



Korrelationsmatrix

Kurt Holm

Der allgemeine Korrelationskoeffizient Groß-Gamma

Heinrich Potuschak

Almo Statistik-System

www.almo-statistik.de

kurt.holm@jku.at

Autoren: em. Prof. Dr. Kurt Holm, Universität Linz, Österreich
Dr. Heinrich Potuschak, 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 Linearem 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.PDF (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

Kurt Holm

Korrelationsmatrix, Kovarianzmatrix, Quadratsummenmatrix für Variable beliebigen Messniveaus	4
P19.0 Einführung	4
P19.0.1 Messniveaus	4
P19.0.2 Die Korrelationskoeffizienten.....	6
P19.0.3 Der Groß-Gamma-Kalkül	7
P19.0.4 Hermann Denz: Die Einbeziehung ordinaler Variablen.....	7
P19.1 Die Almo-Programme und ihre Ergebnisse	8
P19.1.1 Programm-Maske für die Korrelation quantitativer Variabler Prog19am	8
P19.1.3 Ausgabe der Ergebnisse	12
P19.1.4 Programm-Maske für den allgemeinen Fall Prog19bm	20
P19.1.5 Exkurs: Das "paarweise Ausscheiden" zur Lösung des Kein-Wert-Problems	36
P19.1.7 Ausgabe der Ergebnisse aus Prog19bm	45
P19.1.8 Korrelationsmatrix mit Einschluß von Rangvariablen.....	48
P19.2 Korrelations-Matrix bei Vorhandensein nominaler Variabler	51
P19.3 Die Erzeugung einer Matrix partieller Streuungen.....	54
19.3.1 Errechnung einer Partialmatrix aus einer eingelesenen Korrelationsmatrix.	58
P19.4 Erzeugen einer Datei korrelierter Zufallsvariablen	59
P19.4.1 Programm-Maske Prog19cm: Korrelierte Zufallsvariablen.....	59
P19.4.2 Zum Kalkül	63
Literatur	63
Heinrich Potuschak: Der allgemeine Korrelationskoeffizient Groß-Gamma	64
Literatur zum Groß-Gamma-Koeffizienten.....	76

Kurt Holm

Korrelationsmatrix, Kovarianzmatrix, Quadratsummenmatrix für Variable beliebigen Messniveaus

P19.0 Einführung

Almo-Programm 19 rechnet eine Korrelationsmatrix für quantitative, ordinale und nominale Variable. Wahlweise kann auch eine Kovarianz- oder Quadratsummenmatrix berechnet werden. Das Programm wurde von **Kurt Holm** programmiert, der auch den vorliegenden Text verfasst hat. **Hermann Denz** hat jene Programmteile entwickelt, die es ermöglichen, ordinale Variable einzuführen. Er stellt dies in Abschnitt P19.0.4 dar.

Die Korrelationsmatrix wird nach einem Kalkül berechnet, der gelegentlich Groß-Gamma-Kalkül genannt wurde. Siehe dazu die ausführliche Darstellung von Hermann Denz in Abschnitt 19.0.4 und insbesondere von **Heinrich Potuschak** im Anhang.

Abhängig vom Messniveau der Variablen berechnet Almo folgende Korrelationskoeffizienten

Variable i	Variable k				
	quantitativ	Rang	ordinal	nominal-dichotom	nominal-polytom
quantitativ	r	namenlos	namenlos	punktbiserialer r	punktbiserialer r
Rang		Rho	namenlos	namenlos	namenlos
ordinal			tau-b	biserialer tau_b	namenlos
nominal-dichotom				Phi	Phi'
nominal-polytom					Cramers V

Die als "namenlos" bezeichneten Koeffizienten werden (wie alle anderen auch) nach dem "Groß-Gamma-Kalkül" berechnet. Alle quadrierten Korrelationskoeffizienten sind "proportional reduction of error"-Koeffizienten (PRE-Koeffizienten). Sie drücken den Anteil aus, um den sich die Fehlerstreuung in der Variablen k reduziert, wenn i als erklärende Variable eingeführt wird.

P19.0.1 Messniveaus

Almo unterscheidet die 3 Messniveaus: quantitativ, ordinal und nominal.

Nominale (oder qualitative) Variable

Beispiel: Beruf

	Code
Arbeiter	1
Angestellte	2
Beamte	3
Bauern	4
Selbständige	5

Den Ausprägungen werden Ziffern zugeordnet. Die Ziffern drücken keine Ordnungsrelation aus. Sie sind lediglich Kennziffern für die jeweilige Ausprägung. Die Zuordnung der Ziffern ist beliebig. So könnte etwa umgekehrt "Angestellter" mit 1 und "Arbeiter" mit 2 kodiert werden.

Bei den nominalen Variablen unterscheidet Almo im Korrelationsprogramm zwischen

- dichotomen Variablen (2 Ausprägungen)
- und polytomen Variablen (mehr als 2 Ausprägungen)

Ordinale Variable

Beispiel: Schulbildung:

	<u>Code</u>
Volksschulabschluß	1
Hauptschulabschluß	2
Gymnasium	3
Fachschule	3
Universitätsabschluß	4

Die Ziffern 1 bis 4, die den Ausprägungen der Schulbildung zugeordnet werden, drücken die Rangordnung im Bildungsniveau aus. 4 ist mehr als 3 und 3 ist mehr als 2 und 2 ist mehr als 1 - um wieviel mehr ist nicht bekannt. Die Differenzen zwischen den Rangziffern sind nicht bekannt. Die Ziffern drücken also die Relation "mehr" oder "weniger" oder "gleich" aus. Beachte: Gymnasium und Fachschule wurden oben gleichrangig mit 3 eingestuft.

Spezialfall: Rangvariable

Der Vollständigkeit halber wollen wir auch den Spezialfall der Rangvariablen (oder Rangwert-Variablen) darstellen. Wir werden auf diese in Abschnitt P19.1.8. ausführlich eingehen.

Betrachten wir ein Beispiel:

Die Variable "sportliche Leistung" sei eine ordinale Variable. Sie wird in folgender Weise kodiert:

		<u>Code</u>
Leistung:	hervorragend	1
	gut	2
	mittel	3
	schwach	4

7 Personen wurden in ihrer Leistung gemessen. In nachstehender Tabelle ist angegeben welche Werte sie erzielen konnten und welchen Rangplatz sie damit einnahmen.

Person	Wert in der ordinalen Variablen Leistung		Wert in der Rangvariablen
-----	-----		-----
1	hervorragend	1	1
2	gut	2	2.5
3	gut	2	2.5
4	mittel	3	5
5	mittel	3	5
6	mittel	3	5
7	schwach	4	7

Der "Wert in der Rangvariablen" ist sehr einfach der Rangplatz der Person, wenn

alle Personen nach ihrer Leistung hintereinander gestellt werden. Da manche Personen dieselbe Leistung erbringen, wie z.B. die Personen 2 und 3 wird eine "Rangteilung" vorgenommen. Person 2 und 3 teilen sich die Rangplätze 2 und 3. Der mittlere Wert ist 2.5. Die Personen 4, 5 und 6 teilen sich die Rangplätze 4, 5, 6. Der Wert in der Mitte ist 5. Siehe dazu Almo-Dokument Nr.3 "Nichtparametrische Verfahren", Abschnitt P8.2.7.

Quantitative Variable

Beispiel: Lebensalter

<u>Person</u>	<u>Code</u>
Gernot	24
Ariane	18
Roland	16

Die zugeordneten Zahlen drücken nicht nur eine Rangordnung aus wie bei den ordinalen Zahlen, sie geben auch die Distanz zwischen den Messobjekten an. Gernot ist 24 Jahre alt, Ariane 18 und Roland 16. Gernot ist also 6 Jahre älter als Ariane und 8 Jahre älter als Roland.

In den verschiedenen Almo-Programmen werden für die 3 Messniveaus teilweise unterschiedliche Koeffizienten berechnet.

Bei einigen Programmen, etwa im Masken-Programm Prog05m1, kann auch ein und dieselbe Variable als quantitativ und als nominal und als ordinal angegeben werden. Der Benutzer erhält dann für diese Variable die Koeffizienten, die Almo für diese 3 Messniveaus ermittelt.

P19.0.2 Die Korrelationskoeffizienten

Die meisten Korrelationskoeffizienten zwischen einer Variablen 1 und einer Variablen 2, die sich durch die Kombination der drei verschiedenen Messniveaus ergeben, sind bereits dadurch bekannt, dass sie unabhängig vom allgemeinen Groß-Gamma-Kalkül entwickelt wurden; dies trifft aber nicht für alle zu:

Variable 1	Variable 2		
	dichotom	ordinal	quantitativ
dichotom	1	2	3
ordinal		4	5
quantitativ			6

Siehe auch die Tabelle der Korrelationskoeffizienten in Abschnitt P19.1.8.

- 1: Der Groß-Gamma-Kalkül ergibt den Phi-Koeffizienten
- 2: Der Groß-Gamma-Kalkül ergibt den biserialen tau-b-Koeffizienten
- 3: Der Groß-Gamma-Kalkül ergibt den punktbiserialen Korrelationskoeffizienten
- 4: Der Groß-Gamma-Kalkül ergibt den tau-b-Koeffizienten
- 5: Der Groß-Gamma-Kalkül ergibt einen Koeffizienten, der keine eigene Bezeichnung besitzt – wir bezeichnen ihn einfach als Groß-Gamma-Koeffizienten
- 6: Der Groß-Gamma-Kalkül ergibt den Produkt-Moment-Korrelationskoeffizienten r .

In Abschnitt P19.1.8 werden wir die obige Tabelle noch durch eine Zeile/Spalte für Rangvariable erweitern.

Nominale **polytome** Variable, d.h. nominale Variable mit 3 und mehr Ausprägungen, werden zunächst in so viele dichotome Variable, d.h. 0-1 kodierte

Dummies, aufgelöst, wie sie Ausprägungen besitzen. Zur Auflösung nominaler Variablen in Dummies siehe Abschnitt 19.1.4 Erläuterung zu Eingabebox 7.

In einem 2. Rechenschritt werden diese dann über eine kanonische Korrelationsanalyse zusammengefasst. Dabei entstehen wieder aus der Literatur bekannte Korrelationskoeffizienten. Siehe Abschnitt P19.2.

P19.0.3 Der Groß-Gamma-Kalkül

Almo verwendet für die Berechnung aller in obiger Tabelle angegebenen Korrelationskoeffizienten den Groß-Gamma-Kalkül. Im Anhang geben wir eine Arbeit von Heinrich Potuschak wider, in der der allgemeine Groß-Gamma-Korrelationskoeffizient und seine Signifikanz dargestellt wird. Einen kurzen "Einblick" in die Konstruktion des Groß-Gamma-Korrelationskoeffizienten gibt Hermann Denz im nachfolgenden Abschnitt.

P19.0.4 Hermann Denz: Die Einbeziehung ordinaler Variablen

Der Programmteil in unserem Computer-Programm, der es ermöglicht ordinale Variable einzuführen wurde von Hermann Denz programmiert.

Sobald sich auch nur eine ordinale Variable auf Seiten der unabhängigen oder der abhängigen Variablen befindet, entspricht unsere Vorgangsweise ungefähr der Berechnungsweise des Kendall'schen tau-b.

Wir bilden Paare von Untersuchungspersonen und zwar so, dass jede Untersuchungsperson mit allen anderen $(N-1)$ Untersuchungspersonen zusammengebracht wird. Es entstehen so $(N^2-N)/2$ Paare. Für jedes Paar wird ermittelt, welche Differenz es in der jeweiligen Variablen aufweist.

Unsere neue Datenmatrix besteht nun nicht mehr aus individuellen Untersuchungspersonen, sondern aus den Paaren (als Zeilen) und aus den Variablen (als Spalten), wobei für diese nicht mehr die Werte der individuellen Untersuchungspersonen, sondern die Wertedifferenz für ein Paar eingetragen wird.

Der in Almo verwendete Berechnungs-Algorithmus erfordert es allerdings nicht, die Paare tatsächlich zu bilden. Es wird eine Zeit- und Speicherplatzsparende Berechnungsmethode verwendet.

Die Wertedifferenz für ein Paar ij wird nun bei **ordinalen** Variablen in folgender Weise ermittelt: Ist die Untersuchungsperson i in der betreffenden Variablen größer als die Untersuchungsperson j , dann wird dem Paar ij der Wert $+1$ zugewiesen. Ist die Untersuchungsperson i kleiner als j , dann erhält das Paar ij den Wert -1 . Sind i und j gleich groß, dann erhält das Paar ij den Wert 0 .

Bei **quantitativen** Variablen wird die Differenz des Werts der Untersuchungsperson i in der betreffenden Variablen minus dem Wert der Untersuchungsperson j verwendet.

Nominale Variable werden zuerst in 0-1 kodierte dummy Variable aufgelöst. Die letzte Dummy ist redundant und wird nicht verwendet. Die Dummies werden dann wie quantitative Variable behandelt. Ein Paar ij besitzt in einer dummy Variablen dann einen der drei Werte $1, 0, -1$. Für polytome Variable entstehen so mehrere Dummies und demgemäß auch mehrere Koeffizienten. Soll ein einziger Korrelationskoeffizient ermittelt werden, dann muss noch "Pillais Spur" berechnet werden. Siehe dazu ausführlich Abschnitt P19.2.

Wird eine in dieser Weise gebildete Korrelationsmatrix einer Regressionsanalyse unterzogen, dann gilt folgendes: Die Interpretation der Regressionsgleichung (mit

mindestens einer unabhängigen ordinalen Variablen) richtet sich immer nach dem Niveau der abhängigen Variablen:

Wenn die abhängige Variable quantitativ ist, dann ist der durch die rechte Gleichungsseite prognostizierte Wert der abhängigen Variablen Y' eine Differenz zwischen zwei Untersuchungspersonen i und j .

Ist die abhängige Variable, die 0-1 kodierte Dummy-Variablen einer nominalen Variablen, dann gibt der prognostizierte Wert Y' , der zwischen 0 und 1.0 liegen wird, die Wahrscheinlichkeit an, dass die beiden Untersuchungspersonen i und j in der abhängigen Variablen ungleich sind. Es sei daran erinnert, dass 0 Gleichheit des Paares bedeutet.

Ein Wert von beispielsweise $Y' = 0.7$ bedeutet, dass i und j mit 70%-iger Wahrscheinlichkeit in den abhängigen Dummy-Variablen verschieden sind. Die Regressionsgleichung ist als lineare Wahrscheinlichkeitsfunktion zu interpretieren.

Ist die abhängige Variable eine 1,0,-1 kodierte ordinale Variable, dann wird der prognostizierte Wert Y' zwischen +1 und -1 liegen. Er ist als Wahrscheinlichkeit des Abweichens von 0, d.h. der Wahrscheinlichkeit der Ungleichheit zu interpretieren, wobei das Vorzeichen die Richtung der Ungleichheit angibt. Ein Wert von beispielsweise $Y' = +0.7$ besagt, dass i mit 70%-iger Wahrscheinlichkeit in y größer ist als j . Ein Wert von $Y' = -0.3$ besagt, dass i mit 30%-iger Wahrscheinlichkeit kleiner ist als j . Auch in diesem Fall ist die Regressionsgleichung eine lineare Wahrscheinlichkeitsfunktion.

Die Interpretation der Regressionskoeffizienten ist folgende: Sie geben an, wie stark die Ungleichheit zwischen den Untersuchungspersonen i und j in der unabhängigen Variablen die Ungleichheit von i und j in der abhängigen Variablen bestimmt.

Beachte: Die Berechnung parametrischer Signifikanztests (F-Test, t-Test) wie sie in der Regressionsanalyse durchgeführt werden, ist bei unabhängigen ordinalen Variablen problematisch. Siehe dazu die Ausführungen von Potuschak.

P19.1 Die Almo-Programme und ihre Ergebnisse

Wir werden zunächst eine Programm-Maske für den einfachen Fall, dass alle zu korrelierenden Variablen quantitativ sind vortragen und danach dann eine Programm-Maske für den allgemeinen Fall, die dann auch noch eine Fülle von Optionen anbietet.

P19.1.1 Programm-Maske für die Korrelation quantitativer Variabler Prog19am

Sie finden das Programm durch Klick auf den Knopf „Verfahren“ und Auswahl von „Korrelation“.

Prog19am.Msk
 Korrelationsmatrix für quantitative Variable Hilfe
 Berechnet wird die Produkt-Moment-Korrelation r
 Siehe Handbuch, Abschnitt P19

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

 zeige = Namensdatei in Output zeigen
leer = nicht

4 Freie Namensfelder Hilfe

 erzeuge zusätzliche Namensfelder

5 Datei aus der gelesen wird Hilfe bei Datei-Problemen

 Format der Daten Hilfe
 der Datensatz enthält diese Variablen
Bei Format DIREKT schreiben Sie: alle_U

6 Wenn Dateiformat FIX oder Nicht-Standard-FREI Hilfe

7 zu korrelierende quantitative Variable Hilfe

8 Option: Umkodierungen und Kein-Wert-Angaben

9 Programmende

Erläuterungen zu den Eingabeboxen:

Eingabebox 1: Speicher für x Variable

Siehe dazu Almo-Dokument Nr. 0 "Arbeiten mit Almo", Abschnitt P0.1. Die Zahl der vereinbarten Variablen muss mindestens so hoch sein, wie die höchste im gesamten Programmtext vorkommende Variablennummer. Normalerweise ist dies die Nummer der letzten Variable des eingelesenen Datensatzes. Sie können die Zahl der vereinbarten Variablen aus Sicherheitsgründen auch höher setzen. Beispiel: Ein Datensatz aus Ihrer Datei umfasst 40 Variable. Dann geben Sie mindestens 40 an

Eingabebox 2: Option: Weitere Vereinbarungen

Weitere besondere Vereinbarungen sind sehr selten notwendig. Die Optionsbox wird nur geöffnet, wenn Almo dazu auffordert.

Siehe Dokument Nr. 0 "Arbeiten mit Almo.PDF", Abschnitt P0.2.

Eingabebox 3: Datei der Variablennamen

Variablennamen können - müssen aber nicht - in eine Datei gespeichert sein. Die Datei kann hier angefordert werden. Siehe P0.3.

Eingabebox 4: Freie Namensfelder

Einzelnen Variablen können hier Namen gegeben werden. Das ist nicht obligatorisch. Siehe P0.3.

Eingabebox 5: Datei aus der gelesen wird

Eingabebox 6: Wenn Dateiformat FIX oder Nicht-Standard-FREI

The screenshot shows a dialog box with three main sections. The top section is titled "Datei aus der gelesen wird" and contains a file path input field with the text ".\Testdat\TESTDAT.FRE" and a "Hilfe" button. Below this is a section for "Format der Daten" with a dropdown menu set to "frei" and another "Hilfe" button. The bottom section is titled "Wenn Dateiformat FIX oder Nicht-Standard-FREI" and contains a "Hilfe" button. The text "bei Datei-Problemen" is visible between the top and middle sections, and "der Datensatz enthält diese Variablen Bei Format DIREKT schreiben Sie: alle_v" is visible between the middle and bottom sections.

Siehe Dokument Nr. 0 "Arbeiten mit Almo.PDF", Abschnitt P0.4.

Geben Sie hier den Namen der Datei an, in dem sich die zu analysierenden Daten befinden. Schreiben Sie den vollen Pfad- und Dateinamen. Zulässig ist auch die Windows Kurzform, bei der bis einschliesslich Almo der Pfad durch einen Punkt vertreten werden kann, z.B. so ".\Testdat\Testdat.fre".

Die häufigsten Probleme, die Benutzer mit Almo haben treten an dieser Stelle auf. Lesen Sie insbesondere die Hilfe zum "Format". Klicken Sie dazu auf den Hilfeknopf bei der Format-Anweisung.

Eingabebox 7: Zu korrelierende quantitative Variable

The screenshot shows a dialog box with a title bar "zu korrelierende quantitative Variable" and a "Hilfe" button. Below the title bar is an input field containing the text "Leistung,Alter,Einkommen,Kinderzahl".

