sie einen neuen Dateinamen
                          ohne Erweiterung an            Hilfe
                          Almo erzeugt 3 Dateien:        Hilfe

                          1. eine nicht lesbare Almo-Arbeitsdatei
                             mit der Erweiterung ___.dir

                          2. eine anschaubare Datei im freien Format
                             mit der Erweiterung ___.fre

                          3. eine Datei der Variablennamen
                             mit der Erweiterung ___.nam

In den unter 1. und 2. angegebenen neuen Dateien sind nun enthalten:
  - die unveränderten Variablen aus der alten Datei
  - die Diskriminanzwert-Variablen, die als letzte Variablen
    hinter die Variablen der alten Datei gestellt wurden
                          was wird gespeichert ? ------> Hilfe

In der unter 3. angegebenen neuen Datei der Variablennamen sind nun
enthalten:
  - die Variablennamen aus der alten Datei einschliesslich der
    in der Box "Freie Namensfelder" angegebenen Namen
  - die Namen "Dskmwert.." für die neuen angehängten Diskriminanzwert-
    variable. Anstelle der 2 Punkte schreibt Almo die Variablen-
    numern der neuen Variablen, also z.B. "Dskmwert21", "Dskmwert22"
                                   bitte lesen -----> Hilfe
```

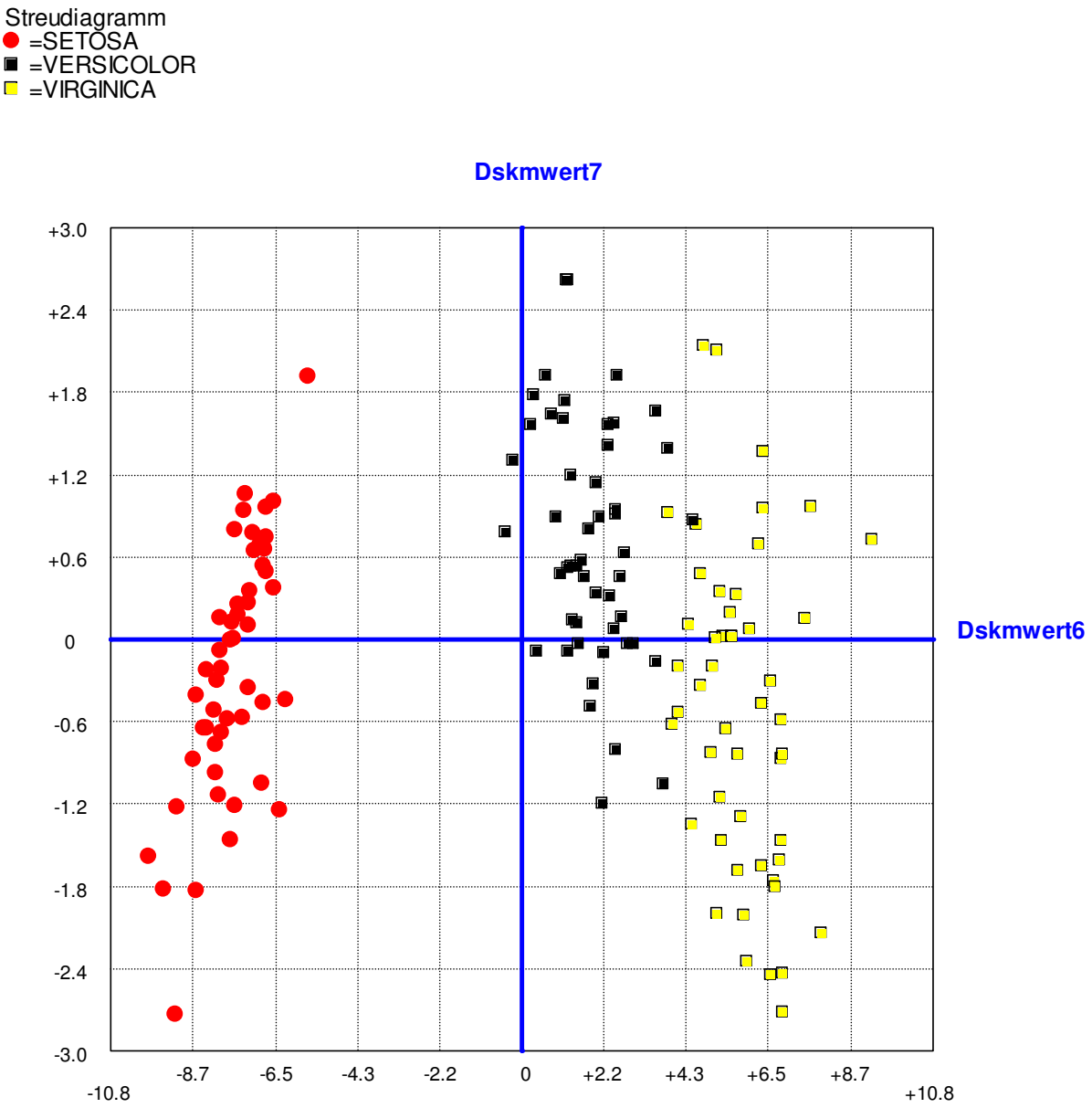
Die um die die Diskriminanzwert-Variable verlängerten Datensätzet werden in eine neue Datei geschrieben. Wollen Sie keine neue Datei anlegen, dann schließen Sie die Box wieder.
Geben Sie den vollen Pfad- und Dateinamen an.

Box: 2- bzw. 3-dimensionales Streudiagramm

```
                 2- bzw. 3-dimensionales Streudiagramm

1                    =1  Streudiagramm zeichnen
                     =0  nicht
```

Wenn Sie hier "1" einsetzen, dann zeichnet Almo ein Streudiagramm. Im Lilien-Beispiel von Fisher fällt dieses besonders schön aus:



Die 3 Lilienarten werden als rote, schwarze und gelbe Punkte dargestellt. Die roten Punkte (die Lilienart "Setosa") sind deutlich von den beiden anderen entfernt. Auch schwarze und rote Punkte sind getrennt. Allerdings überdecken sich die beiden Punktewolken etwas.

P29.2.9 Ausgabe aus Prog29mb

Wir wollen die Ergebnisse aus dem Diskriminanzwerte-Programm wieder an einem sehr einfachen Beispiel darstellen und erläutern. Dazu verwenden wir wieder (wie oben in Abschnitt P29.2.4) die Daten von Tatsuoka (1971, S.180). Die bereits abgebildete Programm-Maske zur Ermittlung der Diskriminanzwerte Prog29mb wird entsprechend ausgefüllt. Die so ausgefüllte Maske ist zu finden unter dem Menü "Almo/Liste aller Almo-Programme/**Tatsuoka3.Alm**". Das Programm liegt auch als Syntax-Programm vor. Zu finden unter "Almo/Liste aller Almo-Programme/Tatsuoka.Alm".

Die Ausgabe besteht aus mehreren Blöcken. Der Benutzer muss die Ergebnisdatei bis zum Ende durchschauen.

- Im 1. Block werden die Ergebnisse der Diskriminanzanalyse (wie bereits oben in Abschnitt P29.2.4 gezeigt) ausgegeben. Dabei wurden die Diskriminanz-koeffizienten in eine Almo-interne Datei gespeichert.

- In einem weiteren Ausgabe-Block wurden die Diskriminanzkoeffizienten aus der internen Datei wieder gelesen und die Diskriminanzwerte für die Untersuchungs-objekte errechnet und in die Datei ".\Progs\Dskmwert" (in unserem Beispiel) gespeichert sowie deren vom Modell prognostizierte Gruppenzugehörigkeit ausgegeben. Die Datei kann vom Benutzer geladen und angeschaut werden.

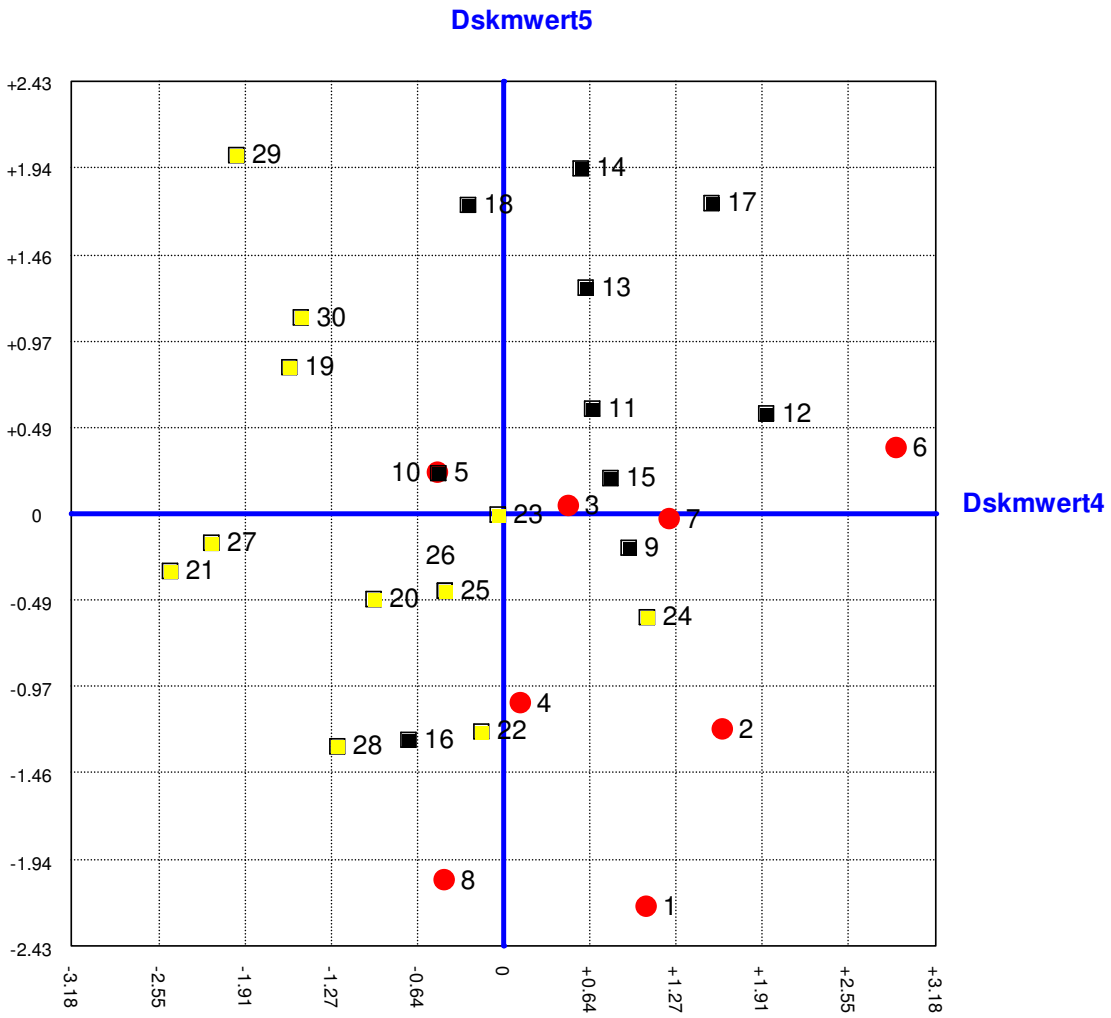
  Die Diskriminanzwerte sind folgende

```
V1 V2 V3  kanFak1    kanFak2
-- -- -- --------  ---------
20  6 1  1.051424  -2.210103
21 10 1  1.621995  -1.216696
15 12 1  0.477697   0.046936
15  8 1  0.127062  -1.064460
11 11 1 -0.489706   0.241043
24 17 1  2.895414   0.374280
 .  . .   .          .
 .  . .   .          .
 .         .    .             .                                    .
```

- Im "selbst geschriebenen" Syntax-Programm "Tatsuoka.Alm" (aber nur in diesem) werden in einem 3. Block die Diskriminanzwerte interkorreliert und ausgegeben

- In einem weiteren Block wird das folgende 2-dimensionales Streudiagramm ausgegeben

Streudiagramm
● =Typ 1
■ =Typ 2
□ =Typ 3



Die 2 kanonischen Diskriminanzfaktoren werden in der Grafik mit Dskmwert4 und Dskmwert5 bezeichnet. Die Nummern 4 und 5 deuten an, dass diese beiden Faktoren in der Datei "Dskmwert" als 4. und 5. Variable gespeichert wurden.
Die 3 identifizierten Typen werden als rote, schwarze und gelbe Punkte dargestellt. Die roten, schwarzen und gelben Punkte sind zwar erkennbar voneinader getrennt. Allerdings überdecken sich die drei Punktewolken etwas.

P29.2.9.1 Ermitteln der Gruppenzugehörigkeit

Almo gibt u.a. aus

```
Als Beispiel wird die kanonische Faktorwert-Berechnung fuer den
1.Datensatz gezeigt

Wurde die Korrelations- oder Kovarianzmatrix analysiert, dann wird
jede einzelne Variable standardisiert und mit dem
unstandardisierten kanonischen Diskriminanzkoeffizienten multipliziert

Die Formel ist folgende:
```

```
    (Variablenwert - Mittelwert) * Diskrimkoeff / Standabwg

Wurde die Kovarianzmatrix analysiert, dann wird in obiger
Formel Standabwg = 1 gesetzt

V1                              (20 - 13.2) * 0.219936 / 1
V2                            + (6 - 11.0667) * 0.0876587 / 1
kanonische Diskriminanzwert-Variable V4 = 1.05142


V1                              (20 - 13.2) * -0.117989 / 1
V2                            + (6 - 11.0667) * 0.277849 / 1
kanonische Diskriminanzwert-Variable V5 = -2.2101


           Die Gruppe mit maximaler Wahrscheinlichkeit ist mit * markiert

           Die tatsaechliche Gruppenzugehoerigkeit wird hinter
           der Datensatznummer in Klammern angegeben
```

Datensatz	Wahrscheinlichkeit der Zugehoerigkeit zu Gruppe 1	2	3	Bayes Wahrscheinlichkeit der Zugehoerigkeit zu Gruppe 1	2	3	kanonische Diskriminanzwert-Variable V4	V5
1 (1)	0.329*	0.013	0.013	0.927*	0.037	0.036	1.051	-2.210
2 (1)	0.641*	0.092	0.017	0.855*	0.122	0.023	1.622	-1.217
3 (1)	0.695	0.811*	0.334	0.378	0.441*	0.181	0.478	0.047
4 (1)	0.752*	0.197	0.328	0.590*	0.154	0.257	0.127	-1.064
5 (1)	0.267	0.526	0.834*	0.164	0.323	0.513*	-0.490	0.241
6 (1)	0.061	0.061*	0.000	0.497	0.499*	0.004	2.895	0.374
7 (1)	0.712*	0.616	0.085	0.504*	0.436	0.060	1.225	-0.029
8 (1)	0.192*	0.014	0.122	0.586*	0.043	0.372	-0.444	-2.058
9 (2)	0.854*	0.636	0.159	0.518*	0.386	0.096	0.918	-0.189
10 (2)	0.267	0.526	0.834*	0.164	0.323	0.513*	-0.490	0.241
etc.								

Die Gruppenzugehörigkeit wird in folgender Weise geschätzt:

Betrachten wir die Untersuchungseinheit 1 (Datensatz 1). Wir wollen im folgenden ihre Wahrscheinlichkeit der Gruppe 1 anzugehören bestimmen.

Sie hat einen Diskriminanzwert

```
für den 1. kanonischen Faktor von V4 =  1.051
für den 2. kanonischen Faktor von V5 = -2.210
```

Die Gruppen-Zentroide werden nun von den Diskriminanzwerten subtrahiert. Es entstehen folgende "Abweichungsdiskriminanzwerte"

für Gruppe 1: $1.051 - 0.808 = 0.243$
$-2.210 + 0.739 = -1.471$

Wir formieren aus diesen beiden Werten einen Zeilenvektor **w'** und einen Spaltenvektor **w** und errechnen folgenden Chi-Quadrat-Wert.

$$\text{Chi} = \mathbf{w'} \,.\, \mathbf{D_1^{-1}} \,.\, \mathbf{W}$$

$\mathbf{D_1^{-1}}$ = Inverse der "within-group"-Kovarianzmatrix der beiden Diskriminanzwert-Variablen für Gruppe 1

Da wir im Rahmen der kanonischen Korrelationsanalyse die gruppenspezifischen **D**-Matrizen nicht errechnen, müssen wir die "gepoolte within-groups"-Kovarianzmatrix der Diskriminanzfunktionen verwenden. Diese ist jedoch eine Einheitsmatrix - wodurch sich obige Gleichung vereinfacht auf

$$Chi = \mathbf{w}'.\mathbf{w}$$

d.h. der Chi-Quadrat-Wert ergibt sich als Summe der quadrierten "Abweichungs-Diskriminanzwerte". Für unser Beispiel erhalten wir so

$$Chi = 0.2432^2 + (-1.471)^2 = 2.223$$

Der p-Wert für diesen Chi-Quadrat-Wert mit k Freiheitsgraden (k=Zahl der kanonischen Faktoren, also 2) ist

$$p = 0.329$$

Damit ist die Wahrscheinlichkeit der 1. Untersuchungseinheit der Gruppe 1 anzugehören mit p=0.329 ermittelt. Dies ist der 1. Wert in obiger Almo-Ausgabe.

Nun besteht die Möglichkeit, eine vorgegebene Wahrscheinlichkeit $\mathbf{p_v}$ der Gruppe 1 anzugehören als Information miteinzubeziehen und die "Bayes-Wahrscheinlichkeit" $\mathbf{p_b}$ zu berechnen. Als vorgegebene Wahrscheinlichkeit $\mathbf{p_v}$ kann z.B. der Anteil der Untersuchungseinheiten, die sich in Gruppe 1 befinden, verwendet werden. In unserem

Beispiel sind dies 8 von 30, also $\mathbf{p_v}$=8/30 = 0.2666.

Die vorgegebene Wahrscheinlichkeit kann vom Benutzer über die Anweisung