Geben Sie hier die Variablen an, die Sie interkorrelieren wollen. Wenn Sie den Variablen keine Namen gegeben haben, dann schreiben Sie hier einfach die Nummern (mit einem 'V' davor), also:

V5, V6, V7, V8

oder kurz: V5:8

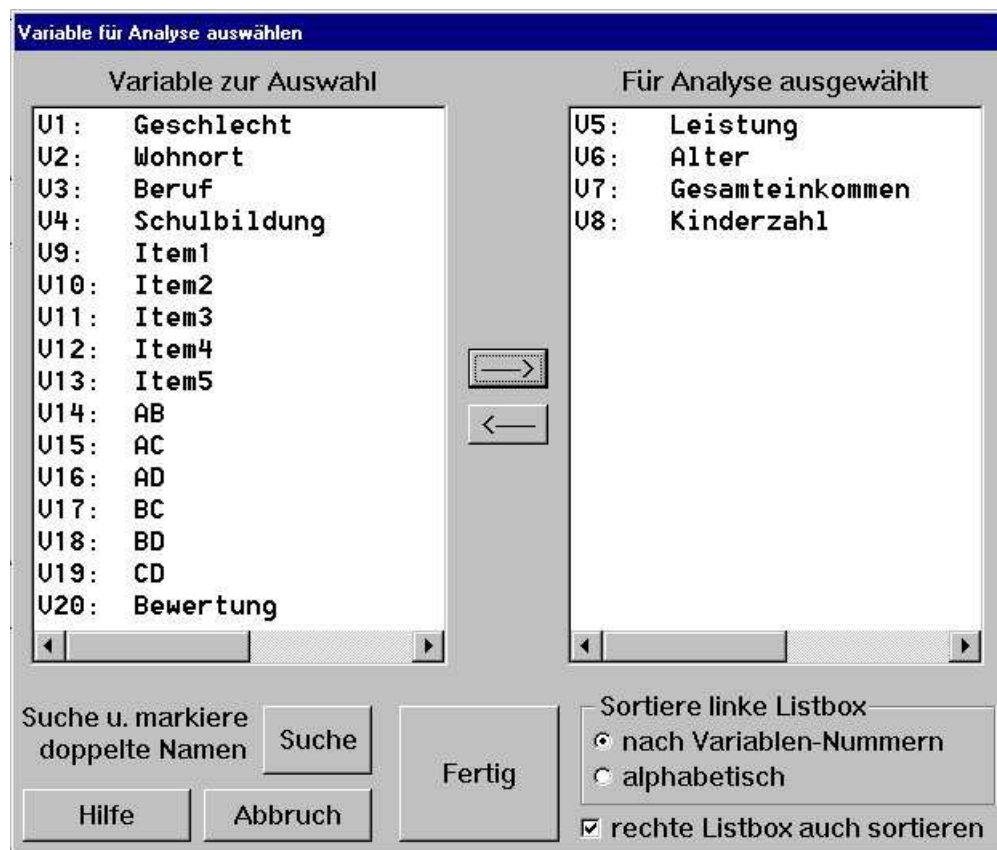
Die Variablen werden von Almo als quantitative behandelt. Almo errechnet den Produkt-Moment-Korrelationskoeffizienten r .

Hinweis:

Sie können hier auch dichotome Variable eingeben. Der Kalkül der Produkt-Moment-Korrelation liefert für 2 dichotome Variable den Phi-Koeffizienten und für die Korrelation zwischen einer quantitativen und einer dichotomen Variablen den punktbiserialen Korrelationskoeffizienten.

Eingabe der zu korrelierenden quantitativen Variablen durch Mausclick

Wenn Sie auf den Knopf mit den 2 Fenstersymbolen klicken, dann wird die Eingabebox "Variablen für Analyse auswählen" geöffnet. In Ihr geben Sie an, welche Variable korreliert werden sollen.



Almo bildet bei dieser Eingabe die Korrelationsmatrix der Variablen Leistung, Alter, Gesamteinkommen, Kinderzahl.

Wie man die Dialogbox "Variablen für Analyse auswählen" bedient

Klicken Sie auf eine Variable in der linken Listbox „Variable zur Auswahl“. Dann klicken Sie auf den Pfeilknopf. Die Variable wird dann in die rechte Listbox "Für Analyse ausgewählt" transportiert. Der „Transport“ kann auch in der umgekehrten Richtung erfolgen.

Die Knöpfe am unteren Rand der Dialogbox haben folgende Bedeutung:

SORTIERE linke Listbox nach Variablennummern

Die Variablen in der linken Listbox werden nach aufsteigenden Nummern hintereinander gestellt.

SORTIERE linke Listbox alphabetisch

Die Variablen in der linken Listbox werden alphabetisch hintereinander gestellt. Variable, die keine Namen besitzen werden an das Ende gestellt.

Rechte Listbox auch sortieren

Die Variablen in der rechten Listbox werden nach der gleichen Sortiermethode wie in der linken Listbox hintereinander gestellt.

Knopf **FERTIG**

Wenn Sie abschließend auf den Knopf **FERTIG** klicken, dann werden die Variablen in der rechten Listboxen in das Eingabefeld der Programm-Maske eingesetzt. Wenn die hintereinander gestellten Variablennamen zu lang würden, dann verwendet Almo automatisch Variablennummern.

Knopf **SUCHE**

Variablennamen müssen eindeutig sein. Sie dürfen nicht doppelt vorhanden sein. Mit Klick auf den Knopf **SUCHE** prüft Almo, ob Namen doppelt oder sogar mehrfach vorkommen. Diese Variablennamen werden dann durch 2 vorausgehende Unterstriche markiert, z.B. so:

V25: __Geschlecht

Diese Variablennamen dürfen dann nicht für die Analyse ausgewählt werden.

Eingabebox 8: Umkodierungen und Kein-Wert-Angabe

Die Variablen, die korreliert werden sollen, können umkodiert werden. Siehe Dokument Nr. 0 "Arbeiten mit Almo.PDF", Abschnitt P0.5.

Anmerkung zur "Kein-Wert-Angabe"

Betrachten wir ein Beispiel: In einer Umfrage wurden Personen u.a. nach ihrem Einkommen befragt. Viele haben die Antwort verweigert. Für die Variable des Einkommens liegt somit kein Wert vor. Beim Schreiben der Daten wurde das mit -1 kodiert. Das muss Almo mitgeteilt werden. In die geöffnete Optionsbox wird geschrieben:

Einkommen (-1=KeinWert)

Das Wort "KeinWert" ist ein Schlüsselwort das von Almo in entsprechender Weise verstanden wird. "KeinWert" kann groß oder klein oder mit Unterstrich (z.B. so: "Kein_Wert") geschrieben werden. Auch die Kurzform "KW" oder "kw" ist zulässig.

Eingabebox 9: Grafik-Optionen

Siehe P0.10.

P19.1.3 Ausgabe der Ergebnisse

Almo liefert folgende Ergebnisse:

Fuer Analyse ausgewaehlte Variable

V5 Leistung
V6 Alter
V7 Einkommen
V8 Kinderzahl

Zahl der insgesamt eingelesenen Einheiten 61
Zahl der in die Analyse einbezogenen Einheiten 61
=====

Zahl der Einheiten, die in die Analyse eingegangen sind
je Zelle der Streuungsmatrix

		Leistung	Alter	Einkomme	Kinderza
		V5	V6	V7	V8
Leistung	V5	61	61	61	61
Alter	V6	61	61	61	61
Einkomme	V7	61	61	61	61
Kinderza	V8	61	61	61	61

Die Zahl der Einheiten, je Zelle der Korrelationsmatrix ist gleich
Es sind keine Kein-Wert-Faelle aufgetreten

als Fallzahl wird verwendet: 61

******* Erläuterung:**

Die Zahl 61 im rechten oberen Eck bedeutet, dass 61 Untersuchungseinheiten für die Korrelation zwischen Leistung und Kinderzahl verwendet wurden. In unserem Beispiel sind für alle Korrelationen 61 Personen verwendet wurden. Hätten aber beispielsweise 2 Personen ihr Einkommen nicht angegeben, dann würde für die Korrelation von Einkommen mit den anderen Variablen nur 59 Personen zur Verfügung stehen.

Wären in obiger Matrix unterschiedliche Besetzungszahlen enthalten, dann errechnet also das harmonische Mittel dieser Zahlen. Die Ausgabe wäre dann beispielsweise folgende:

Die Zahl der Einheiten, je Zelle der Korrelationsmatrix ist verschieden
Es sind Kein-Wert-Faelle aufgetreten

als gemeinsame Fallzahl wird verwendet:
das harmonische Mittel

aus dem unteren Dreieck der obigen Tabelle der Zahl der Einheiten 52
=====

Standardabweichungen
(Standardabweichung ist mit n
nicht mit n-1 dividiert)

Leistung	V5	1.9086
Alter	V6	2.2093
Einkomme	V7	2.4850
Kinderza	V8	2.5278

******* Erläuterung:**

Wenn der Benutzer die Standardabweichungen mit n-1 im Nenner haben will, dann muß er die Programm-Maske Prog19bm in Abschnitt P19.1.4 verwenden und eine entsprechende Option setzen.

Mittelwerte

Leistung	V5	3.8852
Alter	V6	3.9344
Einkomme	V7	3.7049
Kinderza	V8	4.6557

Korrelations-Matrix

		Leistung	Alter	Einkomme	Kinderza
		V5	V6	V7	V8
Leistung	V5	1.0000	0.0371	-0.0244	-0.1645
Alter	V6	0.0371	1.0000	-0.1469	0.0371
Einkomme	V7	-0.0244	-0.1469	1.0000	-0.0997
Kinderza	V8	-0.1645	0.0371	-0.0997	1.0000

******* Erläuterung:**

Da alle Variable bei der Eingabe als quantitativ angegeben wurden, sind die berechneten Koeffizienten Produkt-Moment-Korrelationskoeffizienten.

Um die Koeffizienten inhaltlich zu erklären gehen wir eine Zeile nach der anderen durch. Betrachten wir zuerst die Zeile mit der Variablen "Leistung".

		Leistung	Alter	Einkomme	Kinderza
Leistung	V5	1.0000	0.0371	-0.0244	-0.1645

Die Leistung korreliert mit sich selbst natürlich mit 1, mit dem Alter sehr schwach positiv, mit dem Einkommen sehr schwach negativ und mit der Kinderzahl schwach negativ. Wie man nachfolgender Tabelle entnehmen kann, liegen alle Korrelationen weit unter der üblichen Signifikanzgrenze von 95 %.

Mindestgroesse des Produkt-Moment-Korrelationskoeffizienten r	bei Signifikanz (1-p)*100	fuer df=n-2=61-2=59
0.1664	80	
0.1866	85	
0.2127	90	
0.2296	92.5	

0.2520	95	

0.2869	97.5	
0.3267	99	
0.4139	99.9	

******* Erläuterung:**

Beispiel: Korrelationskoeffizienten, die mindestens eine Größe von 0.2520 haben sind mit mindestens 95 % signifikant.

zweiseitige Signifikanz $100 \cdot (1-p)$ der Korrelationen
(t-verteilt geprueft)

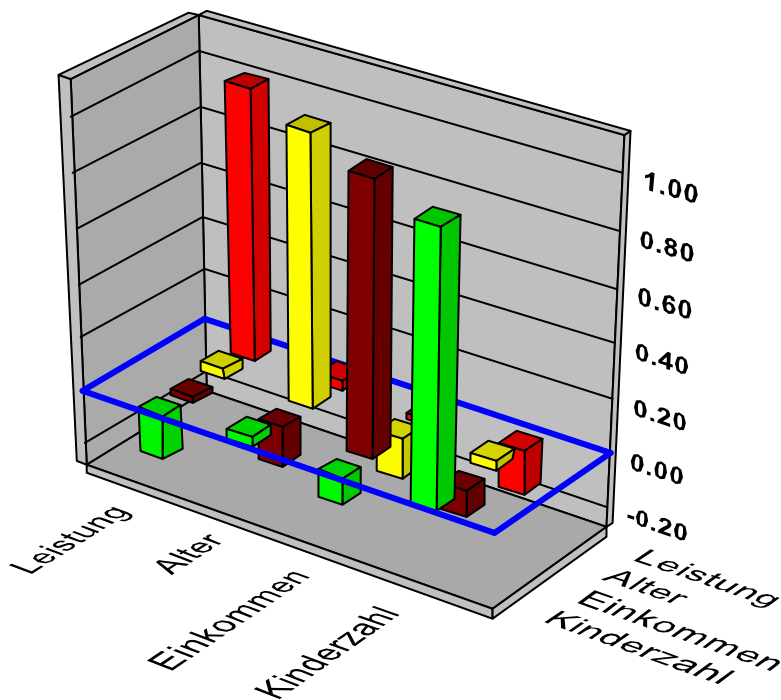
		Leistung	Alter	Einkommen	Kinderzah
		V5	V6	V7	V8
Leistu	V5	-	22.3452	14.8203	79.4788
Alter	V6	22.3452	-	74.1257	22.3217
Einkom	V7	14.8203	74.1257	-	55.5351
Kinder	V8	79.4788	22.3217	55.5351	-

******* Erläuterung:**

Beruhen die Korrelationskoeffizienten auf ungleichen Häufigkeiten (weil Werte fehlten und das "paarweise Ausscheiden" angewandt wurde), dann werden die Signifikanzen korrekt aus den jeweils vorhandenen Datenpaaren errechnet. Die tatsächliche Zahl der Untersuchungsobjekte je Zelle der Tabelle wird für die Ermittlung der Signifikanzen verwendet.

Almo erzeugt nun folgende Grafik:

Korrelations-Matrix



Almo zeichnet ein Balkendiagramm

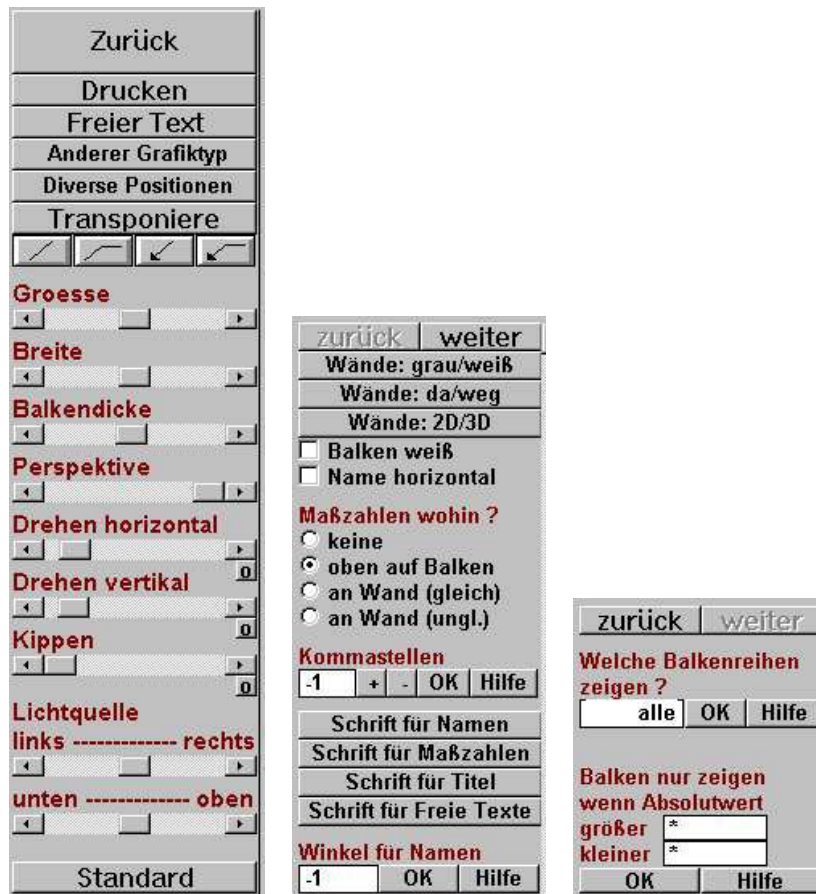
In die Grafik ist eine blau umrandete Ebene eingezeichnet. Dies ist die "Null-Ebene". Negative Korrelationen sind durch einen nach unten weisenden Balken dargestellt. Besonders übersichtlich ist dieser Versuch, eine Korrelationsmatrix grafisch zu verdeutlichen, nicht gelungen. Wir versuchen die Grafik zu "verbessern". Oberhalb der Grafik ist ein großer Knopf mit der Aufschrift "Grafik". Wird auf ihn geklickt, dann öffnet Almo seinen Grafik-Editor.

Links und rechts der Grafik befinden sich je eine Leiste mit Knöpfen und Schiebern.

linke Leiste

rechte Leiste

dritte Leiste



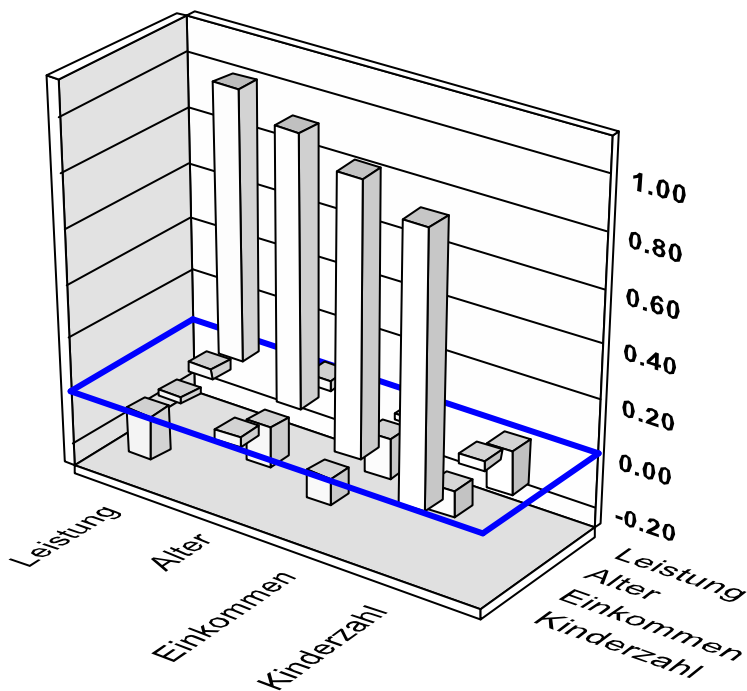
Wenn man in der rechten Leiste oben auf den Knopf „weiter“ klickt, dann wird noch eine 3. Leiste sichtbar, die wir oben auch abgebildet haben.

Die verschiedenen Elemente dieser Leisten werden in Teil 1 des Handbuchs (Bedienungsanleitung) in Abschnitt 10 erläutert.

Wir wollen hier nun einige vorteilhafte Manipulationen der Grafik vortragen:

Wir verschönern die Grafik noch, indem wir in der linken Leiste 1 x in den Schieber „Perspektive“ klicken. Wenn nicht farbig ausgedruckt werden kann, klickt man in der rechten Leiste noch auf "Wände grau/weiß" und "Balken weiß“. Wir sehen nun folgende Grafik.

Korrelations-Matrix



Die Grafik soll ausgedruckt werden. Das erreichen wir durch Klick auf den Knopf „Drucken“ in der linken Leiste. Es entsteht ein Druckbild höchster Qualität.

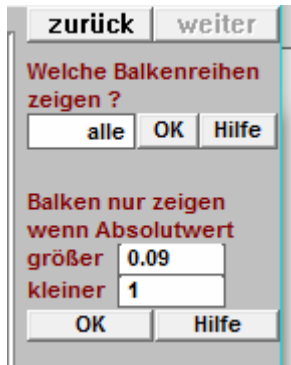
Es besteht auch die Möglichkeit das Bild nach MS-Word (bzw. eine andere kompatible Textverarbeitung) zu exportieren. Siehe dazu Handbuch, Teil 1, Bedienungsanleitung, Abschnitt 10.4.