```
p_Vorgabe=....
```

gewählt werden. Dabei gibt es folgende Möglichkeiten:

`p_Vorgabe=0;`	Es wird keine Wahrscheinlichkeit vorgegeben (Voreinstellung)
`p_Vorgabe=1;`	Anteilsmäßige Verteilung der Untersuchungseinheiten auf die Gruppen vorgeben.
`p_Vorgabe=x1,x2,x3;`	Frei gewählte Prozentwerte x für die 3 Gruppen als Wahrscheinlichkeiten vorgeben, z.B. 17,23,60 **Beachte**: 1. Es müssen soviel Werte, wie Gruppen vorhanden sind, angegeben werden. 2. Die Werte müssen sich exakt zu 100 summieren. 3. Kommastellen dürfen nicht angegeben werden, bzw. werden von Almo negiert.

Die "Bayes-Wahrscheinlichkeit" $\mathbf{p_b}$ der Gruppe 1 anzugehören wird gemäß folgender Gleichung bestimmt:

$$d_i = Chi_i + a + b$$

$a = \ln(\det(D_1))$	wenn die within-groups-Kovarianzmatrix D1 für Gruppe 1 bekannt ist
$= 0$	wenn sie nicht bekannt ist, wie dies bei uns der Fall ist
$b = -\ln(p_{v_1})$	wenn die vorgegebenen Wahrscheinlichkeiten für die Gruppen verschieden sind
$= 0$	wenn sie alle gleich sind, bzw. keine Wahrscheinlichkeiten vorgegeben werden

i	= Index für Gruppe
d_i	= Zwischenwert für Gruppe i
ln	= Logarithmus zur Basis e
det	= Determinante
p_{v_1}	= für Gruppe 1 vorgegebene Wahrscheinlichkeit
Chi_i	= Chi-Quadrat-Wert für Gruppe i - gemäß obiger Gleichung

Für unser Beispiel ergeben sich, wenn wir keine Wahrscheinlichkeiten p_v vorgegeben und D1, D2, D3 als nicht bekannt gelten muß:

$d_1 = 2.223+0+0$

$p_{b1} = e^{-0.5*d1} / \text{Summe}(e^{-0.5*di})$ für Summe von i=1 bis i=3

Der Chi-Quadrat-Wert für Gruppe 2 und 3 ist 8.686. So gilt

d2 = 8.686 + 0 + 0 = 8.686
d3 = 8.686 + 0 + 0 = 8.686

eingesetzt in obige Gleichung für p_{b1}

$p_{b1} = e^{-0.5*2.223} / (e^{-0.5*2.223} + e^{-0.5*8.686} + e^{-0.5*8.686})$
= 0.329 / (0.329 + 0.013 + 0.013)
= 0.927

Chi = 8.686
Signifikanz p = 0.013

Dieser Wert von 0.927 ist in obiger Almo-Ausgabe als Bayes-Wahrscheinlichkeit für Gruppe 1 angegeben.

Die Bayes-Wahrscheinlichkeiten der 3 Gruppen addieren sich zu 1.0. Werden keine Wahrscheinlichkeiten vorgegeben, dann sind sie mit den "normalen" Wahrscheinlichkeiten gleich - nur eben auf Summe 1.0 normiert.

Im 3. Anfang-Ende-Block werden die Diskriminanzwerte mit Programm 19 korreliert. Das geschieht nur um folgendes zu zeigen:

1. V4, die 1. Diskriminanzwert-Variable aus der 1. Variablengruppe korreliert mit V6, der 1. Diskriminanzwert-Variablen aus der 2. Variablengruppe mit 0.6545 - das ist die 1. kanonische Korrelation
2. V5, die 2. Diskriminanzwert-Variable aus der 1. Variablengruppe korreliert mit V7, der 2. Diskriminanzwert-Variablen aus der 2. Variablengruppe mit 0.5042 - das ist die 2. kanonische Korrelation

P29.2.11 Klassifikation bei unbekannter Gruppenzugehörigkeit

Der eigentliche Zweck der Klassifikation besteht darin, für eine Untersuchungseinheit, deren Gruppenzugehörigkeit nicht bekannt ist, diese zu prognostizieren. Dies ist möglich, wenn wir die Werte dieser Untersuchungseinheit in den unabhängigen quantitativen Variablen kennen und wenn wir aus einer vorausgehenden Diskriminanzanalyse (über möglichst viele Untersuchungseinheiten) die Diskriminanzkoeffizienten und die Gruppen-Zentroide kennen.

Mit Programm 27 können wir dann die Diskriminanzwerte und die Wahrscheinlichkeit, bzw. "Bayes-Wahrscheinlichkeit" der Zugehörigkeit zu Gruppe i für diese Untersuchungseinheit berechnen. Siehe dazu die ausführliche Darstellung in P27.

P29.2.12 Nominale Variable als unabhängige Variable in der Diskriminanzanalyse

Für die Diskriminanzanalyse wird üblicherweise gefordert, dass die unabhängigen Variablen quantitativ sein müssen. Siehe etwa Urban (1993, S. 16). Wir wollen hier nicht darüber diskutieren, ob diese Forderung berechtigt ist. Wir verweisen allerdings auf die Verwandtschaft der Diskriminanzanalyse mit der Korrespondenzanalyse. Siehe dazu insbesondere unsere Ausführungen in Abschnitt P29.3.10.

Will der Benutzer (neben quantitativen auch) nominale Variable als unabhängige Variable verwenden, so wird ihm das in Almo ermöglicht - wenn auch auf eine etwas umständliche Art.

Die unabhängigen nominalen Variablen müssen in 0-1 kodierte Dummies aufgelöst werden. Diese Dummies werden dann wie quantitative Variable behandelt.

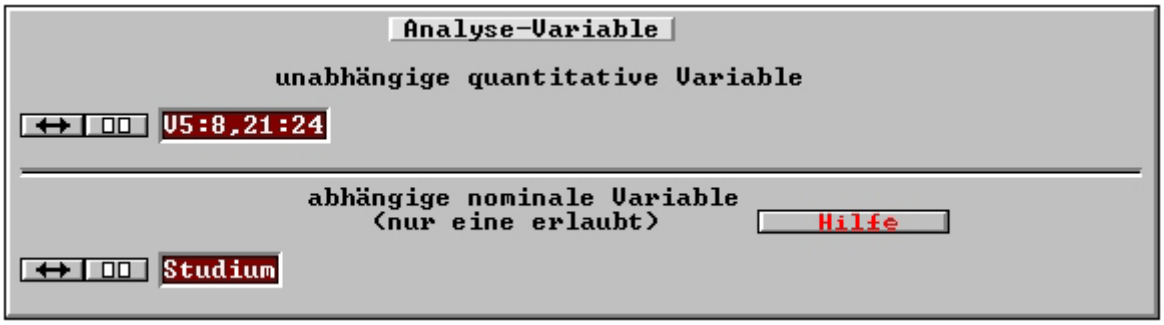
Wir wollen zeigen, wie beim Maskenprogramm Prog29m3 bzw Prog29m4 zu verfahren ist. Aus unseren Beispieldaten "Testdat.fre" verwenden wir V14 und V15 als unabhängige nominale Variable. Beide Variable besitzen 3 Ausprägungen. Sie werden je in 2 Dummies aufgelöst. Die 3. Ausprägung wird nicht in eine Dummy-Variable überführt - um keine linearen Abhängigkeiten zu erzeugen.

Das Programm ist als Beispielprogramm unter dem Namen „DisUnom.Alm" in Almo enthalten.
Es ist identisch mit dem Maskenprogramm Prog29m4. Wir erläutern deshalb nur die Boxen, die sich auf die unabhängigen Variablen bzw. ihre Dummies beziehen. Die anderen Boxen wurden in Abschnitt P29.2.2 und P29.2.8 erläutert.

In der **Box** "**Freie Namensfelder**" geben wir den Dummies Namen

Wir verwenden für die Namensgebung die freien Variablennummern V21,22 für die Dummies von V14 und V23,24 für die Dummies von V15.

In der **Box** "**Analyse-Variable**" geben wir neben den unabhängigen quantitativen auch die Dummies in das 1. Eingabefeld ein.



Da die Variablennamen zusammen zu lang sind und nicht in das Eingabefeld passen, schreiben wir nur die Variablennummern. V5:8 sind die quantitativen Variablen und V21:24 sind die Dummies von V14 und V15.

In der **Box Umkodierungen** ...." werden die beiden nominalen Variablen V14, 15 in Dummies aufgelöst. Almo ermöglicht es hier eine kurze und elegante Umkodierungsanweisung zu verwenden. Siehe dazu Handbuch Teil 2, Abschnitt 16.7

Betrachten wir die Anweisung

```
DumV14_1, DumV14_2 ( Dummy V14 )
```

In die Klammer hinein wird geschrieben "Dummy V14". Damit wir angeordnet, daß V14 in Dummies aufgelöst werden soll. Vor die Klammer werden die Variablen-Nummern oder die Variablen-Namen geschrieben, die für die Dummies vorgesehen sind.

P29.3 Bivariate Korrespondenzanalyse

P29.3.0 Einleitung

Wir unterscheiden zwischen bivariater und multipler Korrespondenzanalyse. Letztere stellen wir in Abschnitt P30.8.2 dar. Von "bivariater Korrespondenzanalyse" sprechen wir, wenn der Zusammenhang zwischen zwei nominalen Variablen untersucht wird, von "multipler", wenn mehr als zwei nominale Variable analysiert werden.

Sinnvoll wäre auch die begriffliche Trennung in "kanonische" Korrespondenzanalyse (nur für 2 nominale Variable) und "faktorenanalytische" Korrespondenzanalyse (für beliebig viele, also 2 und mehr nominale Variable).

Die Korrespondenzanalyse wurde ursprünglich als selbständiges Verfahren entwikkelt. Zur historischen Entwicklung dieses Verfahrens siehe Greenacre, 1984, S.7. Sehr bald wurde jedoch erkannt, daß die bivariate Korrespondenzanalyse identisch ist mit der kanonischen Korrelationsanalyse - angewendet auf die in 2 Sätze von Dummies aufgelösten beiden nominalen Variablen (siehe dazu Greenacre, 1984, S.108ff, S.121ff und Lebart, Morineau, Warwick, 1984, S.79ff).

Wir werden im folgenden die bivariate Korrespondenzanalyse als einen besonderen Anwendungsfall der kanonischen Korrelationsanalyse darstellen und dabei auch überwiegend die Terminologie dieses Verfahrens und seltener die ungewöhnliche Terminologie der Korrespondenzanalyse verwenden.

P29.3.1 Eingabe in Maskenprogramm Prog29m2

```
                        Prog29m2.Msk
                Bivariate Korrespondenzanalyse
                 (mit 2 nominalen Variablen)

        gerechnet als kanonische Korrelation mit den
        Dummies der 2 nominalen Variablen

    Zum Verhältnis der bivariaten zur
    multiplen Korrespondenzanalyse    --> [ Hilfe ]

Beispiel: Die nominalen Variablen, die analysiert werden
sollen, sind:

Autokauf: Porsche,  Mercedes, VW,
Fahrstil: aggresiv, normal,   zurückhaltend

Eine Korrespondenzanalyse über diese 2 nominalen Variablen
könnte zu einem 2-dimensionalen Raum führen, in dem z.B.
folgende Punkte dicht beieinander sind:

       Mercedes, normaler Fahrstil
       Porsche,  aggresiver Fahrstil
       VW,       zurückhaltend

 Grafik: 2-dimensionales Koordinatensystem
                          Siehe Handbuch, Abschnitt P29.4

                Programm-Bedienung ---> [ Hilfe ]
```

```
                Speicher fuer x Variable [ Hilfe ]

Vereinbare Variable= 20
```

```
   Option: Weitere Vereinbarungen - nur wenn Almo dazu auffordert
```

Variablennamen

Datei der Variablennamen — Hilfe

"C:\Almo15\Testdat\Varnamen.nam"

zeige

zeige = Namensdatei in Output zeigen
leer = nicht zeigen

Freie Namensfelder — Hilfe

Leere alle Eingabefelder dieser Sub-Box

Name1=Auto:Porsche,Mercedes,VW
Name3=Fahrstil:aggresiv,normal,zurückhaltend

erzeuge zusätzliche Namensfelder

Variablennamen in Datei speichern — Hilfe
Eingabefeld leer = nicht speichern

Datei aus der gelesen wird — Hilfe bei Datei-Problemen

"C:\Almo15\Testdat\Auto.fre"

frei — Format der Daten — Hilfe

V1:3 — der Datensatz enthält diese Variablen
Bei Format DIREKT schreiben Sie: alle_V

Wenn Dateiformat FIX oder Nicht-Standard-FREI — Hilfe

Analyse-Variable — Hilfe

Fahrstil — die 1. nominale Variable — Hilfe

Auto — die 2. nominale Variable — Hilfe

Option: Ein- und Ausschliessen von Untersuchungseinheiten

```
Option: Umkodierungen und Kein-Wert-Angaben

Option: Spezielle Kein-Wert-Behandlung

Option: Untersuchungseinheiten gewichten

X  Loesche wieder diese Box

                    Optionen

Kovarianz           Matrix
                    der Kalkül der kanon.Korr. wird angewandt auf
                    = Korrelation       Korrelationsmatrix
                    = Kovarianz         Kovarianzmatrix
                    Voreinstellung ist "Korrelation"
                    Bei der kanonischen Korrespondenzanalyse
                    sollte "Kovarianz" eingesetzt werden

0                   Eigenwert-Verfahren
                    =0 die Faktoren werden nach einem
                       Tridiagonal-QR-Algorithmus extrahiert
                       (Voreinstellung)
                    =1 nach dem v. Mises-Verfahren
                    =2 nach dem Jacobi-Verfahren

                    Faktorenzahl
                    Zahl der Faktoren eventuell beschränken
                    keine Angabe = maximal mögliche Faktorenzahl

0                   Zwischenergebnisse ausgeben
                    = 1   Zwischenergebnisse werden ausgegeben
                    = 0   nicht

Option: "Aussehen" der auszugebenden Tabelle bzw. Matrix

Grafik-Optionen

Basisstatistiken ausgeben

0                   1= Basisstatistiken ausgeben
                    0= nicht
```

P29.3.2 Erläuterungen zu den Boxen

Box: Vereinbare Variable
Siehe Dokument 0 "Arbeiten mit Almo.PDF", Abschnitt P0.1.

Box: Option: Weitere Vereinbarungen - nur wenn Almo dazu auffordert
Siehe P0.2.

Box: Variablennamen
Siehe P0.3.

Box: Datei aus der gelesen wird
Siehe P0.4.

Box: Wenn Dateiformat FIX oder nicht Standard-FREI
Siehe P0.4.

Box: Analyse-Variable

```
                    Analyse-Variable
Fahrstil            die 1. nominale Variable    Hilfe

Auto                die 2. nominale Variable    Hilfe
```

Geben Sie die beiden nominalen Variablen an. Dabei ist es gleichgültig, welche Sie als 1. und welche Sie als 2. Variable angeben.

Box: Option: Ein- und Ausschliessen von Untersuchungseinheiten
Siehe P0.7.

Box: Kein_Wert-Angabe und Umkodierungen
Siehe P0.5.

Box: Option: Spezielle Kein-Wert-Behandlung
Siehe dazu Abschnitt P29.1.1.2, Erläuterung zur Box 10

Box: Option: Untersuchungseinheiten gewichten
Siehe P0.8.

Box: Optionen

```
verschiedene Programm-Optionen
```

Optionsbox geöffnet:

```
Loesche wieder diese Box
                    Optionen
Kovarianz
                    der Kalkül der kanon.Korr. wird angewandt auf
                    = Korrelation       Korrelationsmatrix
                    = Kovarianz         Kovarianzmatrix
                    Voreinstellung ist "Korrelation"
                    Bei der kanonischen Korrespondenzanalyse
                    sollte "Kovarianz" eingesetzt werden

                    Faktorenzahl
                    Zahl der Faktoren eventuell beschränken
                    keine Angabe = maximal mögliche Faktorenzahl