Die Korrelationsmatrix ist symmetrisch. Es ist also kein Verlust, dass die Balken hinter dem hohen Balken (mit 1.0) in der Diagonale unsichtbar bleiben. Wenn man will, dann kann man die höchsten Balken oder den Balken für eine besonders wichtige Korrelation farblich hervorheben. Man klickt mit der linken Maustaste auf den Balken. Es erscheint dann eine kleine Auswahlbox mit dem Inhalt

„Andere Balkenfarbe“
„Nichts tun“

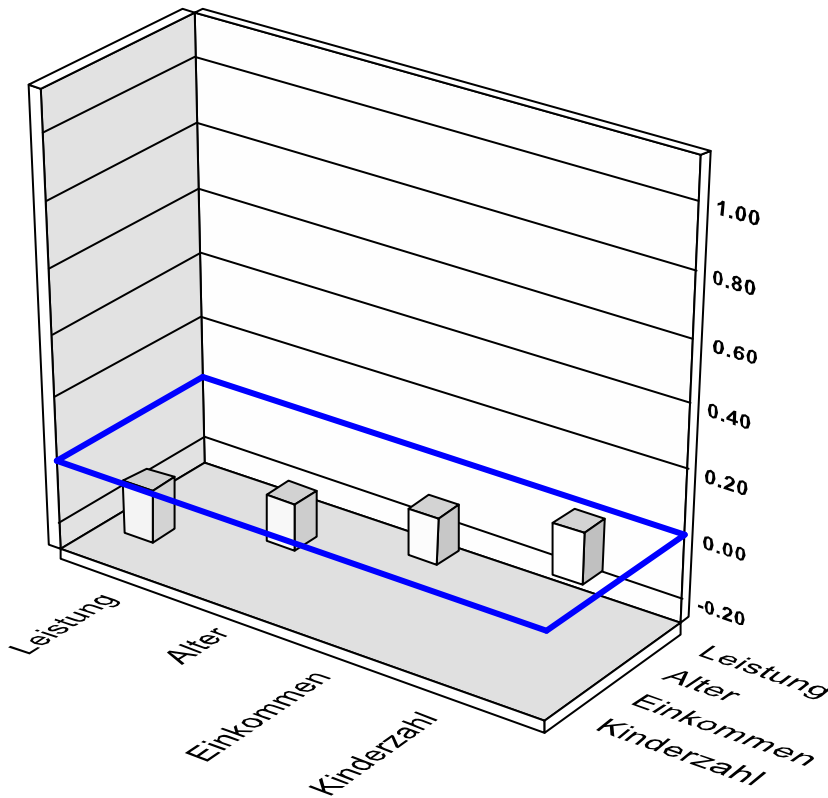
Erscheint diese Auswahlbox nicht, dann muß man an einer anderen Stelle auf den Balken klicken. Nach Klick auf „Andere Balkenfarbe“ wird die Farbauswahl-Box präsentiert, in der man sich dann für eine Farbe entscheidet.

In der 3. Grafikleiste (die man über Klick auf den Knopf „weiter“ in der rechten Grafikleiste sichtbar macht) findet man 2 Eingabefelder mit der Beschriftung „Balken nur zeigen, wenn Absolutwert größer“. Hier bietet es sich an, nur Balken für signifikante Korrelationen zu zeigen. Wir müßten also 0.252 eingeben. Unglücklicherweise sind in unserem Beispiel alle Korrelationen nicht signifikant. Wir tragen die Wert 0.09 und 1.0 ein und klicken auf OK. Damit erreichen wir, dass nur Korrelationen zwischen 0.1 und kleiner 1.0 abgebildet werden. Die Diagonalglieder bleiben dadurch leer.

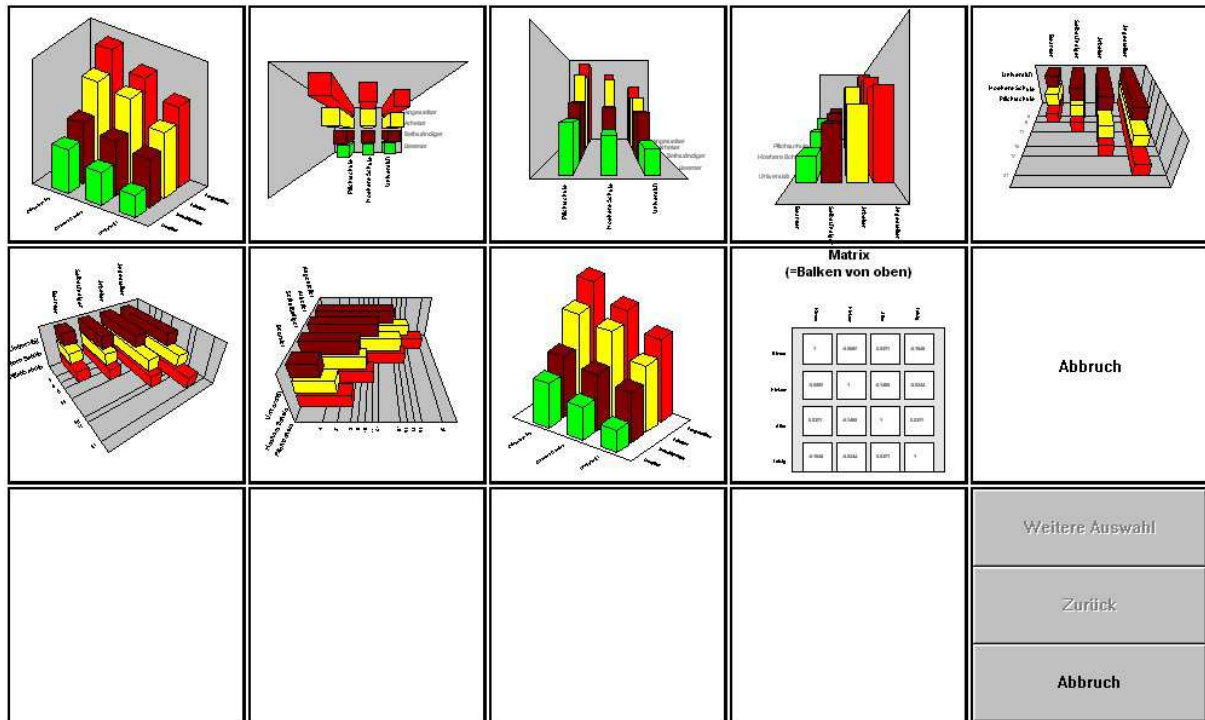


Es entsteht nun folgende Grafik:

Korrelations-Matrix



Die Gestaltungsmöglichkeiten des Almo-Grafik-Editors sind noch nicht erschöpft. Klicken Sie in der linken Leiste auf den Knopf „Diverse Positionen“. Es erscheint dann folgende Auswahl.



Klicken Sie auf das Bild „Matrix (= Balken von oben)“. Also erzeugt dann Balken, die (ohne Perspektive) von oben betrachtet werden. Es entsteht dabei der graphische Eindruck einer Matrix. Sie sehen folgendes:

Korrelations-Matrix

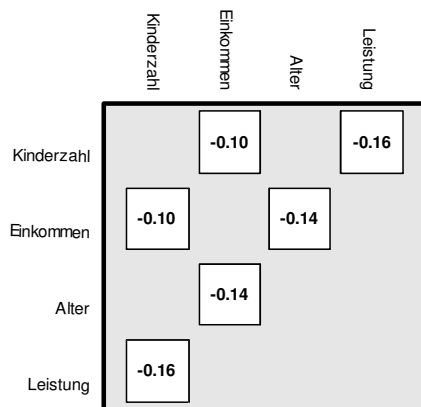
	Kinderzahl	Einkommen	Alter	Leistung
Kinderzahl	1.00	-0.10	0.04	-0.16
Einkommen	-0.10	1.00	-0.14	-0.02
Alter	0.04	-0.14	1.00	0.04
Leistung	-0.16	-0.02	0.04	1.00

Es ist sinnvoll mit dem Schieber „Größe“ und dem Schieber „Breite“ in der linken Leiste die Bildgröße anzupassen.

Wenn eine Korrelationsmatrix aus sehr vielen Variablen besteht, dann muß das Bild so stark vergrößert werden, dass es am linken und rechten Rand, aber auch oben und unten verschwindet. Das Bild kann jedoch gescrollt werden. Beim Ausdrucken werden die nicht sichtbaren Teile, sofern sie überhaupt noch auf das Papier passen, gedruckt. Auch der Titel „Korrelationsmatrix“ im linken oberen Eck wandert beim Drucken ganz nach oben auf das linke obere Eck des Papiers.

Wieder bestünde nun die Möglichkeit in der 3. Leiste, nur Balken zu zeigen mit einem Absolutwert größer 0.09 und kleiner 1.0. Es entsteht folgende Grafik:

Korrelations-Matrix



Beachten Sie, dass die Korrelationsmatrix symmetrisch ist. Die Werte unter- und oberhalb der Diagonalen entsprechen sich.

P19.1.4 Programm-Maske für den allgemeinen Fall Prog19bm

Nun ist es auch möglich, Variable mit verschiedenem Messniveau zu interkorrelieren. Die nominalen Variablen dürfen dabei auch polytom sein (d.h. 3 oder mehr Ausprägungen besitzen).

Sie finden das Programm durch Klick auf den Knopf „Verfahren“ und Auswahl von „Korrelation“.

Prog19bm.Msk
**Korrelationsmatrix für quantitative, ordinale und nominale Variable
mit Optionen**

Abhängig vom Messniveau der Variablen berechnet Almo folgende
Korrelationskoeffizienten:

	quant.	ordinal	nominal- dichotom	nominal- polytom
quantitativ	r	Groß-Gamma	punktbiser.r	Eta
ordinal		tau-b	biser. tau-b	Groß-Gamma
nominal-dichotom			Phi	Phi'
nominal-polytom				Cramers U

r = Produkt-Moment-Korrelation
Groß-Gamma = siehe Handbuch Teil 3, Abschnitt P19.0.3

Nominale Variable werden in "Dummies" aufgelöst. Das bedeutet:
Eine nominale Variable wird in ihre Ausprägungen aufgelöst. Diese
werden dann als 0, 1 kodierte "nominal-dichotome" Variable behandelt

Sind nominale Variable vorhanden dann berechnet Almo eine
2. Korrelationsmatrix. Dabei werden die Dummies der nominalen
Variablen über eine kanonische Korrelationsanalyse zusammengefasst.

Alle (quadrierten) Korrelationskoeffizienten zwischen i und k sind
"proportional reduction of error"-Koeffizienten (PRE-Koeffizienten)
Sie drücken den Anteil aus, um den sich die Fehlerstreuung in der
Variablen k reduziert, wenn i als erklärende Variable eingeführt
wird - oder umgekehrt (da die Korrel.koeffizienten symmetrisch sind)

Zusätzlich kann auch eine Matrix partieller Korrelationen gebildet
werden

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

1
Vereinbare Variable= ;

2 Option: Weitere Vereinbarungen - nur wenn Almo dazu auffordert

3
 "C:\Almo7\TESTDAT\Uarnamen.nam"
 zeige zeige = Namensdatei in Output zeigen
 leer = nicht

4

 erzeuge zusätzliche Namensfelder

5

bei Datei-Problemen

Format der Daten

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

6

Wenn Dateiformat FIX oder Nicht-Standard-FREI

7

quantitative Variable

nominale Variable
 sie werden in Dummies aufgelöst

aus den nominalen Variablen werden auch
 Interaktionsvariable x. Ordnung gebildet
 0 = Keine Interaktionsvariable bilden

ordinale Variable

8

- 9 Option: Nenner für Standardabweichung und Kovarianzmatrix
- 10 Option: Ein- und Ausschliessen von Untersuchungseinheiten
- 11 Option: Umkodierungen und Kein-Wert-Angaben
- 12 Option: Spezielle Kein-Wert-Behandlung
- 13 Option: Untersuchungseinheiten gewichten
- 14 Option: Partielle Korrelationsmatrix bilden
- 15 Option: Schreibe errechnete Matrix in Datei
- 16 Option: "Aussehen" der auszugebenden Tabelle bzw. Matrix
- 17 Option: Nur Koeffizienten ab einer bestimmten Größe zeigen
- 18 Option: Programm-Optionen lt. Handbuch
- 19 Grafik-Optionen
- 20 [Programmende](#)

Erläuterungen zu den Eingabeboxen:

Eingabebox 1: Speicher für x Variable

Siehe Dokument Nr. 0 "Arbeiten mit Almo.PDF", Abschnitt P0.1.

Eingabebox 2: Option: Weitere Vereinbarungen

Siehe Dokument Nr. 0 "Arbeiten mit Almo.PDF", Abschnitt P0.2.

Eingabebox 3: Datei der Variablennamen

Siehe P0.3.

Eingabebox 4: Freie Namensfelder

Siehe P0.3.

Eingabebox 5: Datei aus der gelesen wird

Eingabebox 6: Datei aus der gelesen wird. Teil II

Siehe P0.4.

Eingabebox 7: Analyse-Variable: Zu korrelierende Variable

Zu korrelierende Variable Hilfe

quantitative Variable

↔ □□ **Alter, Einkommen**

nominale Variable Hilfe
sie werden in Dummies aufgelöst

↔ □□ **Geschlecht, Beruf**

↑↓ 0 Hilfe

aus den nominalen Variablen werden auch
Interaktionsvariable x. Ordnung gebildet
0 = Keine Interaktionsvariable bilden

ordinale Variable Hilfe

↔ □□ **Leistung**

Geben Sie hier die Variablen an, die Sie interkorrelieren wollen. Sie können eingeben:

im *Eingabefeld 1*: Quantitative Variable

im *Eingabefeld 2*: Nominale Variable

im *Eingabefeld 3*: Interaktionsvariable (gebildet aus den nominalen Variablen)

im *Eingabefeld 4*: Ordinale Variable

Die Unterschiede zwischen den 3 Messniveaus quantitativ, nominal und ordinal haben wir bereits oben in P19.0.1 dargestellt.

Sie können die Analyse-Variablen „von Hand“ in die Eingabefelder schreiben oder Sie klicken auf den Knopf mit den 2 kleinen symbolischen Fenstern. Almo öffnet dann die Dialogbox „Variable für Analyse auswählen“. In dieser können Sie die Variable, die in die Eingabefelder eingeschrieben werden sollen per Mausklick selektieren. Siehe die ausführliche Beschreibung dieser Dialogbox in P0.11.

Zu Eingabefeld 2:

Die nominalen Variablen (und ihre Interaktionen) werden in 0-1 kodierte Dummies aufgelöst. Jede einzelne Ausprägung wird zu einer Dummy-Variablen.

Zum Begriff der „Dummy-Variablen“

Betrachten wir ein Beispiel: Die nominale Variable sei "Beruf" mit den Ausprägungen Arbeiter, Angestellter, Beamter, Selbständiger.

Sie wird in vier Nominaldummies a1 , a2 , a3 , a4 aufgelöst. Dafür gibt es 2 Verfahren:

(1) die 0,1-Kodierung und

(2) die 0,1,-1 -Kodierung, die wir hier nicht darstellen. Siehe dazu das Almo-Dokument Nr. 13 "Allgemeines lineares Modell", Teil 1, Abschnitt P20.3.

	0,1 - Kodierung			
	dummy Variable			
	a1	a2	a3	a4
Arbeiter	1	0	0	0
Angestellter	0	1	0	0
Beamter	0	0	1	0
Selbständiger	0	0	0	1

Eine Untersuchungsperson sei Arbeiter. Wir müssen ihr nun in den 4 Dummies Werte zuweisen. Bei der 0,1-Kodierung erhält sie in der Dummy a1 den Wert 1. In den Dummies a2 , a3 , a4 erhält sie den Wert 0. Die Auflösung der nominalen Variablen in Dummy-Variable wird automatisch von unserem Computer-Programm besorgt.

Die Auflösung der Interaktionen in Dummy-Variable

Wir wollen annehmen, wir hätten nach dem Beruf eine zweite unabhängige nominale Variable, die Schulbildung, mit folgenden Ausprägungen und folgenden Dummy Variablen b1, b2, b3 .

	0,1 - Kodierung		
	b1	b2	b3
einfache Schulbildung	1	0	0
höhere Schulbildung	0	1	0
Universität	0	0	1

Wenn wir nun die Interaktion zwischen Beruf und Schulbildung in dummy Variable auflösen wollen, dann müssen wir zunächst die beiden Sätze von dummy Variablen gegeneinander tabellieren. Auf diese Weise entstehen die "multiplikativen Dummies" der Interaktion AB.

	b1	b2	b3
a1	a1b1	a1b2	a1b3
a2	a2b1	a2b2	a2b3
a3	a3b1	a3b2	a3b3
a4	a4b1	a4b2	a4b3

Die Kodierungszahl der jeweiligen multiplikativen Dummy $a_i b_j$ erhalten wir sehr einfach aus der Multiplikation der Kodierungszahl der Dummies a_i und b_j , also $a_i * b_j$. Betrachten wir ein Beispiel: Ein Befragter sei Arbeiter und besitze eine einfache Schulbildung. Dann hat er in der multiplikativen Dummy $a_1 b_1$ den Wert 1 und in allen anderen den Wert 0.

In entsprechender Weise werden auch die multiplikativen Dummies von 3er-Interaktionen und Interaktionen höherer Ordnung gebildet.

Almo errechnet dann noch eine **zweite Korrelationsmatrix**, bei der die Dummies der nominalen Variablen (nicht aber der Interaktionen) über eine kanonische Korrelationsanalyse zusammengefaßt werden. Bei der Interpretation der Almo-Ausgabe werden wir das ausführlich darstellen.

Hinweis: Sie können dichotome nominale Variable auch als quantitative eingeben. Der Kalkül der Produkt-Moment-Korrelation, der in Almo bei quantitativen Variablen verwendet wird, liefert für 2 dichotome Variable den Phi-Koeffizienten und für die Korrelation zwischen einer quantitativen und einer dichotomen Variablen den punktbiserialen Korrelationskoeffizienten. Interaktionen sind dann allerdings nicht möglich.

Zu Eingabefeld 3

Eine Besonderheit von Prog19bm ist es, dass auch die Interaktionen zwischen den nominalen Variablen korreliert werden können. Das ist allerdings nicht immer sinnvoll.

Betrachten wir ein Beispiel mit 3 nominalen Variablen A, B, C. Aus unseren Testdaten verwenden wir dafür die Variablen die Variablen V1, V3, V4.

Wird als Interaktion in das Eingabefeld 3 geschrieben, dann werden alle Interaktionen bis zur 3. Ordnung gebildet, also

Interaktionen 2. Ordnung: AB, AC, BC
 Interaktionen 3. Ordnung: ABC

wird in das Eingabefeld 2 geschrieben, dann werden die Interaktionen höherer Ordnung, also 3. Ordnung, nicht gebildet.

Es besteht auch die Möglichkeit nur einige ausgewählte Interaktionen zu bilden, z.B. die Interaktionen

AC BC

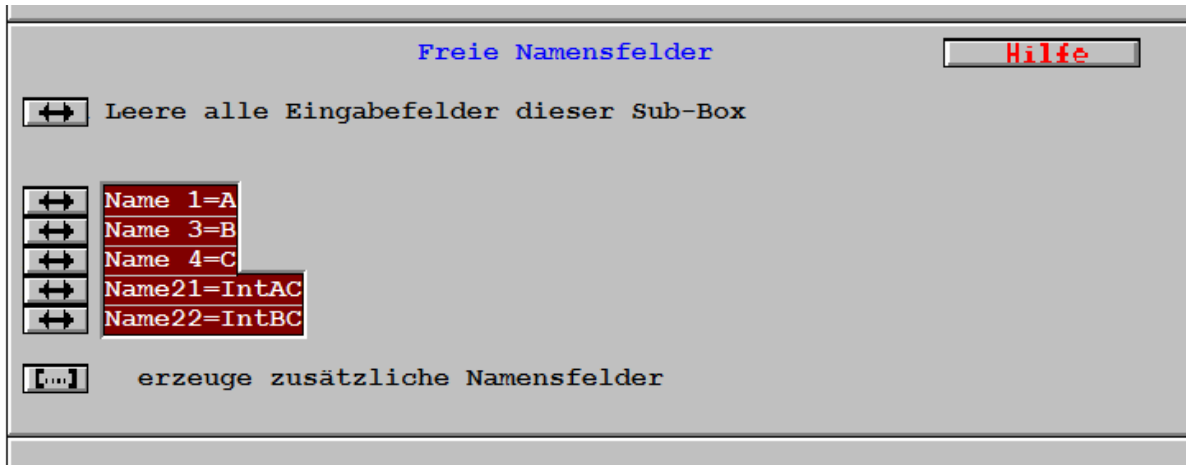
Die Vorgehensweise ist dann (etwas kompliziert) folgende:

1. In der Eingabebox "Freie Namensfelder" schreiben Sie die 2 Namen

```
Name 1 =A;          # Name für V1 #  
Name 3 =B;          #           V3 #  
Name 4 =C;          #           V4 #  
Name 21 =IntAC;     # Name für Interaktion A mal C #  
Name 22 =IntBC;     #           B mal C #
```

(Die Namen können Sie wählen, wie Sie wollen.) Für die Interaktionsvariablen verwenden Sie dabei hinter "Name ..." Variablennummern die frei sind (am besten Nummern, die höher sind als die Nummer der letzten eingelesenen Variablen).

Beachte: In der 1. Eingabebox der Programm-Maske "Vereinbare Variable" müssen Sie dann mindestens 22 angeben.



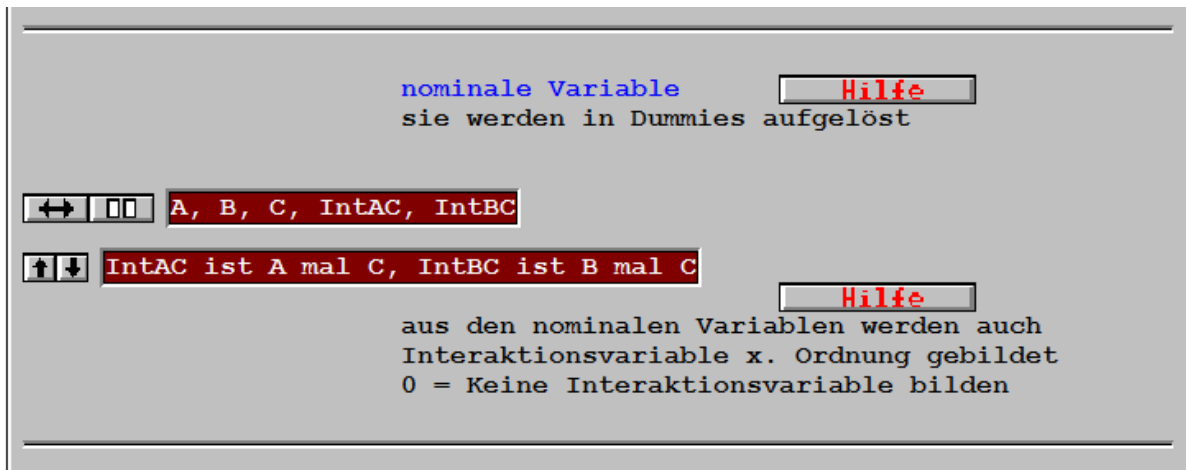
2. In der Eingabebox "Analysevariable: Zu korrelierende Variable" geben Sie als nominale Variable zusätzlich IntAC und IntBC an. Im Eingabefeld steht dann (in unserem Beispiel)

A, B, C, IntAC, IntBC

3. Im Eingabefeld für die Interaktionen schreiben Sie:

IntAC ist A mal C, IntBC ist B mal C

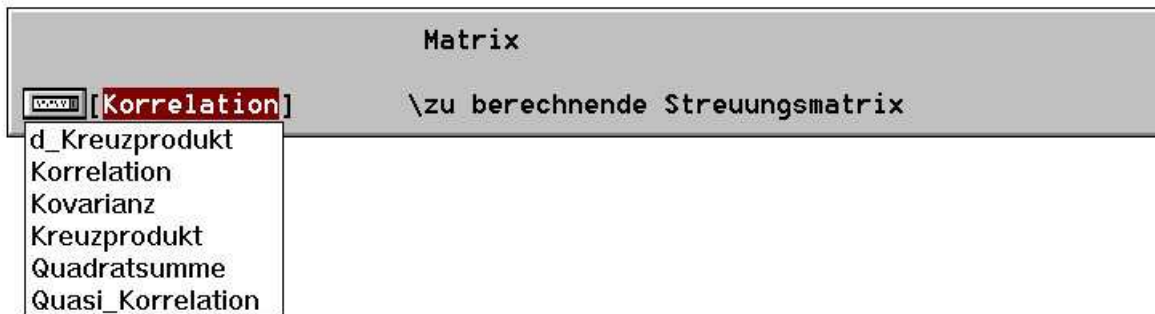
Sie müssen also angeben, aus welchen nominalen Variablen die Interaktionsvariablen gebildet werden sollen. Sie können das Wort "mal" auch durch den Stern * ersetzen, also z.B. IntAC ist A*C



Zu Eingabefeld 4:

Sobald auch nur eine ordinale Variable in die Analyse eingeführt wird, werden die quantitativen Variablen als diskrete Variable behandelt. Für den Benutzer nicht sichtbar, werden ALMO-intern alle Variable (also auch die quantitativen) in so viele Dummy-Variable aufgelöst, wie sie Ausprägungen besitzen. Die ALMO-internen Matrizen sind in ihrer Größe also nicht mehr durch die Zahl der Variablen sondern durch die Zahl der aufsummierten Ausprägungen bestimmt. *Der Speicherbedarf wird somit drastisch erhöht!*

Eingabebox 8: Zu berechnende Streuungsmatrix



Mit Programm-Maske Prog19bm wird man normalerweise eine Korrelationsmatrix errechnen. Es ist aber auch möglich anstelle dieser zu ermitteln.

- oder die Matrix der Quasi_Korrelationen
- oder die Matrix der Kovarianzen
- oder die Matrix der Quadratsummen
- oder die Matrix der Kreuzprodukte
- oder die Matrix der d_Kreuzprodukte

Sind ordinale Variable vorhanden, sind die Matrizen „Kreuzprodukt“ und „d_Kreuzprodukt“ nicht möglich.

Klicken Sie auf den Knopf vor dem Eingabefeld und selektieren Sie eine Angabe.

Zur Matrix der **Quasi_Korrelationen**

Diese Matrix entsteht, wenn die zu korrelierenden Variablen fehlende Werte aufweisen und der Benutzer wählt

- (1) als Matrix: Quasi-Korrelation
- (2) als Kein-Wert-Behandlung: Paarweises Ausscheiden. Siehe dazu die ausführliche Darstellung im nächstenAbschnitt

Beim „paarweisen Ausscheiden“ als Methode fehlende Werte zu behandeln lautet die Formel für die Korrelationsmatrix:

$$r_{x,y} = \frac{\text{cov}(xy)}{s_x * s_y}$$

mit:

$\text{cov}(xy)$ = Kovarianz von x und y

s_x, s_y = Standardabweichung von x bzw. y

Beim paarweisen Ausscheiden wird die Standardabweichung s_x und s_y nur aus den Daten ermittelt, die auch in die Berechnung der Kovarianz eingehen. Im Falle der **Quasikorrelationsmatrix** setzen wir in obiger Formel für s_x und s_y die Standardabweichung der Variablen x bzw. y ein, wie wir sie aus allen für x vorhandenen Daten (ohne Berücksichtigung der Kein-Wert-Fälle in y) erhalten bzw. wie wir sie aus allen für y vorhandenen Daten (ohne Berücksichtigung der Kein-Wert-Fälle in x) erhalten. Da $\text{cov}(xy)$ einerseits, s_x und s_y andererseits aus verschiedenen Datenmengen berechnet werden, ist r nicht exakt normiert. In extremen Ausnahmefällen kann r größer als 1.0 werden. Wir nennen ihn deswegen "Quasi-Korrelationskoeffizient". Siehe auch unsere ausführliche Darstellung im Handbuch zum Almo-Data-Mining, Abschnitt P45.12.4.

Siehe dazu das Handbuch zum Allgemeinen linearen Modell, Abschnitt P20.8.5.3.1

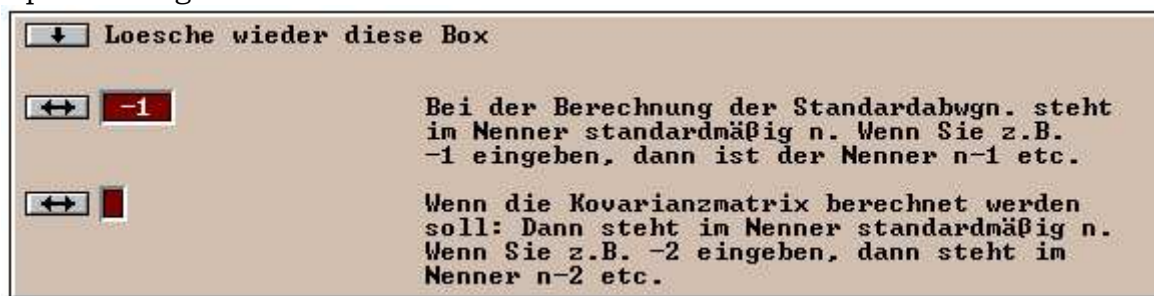
Zur Matrix der "**d_Kreuzprodukte**"

Die "d_Kreuzprodukte" sind die *durchschnittlichen* Kreuzprodukte, also die durch n dividierten Kreuzprodukte

Eingabebox 9: Option: Nenner für Standardabweichung und Kovarianzmatrix



Optionsbox geöffnet:



Bei der Berechnung der Standardabweichung bzw. der Varianz setzt Almo im Nenner n ein (n=Gesamtzahl aller Untersuchungseinheiten). Soll im Nenner n-1 stehen, dann schreiben Sie in das Eingabefeld -1. Wenn Sie z.B. -2 schreiben, dann ist der Nenner n-2 etc. Wenn Sie 0 schreiben oder das Editfeld leeren, dann ist der Nenner n. Für den Nenner der Kovarianzmatrix gilt entsprechendes.

Eingabebox 10: Option: Ein- und Ausschließen von Untersuchungseinheiten

Siehe P0.7.

Eingabebox 11: Umkodierungen und Kein-Wert-Angabe

Variable können umkodiert werden. Siehe Dokument Nr. 0 "Arbeiten mit Almo", Abschnitt P0.5.

Anmerkung zur "Kein-Wert-Angabe"

Betrachten wir ein Beispiel: In einer Umfrage wurden Personen u.a. nach ihrem Einkommen befragt. Viele haben die Antwort verweigert. Für die Variable des Einkommens liegt somit kein Wert vor. Beim Schreiben der Daten wurde das mit -1 kodiert. Das muss Almo mitgeteilt werden. In die geöffnete Optionsbox wird geschrieben:

Einkommen (-1=KeinWert)

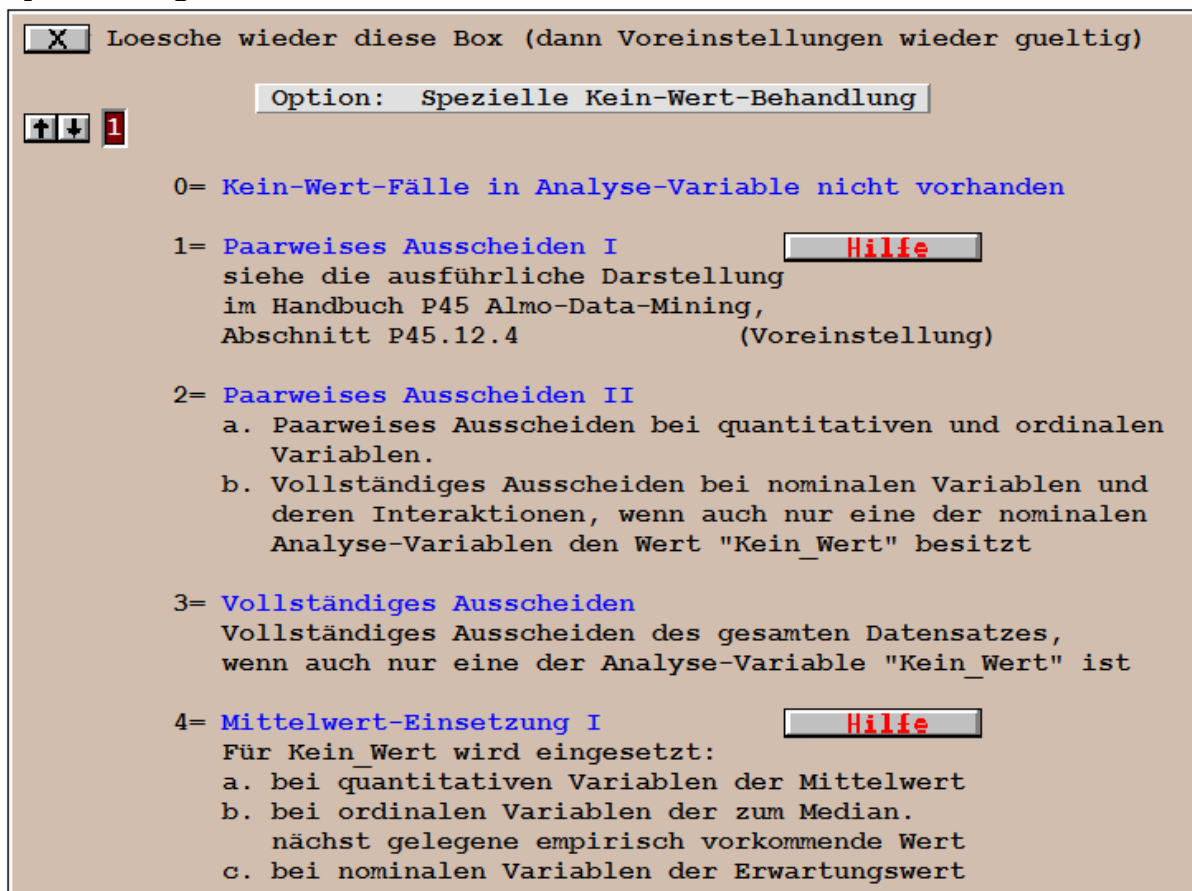
Das Wort "KeinWert" ist ein Schlüsselwort das von Almo in entsprechender Weise verstanden wird. "KeinWert" kann groß oder klein oder mit Unterstrich (an beliebiger Stelle) geschrieben werden. Auch die Kurzform "KW" oder "kw" ist zulässig.

Eingabebox 12: Option: Spezielle Kein-Wert-Behandlung



Wenn Almo im Programm 19 auf eine Variable stößt, die keinen Wert besitzt, dann wird beim Korrelieren standardmäßig das "paarweise Ausscheiden" durchgeführt. Wenn Sie das akzeptieren, dann brauchen Sie diese Optionen-Box nicht zu aktivieren. Wenn nicht, dann stehen Ihnen folgende 7 Vorgehensweisen zur Verfügung:

Optionsbox geöffnet:



.
. .
. .
. .
. .

5= Mittelwert-Einsetzung II

Hilfe

Für Kein_Wert wird eingesetzt:

- bei quantitativen Variablen der zum Mittelwert nächste empirisch vorkommende Wert
- bei ordinalen der Median (wie bei 4)
- bei nominalen Variablen der Erwartungswert (wie bei 4)

6= Mittelwert-Einsetzung III

Hilfe

Für Kein_Wert wird eingesetzt:

- bei quantitativen Variablen der Mittelwert +/- einem normalverteilten Zufallswert mit Mittelwert=0 und Standardabweichung der Variablen
- 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
- bei nominalen der wahrscheinlichste Ausprägungswert

7= Mittelwert-Einsetzung IV

Hilfe

- bei quantitativen Variablen zunächst wie bei 6
Der nächst gelegene empirische Skalenwert wird dann eingesetzt
- bei ordinalen der Median (wie bei 6)
- bei nominalen Variablen (wie bei 6)

BEACHTEN: Sie können auch das "Imputationsverfahren" für fehlende Werte rechnen. Klicken Sie auf den Knopf "Data Mining", dann Kapitel 3 "Daten bereinigen"



1

- nur relevant für Allgemeines Lineares Modell (ALM) !!
- 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
 - gewählte Kein-Wert-Behandlung gilt auch für abhängige Variable



123457

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

Hilfe



1

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

Hilfe

Es ist nahezu normal, dass manche Untersuchungseinheiten in manchen Variablen keine Werte besitzen. Bei Befragungen wird beispielsweise von einem relativ hohem Prozentsatz der Befragten die Frage nach dem Einkommen nicht beantwortet. In diesem Falle wird man dann etwa -1 als Einkommenshöhe kodieren. Anders formuliert:

-1 ist der "Kein-Wert-Code" für die Einkommensvariable

In der Eingabebox 9 "Umkodierungen und Kein-Wert-Angabe" muß dann stehen

```
Einkommen (-1 = Kein_Wert)
```

Almo überführt dann den Wert -1 in einen internen Almo-Code.

Der interne Almo-Code für "Kein_Wert" ist die riesige negative Zahl -10 hoch 40 (Stand Almo15). Aber das ist für den Benutzer irrelevant. Der Vorgang ist also folgender: Die Zahl -1 wird in die Zahl -10 hoch 40 umkodiert. Diese Zahl wird von allen Almo-Programmen als "Kein-Wert-Code" begriffen. Wenn Almo auf diese Zahl stößt, dann weiß es, dass es diese auf eine besondere Weise behandeln muß.

Wenn Almo im Programm 19 auf diesen internen Kein_Wert-Code stößt, dann wird beim Korrelieren standardmäßig das "paarweise Ausscheiden" durchgeführt. Wenn Sie das akzeptieren, dann brauchen Sie diese Optionen-Box nicht zu aktivieren. Wenn nicht, dann stehen Ihnen folgende 7 Vorgehensweisen zur Verfügung:

Eingabe: 0 Kein-Wert-Fälle nicht vorhanden

Der Benutzer ist sich sicher, dass keine Kein-Wert-Fälle vorliegen. Almo kann die Korrelationsmatrix dann schneller und speicherplatzsparend errechnen. Der Zeitgewinn ist bei kleinerer Variablenzahl und kleineren Datenmengen kaum spürbar.

Eingabe 1: Paarweises Ausscheiden

Almo führt das "paarweise Ausscheiden" durch. Betrachten wir ein Beispiel: Es sollen die 3 Variablen x, y, z korreliert werden. Bei der 12. Untersuchungseinheit besitzt x keinen Wert. Dann werden die Kreuzprodukte etc. zwischen x einerseits und y sowie z andererseits für diese Untersuchungseinheit nicht berechnet. Für die 12. Untersuchungseinheit liegen jedoch valide Werte für y und z vor, so dass die Kreuzprodukte etc. zwischen y und z berechnet werden können. Almo merkt sich dann, dass die Korrelation xy und xz auf einer um 1 verringerten Zahl von Untersuchungseinheiten beruht. Die Zahl der Untersuchungseinheiten, auf der die jeweilige Korrelation zwischen dem Variablenpaar ik beruht, wird in der Ergebnis-Ausgabe mitgeteilt.

Das "paarweise Ausscheiden" wird auch durchgeführt, wenn die Variable x, für die ein Wert fehlt eine nominale ist. Da die nominalen Variablen in Dummies aufgelöst werden, werden in diesem Falle die Kreuzprodukte etc. aller Dummies von x mit y und z nicht berechnet.

Die besondere Problematik des paarweisen Ausscheidens wird in nachfolgendem Exkurs ausführlich beschrieben.

Eingabe 2: Paarweises Ausscheiden II

Liegt für eine Untersuchungseinheit auch nur für eine nominale Variable A kein Wert vor, dann werden die Werte aller nominaler Variablen nicht berücksichtigt, auch wenn für die anderen nominalen Variablen B, C, ... Werte vorhanden sind. Zwischen den quantitativen / ordinalen Variablen hingegen wird das gewohnte "paarweise Ausscheiden" durchgeführt.

Eingabe 3: Vollständiges Ausscheiden

Liegt für eine Untersuchungseinheit auch nur in einer Variablen kein Wert vor, dann wird die gesamte Untersuchungseinheit aus der Analyse ausgeschlossen. Diese Vorgehensweise wird gelegentlich "listenweises Ausscheiden" genannt.

Eingabe 4: Mittelwert-Einsetzung I

Almo ermittelt zuerst Mittelwerte (für quantitative Variable), Median (für ordinale Variable) und den Erwartungswert (für nominale Variable).

Almo gibt diese Werte aus.