0                   Zwischenergebnisse ausgeben
                    = 1   Zwischenergebnisse werden ausgegeben
                    = 0   nicht
```

Eingabefeld 1: Auf welche Matrix soll der Kalkül der Korrespondenzanalyse angewendet werden. Es sollte die Kovarianzmatrix eingesetzt werden.

Eingabefeld 2: Eigenwert-Verfahren. Die Verfahren sind gleichwertig. Es können Vorzeichen-Umkehrungen auftreten, die geometrisch als Spiegelung zu

interpretieren sind und damit bedeutungslos sind.

Eingabefeld 3: Bleibt das Eingabefeld leer, dann extrahiert Almo die maximal mögliche Zahl von Faktoren. Diese ist gleich der kleineren Zahl der Ausprägungen der 1. oder der 2. Variablen (minus 1). Der Benutzer kann aber diese Faktorenzahl einschränken.

Eingabefeld 3: Es können Zwischenergebnisse angefordert werden.

Box: Option: "Aussehen" der auszugebenden Tabelle bzw. Matrix
Siehe P0.9.

Box: Grafik-Optionen
Siehe P0.10.

Box: Basisstatistiken ausgeben
Siehe dazu Abschnitt P29.1.1.2, Erläuterung zur Box 15.

P29.3.3 Maskenprogramm Prog29m6 mit Eingabe einer fertigen Tabelle

Liegt eine bereits gebildete 2-dimensionale Tabelle vor, dann kann man folgendes Maskenprogramm Prog29m6 verwenden:

Prog29m6.Msk
Bivariate Korrespondenzanalyse
(mit 2 nominalen Variablen)
mit eingegebener fertiger Tabelle

Zum Verhältnis der bivariaten zur
multiplen Korrespondenzanalyse --> **Hilfe 79**

Beispiel: Die nominalen Variablen, die analysiert werden sollen, sind:

Autokauf: Porsche, Mercedes, VW
Fahrstil: Aggresiv, normal, zurückhaltend

Es liegt z.B. folgende Tabelle vor:

		Fahrstil		
		aggresiv	normal	zurückh.
Auto	Porsche	8	2	1
	Mercedes	1	7	3
	VW	1	2	9

Die Tabelle muß in folgender Form geschrieben werden:

Auto	Fahrstil	Häufigkeit
Porsche	aggressiv	8
Porsche	normal	2
Porsche	zurückhalt	1
Mercedes	aggressiv	1
Mercedes	normal	7
Mercedes	zurückhalt	3
VW	aggressiv	1
VW	normal	2
VW	zurückhalt	9

Diese Tabelle steht am Programmende, wobei selbstverständlich die Ausprägungen in Ziffern geschrieben sind

Eine Korrespondenzanalyse über diese 2 nominalen Variablen könnte zu einem 2-dimensionalen Raum führen, in dem z.B. folgende Punkte dicht beieinder sind:
Mercedes, normaler Fahrstil
Porsche, aggresiver Fahrstil
VW, zurückhaltend

Grafik: 2-dimensionales Koordinatensystem
Siehe Handbuch, Abschnitt P29.4

Was ist ein Kurzprogramm ? --> Hilfe
Bedienung --> Hilfe

1

Namen für Variable

Name 1 = Fahrstil:aggressiv,normal,zurückhaltend;
Name 2 = Auto:Porsche,Mercedes,VW;

2

Analyse-Variable

Fahrstil — die 1. nominale Variable — Hilfe

Auto — die 2. nominale Variable — Hilfe

3,3 — Zahl der Ausprägungen der 1. und der 2. nominalen Variablen

3

Optionen

Faktorenzahl
Zahl der Faktoren eventuell beschränken
keine Angabe = maximal mögliche Faktorenzahl

0 — Zwischenergebnisse
=1 Zwischenergebnisse werden ausgegeben
=0 nicht

4

Option: "Aussehen" der auszugebenden Tabelle bzw. Matrix

5

Grafik-Optionen

6

Schreiben der Tabellenwerte

Schreiben Sie hier dahinter die Tabelle
In den vorderen Spalten stehen die Ausprägungen der Variablen
In der letzten Spalte stehen die Häufigkeiten
Zeilen mit Häufigkeit 0 müssen nicht geschrieben werden

Schalten Sie dazu die Schreibsperre aus

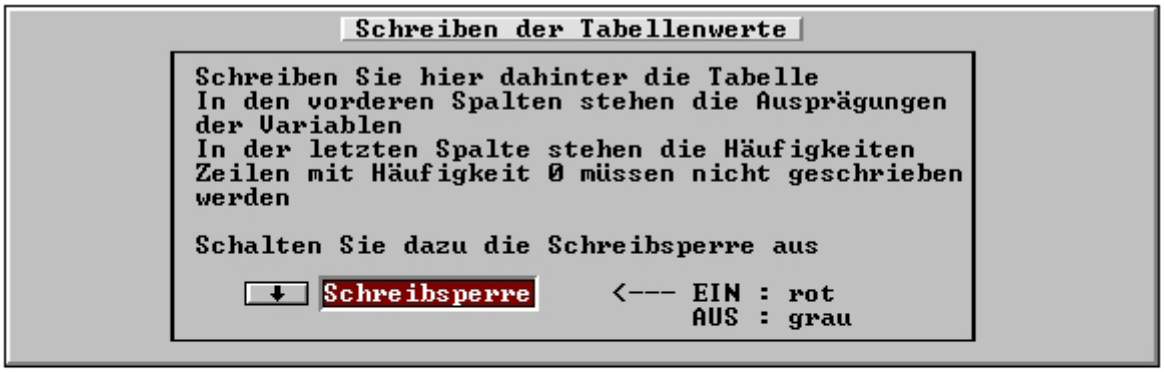
Schreibsperre <--- EIN : rot
AUS : grau

1	1	8
1	2	2
1	3	1
2	1	1
2	2	7
2	3	3
3	1	1
3	2	2
3	3	9

P29.3.4 Erläuterungen zu den Boxen

Die Boxen sind dieselben wie die oben in Abschnitt P29.3.2 bereits erläuterten. Wir erläutern hier nur die letzte Box

Box: Eingabe der Tabelle




```
1  1  8
1  2  2
1  3  1
2  1  1
2  2  7
2  3  3
3  1  1
3  2  2
3  3  9
```

Hinter der Box wird die Tabelle geschrieben. Der Benutzer muß zu diesem Zwecke zuerst in der Box auf den Knopf "Schreibsperre EIN/AUS" klicken.

Es liegt z.B. folgende Tabelle vor:

	Fahrstil		
	aggresiv	normal	zurückh.
	1	2	3
Auto Porsche 1	8	2	1
Mercedes 2	1	7	3
VW 3	1	2	9

Die Tabelle muß in folgender Form geschrieben werden:

Auto	Fahrstil	Häufigkeit
1 Porsche	1 aggressiv	8
1 Porsche	2 normal	2
1 Porsche	3 zurückhalt	1
2 Mercedes	1 aggressiv	1
2 Mercedes	2 normal	7
2 Mercedes	3 zurückhalt	3
3 VW	1 aggressiv	1
3 VW	2 normal	2
3 VW	3 zurückhalt	9

Selbstverständlich müssen die Ausprägungen in Ziffern geschrieben werden. D.h. die zu schreibende Tabelle ist folgende:

```
1  1  8
1  2  2
1  3  1
2  1  1
2  2  7
2  3  3
3  1  1
3  2  2
3  3  9
```

P29.3.8 Ausgabe aus Prog29m2 bzw. Prog29m6

Wir wollen die Ergebnis-Ausgabe aus den Bivariaten Korrespondenzanalyse an einem Beispiel von Greenacre (1984, S.55, S.123) zeigen. Siehe das Beispielprogramm **Greenac3.Alm**. Man findet es über das Menü "Almo/Liste aller Almo-Programme". Greenacre untersucht den Zusammenhang zwischen dem Status von Angestelltem und ihrem Verhalten als Raucher. Er erhält folgende Tabelle

		Raucher				Summe
		Nicht	Leicht	Mittel	Schwer	
Beruf	sen.Mana	4	2	3	2	11
	jun.Mana	4	3	7	4	18
	sen.Empl	25	10	12	4	51
	jun.Empl	18	24	33	13	88
	Secretar	10	6	7	2	25
Summe		61	45	62	25	193

Wir rechnen für diese Tabelle zuerst mit Maskenprogramm Prog10m3 (siehe Handbuch, Teil 3, Grundlegende Verfahren, Abschnitt P10 oder Dokument 1 Zwei- und drei-dimensionale Tabellierung.PDF) eine 2-dimensionale Tabellenanalyse. Sie liefert uns u.a. die zeilenweise prozentuierte Tabelle, die in der Korrespondenzanalyse "Zeilenprofile" genannt wird.

Tabelle zeilenweise prozentuiert

		Raucher				Summe
		Nicht	Leicht	Mittel	Schwer	
Beruf	sen.Mana	36.36	18.18	27.27	18.18	100.00
	jun.Mana	22.22	16.67	38.89	22.22	100.00
	sen.Empl	49.02	19.61	23.53	7.84	100.00
	jun.Empl	20.45	27.27	37.50	14.77	100.00
	Secretar	40.00	24.00	28.00	8.00	100.00
Summe		31.61	23.32	32.12	12.95	100.00

Prog 10m3 gibt auch die spaltenweise prozentuierte Tabelle aus, die in der Korrespondenzanalyse "Spaltenprofile" genannt wird.

Tabelle spaltenweise prozentuiert

		Raucher				Summe
		Nicht V2-1	Leicht V2-2	Mittel V2-3	Schwer V2-4	
Beruf	sen.Mana	6.56	4.44	4.84	8.00	5.70
	jun.Mana	6.56	6.67	11.29	16.00	9.33
	sen.Empl	40.98	22.22	19.35	16.00	26.42
	jun.Empl	29.51	53.33	53.23	52.00	45.60
	Secretar	16.39	13.33	11.29	8.00	12.95
Summe		100.00	100.00	100.00	100.00	100.00

Prog 10m3 ermittelt eine Signifikanz des Zusammenhangs zwischen den beiden Variablen von

```
Chi-Quadrat           = 16.4416
df                    = 2
Signifikanz(1-p)100   = 82.837%
```

Die Korrelation ist

```
Cramers V = 0.1685
```

Unser Programm zur Bivariaten Korrespondenzanalyse „Greenac3.Alm“ liefert folgende Ausgabe:

```
Faktor  Kanonische  Eigenwert   Wilks' Lambda  Chi-Quadrat  df  Signifikanz
        Korrelation (=Inertia)                                  (1-p)*100

 1       0.27342     0.07476       0.91559      16.57831    12   83.31271 %
 2       0.10009     0.01002       0.98957       1.97049     6    7.80784 %
 3       0.02034     0.00041       0.99959       0.07777     2    4.92449 %
-------------------------------------------------------------------------
Summe                0.08519
(=Pillais Spur)