Für Kein_Wert wird eingesetzt:

- a. bei quantitativen Variablen der Mittelwert
- b. bei ordinalen Variablen der Median (=der mittlere Wert) liegt der Median nicht auf einem empirischen Wert, sondern zwischen 2 empirischen Werten, dann wird der nächst gelegene Nachbarwert als KW-Einsetzungswert verwendet.
- c. 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			1.00	2.10

Der Erwartungswert ist 2.1

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

Eingabe 5: Mittelwert-Einsetzung II

Für Kein_Wert wird eingesetzt:

- a. bei quantitativen Variablen der zum Mittelwert nächste empirisch 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

Eingabe 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

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.

Eingabefeld: Startwert für Zufallsgenerator

Die Zufallswerte, die in den oben beschriebenen Kein-Wert-Behandlungen 6 und 7 benötigt werden, erzeugt Almo mit einem "Zufallsgenerator". Wenn die Startzahl nicht verändert wird, dann werden bei einem 2. und jedem weiteren Lauf des Programms Prog45mo immer dieselben Zufallszahlen und damit dieselben KW-Einsetzungswerte erzeugt. Ist dies jedoch nicht erwünscht, dann muß der Benutzer die Startzahl ändern. Verwenden Sie eine 6-stellige ungerade Zahl.

Eingabefeld: Gemeinsame Fallzahl

Wird als Kein-Wert-Behandlung=1 oder =2, das "paarweise Ausscheiden" gewählt, dann liefert Almo folgende Tabelle - sofern Kein-Wert-Fälle auch tatsächlich auftreten:

Besipiel:

Zahl der Einheiten, die in die Analyse eingegangen sind
je Zelle der Streuungsmatrix bei "paarweisem Ausscheiden"

	x1	x2	x3
x1	90	81	73
x2	81	90	75
x3	73	75	80

Für das Variablenpaar x1 x2 werden nur die Einheiten, die in beiden Variablen valide Daten besitzen, ausgewertet. Das sind in unserem Beispiel 81.

Für das Variablenpaar x1 x3 werden nur 73 Einheiten ausgewertet und für das Variablenpaar x2 x3 nur 75.

Almo benötigt für den weiteren Rechengang eine einzige

"gemeinsame Fallzahl"

Diese wird für die Ermittlung der Signifikanzen der Korrelationen gebraucht.

Almo bietet für die "gemeinsame Fallzahl" 4 Alternativen an:

0 = die kleinste Fallzahl, aus der die Korrelationen zwischen je 2 Variablen i und k errechnet wurden

1 = das harmonisches Mittel aus den Fallzahlen, aus denen die Korrelationen zwischen den Variablen errechnet wurden

2 = die Zahl der Fälle, die in allen Analysevariablen valide Werte besitzen

3 = die Zahl der eingelesenen Fälle

Die vorsichtigste Alternative ist =2.

Für die Signifikanztests im Korrelationsprogramm werden nur so viele Fälle verwendet, wie sie beim "vollständigen Ausscheiden" vorhanden wären.

Die optimistischste Alternative ist =3.

Hier werden alle eingelesenen Fälle für die Signifikanztests verwendet. 1 und 2 liegen dazwischen.

Vergleich

Beim "paarweisen Ausscheiden" werden die Daten optimal "ausgenützt". Der Nachteil dieses Verfahrens ist jedoch, dass die Korrelationen zwischen den Variablen auf unterschiedlichen Häufigkeiten beruhen. Sind die Unterschiede nicht zu groß, dann kann man diesen Nachteil ignorieren.

Das "listenweises Ausscheiden" ist korrekt, die Menge an Daten, die nicht benutzt werden, kann jedoch beträchtlich sein.

P19.1.5 Exkurs: Das "paarweise Ausscheiden" zur Lösung des Kein-Wert-Problems

Nicht selten sind Daten unvollständig. Für manche Untersuchungseinheiten fehlen Werte in der Variablen i für andere in der Variablen j und wieder für andere in der Variablen k etc. Wie sollen diese Variable miteinander korreliert werden.

Das "paarweise Ausscheiden" kann bei allen Verfahren verwendet werden, bei denen eine Korrelationsmatrix oder allgemein: eine Streuungsmatrix (als Endergebnis oder als Zwischenschritt) errechnet werden muß.

Für unsere folgende Darstellung verwenden wir folgende Beispiel-Daten

Tabelle 1: Beispieldaten

Person	x1	x2	x3
1	kw	5	6
2	1	kw	3
3	9	9	kw
4	4	7	kw
5	4	7	5
6	1	1	1
7	5	6	5
8	5	4	4

9	3	5	4
10	2	5	3

Die Datei umfasst 10 Untersuchungseinheiten (Personen) und die 3 Variablen x1, x2, x3. Fehlende Werte wurden mit "kw" (=KeinWert) symbolisiert.

Die 1. Person besitzt in x1 keinen Wert.

Die 2. Person besitzt in x2 keinen Wert.

Die 3. und die 4. Person besitzen in x3 keinen Wert.

Dies sind extreme Daten. Man wird wohl kaum eine Korrelationsmatrix aus nur 10 Datensätzen errechnen wollen. Die Beispieldaten eignen sich jedoch sehr gut um die Problematik des paarweisen Ausscheidens darzustellen.

Diese Daten sind unter dem Namen "PaarwKW.fre" und "PaarwKW.dir" sowie als SPSS-File unter "PaarwKW.sav" im Almo-Ordner TESTDAT enthalten. Der Benutzer kann mit diesen Dateien unsere folgenden Berechnungen nachvollziehen bzw. selbst experimentieren.

P19.1.5.1 Die Berechnung einer Korrelationsmatrix

Wenn wir als Kein-Wert-Behandlung =3, das "vollständige Ausscheiden" wählen dann werden nur die Personen 5 bis 10 ausgewertet, also 6 Personen.

Wir erhalten folgendes Ergebnis:

Tabelle 2: Korrelationsmatrix bei "vollständigem Ausscheiden"

	x1	x2	x3
	v1	v2	v3
x1	1.0000	0.6325	0.8677
x2	0.6325	1.0000	0.9218
x3	0.8677	0.9218	1.0000

Wird als Kein-Wert-Behandlung =1, das "paarweise Ausscheiden" gewählt, dann erhalten wir folgendes Ergebnis:

Tabelle 3: Zahl der Einheiten, die in die Analyse eingegangen sind je Zelle der Streuungsmatrix bei "paarweisem Ausscheiden"

	x1	x2	x3
x1	9	8	7
x2	8	9	7
x3	7	7	8

Für das Variablenpaar x1 x2 werden nur die Personen, die in beiden Variablen valide Daten besitzen, ausgewertet. Das sind die Personen 3 bis 10, also 8 Personen. Für das Variablenpaar x1 x3 werden nur die Personen 2, 5 bis 10, also 7 Personen ausgewertet Für das Variablenpaar x2 x3 werden nur die Personen 1, 5 bis 10, also 7 Personen ausgewertet

Tabelle 4: Korrelations-Matrix bei "paarweisem Ausscheiden"

	x1	x2	x3
x1	1.0000	0.7790	0.8264
x2	0.7790	1.0000	0.8101
x3	0.8264	0.8101	1.0000

Jede einzelne Korrelation r_{ik} zwischen den Variablen i und k wird nur aus den Personen errechnet, die in beiden Variablen einen validen Wert besitzen.

Die Berechnungsformel ist

$$(1) \quad r_{ik} = \text{cov}_{ik} / (s_i * s_k)$$

cov_{ik} = Kovarianz zwischen Variablen i und k

s_i = Standardabweichung der Variablen i berechnet aus den Personen, die in i und k einen validen Wert besitzen

s_k = Standardabweichung der Variablen k berechnet aus den Personen, die in i und k einen validen Wert besitzen

Almo errechnet folgende Standardabweichungen:

Tabelle 5: Standardabweichungen der Variablen bei "paarweisem Ausscheiden"

	x1	x2	x3
Standabwg. der Variablen x1	-	2.26039	1.60357
Standabwg. der Variablen x2	2.23607	-	1.74964
Standabwg. der Variablen x3	1.29363	1.51186	-

In der 1. Zeile stehen die Standardabweichungen der Variablen x1.

Im Variablenpaar x1 x2 beispielsweise ist die Standardabweichung von x1= 2.26039

In der 2. Zeile stehen die Standardabweichungen der Variablen x2.

Im Variablenpaar x1 x2 beispielsweise ist die Standardabweichung von x2= 2.23607

Die Standardabweichungen sind mit n_{ik} , nicht mit $n_{ik}-1$ dividiert. Das kann aber über eine Option umgestellt werden.

Für jeden Korrelationskoeffizient kann dann noch die Signifikanz ermittelt werden, bei deren Berechnung die jeweils verschiedenen Fallzahlen verwendet werden.

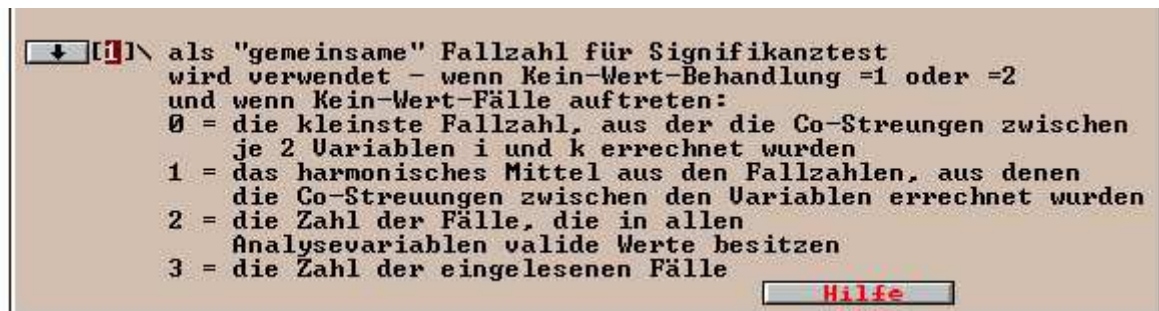
Tabelle 6: Signifikanzen p bei "paarweisem Ausscheiden"

	x1	x2	x3
x1	-	0.0223	0.0220
x2		-	0.0271
x3			-

Wir erkennen, dass die Korrelation $r_{x1x2}=0.779$ mit $p=0.0223$ eine bessere Signifikanz aufweist als die höhere Korrelation $r_{x2x3}=0.8101$ mit $p=0.0271$. Im Prinzip ist es möglich, dass eine Variable i mit der Variablen j eine kleine Korrelation aufweist, die signifikant ist und mit der Variable k eine hohe Korrelation

aufweist, die nicht signifikant ist - eben weil sie auf unterschiedlichen Häufigkeiten beruhen. Die Korrelationskoeffizienten sind also nicht vergleichbar. Das ist nicht sinnvoll und macht das "paarweise Ausscheiden" im Prinzip unbrauchbar.

Man wird so verfahren müssen, dass man doch eine Vergleichbarkeit unterstellt und für die Ermittlung der Signifikanz eine gemeinsame Häufigkeit einsetzt. Also bietet 4 Möglichkeiten an. Um sie zu nutzen muß man im Prog45m6 die Optionsbox 9 „Spezielle Kein-Wert-Behandlung“ öffnen. Man sieht dann eine sehr große Eingabebox, die wir bereits in Abschnitt P45.7.3, Eingabebox 9 (für die Methode 1 bis 3) und in P45.6.1, Eingabebox 7 (für die Methode 4 bis 7) dargestellt haben. Im unteren Teil dieser großen Eingabebox ist folgendes zu sehen.



Als gemeinsame Fallzahl für Signifikanztests kann verwendet werden:

Möglichkeit

- 0 das kleinste n_{ik} aus dem unteren oder oberen Dreieck der obigen Tabelle 3 der Zahl der Einheiten, die in die Analyse eingegangen sind (ohne Diagonale)
- 1 das harmonisches Mittel aus der obigen Tabelle 3 der Zahl der Einheiten
- 2 die Zahl der Fälle, die in allen Analysevariablen valide Werte besitzen. Diese Zahl ist identisch mit der Zahl der Fälle, wie sie beim "vollständigen Ausscheiden" vorhanden wären.
- 3 die Zahl der eingelesenen Fälle

Eine unseres Erachtens sichere Vorgehensweise ist es, die Möglichkeit 2 zu wählen. Dann werden die Korrelationen aus den paarweise vorhandenen Daten errechnet, die Daten also optimal ausgeschöpft. Dabei wird unterstellt, dass die Korrelationskoeffizienten alle aus dem n errechnet wurden, das beim "vollständigen Ausscheiden" entstehen würde. Damit sind sie vergleichbar.

Fehlen in einigen Variablen nur wenige in anderen, aber sehr viele, dann ist es eher sinnvoll die Möglichkeit 1, das harmonische Mittel zu wählen.

Mit diese Unterstellung ist das "paarweise Ausscheiden", wenn es um die Berechnung einer Korrelationsmatrix geht, eine plausible Vorgehensweise.

P19.1.5.2 Die Berechnung einer "Quasi-Korrelationsmatrix" bei "paarweisem Ausscheiden"

Wird als Streuungsmatrix "Quasi_Korrelation" eingestellt und als Kein-Wert-Behandlung =1, das "paarweise Ausscheiden", dann errechnet Also die einzelnen Korrelationen ebenfalls nach der oben unter (1) angegebenen Formel. Die Kovarianz cov_{ik} wird aus den Personen ermittelt, die in beiden Variablen i und k einen validen Wert haben. Das geht nicht anders. Die Standardabweichungen s_i und s_k sind nun anders definiert:

- s_i = Standardabweichung der Variablen i
 berechnet aus den Personen, die in i (und nur in i) einen validen Wert besitzen
- s_k = Standardabweichung der Variablen k
 berechnet aus den Personen, die in k (und nur in k) einen validen Wert besitzen

Die Daten werden also besser ausgenutzt. Die Standardabweichung der Variablen i wird aus allen Personen, die in der Variablen i einen validen Wert besitzen, ermittelt und nicht nur aus jenen, die in i und k einen validen Wert besitzen.

Almo ermittelt folgende Standardabweichungen:

Standardabweichung von Variable x_1 : 2.3465
 Standardabweichung von Variable x_2 : 2.1140
 Standardabweichung von Variable x_3 : 1.4524

Die Standardabweichungen sind mit n , nicht mit $n-1$ dividiert. Das kann aber über eine Option umgestellt werden.

Tabelle 7: Quasi-Korrelationsmatrix

	x1	x2	x3
x1	1.0000	0.7938	0.5030
x2	0.7938	1.0000	0.6979
x3	0.5030	0.6979	1.0000

Die Unterschiede zur Korrelationsmatrix in Tabelle 4 sind teilweise erheblich. Das erklärt sich dadurch, dass wir nur 10 Untersuchungseinheiten haben und im Verhältnis zu diesen relativ viele Kein-Wert-Fälle. Bei größeren Datenmatrizen und prozentual weniger Kein-Wert-Fällen werden die Unterschiede minimal.

Bei der Quasi-Korrelationsmatrix werden die Daten besser ausgenutzt. Die Standardabweichung der Variablen i wird aus allen Personen, die in der Variablen i einen validen Wert besitzen, ermittelt und nicht nur aus jenen, die in i und k einen validen Wert besitzen.

Dafür muß allerdings ein Preis bezahlt werden: Es können Korrelationskoeffizienten entstehen, die außerhalb des Bereichs 0 - 1 liegen. Deswegen sprechen wir auch von "Quasi-Korrelation".

Bei empirischen Daten wird dies sehr selten sein. Wir haben das noch nie erlebt.

Einige Autoren weisen auch darauf hin (so Little/Rubin 1990, S. 379), dass die Korrelationsmatrix mit "paarweisem Ausscheiden" und die Quasi-Korrelationsmatrix so gestaltet sein können, dass sie nicht invertierbar sind, mit der Folge, dass der Kalkül des ALM nicht anwendbar ist. Auch das haben wir noch nie erlebt. Bei empirischen Daten wird dieser Fall nicht auftreten.

Wir können folgendes Fazit ziehen:

Für das "paarweise Ausscheiden" gibt es Varianten. Es ist durchaus eine plausible Vorgehensweise, wenn Kein-Wert-Fälle vorliegen. Es ist aber keinesfalls frei von Problemen.

P19.1.5.3 Die Berechnung einer Quadratsummen- oder einer Kovarianzmatrix bei "paarweisem Ausscheiden"

Wir rechnen nun mit denselben Daten eine Quadratsummenmatrix, zuerst wieder mit Kein-Wert-Behandlung =3, dem "vollständige Ausscheiden". Dabei werden nur die Personen 5 bis 10 ausgewertet, also 6 Personen. Wir erhalten folgendes Ergebnis:

Tabelle 8: Quadratsummen-Matrix bei "vollständigem Ausscheiden"

	x1	x2	x3
	v1	v2	v3
x1	13.3333	10.6667	10.6667
x2	10.6667	21.3333	14.3333
x3	10.6667	14.3333	11.3333

Wird als Kein-Wert-Behandlung =1, das "paarweise Ausscheiden" gewählt, dann erhalten wir zunächst wieder die Matrix der

Zahl der Einheiten, die in die Analyse eingegangen sind
je Zelle der Streuungsmatrix

die wir bereits oben in Tabelle 3 abgebildet haben. Also benötigt nun für den weiteren Rechengang ein einziges gemeinsames n. Wir bezeichnen es mit n_g . Dieses n_g wird für die Ermittlung der Signifikanzen in Programmen gebraucht, in die die Streuungsmatrix als Eingabe eingeht - z.B. das ALM. Also bietet hier 4 Alternativen an. Als gemeinsame Fallzahl wird verwendet:

Möglichkeit

- 0 das kleinste n_{ik} aus dem unteren oder oberen Dreieck der obigen Tabelle 3 der Zahl der Einheiten, die in die Analyse eingegangen sind (ohne Diagonale)
- 1 das harmonisches Mittel aus der obigen Tabelle 3 der Zahl der Einheiten
- 2 die Zahl der Fälle, die in allen Analysevariablen valide Werte besitzen. Diese Zahl ist identisch mit der Zahl der Fälle, wie sie beim "vollständigen Ausscheiden" vorhanden wären.
- 3 die Zahl der eingelesenen Fälle

Die vorsichtigste Alternative ist =2. Für die Signifikanztests (etwa im Rahmen des ALM) werden nur so viele Fälle verwendet, wie sie beim "vollständigen Ausscheiden" vorhanden wären. Die optimistischste Alternative ist =3. Hier werden alle eingelesenen Fälle für die Signifikanztests verwendet. 1 und 2 liegen dazwischen.

Wenn wir die Alternative 0 wählen, dann ist $n_g =$

0	7
1	7
2	6
3	10

Beim Maskenprogramm kann diese Option in der entsprechend Eingabebox eingesetzt werden. Ist die Eingabebox nicht vorhanden, dann verwendet Almo in der Regel =1. Beim in der Almo-Programmiersprache "selbst geschriebenen" Programm

kann eine dieser 4 Möglichkeiten über die Option 78 eingegeben werden.

Almo berechnet zuerst für jedes Variablenpaar die Kreuzprodukte.

Tabelle 9: Matrix der Kreuzprodukte bei "paarweisem Ausscheiden"

	x1	x2	x3
x1	178	213	87
x2	213	308	137
x3	87	147	-

Für das Variablenpaar x1 x2 beispielsweise werden nur die Personen, die in beiden Variablen valide Daten besitzen, ausgewertet. Das sind die Personen 3 bis 10, also 8 Personen. Die Matrix ist selbstverständlich symmetrisch. Entsprechend wird für die anderen Variablenpaare verfahren.

In der Diagonale der Matrix stehen die Kreuzprodukte der Variablen "mit sich selbst". Sie werden aus den Personen gebildet, die in der Variablen valide Werte besitzen, bei x1 und x2 sind das 9 Personen, bei x3 sind es 8.

Almo berechnet dann noch für jedes Variablenpaar die Wertesummen.

Tabelle 10: Matrix der Wertesummen bei "paarweisem Ausscheiden"

	x1	x2	x3
Wertesumme der Variablen x1	34	33	21
Wertesumme der Variablen x2	44	49	33
Wertesumme der Variablen x3	25	28	31

In der 1. Zeile stehen die Wertesummen der Variablen x1. In der Zelle x1 x2 steht 33. Das ist die Summe der Variablenwerte der Variablen x1, wenn sie mit x2 gepaart wird. Für diese Summe werden nur die x1-Werte verwendet, für die x1 und x2 je einen validen Wert besitzen. Das sind die Personen 3 bis 10.