Koeffizienten fuer Gesamtmodell
-------------------------------
multiple Korrelation                    0.168513
  beruhend auf Pillais Spur
  siehe Handbuch P 29.1.2,(10b),(10c)
F-Wert                                  1.373648
Freiheitsgrade Nenner =   12
               Zaehler=  564
Signifikanz: p                          0.173586
Signifikanz: (1-p)*100                 82.641437 %
Teststaerke von F                       0.765867
-----------------------------------------------
```

Die Zahl der Faktoren ist gleich der kleineren Ausprägungszahl minus 1.

Die Frage ist nun: Wieviele Faktoren sollen für die inhaltliche Interpretation der Ergebnisse verwendet werden? Wilks Lambda und die Signifikanz helfen hier kaum weiter. Bei den Praktikern der Korrespondenzanalyse ist hier die Neigung festzustellen, sich auf 2 Faktoren zu beschränken - weil man dann die Ergebnisse graphisch in 2 Dimensionen darstellen kann. Siehe dazu auch die Ausführungen in P30.8.2.2 ("Modellprüfgrößen für die Korrespondenzanalyse ").

Die multiple Korrelation (beruhend auf Pillais Spur) mit 0.1685 ist identisch mit Cramers V, wie wir es aus dem Tabellierungsprogramm Prog10m3 erhalten haben. Siehe oben, Abschnitt P29.3.5. Der geringfügige Unterschied bei der Signifikanz entstand durch unterschiedliche approximative Berechnungsverfahren in Prog10m3 und der Bivariaten Korrespondenzanalyse.

ALMO liefert folgende weiteren Ergebnisse:

Kanonische Gewichtszahlen fuer 1.(unabhaengige) Variablengruppe
(unstandardisiert)

			Faktor 1	Faktor 2	Faktor 3
Beruf	sen.Mana	V1-1	0.2405	1.9357	-3.4903
Beruf	jun.Mana	V1-2	-0.9471	2.4310	1.6574
Beruf	sen.Empl	V1-3	1.3920	0.1065	0.2535
Beruf	jun.Empl	V1-4	-0.8519	-0.5769	-0.1625
Beruf	Secretar	V1-5	0.7354	-0.7884	0.3973

Die unstandardisierten kanonischen Gewichtszahlen können in verschiedener Weise normalisiert werden. Diese Normalisierungen werden dadurch vorgenommen, daß je Faktor mit einer Konstanten multipliziert wird - insofern ist sie banal.

Gewichtszahlen aus Korrespondenzanalyse für unabhängige Dummies
("canonical normalization")
(=Kanon. Gewichtszahl * Wurzel aus kanon. Korrelation)

			Faktor 1	Faktor 2	Faktor 3
Beruf	sen.Mana	V1-1	0.1257	0.6123	-0.4977
Beruf	jun.Mana	V1-2	-0.4952	0.7690	0.2363
Beruf	sen.Empl	V1-3	0.7278	0.0337	0.0361
Beruf	jun.Empl	V1-4	-0.4455	-0.1825	-0.0231
Beruf	Secretar	V1-5	0.3845	-0.2494	0.0566

Gewichtszahlen aus Korrespondenzanalyse für unabhängige Dummies
("principal normalization")
(=Kanon. Gewichtszahl * kanon. Korrelation)

			Faktor 1	Faktor 2	Faktor 3
Beruf	sen.Mana	V1-1	0.0657	0.1937	-0.0709
Beruf	jun.Mana	V1-2	-0.2589	0.2433	0.0337
Beruf	sen.Empl	V1-3	0.3805	0.0106	0.0051
Beruf	jun.Empl	V1-4	-0.2329	-0.0577	-0.0033
Beruf	Secretar	V1-5	0.2010	-0.0789	0.0080

Standardisierte kanonische Gewichtszahlen
der 1.(unabhaengigen) Variablengruppe
(=kanon.Gew.zahl * Wurzel aus Diagonalglied der Streuungsmatrix)

			Faktor 1	Faktor 2	Faktor 3
Beruf	sen.Mana	V1-1	0.0557	0.4487	-0.8091
Beruf	jun.Mana	V1-2	-0.2754	0.7069	0.4819
Beruf	sen.Empl	V1-3	0.6137	0.0469	0.1117
Beruf	jun.Empl	V1-4	-0.4243	-0.2873	-0.0809
Beruf	Secretar	V1-5	0.2469	-0.2647	0.1334

Kanonische Strukturkoeffizienten
der 1.(unabhaengigen) Variablengruppe
(=Korrelation der Variablen mit kanonischen Faktoren)

			Faktor 1	Faktor 2	Faktor 3
Beruf	sen.Mana	V1-1	0.0591	0.4758	-0.8580
Beruf	jun.Mana	V1-2	-0.3037	0.7796	0.5315
Beruf	sen.Empl	V1-3	0.8342	0.0638	0.1519
Beruf	jun.Empl	V1-4	-0.7799	-0.5281	-0.1488
Beruf	Secretar	V1-5	0.2837	-0.3041	0.1532

Prozent erklärte (standardisierte) Varianz
in der 1.(unabhängigen) Variablengruppe
durch eigene kanonische Faktoren erklärt

Faktor 1	29.6101 %
Faktor 2	24.1971 %
Faktor 3	21.7511 %

durch kanonische Faktoren der anderen Variablengruppe erklärt

Faktor 1	2.2136 %
Faktor 2	0.2424 %
Faktor 3	0.0090 %

Kanonische Gewichtszahlen für 2.(abhängige) Variablengruppe
(unstandardisiert)

			Faktor 1	Faktor 2	Faktor 3
Raucher	Nicht	V2-1	1.4385	0.3046	0.0437
Raucher	Leicht	V2-2	-0.3637	-1.4094	-1.0817
Raucher	Mittel	V2-3	-0.7180	-0.0735	1.2617
Raucher	Schwer	V2-4	1.0744	1.9760	-1.2889

Gewichtszahlen aus Korrespondenzanalyse fuer abhaengige Dummies
("canonical normalization")
(=Kanon. Gewichtszahl * Wurzel aus kanon. Korrelation)

			Faktor 1	Faktor 2	Faktor 3
Raucher	Nicht	V2-1	0.7521	0.0963	0.0062
Raucher	Leicht	V2-2	-0.1902	-0.4458	-0.1542
Raucher	Mittel	V2-3	-0.3754	-0.0232	0.1799
Raucher	Schwer	V2-4	0.5618	0.6251	-0.1838

Gewichtszahlen aus Korrespondenzanalyse fuer abhaengige Dummies
("principal normalization")
(=Kanon. Gewichtszahl * kanon. Korrelation)

			Faktor 1	Faktor 2	Faktor 3
Raucher	Nicht	V2-1	0.3933	0.0304	0.0008
Raucher	Leicht	V2-2	-0.0994	-0.1410	-0.0220
Raucher	Mittel	V2-3	0.1963	0.0073	0.0256
Raucher	Schwer	V2-4	-0.2937	0.1977	-0.0262

Standardisierte kanonische Gewichtszahlen der 2.(abhaengigen) Variablengruppe
(=kanon.Gew.zahl * Wurzel aus Diagonalglied der Streuungsmatrix)

			Faktor 1	Faktor 2	Faktor 3
Raucher	Nicht	V2-1	0.6688	0.1416	0.0203
Raucher	Leicht	V2-2	-0.1538	-0.5959	-0.4573
Raucher	Mittel	V2-3	-0.3352	-0.0343	0.5891
Raucher	Schwer	V2-4	-0.3607	0.6635	-0.4327