In der Zelle x1 x3 steht 21. Das ist die Wertesumme der Variablen x1, wenn sie mit x3 gepaart wird.

Die Variable x1 hat also ungleiche Wertesummen - je danach mit welcher anderen Variablen sie gepaart wird. Entsprechendes gilt für die anderen Variablen x2 und x3.

In der Zeile i der Matrix steht also die Variable i, deren Wertesummen wir betrachten. In der Spalte k der Matrix steht die Variable k, mit der i gepaart wird.

In der Diagonale der Matrix stehen die Wertesummen der Variablen.

Die Quadratsumme für die Zelle ik ergibt sich dann gemäß folgendem Ausdruck:

$$(2) \quad Q_{ik} = K_{ik} - W_{ik} * W_{ki} / n_{ik}$$

Für die Zelle x1 x2 beispielweise

$$Q_{12} = 213 - 33 * 44 / 8 = 31.5$$

- Q_{ik} = Quadratsumme für das Variablenpaar x_i x_k
- K_{ik} = Kreuzprodukt für das Variablenpaar x_i x_k
- W_{ik} = Wertesumme der Variablen x_i , wenn sie mit x_k gepaart wird
- W_{ki} = Wertesumme der Variablen x_k , wenn sie mit x_i gepaart wird
- n_{ik} = valide Häufigkeit für das Variablenpaar x_i x_k

Für das Diagonalglied ii ergibt sich

$$(2a) \quad Q_{ii} = K_{ii} - W_{ii} \cdot W_{ii} / n_{ii}$$

So erhalten wir die

Tabelle 11: Quadratsummenmatrix bei "paarweisem Ausscheiden"

	x1	x2	x3
x1	49.56	31.50	12
x2	31.50	40.22	15
x3	12	15	16.88

Die Matrix ist symmetrisch. Die Quadratsumme Q_{ik} in der Zelle ik der Matrix ist nur aus den Personen errechnet worden, die für beide Variable i und k valide Werte besitzen. Die Quadratsummen beruhen also auf unterschiedlichen Häufigkeiten. Die Matrix ist in dieser Form nicht zu gebrauchen.

Wir dividieren nun die einzelnen Quadratsummen durch die entsprechenden Häufigkeiten. So erhalten wir die "durchschnittliche" Quadratsummen-Matrix. Das ist die Kovarianzmatrix. Für die Zelle $x1$ $x2$ rechnen wir also $31.5/8=3.9375$

Tabelle 12: Kovarianz-Matrix bei "paarweisem Ausscheiden"
(Kovarianz ist mit n dividiert)

	x1	x2	x3
x1	5.5062	3.9375	1.7143
x2	3.9375	4.4691	2.1429
x3	1.7143	2.1429	2.1094

Die "durchschnittlichen" Quadratsummen wird mit ng multipliziert. " ng " ist die gemeinsame Fallzahl. Wir haben oben gezeigt, dass es 4 Möglichkeiten gibt, ng zu bestimmen. Wir wählen die Möglichkeit "1", das harmonische Mittel der unterschiedlichen Fallzahlen. Die gemeinsame Fallzahl ng ist dann 7.

Aus der Multiplikation erhalten wir die auf ng Personen hochgerechnete Quadratsummen-Matrix, die Almo schließlich als Ergebnis seiner Berechnungen ausgibt. Wird die Quadratsummen-Matrix als Zwischenschritt für Verfahren wie z.B. das ALM verwendet, dann ist es vollkommen gleichgültig, mit welcher konstanten Zahl multipliziert wird. Man kann auch immer das Multiplizieren unterlassen und die Kovarianz-Matrix verwenden.

Tabelle 13: "Hochgerechnete" Quadratsummen-Matrix bei "paarweisem Ausscheiden" mit dem harmonischen Mittel als gemeinsamer Fallzahl

	x1	x2	x3
	V1	V2	V3
x1	38.5432	27.5625	12.0000
x2	27.5625	31.2840	15.0000
x3	12.0000	15.0000	14.7656

Ende Exkurs: Das "paarweise Ausscheiden" zur Lösung des Kein-Wert-Problems

Eingabebox 13: Option: Untersuchungseinheiten gewichten
Siehe Dokument 0 Arbeiten_mit_Almo.PDF, Abschnitt P0.8.

Eingabebox 14: Option: Partielle Korrelationsmatrix bilden
Wir werden darauf ausführlich im Abschnitt P19.3 eingehen

Eingabebox 15: Option: Schreibe errechnete Matrix in Datei



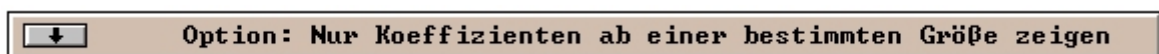
Optionsbox geöffnet:



Geben Sie einen Dateinamen an. Almo schreibt die errechnete Matrix in diese Datei. Almo schreibt nicht nur die Matrix, sondern auch die Variablennummern der korrelierten Variablen und weitere Informationen. Siehe dazu Handbuch zu P20, Abschnitt P20.8.6.

Eingabebox 16: Optionen, die das "Aussehen" der auszugebenden Matrix steuern
Siehe P0.9.

Eingabebox 17: Option: Nur Koeffizienten ausgeben, deren absolute Werte größer sind als ...



Optionsbox geöffnet:



Werden viele Variable korreliert, dann ist die Korrelationsmatrix sehr unübersichtlich. In diesem Falle ist folgende Vorgehensweise sinnvoll: Man bildet zunächst – ohne diese Option zu nutzen – eine Korrelationsmatrix. Almo gibt unterhalb der Matrix eine Tabelle der Signifikanzen der Korrelationskoeffizienten aus. Der Benutzer sieht folgendes:

Mindestgroesse des Produkt-Moment- bei Signifikanz fuer

Korrelationskoeffizienten r	(1-p) *100	df=n-2=61-2=59
0.1664	80	
0.1866	85	
0.2127	90	
0.2296	92.5	

0.2520	95	

0.2869	97.5	
0.3267	99	
0.4139	99.9	

Bei einer Signifikanz von 95% müssen die Korrelationskoeffizienten größergleich 0.2520 sein.

Man kann nun in oben abgebildeter Eingabebox diesen Wert von 0.2520 eingeben und die Analyse ein 2. Mal rechnen. Dann bleiben in der Korrelationsmatrix alle Zellen leer, in denen ein Koeffizient von kleiner 0.2520 enthalten ist. Die Korrelationsmatrix wird dadurch wesentlich übersichtlicher. Man erkennt die bedeutsamen Variablen-zusammenhänge prägnanter.

Man sollte folgende 2 Punkte aber berücksichtigen:

- (1) Signifikanzen zu ermitteln ist nur sinnvoll, wenn die Daten als eine Zufallsstichprobe aus einer definierten Grundgesamtheit stammen.
- (2) Bei großen Datenmengen (Stichproben) sind schon relativ kleine Korrelations-koeffizienten signifikant.

Die Folgerung daraus: Häufig geht es nur darum, Zusammenhänge zu explorieren. Die Korrelationskoeffizienten sind dafür Kennziffern. Es spricht also nichts dagegen, unabhängig von Überlegungen zur Signifikanz, einen Mindestwert festzulegen – z.B. 0.09. Koeffizienten, die größer sind, weisen uns auf einen relevanten Variablenzusammenhang hin.

Eingabebox 18: Grafik-Optionen

Siehe Dokument 0 Arbeiten_mit_Almo.PDF, Abschnitt P0.10.

P19.1.7 Ausgabe der Ergebnisse aus Prog19bm

Almo erzeugt aus der Programm-Maske Prog19bm eine Ergebnisliste, die der in Abschnitt P19.1.3. entspricht. Wir werden deswegen nur jene Ausgabeteile darstellen und erläutern, die hinzu kommen. Die Ausgabe besteht aus zwei Teilen.

1. Teil der Ausgabe: Basis-Statistiken

Median und Quartile der ordinalen Variablen

Variable	1. Quartil	Median	3. Quartil
5 Leistung	2.5000	4.0000	5.0000

Variable	mittlerer Wert	Streuung	Untergrenze	Diverse Werte	Kein-Wert	
					Faelle	abs in %
-----	-----	-----	-----	-----	-----	-----

Nominale Variable Erwartungswert

2 Wohnort	1.5574	-	-	2.0000	0	0.00
-----------	--------	---	---	--------	---	------

3 Beruf	2.0000	-	-	3.0000	0	0.00
Ordinale Variable	Median	Quartilsdiff/2				
-----	-----	-----				
5 Leistung	4.0000	1.2500	-	9.0000	0	0.00
Quantitat. Variable	arithm.Mittel	Stand.abwg.				
-----	-----	-----				
6 Alter	3.9344	2.2093	-	9.0000	0	0.00
7 Einkommen	3.7049	2.4850	-	9.0000	0	0.00

Für die quantitativen Variablen wird als "mittlerer Wert" das arithmetische Mittel und als Streuung die Standardabweichung ausgegeben

Die Standardabweichung der quantitativen Variablen wurde mit n im Nenner gerechnet

Für die nominalen Variablen wird als "mittlerer Wert" der Erwartungswert ausgegeben. Siehe dazu oben die Erläuterungen zu Prog19bm, Eingabebox 12, Eingabe 4

Für die ordinalen Variablen wird als "mittlerer Wert" der Median und als Streuung die halbe Quartilsdifferenz ausgegeben

2. Teil der Ausgabe: Korrelationsmatrix

Fuer Analyse ausgewaehlte Variable

```
V2   Wohnort: Stadt Land
V3   Beruf: Arbeiter Angestellter Selbständiger
V5   Leistung
V6   Alter
V7   Einkommen
```

V2 wird auch bezeichnet mit A
die Auspraegungen (bzw.Dummies) mit
A1 =Stadt
A2 =Land

V3 wird bezeichnet mit B
die Auspraegungen (bzw.Dummies) mit
B1 =Arbeiter
B2 =Angestellter
B3 =Selbständiger

Haeufigkeiten je Auspraegung der nominalen Variablen

```
-----
V2 Wohnort
V2-1 Stadt           27
V2-2 Land           34

V3 Beruf
V3-1 Arbeiter       16
V3-2 Angestellter   29
V3-3 Selbständiger  16
```

***** Erläuterung:

Für die nominalen Variablen teilt Also deren Häufigkeiten je Ausprägung mit

Korrelations-Matrix

Abhängig vom Messniveau der Variablen berechnet also folgende Korrelationskoeffizienten:

	quant.	ordinal	nominal- dichotom	nominal- polytom
quantitativ	r	Groß-Gamma	punktbiser.r	Eta
ordinal		tau-b	biser. tau-b	Groß-Gamma
nominal-dichotom			Phi	Phi'
nominal-polytom				Cramers V

Zum Groß-Gamma-Koeffizienten siehe letztes Kapitel des hier vorliegenden Dokuments.

Sind nominal-polytome Variable vorhanden, dann werden diese zuerst in Dummies aufgelöst. Durch eine kanonische Korrelation werden sie dann wieder zusammengefasst und in einer 2. Korrelationsmatrix ausgegeben.

Dies wird im nachfolgenden Abschnitt P19.2 ausführlich erläutert.

Alle (quadrierten) Korrelationskoeffizienten $r_{(ik)}$ sind "proportional reduction of error"-Koeffizienten. Sie drücken den Anteil aus, um den sich die Fehlerstreuung in der Variablen k reduziert, wenn i als erklärende Variable eingeführt wird.

	Wohnort Stadt A1	Wohnort Land A2	Beruf Arbeit B1	Beruf Angest B2	Beruf Selbst B3	Leistun V5	Alter V6	Einkomm V7
Wohnort Stadt A1	1.0000	-1.0000	0.0689	-0.1213	0.0689	0.0981	-0.1229	-0.0934
Wohnort Land A2	-1.0000	1.0000	-0.0689	0.1213	-0.0689	-0.0981	0.1229	0.0934
Beruf Arbeiter B1	0.0689	-0.0689	1.0000	-0.5676	-0.3556	0.1520	-0.0160	-0.0792
Beruf Angestel B2	-0.1213	0.1213	-0.5676	1.0000	-0.5676	0.0411	-0.1352	-0.1247
Beruf Selbstän B3	0.0689	-0.0689	-0.3556	-0.5676	1.0000	-0.1987	0.1695	0.2208
Leistung V5	0.0981	-0.0981	0.1520	0.0411	-0.1987	1.0000	0.0495	-0.0469
Alter V6	-0.1229	0.1229	-0.0160	-0.1352	0.1695	0.0495	1.0000	-0.1469
Einkomm V7	-0.0934	0.0934	-0.0792	-0.1247	0.2208	-0.0469	-0.1469	1.0000

zweiseitige

Signifikanz $100 \cdot (1-p)$ der Korrelationen

(wenn eine Variable ordinal, dann über z-Wert geprüft)

(sonst t-verteilt geprüft)

	Wohnort Stadt A1	Wohnort Land A2	Beruf Arbeiter B1	Beruf Angestel B2	Beruf Selbstän B3	Leistung V5	Alter V6	Einkomme V7
Wohnort Stadt A1	-	100.00	40.63	65.26	40.63	57.84	65.89	53.06
Wohnort Land A2	100.00	-	40.63	65.26	40.63	57.84	65.89	53.06
Beruf Arbeiter B1	40.63	40.63	-	100.00	99.41	78.70	9.89	46.02
Beruf Angestel B2	65.26	65.26	100.00	-	100.00	26.39	70.48	66.59
Beruf Selbstän B3	40.63	40.63	99.41	100.00	-	89.64	81.07	91.26
Leistung V5	57.84	57.84	78.70	26.39	89.64	-	31.50	29.92
Alter V6	65.89	65.89	9.89	70.48	81.07	31.50	-	74.13
Einkomme V7	53.06	53.06	46.02	66.59	91.26	29.92	74.13	-

***** Erläuterung:

Also berechnet die Varianzen (und daraus die Signifikanzen) der verschiedenen Korrelationskoeffizienten nach dem Groß-Gamma-Kalkül. Das gilt auch für "Mischkoeffizienten" zwischen einer ordinalen Variablen einerseits und einer quantitativen oder dichotomen Variablen andererseits. Bei allen Koeffizienten, bei

denen mindestens eine ordinale Variable beteiligt ist, werden Bindungen berücksichtigt.

Beruhend auf den Korrelationskoeffizienten auf ungleichen Häufigkeiten (weil Werte fehlten und das "paarweise Ausscheiden" angewandt wurde), dann werden die Varianzen (und daraus die Signifikanzen) korrekt aus den jeweils vorhandenen Datenpaaren errechnet.

Almo stellt die Korrelationsmatrix als Balkendiagramm dar. Wir haben das bereits oben in Abschnitt P19.1.3 gezeigt.

Die Dummies nominal-polytomer Variablen werden nun über eine kanonische Korrelationsanalyse zusammengefaßt. Das wird ausführlich im übernächsten Abschnitt P19.2 erläutert. Es entsteht folgende zweite Korrelationsmatrix

Korrelations-Matrix

		Wohnort V2	Beruf V3	Leistung V5	Alter V6	Einkomme V7
Wohnor	V2	1.0000	0.1213	0.0981	0.1229	0.0934
Beruf	V3	0.1213	1.0000	0.2169	0.1760	0.2208
Leistu	V5	0.0981	0.2169	1.0000	0.0495	-0.0469
Alter	V6	0.1229	0.1760	0.0495	1.0000	-0.1469
Einkom	V7	0.0934	0.2208	-0.0469	-0.1469	1.0000

Auch diese Matrix wird als Balkendiagramm dargestellt. Siehe dazu oben Abschnitt P19.1.3.

P19.1.8 Korrelationsmatrix mit Einschluß von Rangvariablen

In Abschnitt P19.0.1 Messniveaus haben wir den Begriff der Rangvariablen definiert. Hier sei nochmals kurz definiert: Die Werte, die eine Rangvariable annimmt, sind die Rangplätze der Untersuchungseinheiten.

Werden 2 ordinale Variable in Rangvariable transformiert und dann wie quantitative Variable korreliert, dann entsteht die "Spearman'sche Rangkorrelation Rho".

Im Rahmen des Groß-Gamma-Korrelationsmodells können nun auch problemlos Rangvariable eingeführt werden. Siehe dazu die Darstellung von Heinrich Potuschak im Anhang.

Abhängig vom Messniveau entstehen nach dem Groß-Gamma-Kalkül folgende Korrelationskoeffizienten:

	quantitativ	Rang	ordinal	nominal-dichotom
quantitativ	r	namenlos	namenlos	punktbiseriales r
Rang		Rho	namenlos	namenlos
ordinal			tau-b	biseriales tau_b
nominal-dichotom				Phi

Mit Programm Prog08m6 können Variable in Rangwerte transformiert werden und dann, zusammen mit Variablen anderer Messniveaus, in eine neue Datei gespeichert werden. In der Regel wird man nur ordinale Variable in Rangvariable transformieren. Werden quantitative Variable in Rangvariable transformiert, dann erleidet man einen Informationsverlust. Nominale Variable in Rangvariable zu transformieren ist unzulässig.

Mit der neuen Datei rechnet man dann eine Korrelationsmatrix (mit Prog19am oder Prog19bm). Dabei werden die in Rangwerte transformierten Variablen als "quantitative Variable" angegeben.

Einfacher ist es die folgende Programm-Maske Prog19dm zu verwenden. Es nimmt die Transformation, für den Benutzer unsichtbar, automatisch vor.

P19.1.8.1 Eingabe in Prog19dm "Korrelationsmatrix mit Einschluß von Rangvariablen"

Prog19dm ist fast identisch mit dem bereits dargestellten Programm Prog19bm. Wir werden deswegen hier nur die Eingabefelder wiedergeben und erläutern, bei denen die Eingabe etwas anders ist.

Eingabefeld 7: Analyse-Variable: Zu korrelierende Variable

The screenshot shows a software interface for selecting variables for correlation analysis. It is divided into four main sections, each with a title, a description, and a list of variables with selection controls.

- Zu korrelierende Variable** (Help button):
 - quantitative Variable
 - Selection controls: left arrow, right arrow, empty box, empty box
 - Selected variables: Alter, Kinderzahl
- nominale Variable** (Help button):
 - sie werden in Dummies aufgelöst
 - Selection controls: left arrow, right arrow, empty box, empty box
 - Selected variables: Beruf, Schulbildung
 - Interaction controls: up arrow, down arrow, empty box, 0
 - Help button: 0
 - Description: aus den nominalen Variablen werden auch Interaktionsvariable x. Ordnung gebildet
0 = Keine Interaktionsvariable bilden
- ordinale Variable** (Help button):
 - Selection controls: left arrow, right arrow, empty box, empty box
 - Selected variables: Bewertung
- Rang-Variable** (Help button):
 - = (ordinale) Variable, die von Almo zuerst in Rangwerte transformiert und danach mit den anderen Variablen interkorreliert werden sollen
 - Selection controls: left arrow, right arrow, empty box, empty box
 - Selected variables: Leistung, Einkommen

Die Eingabefelder 1 bis 4 wurden bereits in Abschnitt P19.1.4 erläutert

Eingabefeld 5: Rangwert-Variable. Geben Sie hier die (ordinalen) Variablen an, die in Rangwerte transformiert werden sollen und danach mit den in den anderen Eingabefeldern angegebenen Variablen interkorreliert werden sollen.

Die Werte einer Rangvariablen (auch "Rangwertvariable" genannt) sind die Rangplätze, die Untersuchungsobjekte in einer Messdimension hintereinander einnehmen.

Beispiel: Bei einem Marathon-Lauf trifft ein Läufer nach dem anderen im Ziel ein (dabei können auch 2 oder mehr gleichzeitig eintreffen). Ihre Rangplätze bilden die Rangvariable.

Eine Rangvariable kann auch durch eine einfache Transformation aus einer ordinalen oder quantitativen Variablen hervorgehen.

Betrachten wir folgende ordinale Variable

Schulbildung	Codeziffer
Volksschule	1
Hauptschule	2
Gymnasium	3
Fachschule	3
Universität	4

Die Codeziffern 1 bis 4 drücken eine Rangordnung im Bildungsniveau aus. 4 ist mehr als 3 und 3 ist mehr als 2 etc. Um wieviel mehr ist allerdings unbekannt. Die Differenzen zwischen den Bildungsstufen sind nicht bekannt. Die Ziffern drücken die Relation "mehr" oder "weniger" oder "gleich" aus.

Beachte:

Gymnasium und Fachschule wurden gleichrangig betrachtet und deswegen beide mit 3 kodiert.

Unterschied zwischen ordinaler und Rang-Variabler:

Bei der ordinalen Variablen werden den Ausprägungen der Variablen Rangplätze zugewiesen

Bei der Rang-Variablen werden den Untersuchungseinheiten Rangplätze zugewiesen

Aus der ordinalen Variablen "Schulbildung" kann nun eine Rangvariable gebildet werden.

Von 7 Personen kennen wir die Schulbildung in Form ordinaler Codeziffern.

Person	Schulbildung	Wert in der ordinalen Variablen	Wert in der Rangvariablen
1	Volksschule	1	1
2	Hauptschule	2	2.5
3	Hauptschule	2	2.5
4	Gymnasium	3	5
5	Gymnasium	3	5
6	Fachschule	3	5
7	Universität	4	7

Der Wert der Rangvariablen ist sehr einfach der Rangplatz der Person, wenn alle Personen ihrer Schulbildung nach hintereinander gestellt werden. Da manche Personen dieselbe Schulbildung besitzen wie z.B. Person 2 und 3 bzw. eine als gleichrangig erachtete Schulbildung besitzen, wie z.B. Person 6 mit 4 und 5, wird

eine "Rangteilung" vorgenommen. Person 2 und 3 teilen sich die Plätze 2 und 3. Der mittlere Wert ist 2.5. Person 4,5,6 teilen sich die Plätze 4,5,6. Der Wert in der Mitte ist 5.

Von "Bindung" wird gesprochen, wenn 2 oder mehr Personen denselben Rangplatz einnehmen. Wie am Beispiel gezeigt, wird dann üblicherweise eine "Rangteilung" vorgenommen.

In entsprechender Weise kann natürlich auch eine quantitative Variable in eine Rangvariable überführt werden (wobei allerdings ein Informationsverlust eintritt). Die Untersuchungsobjekte werden entsprechend ihren Werten in der quantitativen Variablen hintereinander gestellt. Bei gleichem Wert wird eine Rangteilung vorgenommen.

Werden 2 Rangvariable nach dem Kalkül des Produkt-Moment-Korrelationskoeffizienten korreliert (d.h. werden sie wie quantitative Variable behandelt), dann entsteht der Spearman'sche Rangkorrelationskoeffizient ρ .

Wird für eine unabhängige nominal-dichotome Variable und eine abhängige Rangvariable ein Allgemeines Lineares Modell gerechnet (in diesem Fall also eine Varianzanalyse), dann entspricht dies dem U-Test nach Mann-Whitney. Bei dieser Varianzanalyse wird die Rangvariable also so behandelt, wie wenn sie quantitativ wäre. Siehe Almo-Dokument Nr. 3 "Nicht-parametrische Verfahren".

Eingabebox 10: Kein_Wert-Angabe und Umkodierungen

Wird hier eine Kein-Wert-Deklaration oder eine Umkodierung für eine Rangvariable angegeben, dann wird diese vor der Transformation in Rangwerte durchgeführt. Die Rangwert-Transformation wird also auf die umkodierte Variable angewendet.

P19.1.8.1 Ausgabe aus Prog19dm "Korrelationsmatrix mit Einschluß von Rangvariablen"

Die Ausgabe ist dieselbe wie bei Prog19bm. Der Korrelationskoeffizient zwischen den Rangvariablen ist identisch mit dem Spearman'schen Rangkorrelationskoeffizienten ρ . Rechnen Sie zum Vergleich mit Prog10m1 eine 2-dimensionale Tabelle. Hier wird ebenfalls ρ und seine Signifikanz berechnet.

P19.2 Korrelations-Matrix bei Vorhandensein nominaler Variabler

Wenn eine Variable nominal ist, dann wird sie von Almo in Dummies aufgelöst. In der Korrelationsmatrix stehen dann die Korrelationskoeffizienten zwischen den Dummies. In unserem Beispiel werden die nominalen Variablen Wohnort und Beruf korreliert. Das Beispiel ist insofern sehr einfach, weil Wohnort dichotom ist und die Dummies Stadt und Land mit -1 korrelieren.

Wir wollen ein komplexeres Beispiel betrachten, das allerdings nicht aus unseren Daten hervorgegangen ist. Die Variable Beruf (mit 3 Ausprägungen) und die Variable Schulbildung (mit 3 Ausprägungen) werden korreliert.

Almo ermittelt folgende Korrelationsmatrix der Dummies:

		Beruf Arbeit	Beruf Angest	Beruf Selbst	Schulbil Hauptsch	Schulbil Gymnasiu	Schulbil Uni
Beruf	Arbeit	1.00	-0.73	-0.24	-0.27	0.11	0.21
Beruf	Angest	-0.73	1.00	-0.48	0.06	-0.01	-0.06
Beruf	Selbst	-0.24	-0.48	1.00	0.25	-0.12	-0.17
Schulb	Haupts	-0.27	0.06	0.25	1.00	-0.71	-0.34
Schulb	Gymnas	0.11	-0.01	-0.12	-0.71	1.00	-0.40
Schulb	Uni	0.21	-0.06	-0.17	-0.34	-0.40	1.00

Wir benötigen einen Koeffizienten, der in einer einzigen Zahl ausdrückt, wie Beruf und Schulbildung miteinander korrelieren.

Almo rechnet nun in einem ersten Schritt ein multivariates Allgemeines Lineares Modell. Die Dummies der einen nominalen Variablen sind dabei die unabhängigen Variablen und die der anderen nominalen Variablen sind die abhängigen.

Dabei entsteht "Pillais Spur".

Siehe dazu auch das Almo-Dokument Nr 13a "ALM Allgemeines lineares Modell II", Abschnitt P20.9.4.1. Aus dieser kann dann der gewünschte Korrelationskoeffizient (den wir "Pillais Korrelation" nennen) abgeleitet werden:

$$\text{Pillais Korrelation} = \sqrt{\text{Pillais Spur} / t}$$

t = wenn beide Variable nominal-polytom sind, dann ist t die Zahl der Ausprägungen minus 1 der einen oder der anderen Variablen. Die kleinere Zahl wird verwendet.

t = 1, wenn eine der beide Variable nominal-dichotom ist

t = 1, wenn eine der beide Variable quantitativ oder ordinal ist

Wenn zwei in Dummies aufgelöste nominale Variable vorliegen, dann ist auch eine alternative Berechnung möglich:

Es wird eine kanonische Korrelation zwischen den Dummies der beiden nominalen Variablen gerechnet. Diese ist identisch mit der bivariaten Korrespondenzanalyse. Siehe dazu Almo-Dokument Nr. 4 "Kanonische Analysen", Abschnitt P29.3. und die Programm-Maske P29m2.

Die Summe der Eigenwerte der kanonischen Faktoren ist identisch mit Pillais Spur. Wie man aus der Theorie der Korrespondenzanalyse weiß, ist sie weiterhin identisch mit dem durchschnittlichen Chi-Quadrat-Wert, wie wir ihn etwa im Rahmen einer 2-dimensionalen Tabellierung mit Programm-Maske P10m1 erhalten (siehe Almo-Dokument 1b "Zwei- und drei-dimensionale Tabellierung"). Wir haben also folgende Identitäten:

$$\text{Pillais Spur} = \text{Summe der Eigenwerte der kanonischen Faktoren} = \text{Chi-Quadrat} / n$$

In unserem obigem Beispiel rechnet Almo Pillais Korrelation aus einem multivariaten Allgemeinen Linearen Modell. Es entsteht die Korrelation = 0.14. Wir verfügen damit über eine einzige Zahl, die die Korrelation zwischen Beruf und Schulbildung ausdrückt.

Dieser Korrelationskoeffizient ist vorzeichenlos bzw. immer positiv.

Die Vorgehensweise in Almo ist nun folgende:

1. Also rechnet nach dem "Groß-Gamma-Kalkül eine Korrelationsmatrix. Nominale Variable werden dabei in Dummies aufgelöst.
2. Sind keine nominalen Variable vorhanden, dann ist der Kalkül beendet. Sind welche vorhanden, dann berechnet Also eine zweite Korrelationsmatrix. Sind beide zu korrelierende Variable nicht nominal, dann wird der Koeffizient aus der ersten Matrix übernommen. Ist mindestens ein der beiden zu korrelierenden Variablen nominal, dann errechnet Also über ein multivariates Allgemeines Lineares Modell "Pillais Spur" und daraus dann "Pillais Korrelation". Die nominale Variable wird dabei als abhängige Variable betrachtet. Ist die als abhängige Variable betrachtete nominale Variable dichotom, dann vereinfacht sich der Kalkül auf ein univariates Allgemeines Lineares Modell.

Haben beide nominale Variable nur 2 Ausprägungen, dann ist Pillais Korrelation identisch mit dem Phi-Korrelationskoeffizienten.

Haben beide nominale Variable mehr als 2 Ausprägungen, dann ist Pillais Korrelation identisch mit Cramers V (siehe Bortz, Lienert, Boehnke, 1990, S.357; dort "Cramers Index" genannt)

Ist die eine nominale Variable dichotom und die ander polytom, dann ist Pillais Korrelation identisch mit Phi' (siehe Bortz, Lienert, Boehnke, 1990, S. 342).

Wird eine nominale-polytome Variable mit einer quantitativen Variable korreliert, dann geht Also im Prinzip genau so vor. Der dabei aus Pillais Spur entstehende Korrelationskoeffizient ist dann identisch mit dem Eta-Korrelationskoeffizienten (wie er im Rahmen der Varianzanalyse mit der quantitativen Variablen als abhängiger und der nominalen Variablen als unabhängiger Variablen berechnet wird). Er entspricht auch exakt der multiplen Korrelation R zwischen den nicht-redundanten Dummies der nominalen Variablen und der quantitativen Variablen.

Ist die eine Variable nominal-dichotom und die andere quantitativ, dann ist Pillais Korrelation identisch mit der punktbiserialen Korrelation. Dabei gilt dann folgende Identität:

punktbiserialer Korrelation = Eta = Pillais Korrelation

Also errechnet also, abhängig vom Messniveau der Variablen folgende Koeffizienten:

	quantitativ	Rang	ordinal	nominal-dichotom	nominal-polytom
quantitativ	r	namenlos	namenlos	punktbiserialer r	punktbiserialer r
Rang		Rho	namenlos	namenlos	namenlos
ordinal			tau-b	biserialer tau_b	namenlos
nominal-dichotom				Phi	Phi'
nominal-polytom					Cramers V

Also errechnet mit demselben Algorithmus auch einen Korrelationskoeffizient zwischen einer ordinalen Variablen und einer nominal-polytomen Variablen. In der statistischen Literatur gibt es dafür kein Äquivalent. Da auch die Signifikanz dieses Korrelationskoeffizienten nicht ermittelbar ist, raten wir hier eher zur Vorsicht. Wenn es jedoch darum geht, die Stärke des Zusammenhangs zwischen einer Menge

von Variablen relativ grob zu explorieren, dann darf er sicherlich als ein brauchbarer Indikator betrachtet werden.

P19.3 Die Erzeugung einer Matrix partieller Streuungen

Die folgenden Ausführungen gelten nicht nur für Partialmatrizen von Korrelationen sondern auch für Partialmatrizen von Kovarianzen und Quadratsummen. Die Matrix für die Almo eine Partialmatrix berechnen soll, kann also eine Quadratsummen- oder Kovarianz- oder Korrelations- oder Quasi-Korrelations-Matrix sein - je nach Eingabe des Benutzers.

Dabei können die Variablen quantitativ und auch nominal sein. Nominale Variable werden automatisch in Dummies aufgelöst. Ordinale Variable können zwar, dadurch dass der Großgamma-Kalkül verwendet wird, in die Korrelationsmatrix eingeführt werden - ob es aber zulässig ist, aus ihr eine Partialmatrix abzuleiten, ist ungeklärt.

Betrachten wir als Beispiel die *Korrelations-Partialmatrix*. Mit den Variablen

V1	Alter	(V11)
V2	Herkunft	(V5)
V3	Leistung	(V6)
V4	Arbeit	(V10)
V5	Zufriedenheit	(V8)
V6	Stress	(V12)

aus der Datei ".Testdat\Adat.fre" bilden wir eine Korrelationsmatrix. Die originalen Variablennummern stehen hinter den Namen in Klammern. Wir haben auf fortlaufende Nummern umnummeriert um übersichtlicher darstellen zu können. Verwendet wird die Programm-Maske "Prog19mb.Msk". Es entsteht folgende Korrelationsmatrix.

Korrelations-Matrix

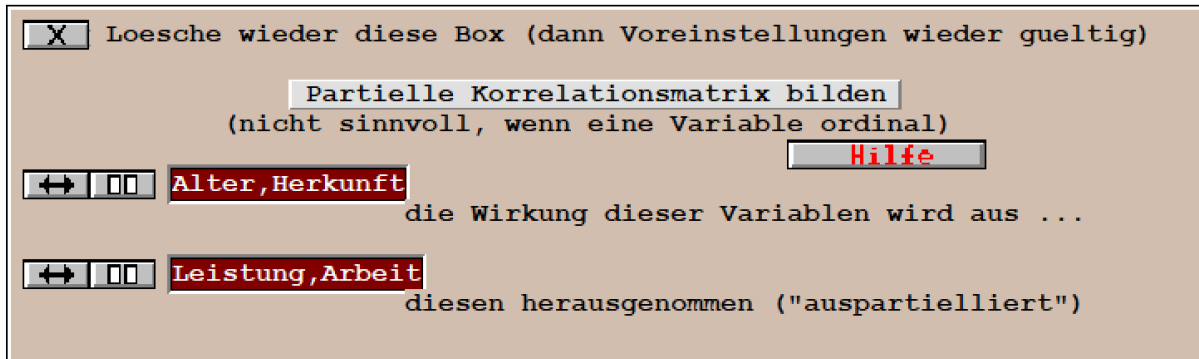
		A		B		C	
		Alter V1	Herkunft V2	Leistung V3	Arbeit V4	Zufried V5	Stress V6
A	Alter V1	1.000	0.572	0.900	0.770	0.863	0.929
	Herkunft V2	0.572	1.000	0.543	0.174	0.625	0.445
B	Leistung V3	0.900	0.543	1.000	<u>0.674</u>	0.700	0.759
	Arbeit V4	0.770	0.174	<u>0.674</u>	1.000	0.560	0.873
C	Zufrieden V5	0.863	0.625	0.700	0.560	1.000	0.862
	Stress V6	0.929	0.445	0.759	0.873	0.862	1.000

Die Korrelations-Submatrizen für V5 und V6 mit den anderen Variablen V1 bis V4 haben wir graphisch abgetrennt, da zunächst die gegenseitigen Korrelationen der Variablen V1 bis V4 betrachtet werden sollen. Wir vermuten, dass beide Variable, V3 Leistung und V4 Arbeit durch V1 Alter und V2 Herkunft stark determiniert werden und die hohe Korrelation zwischen V3 und V4 mit $r_{34} = 0.674$ wesentlich darauf zurück zu führen ist. Wir rechnen deswegen zusätzlich eine Partial-Korrelationsmatrix. In der Programm-Maske "Partmat_Adat.Alm" geschieht dies dadurch dass die Optionsbox "Partielle Korrelationsmatrix bilden" geöffnet und entsprechend ausgefüllt wird. Das Programm findet man durch Klick auf den Knopf

"alle Progs" am Oberrand des Almo-Fensters.



Optionsbox geöffnet:



Die übrigen Variablen also V5 Zufriedenheit und V6 Stress werden in der Optionsbox nicht eingetragen. ALMO ermittelt sie sich selbst als "Restmenge". Sie werden als Variable der Gruppe C bezeichnet.

Die Zahl der Variablen und die Reihenfolge innerhalb der beiden Eingabefelder ist beliebig.

Almo liefert folgendes Ergebnis:

Partialmatrix - Korrelations-Matrix

Diese Variablen - **Gruppe A**

- V1 Alter
- V2 Herkunft

werden aus diesen Variablen herausgenommen ("auspartielliert") - **Gruppe B**

- V3 Leistung
- V4 Arbeit

Die Korrelationen zwischen den Variablen von Gruppe A zu B sind dann .0
Die Korrelationen zwischen den Variablen der Gruppe B ändern sich

Unbeteiligt sind diese Variablen - **Gruppe C**

- V5 Zufriedenheit
- V6 Stress

jedoch ändert sich ihre Korrelation mit den Variablen der Gruppe B

Partial-Korrelations-Matrix

		A		B		C	
		Alter V1	Herkunft V2	Leistung V3	Arbeit V4	Zufried V5	Stress V6
A	Alter V1	1.000	0.572	0	0	0.863	0.929
	Herkunft V2	0.572	1.000	0	0	0.625	0.445
B	Leistung V6	0	0	1.000	<u>-0.037</u>	-0.190	-0.171
	Arbeit V1	0	0	<u>-0.037</u>	1.000	-0.097	0.224
C	Zufriede V8	0.863	0.625	-0.190	-0.097	1.000	0.862
	Stress V1	0.929	0.445	-0.171	0.224	0.862	1.000

Man erkennt: Die *bivariate* Korrelation zwischen V3 und V4 mit $r_{34} = 0.674$ wird auf die *partielle* Korrelation $r_{34.12} = -0.037$ reduziert.

Betrachten wir kurz die Theorie der Partialmatrix.

Wir unterteilen die Variablen in 3 Mengen und die Partialmatrix in 9 Submatrizen:

		Gruppe A		Gruppe B		Gruppe C	
		V1	V2	V3	V4	V5	V6
Gruppe A	V1 V2	QAA		QAB		QAC	
Gruppe B	V3 V4	QBA		QBB		QBC	
Gruppe C	V5 V6	QCA		QCB		QCC	

Wir betrachten zunächst den einfachen Fall mit 2 Gruppen **A** und **B**

1. Die Variablenmenge **A** (in unserem Beispiel V1,2). Das sind die Variablen, deren Wirkungen auf **B** "herausgenommen" werden.
2. Die Variablenmenge **B** (in unserem Beispiel V3,4). Das sind die Variablen, aus denen die Variablen der Menge **A** "herausgenommen" werden.

Wir unterteilen die Korrelationsmatrix in der oben angegebenen Weise. Gesucht ist die neue Korrelationsmatrix **Q*_{BB}**, die die veränderte Korrelation $r_{34.12}$ zwischen V3,V4 enthält - nachdem die Einflüsse von V1 und V2 herausgenommen wurden.

Wir erhalten **Q*_{BB}** durch folgende Gleichung:

$$(1) Q^*_{BB} = Q_{BB} - Q_{BA} \cdot Q_{AA}^{-1} \cdot Q_{AB}$$

Wie oben bereits ausgeführt gilt unsere Darstellung (und auch diese Gleichung) nicht nur für die Partialmatrix der Korrelationen sondern auch für die Partialmatrix der Kovarianzen und der Quadratsummen.

Im Falle der Korrelations-Partialmatrix wird **Q*_{BB}** noch nach folgender Formel zu **Q**_{BB}** normiert:

$$(1a) Q^{**}_{BB} = [\text{diag}(Q^*_{BB})]^{-1/2} \cdot Q^*_{BB} \cdot [\text{diag}(Q^*_{BB})]^{-1/2}$$

Die Matrizen **Q_{AB}** und **Q_{BA}** werden durch die Auspartiellierung = 0. Die beiden Variablensätze korrelieren nicht mehr. Die Matrix **Q_{AA}** bleibt unverändert.

Nun betrachten wir den Fall, dass noch eine 3. Gruppe **C** von Variablen vorhanden ist. In unserem Beispiel wird sie von den Variablen V5,6 gebildet.