Kanonische Strukturkoeffizienten der 2.(abhängigen) Variablengruppe
(=Korrelation der Variablen mit kanonischen Faktoren)

			Faktor 1	Faktor 2	Faktor 3
Raucher	Nicht	V2-1	0.9778	0.2071	0.0297
Raucher	Leicht	V2-2	-0.2005	-0.7771	-0.5964
Raucher	Mittel	V2-3	-0.4939	-0.0505	0.8680
Raucher	Schwer	V2-4	-0.4144	0.7622	-0.4971

Prozent erklärte (standardisierte) Varianz in der 2.(abhängigen) Variablengruppe durch

eigene kanonische Faktoren erklärt

Faktor 1 35.3060 %
Faktor 2 30.7617 %
Faktor 3 33.9322 %

durch kanonische Faktoren der anderen Variablengruppe erklärt

Faktor 1 2.6394 %
Faktor 2 0.3081 %
Faktor 3 0.0140 %

Gemeinsame Matrix aller Variablen
der unstandardisierten, nicht-normalisierten Gewichte

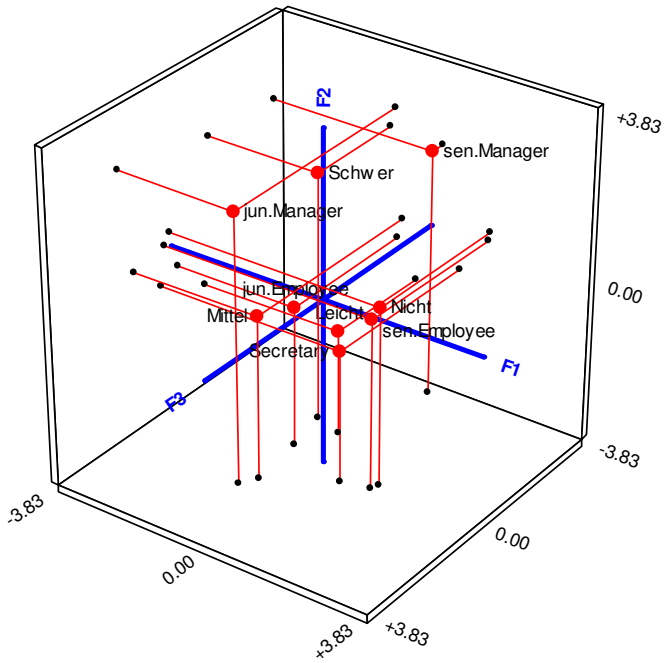
			Faktor 1	Faktor 2	Faktor 3
Beruf	sen.Mana	V1-1	0.2405	1.9357	-3.4903
Beruf	jun.Mana	V1-2	-0.9471	2.4310	1.6574
Beruf	sen.Empl	V1-3	1.3920	0.1065	0.2535
Beruf	jun.Empl	V1-4	-0.8520	-0.5769	-0.1625
Beruf	Secretar	V1-5	0.7355	-0.7884	0.3974
Raucher	Nicht	V2-1	1.4385	0.3047	0.0438
Raucher	Leicht	V2-2	-0.3637	-1.4094	-1.0817
Raucher	Mittel	V2-3	-0.7180	-0.0735	1.2617
Raucher	Schwer	V2-4	-1.0744	1.9760	-1.2889

********** MITTEILUNG
Beachte: Das Vorzeichen in einer Spalte k (=Faktor k) der gemeinsamen Matrix kann umgedreht werden. Dem entspricht geometrisch eine Spiegelung

In dieser Matrix sind die unstandardisierten, nicht-normalisierten Gewichte der unabhängigen Variablen (Beruf) und der abhängigen Variablen (Rauchertyp) zusammengefaßt.

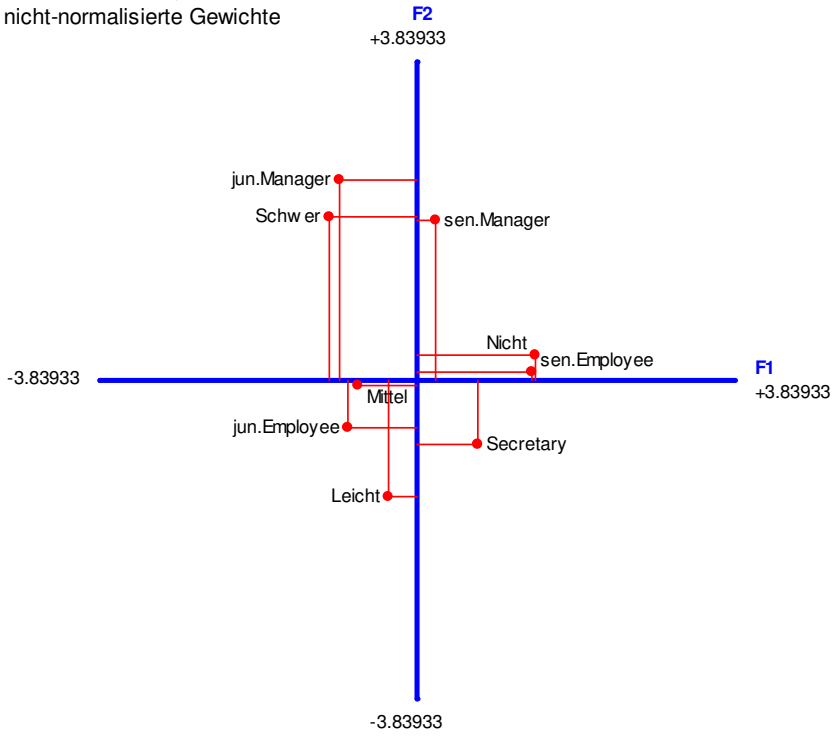
Wir stellen diese Matrix grafisch als 3-dimensionales xyz-Punktediagramm dar:

Korrespondenzanalyse
Unstandardisierte,
nicht-normalisierte Gewichte



Es ist ersichtlich, daß der 3. Faktor keine große Rolle spielt. Wir wählen deshalb eine 2-dimensionale grafische Darstellung. Das können wir auf 2 Wegen erreichen. Wir klicken in der linken Grafikleiste auf den Knopf „Diverse Positionen“. Almo präsentiert dann eine Auswahl von Positionen in die die Grafik transformiert werden kann. Wir wählen „F1-F2“. Das ist in der 2. Reihe das 2. Fenster. Eine schönere Darstellung erhalten wir, wenn wir auf den Knopf „Anderer Grafiktyp“ klicken. Almo präsentiert dann eine Auswahl anderer Darstellungsarten. Wir wählen „2-dim. Koordin.-System“. Das ist das 4. Bild in der 2. Reihe. Almo liefert dann folgende Grafik.

Korrespondenzanalyse
Unstandardisierte,
nicht-normalisierte Gewichte



Gemeinsame Matrix aller Variablen
der kanonisch normalisierten Gewichte

			Faktor 1	Faktor 2	Faktor 3
Beruf	sen.Mana	V1-1	0.1258	0.6124	-0.4977
Beruf	jun.Mana	V1-2	-0.4952	0.7691	0.2364
Beruf	sen.Empl	V1-3	0.7279	0.0337	0.0362
Beruf	jun.Empl	V1-4	-0.4455	-0.1825	-0.0232
Beruf	Secretar	V1-5	0.3846	-0.2494	0.0567
Raucher	Nicht	V2-1	0.7522	0.0964	0.0062
Raucher	Leicht	V2-2	-0.1902	-0.4459	-0.1543
Raucher	Mittel	V2-3	-0.3754	-0.0233	0.1799
Raucher	Schwer	V2-4	-0.5618	0.6251	-0.1838

********** MITTEILUNG
Beachte: Das Vorzeichen in einer Spalte k (=Faktor k) der gemeinsamen Matrix kann umgedreht werden. Dem entspricht geometrisch eine Spiegelung

Auch zu dieser Matrix zeichnet Almo ein xyz-Koordinatensystem, das wir hier nicht zeigen.

Gemeinsame Matrix aller Variablen
der "row principal" normalisierten Gewichte

			Faktor 1	Faktor 2	Faktor 3
Beruf	sen.Mana	V1-1	0.0658	0.1937	-0.0710
Beruf	jun.Mana	V1-2	-0.2590	0.2433	0.0337
Beruf	sen.Empl	V1-3	0.3806	0.0107	0.0052
Beruf	jun.Empl	V1-4	-0.2330	-0.0577	-0.0033
Beruf	Secretar	V1-5	0.2011	-0.0789	0.0081
Raucher	Nicht	V2-1	1.4385	0.3047	0.0438
Raucher	Leicht	V2-2	-0.3637	-1.4094	-1.0817
Raucher	Mittel	V2-3	-0.7180	-0.0735	1.2617
Raucher	Schwer	V2-4	-1.0744	1.9760	-1.2889

********** MITTEILUNG
Beachte: Das Vorzeichen in einer Spalte k (=Faktor k) der gemeinsamen Matrix kann umgedreht werden. Dem entspricht geometrisch eine Spiegelung

Auch zu dieser Matrix zeichnet Almo ein xyz-Koordinatensystem, das wir hier nicht zeigen.

Gemeinsame Matrix aller Variablen
der "column principal" normalisierten Gewichte

			Faktor 1	Faktor 2	Faktor 3
Beruf	sen.Mana	V1-1	0.2405	1.9357	-3.4903
Beruf	jun.Mana	V1-2	-0.9471	2.4310	1.6574
Beruf	sen.Empl	V1-3	1.3920	0.1065	0.2535
Beruf	jun.Empl	V1-4	-0.8520	-0.5769	-0.1625
Beruf	Secretar	V1-5	0.7355	-0.7884	0.3974
Raucher	Nicht	V2-1	0.3933	0.0305	0.0009
Raucher	Leicht	V2-2	-0.0995	-0.1411	-0.0220
Raucher	Mittel	V2-3	-0.1963	-0.0074	0.0257
Raucher	Schwer	V2-4	-0.2938	0.1978	-0.0262

********** MITTEILUNG
Beachte: Das Vorzeichen in einer Spalte k (=Faktor k) der gemeinsamen Matrix kann umgedreht werden. Dem entspricht geometrisch eine Spiegelung

Auch zu dieser Matrix zeichnet Almo ein xyz-Koordinatensystem, das wir hier nicht zeigen.

Gemeinsame Matrix aller Variablen
der "principal" normalisierten Gewichte

			Faktor 1	Faktor 2	Faktor 3
Beruf	sen.Mana	V1-1	0.0658	0.1937	-0.0710
Beruf	jun.Mana	V1-2	-0.2590	0.2433	0.0337
Beruf	sen.Empl	V1-3	0.3806	0.0107	0.0052
Beruf	jun.Empl	V1-4	-0.2330	-0.0577	-0.0033
Beruf	Secretar	V1-5	0.2011	-0.0789	0.0081
Raucher	Nicht	V2-1	0.3933	0.0305	0.0009
Raucher	Leicht	V2-2	-0.0995	-0.1411	-0.0220
Raucher	Mittel	V2-3	-0.1963	-0.0074	0.0257
Raucher	Schwer	V2-4	-0.2938	0.1978	-0.0262

```
                                        ******************** MITTEILUNG
   Beachte: Das Vorzeichen in einer Spalte k (=Faktor k) der gemeinsamen
Matrix kann umgedreht werden. Dem entspricht geometrisch eine Spiegelung
```

Auch zu dieser Matrix zeichnet Almo ein xyz-Koordinatensystem, das wir hier nicht zeigen.

```
Gemeinsame Matrix aller Variablen
der MCA-normalisierten Gewichte
```

			Faktor 1	Faktor 2	Faktor 3
Beruf	sen.Mana	V1-1	0.1919	1.4356	-2.4930
Beruf	jun.Mana	V1-2	-0.7557	1.8029	1.1838
Beruf	sen.Empl	V1-3	1.1107	0.0790	0.1811
Beruf	jun.Empl	V1-4	-0.6798	-0.4279	-0.1161
Beruf	Secretar	V1-5	0.5869	-0.5847	0.2838
Raucher	Nicht	V2-1	1.1478	0.2260	0.0313
Raucher	Leicht	V2-2	-0.2902	-1.0453	-0.7726
Raucher	Mittel	V2-3	-0.5729	-0.0545	0.9012
Raucher	Schwer	V2-4	-0.8573	1.4655	-0.9206

```
                                        ******************** MITTEILUNG
   Beachte: Das Vorzeichen in einer Spalte k (=Faktor k) der gemeinsamen
Matrix kann umgedreht werden. Dem entspricht geometrisch eine Spiegelung
```

Auch zu dieser Matrix zeichnet Almo ein xyz-Koordinatensystem, das wir hier nicht zeigen.

********** **Erläuterung:**
Dieses Ergebnis der Korrespondenzanalyse würden wir erhalten, wenn wir eine (faktorenanalytische) MCA mit Programm 30, z.B. mit dem Maskenprogramm Prog30m5 rechnen würden.

Vergleich zu SPSS:

In "SPSS Categories" (1990, Seite B-47) wird das Beispiel von Greenacre ebenfalls gerechnet. Die Ergebnisse aus Almo und SPSS stimmen selbstverständlich überein - mit dem einen Unterschied, daß beim 1. Faktor die Vorzeichen umgedreht sind. Dies entspricht geometrisch einer Spiegelung um die 2. Koordinatenachse, ist also irrelevant.

P29.3.9 Bivariate Korrespondenzanalyse und Regressionsanalyse

Betrachten wir folgende Häufigkeitstabelle

		Variable B	
		B1	B2
Variable A	A1	10	6
	A2	20	9
	A3	13	3

Wenn wir B als abhängige nominale und A als unabhängige nominale Variable betrachten, dann können wir mit folgendem „selbst geschriebenem“ Syntax-Programm

eine Regressionsanalyse für (in Dummies aufgelöste) nominale Variable rechnen.

```
Vereinbare
Variable = 10;
```

```
Anfang
Name1=A;
Name2=B;
Name3=Faelle;
Programm=20;
U_nominale_V = A;
A_quantitative_V = B;
Untergrenze A,B=1,1;
Obergrenze  A,B=2,2;
Matrix=Kovarianz
Gewichtung=Faelle;
Ende_Programmparameter;
```

```
Lese A,B,Faelle;
GEHE_IN_PROGRAMM
Gehe_zu Lese
Ende
```

```
1 1 10
1 2 6
2 1 20
2 2 9
3 1 13
3 2 3
```

Vergleichen wir die Ergebnisse

Korrespondenzanalyse	Regressionsanalyse
Kanon. Korrelation: 0.15225	multipler Korrelat.Koeff. 0.15225
Signifikanz (1-p)100 über Chi-Quadrat 48.89%	Signifikanz (1-p)100 über F-Wert: 48.90%
kanonische Gewichtszahlen (unstandardisiert):	Effekte von A:
A1 -1.1509	A1 -0.0799
A2 -0.2198	A2 -0.0153
A3 1.5493	A3 0.1076

Die Effekte sind (mit dem Wert 14.4 multipliziert) proportional zu den (unstandardisierten) kanonischen Gewichtszahlen.

Wir können also festhalten: Die Ergebnisse der Regressionsanalyse sind denen der Korrespondenzanalyse äquivalent. Dies gilt allerdings nur für den Fall, daß eine der beiden nominalen Variablen dichotom ist (die dann in der Regressionsanalyse als abhängige Variable betrachtet wird).

P29.3.10 Korrespondenzanalyse und Diskriminanzanalyse

In Abschnitt P29.2.11 "Nominale Variable als unabhängige Variable in der Diskriminanzanalyse" haben wir gezeigt, dass man eine Diskriminanzanalyse auch mit nominalen Variablen als unabhängige Variable rechnen kann - wenn auch gelegentlich darauf hingewiesen wird, dass das unstatthaft sei. Die Korrespondenzanalyse kann nun auch gerechnet werden als eine Diskriminanzanalyse mit einer abhängigen und einer unabhängigen nominalen Variablen. Dabei können 2 Analysen gerechnet werden, wobei die Stellung der beiden Variablen als abhängige bzw. unabhängige vertauscht werden. Die jeweils unabhängige nominale Variable muß dabei in Dummies aufgelöst werden - so wie wir dies im Beispielprogramm "DisUnom.Alm" gezeigt haben.

Die Ergebnisse sind selbstverständlich exakt dieselben, wie wenn eine Korrespondenzanalyse gerechnet würde.

Man kann also die Korrespondenzanalyse auch begreifen als eine Diskriminanzanalyse mit (nur) einer unabhängigen nominalen Variablen.

P29.5 Optimale Skalierung

Die kanonische Korrelationsanalyse kann auch dazu verwendet werden, um nominale Variable zu "skalieren". Gelegentlich wird in diesem Zusammenhang auch der Begriff "optimale Skalierung" verwendet. Hartung/Elpelt (1989, S. 286) verwenden die Bezeichnung "Lancaster-Skalierung" (wobei sie sich auf den Statistiker H.O. Lancaster beziehen). Sie zeigen die Lancaster-Skalierung an folgendem Beispiel (S.283,287):

		Augenfarbe		
		blau	braun	sonstig
Haarfarbe	blond	23	4	9
	anders	17	25	22

Wir rechnen für diese Tabelle mit Prog29m6 eine Korrespondenzanalyse. Siehe Abschnitt P29.3.3. Das Programm ist auch als Beispielprogramm „Opt_Skal.Alm“ in Almo vorhanden. Erreichbar über das Menü "Almo/Liste aller Almo-Programme"

Die Ergebnisse sind u.a.

Faktor	Kanonische Korrelation	Eigenwert (=Inertia)	Wilks' Lambda	Chi-Quadrat	df	Signifikanz (1-p)*100
1	0.38582	0.14886	0.85114	15.63380	2	99.92884 %
Summe (=Pillais Spur)		0.14886				

Gemeinsame Matrix aller Variablen
der unstandardisierten, nicht-normalisierten Gewichte

Haar	blond	V1-1	1.3333
Haar	nichtblo	V1-2	-0.7500
Augen	blau	V2-1	1.1610
Augen	braun	V2-2	-1.1991
Augen	sonstig	V2-3	-0.3762

Die unstandardisierten kanonischen Koeffizienten der beiden nominalen Variablen werden als Skalenwerte verstanden. Die kanonische Korrelation k=0.38582 ist die Korrelation der beiden mit diesen Skalenwerten gebildeten Linearkombinationen H und A

$H = 1.3333*h_1 - 0.7500*h_2$

$A = 1.1610*a_1 - 1.1991*a_2 - 0.3762*a_3$

Siehe hierzu P29. Gleichung 0a und 0b.

A = Augenfarbe
a1, a2, a3 = 0-1 kodierte Dummies für blau, braun, sonstig

H = Haarfarbe
h1, h2 = 0-1 kodierte Dummies für blond, anders.

Folgende Probleme sind zu beachten:

1. Wenn ein 2. oder weitere kanonische Faktoren auftreten, wird die inhaltliche Interpretation der Ergebnisse schwierig.
2. Auch wenn nur ein kanonischer Faktor auftritt ist zu berücksichtigen, daß die Skalenwerte von z.B. Augenfarbe andere sein können, wenn Augenfarbe mit einer anderen nominalen Variablen, z.B. Geschlecht tabelliert und der kanonischen Korrelationsanalyse unterworfen wird.

Die optimale Skalierung bzw. Lancaster-Skalierung kann also keinesfalls als Messmodell für eine Dimension verwendet werden.

Siehe zu diesen Problemen Hartung/Elpelt (1989, Kap.V).

Literatur

a. Literatur zu kanonischer Korrelation:

W.W. Cooley/P.R. Lohnes Multivariate Data Analysis, Wiley, 1971, Kap. 6
Hartung/Eipelt: Multivariate Statistik, München 1989, S.172ff.

b. Literatur zu Diskriminanzanalyse:

W.W. Cooley/P.R. Lohnes: Multivariate Data Analysis, Wiley, 1971
Tatsuoka, M.M.: Multivariate Analysis, Wiley 1971

c. Literatur zur Korrespondenzanalyse:

Greenacre: Theory and Applications of Correspondence Analysis, Academic Press, 1984
M. Greenacre/Jörg Blasius Correspondenzanalysis in the Social Sciences, Academic Press, London, 1994
M.J. Hartung/Elpelt Multivariate Statistik, Oldenbourg Verlag, München, 1989, S.369 ff.
Lebart/Morineau/Warwick: Multivariate descriptive statistical analysis, Wiley, 1984
D.F. Morrison Multivariate Statistical Methods, McGraw-Hill, New York, 1967
J. Reinecke/C. Tarnai (Hg.) Angewandte Klassifikationsanalyse in den Sozialwissenschaften, Waxmann, Münster 2000