Wenn wir V1,2 aus V3,4 herausnehmen, dann müssen sich auch die Korrelationen zwischen V3,4 und V5,6 ändern. Gesucht ist also die Matrix **Q*_{CB}** der veränderten Koeffizienten (bzw. ihre Transponierte **Q*_{BC}**).

Vergleichen wir die beiden Matrizen in unserem Beispiel

		QCB		QCB*	
		aus Korrelationsmatrix		aus Partial-Korrelationsmatrix	
		V3	V4	V3	V4
V5		0.700	0.560	-0.190	-0.097
V6		0.759	0.873	-0.171	0.224

Betrachten wir den Korrelationskoeffizient zwischen V3 aus der Gruppe B mit V5 aus der Gruppe C. Er ist $r_{53}=0.700$. In der Partialmatrix wird er zu einem "teilpartiellen" Korrelationskoeffizienten. Wir bezeichnen ihn mit $r_{5*3.12} = -0.190$. Das Zeichen '*' soll als "mit" gelesen werden. $r_{5*3.12}$ bezeichnet die Korrelation zwischen der Variablen V5 aus der Gruppe C "mit" der "Partialvariablen" V3.12 aus der Gruppe B.

Der Wert eines Probanden in V3 kann gedacht werden als aus 2 Teilen bestehend, einem durch die Auspartiellierung von V1 und V2 aus V3 entstandenen Teil und einem "eigenständigen" Teil von V3. Wie ist das nun zu verstehen?

Wir rechnen eine Regressionsanalyse mit V3 als abhängiger und V1 und V2 als unabhängigen Variablen und ermitteln dabei für V3 die Prognosewerte und Residuen der Probanden. Werden dann für die Probanden die Werte in V5 mit den Residuen für V3 aus der Regressionsanalyse korreliert, dann entsteht ein r von -0.190 . Das Residuum von V3 aus der Regression auf V1 und V2 ist also die Partialvariable V3.12. Die Korrelation zwischen einer Variablen und einer Partialvariablen ist inhaltlich nicht einfach zu interpretieren, so dass in der praktischen Forschung "teilpartielle" Korrelation von geringer Bedeutung sind

Die Formel für die partielle Submatrix Q^*_{CB} lautet

$$(2) Q^*_{CB} = Q_{CB} - Q_{CA} \cdot Q_{AA}^{-1} \cdot Q_{AB}$$

Die Matrizen Q^*_{CB} und Q^*_{BC} werden im Falle der Korrelationsmatrix analog Gleichung (1a) normiert.

Die Matrizen Q_{CA} bzw. Q_{AC} und Q_{CC} bleiben unverändert.

Mit dem selbst geschriebenen **Almo-Syntax-Programm** sind für den "Almo-Spezialisten" noch kompliziertere Konstruktionen möglich.

Wir schreiben nur den Programmparameter-Block:

```

Programm = 19;
Nominale_V = V1,2,3;
Dummy = -1;
Ordinale_V = V4;
Quantitative_V = V5,6,7;

Untergrenze 1:7 = 7*1;
Obergrenze 1:7 = 3*3,5,3*9;

Partial = V1,2 aus V4,5,6/
          V1:6 aus V7;

Matrix = Quadratsumme;

Schreibe Ergebnis_Matrix
in Datei 11
".\Progs\Mat.dat";

Ende_Programmparameter;

```

Die nominalen Variablen V1,2,3 die ordinale Variable V4 und die quantitativen Variablen V5,6,7 werden interkorreliert.

Es sollen 2 aufeinander folgende Auspartiellierungen durchgeführt werden. Schrägstrich als Trennzeichen. Semikolon als Abschluß.

Die partielle Quadratsummen-Matrix soll gebildet werden.

Die Partialmatrix wird in Datei 11 geschrieben.

Beachte beim „selbst geschriebenen“ Almo-Syntax-Programm:

1. Die Anweisung für die Auspartiellierung lautet:

PARTIAL = ... AUS....

Vor AUS werden die Variablen der Menge A, nach AUS, die der Menge B angegeben.

2. ALMO bildet selbst die Menge C als die in der PARTIAL-Anweisung nicht genannten Variablen, die jedoch in NOMINALE_V, ORDINALE_V, QUANTITATIVE_V angegeben sind. In obigem Beispiel sind das in der 1. Auspartiellierung die Variablen V3,7. In der 2. Auspartiellierung ist die Menge C leer.
3. Nach PARTIAL können beliebig viele Auspartiellierungs-Aufgaben angegeben werden. Sie werden durch einen Schrägstrich voneinander getrennt. Zum Schluß folgt ein Semikolon. Beispiel:

```
PARTIAL =   V1,2 AUS V4,5,6 /  
           V1:6 AUS V3,7;
```

4. Werden mehrere Auspartiellierungs-Aufgaben angegeben, dann bezieht sich die nachfolgende auf die Partialmatrix, die die vorausgehenden Auspartiellierungen erzeugt haben. In unserem obigen Beispiel wird in der 1. Auspartiellierung V1,2 aus V4,5,6 herausgenommen. Es entsteht die Partialmatrix Q^*_1 . Die 2. Auspartiellierung, bei der V1 bis 6 aus V3,7 herausgenommen werden soll, wird an der Matrix Q^*_1 vorgenommen. Es entsteht die Partialmatrix Q^*_2 . Würde noch eine 3. Auspartiellierung verlangt werden, so würde sich diese auf Q^*_2 als Ausgangsmatrix richten.
5. Bei den in NOMINALE_V = ... genannten Variablen ist darauf zu achten, ob sie in der PARTIAL-Anweisung (1) vor AUS oder (2) nach AUS angeführt werden. Für diejenigen nominalen Variablen, die vor AUS stehen, werden nur die "notwendigen" Dummies gebraucht, d.h. die jeweils letzte Dummy ist überflüssig. Man sollte also die Anweisung DUMMY=-1; schreiben. Werden (versehentlich) doch *alle* Dummies gefordert, dann eliminiert ALMO automatisch (um lineare Abhängigkeiten zu verhindern) eine Dummy. Die Eliminierung wird dadurch vorgenommen, dass in der Matrix, die Zeile/Spalte der betreffenden Variablen =0 und das Diagonalglied =1 gesetzt wird. Werden Interaktionsvariable verwendet, dann werden diese, wie in Abschnitt P19.1.7 beschrieben, gebildet.
6. Ordinale Variable werden zwar von Almo akzeptiert. Sie werden im Kalkül des Auspartiellierens von Almo aber so behandelt wie wenn sie quantitativ wären. Das ist nicht sinnvoll.
7. Die Option SCHREIBE ERGEBNIS_MATRIX IN DATEI... hat eine gegenüber der Definition in Abschnitt P19.1.7 bzw. P20.8.5 abweichende Bedeutung. Es wird die Partialmatrix Q^* in die angegebene Datei geschrieben.

19.3.1 Errechnung einer Partialmatrix aus einer eingelesenen Korrelationsmatrix.

Betrachten wir folgendes Almo-Syntax-Programm.

```
Vereinbare  
Variable=20;  
Anfang  
Programm = 19;
```

```

Variable=V1:10;
Partial = V1,2 aus V3,4;
Ende_Programmparameter;
Lese Matrix aus Datei 1
".\Testdat\Korrmat";
GeheInProgramm
Ende

```

Aus der Datei "Korrmat" wird eine Korrelationsmatrix eingelesen. Für diese wird eine Partialmatrix errechnet. Die Korrelationsmatrix kann auch direkt hinter dem Wort "Ende" des obigen Algo-Programms von Benutzer geschrieben werden. Wie die Korrelationsmatrix in die Datei zu schreiben ist, wird in Teil 2, Abschnitt 43.1.1 dargestellt.

P19.4 Erzeugen einer Datei korrelierter Zufallsvariablen

Es soll eine Datenmatrix von standard-normalverteilten Zufallsvariablen erzeugt werden, die in einer gewünschten Weise korrelieren.

Beispiel: 6 Variable sollen in folgender Weise korrelieren

	v1	v2	v3	v4	v5	v6
v1	1.0					
v2	0.4	1.0				
v3	0.3	0.7	1.0			
v4	0.1	0.2	0.7	1.0		
v5	0.2	0.3	0.2	0.3	1.0	
v6	0.3	0.2	0.2	0.1	0.8	1.0

Die nachfolgende Programm-Maske erzeugt dann folgende Datei - wobei die 6 Variablen, dann ungefähr so korrelieren wie gewünscht.

Datensatz	v1	v2	v3	v4	v5	v6
1	-0.76	0.49	0.67	0.17	1.48	0.82
2	1.78	1.08	0.09	0.30	2.42	1.69
3	-0.25	-0.69	-0.71	-0.45	0.60	0.53
.
.
.

Die Variable sind standard-normalverteilt, d.h. sie haben einen Mittelwert von ca. 0 und eine Standardabweichung von ca. 1.

P19.4.1 Programm-Maske Prog19cm: Korrelierte Zufallsvariablen

Sie finden das Programm durch Klick auf „Verfahren/Korrelation“.

Prog19cm.Msk

Erzeuge eine Datenmatrix von standard-normalverteilten Zufallsvariablen, die in einer gewünschten Weise korrelieren

Beispiel: 6 Variable sollen in folgender Weise korrelieren

	U1	U2	U3	U4	U5	U6
U1	1.0					
U2	0.4	1.0				
U3	0.3	0.7	1.0			
U4	0.1	0.2	0.7	1.0		
U5	0.2	0.3	0.2	0.3	1.0	
U6	0.3	0.2	0.2	0.1	0.8	1.0

Das Programm erzeugt dann folgende Datei - wobei die 6 Variablen, dann (ungefähr) so korrelieren wie gewünscht

	U1	U2	U3	U4	U5	U6
	-0.76	0.49	0.67	0.17	1.48	0.82
	1.78	1.08	0.09	0.30	2.42	1.69
	-0.25	-0.69	-0.71	-0.45	0.60	0.53
	:	:	:	:	:	:

Die Variable sind standard-normalverteilt

Was ist ein Kurzprogramm ? -->

Bedienung -->

1 **Vereinbare Variable = 20 ; Speicher fuer mindestens 20 Variable**
UA = 6 ; Zeilen der Korrelationsmatrix
IA = 6,6 ; Zeilen und Spalten der Korr.matrix

2 **Option: Weitere Vereinbarungen - nur wenn Almo dazu auffordert**

3 **Zahl der Variablen (Spalten der Datenmatrix)**
 Zahl der Untersuchungseinheiten (Zeilen)

4 **Lese die gewünschte Korrelationsmatrix aus folgender Datei**
 "C:\Almo7\Testdat\Korr.mat"

5 **Die Variable umkodieren**
 etwa um einen anderen Mittelwert und eine andere Standardabweichung
 oder einen anderen Wertebereich zu erhalten

erzeuge zusätzliche Felder für Umkodierungen

6 **Schreibe die Datenmatrix in die folgende Datei**
 "C:\Almo7\Progs\Korrdaten.fre"

Erläuterungen zu den Eingabeboxen:

Eingabebox 1: Vereinbare

Vereinbare Variable =	20	:	Speicher fuer mindestens 20 Variable
UA =	6	:	Zeilen der Korrelationsmatrix
TA =	6,6	:	Zeilen und Spalten der Korr.matrix

Eingabefeld 1: Speicher für mindestens so viele Variable, wie erzeugt werden sollen; mindestens aber 20.

Eingabefeld 2: Zeilen der Korrelationsmatrix = Zahl der Variablen, die erzeugt werden sollen.

Eingabefeld 3: Zeilen und Spalten der Korr.matrix = Zahl der Variablen, die erzeugt werden sollen.

Eingabebox 3: Variable und Untersuchungseinheiten

<input type="text" value="6"/>	Zahl der Variablen <Spalten der Datenmatrix>
<input type="text" value="1000"/>	Zahl der Untersuchungseinheiten <Zeilen>

Geben Sie hier die Zahl der Variablen an, die erzeugt werden sollen, sowie die Zahl der Untersuchungseinheiten (=Zahl der Datensätze), die das Programm in die Datei schreiben soll.

Eingabebox 4: Gewünschte Korrelationsmatrix

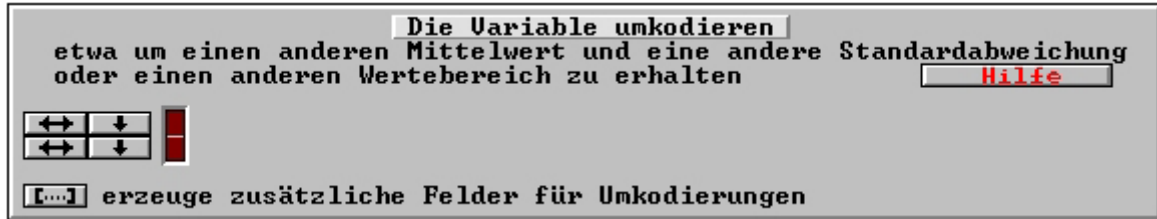
Lese die gewünschte Korrelationsmatrix aus folgender Datei	
<input type="text" value="C:\Almo7\Testdat\Korr.mat"/>	

Die Korrelationen, die zwischen den Variablen bestehen sollen müssen zuvor in eine Datei geschrieben werden. In unserem Beispiel heißt diese Datei "C:\Almo6\Testdakorr.mat". Sie enthält folgende Daten:

```
* * *
1.0
0.4 1.0
0.3 0.7 1.0
0.1 0.2 0.7 1.0
0.2 0.3 0.2 0.3 1.0
0.3 0.2 0.2 0.1 0.8 1.0
* *
```

Die Korrelationsmatrix wird als untere Dreiecksmatrix (inklusive Diagonale) geschrieben. Vor der Matrix müssen 3 Sterne und hinter der Matrix 2 Sterne geschrieben werden. **Beachte:** Die Sterne müssen durch ein Blank getrennt sein!

Eingabebox 5: Variable umkodieren



Almo erzeugt korrelierte, normalverteilte, standardisierte Werte. Die Variable haben also einen Mittelwert von ca. 0 und eine Standardabweichung von ca. 1 und sie sind normalverteilt.

Almo gibt den Mittelwert, die Standardabweichung, die Werte-Unter- und Obergrenze aus.

Wenn der Benutzer einen anderen Mittelwert und eine andere Standardabweichung wünscht, dann kann er das durch Umkodieren leicht erreichen. Beispiel: Die Variablen V4 bis V6 sollen einen Mittelwert von 10 und eine Standardabweichung von 5 haben.

Die Umkodierungsanweisung lautet dann:

`v4:6 (*5; +10)`

Die Variablen werden also zuerst mit der gewünschten Standardabweichung multipliziert. Dann wird der gewünschte Mittelwert hinzuaddiert.

Die Korrelationen zwischen den Variablen ändern sich durch Addieren/Subtrahieren und Multiplizieren/Dividieren nicht.

Die korrelierten standard-normalverteilten Variablen liegen etwa im Wertebereich von -4 bis +4. In unserem Beispiel ist die Untergrenze von V3=-4.0255 und die Obergrenze von V3=3.0906.

Der Benutzer kann nun z.B. wünschen, dass V3 mit seinen Werten bei 0 beginnt und bis 9 reicht. Er muß also zu den standardisierten Werten von V3 zuerst 4.0255 addieren. Dann ist der Wertebereich 0 bis 7.1161. Dann muß er noch mit $9/7.1161=1.26473771$ multiplizieren. Dann ist der Wertebereich 0 bis 9. V3 ist immer noch normalverteilt. Mittelwert und Standardabweichung haben sich geändert.

Der Benutzer muß also in das Eingabefeld schreiben:

`v3 (+4.0255 ; *1.2647377)`

Möglich wäre es auch anschließend zu runden, etwa auf Ganzzahligkeit zu runden:

`v3 (+4.0255 ; *1.2647377 ; runde 1)`

V3 nimmt dann die ganzzahligen Werte 1, 2, 3, 9 an. Die Korrelationen von V3 mit den anderen Variablen werden dadurch allerdings etwas verändert.

Möglich wäre es auch V3 um seinen Mittelwert M (der ja ca. 0 ist) zu dichotomisieren. Almo gibt in unserem Beispiel für V3 einen Mittelwert von 0.0801 aus. Die Umkodierungsanweisung müßte also lauten:

`v3 (-99 : 0.0801 = 0 ; 0.0801 : 99 = 1)`

Die Korrelationen von V3 mit den anderen Variablen werden dadurch erheblich verändert.

Eingabebox 6: Datenmatrix in eine Datei schreiben



Geben Sie den Namen der Datei an, die erzeugt werden soll und in die die Datensätze mit den korrelierten Zufallsvariablen geschrieben werden sollen.

P19.4.2 Zum Kalkül

Almo erzeugt zuerst eine Datenmatrix **D** mit standardnormalverteilten Zufallszahlen. Diese wird dann multipliziert mit der transponierten Cholesky-Matrix **T** der gewünschten Korrelationsmatrix **R**

$$\mathbf{D}' = \mathbf{D} * \mathbf{T}$$

wobei $\mathbf{T}' * \mathbf{T} = \mathbf{R}$

D' = Datenmatrix der korrelierten Zufallsvariablen

D = Datenmatrix mit standardnormalverteilten Zufallszahlen

T = oberes Dreieck der Cholesky-Matrix der gewünschten Korrelationsmatrix **R**

R = die gewünschte Korrelationsmatrix

D' und **D** sind $n*m$ - Matrizen, **T** und **R** sind $m*m$ - Matrizen wobei

n = Zahl der Zeilen (Datensätze)

m = Zahl der Variablen

Literatur

Die Literatur zu Korrelationskoeffizienten ist sehr umfangreich. Wir verweisen hier nur auf das Buch von

Bortz J., Lienert G. A, Boehnke K.: Verteilungsfreie Methoden in der Biostatistik; Springer Verlag: Berlin-Heidelberg 1990

Denz, H.: Regressionsanalyse mit ordinalen Variablen, in: Holm, K. (Hg.): Die Befragung 5, Francke, UTB 435,

Heinrich Potuschak: Der allgemeine Korrelationskoeffizient Groß-Gamma

0. Einführung

Um den Zusammenhang einer zweidimensionalen Zufallsvariablen (X,Y) zu messen, von der eine Stichprobe (X,Y) des Umfangs n gezogen ist, wurde von Daniels (1944)

der verallgemeinerte Korrelationskoeffizient $\Gamma = \frac{\sum_{i=1}^n \sum_{j=1}^n a_{ij} \cdot b_{ij}}{\sqrt{\sum \sum a_{ij}^2 \cdot \sum \sum b_{ij}^2}}$ eingeführt, wobei

a_{ij} und b_{ij} Scores sind, die den n_2 Wertepaaren (x_i, x_j) und (y_i, y_j) zugeordnet werden; ihre einzigen Einschränkungen sind $a_{ij} = -a_{ji}$ und $b_{ij} = -b_{ji}$.

Die Scores werden in folgender Weise kodiert:

- | | |
|--|---|
| 1. $a_{ij} = x_i - x_j$ und $b_{ij} = y_i - y_j$ | quantitative Differenz |
| 2. $a_{ij} = R(x_i) - R(x_j)$ und $b_{ij} = R(y_i) - R(y_j)$ | Rangdifferenz |
| 3. $a_{ij} = \text{sgn}(x_i - x_j)$ und $b_{ij} = \text{sgn}(y_i - y_j)$ | -1, 0, +1
entsprechend dem
Vorzeichen der Differenz |

Kendall (1962) hat Daniels Formel aufgegriffen, aber nur auf den Fall angewendet, dass beide Variable nach der oben unter 3 angegebenen Weise kodiert sind. Es entstand der bekannte ordinale Korrelationskoeffizient τ_b (tau-b) – gelegentlich auch als τ^{**} bezeichnet. Es gilt also τ_b ist gleich $\Gamma_{3,3}$. Mit dem ersten Index von $\Gamma_{3,3}$ wird die Definition von a_{ij} mit dem zweiten die von b_{ij} ausgedrückt. Sehr schnell wurde erkannt, dass $\Gamma_{1,1}$ identisch ist mit dem Pearson-Bravais Korrelationskoeffizient r und $\Gamma_{2,2}$ mit dem Spearman'schen Rangkorrelationskoeffizienten ρ_s .

Nach Kendall haben sich mit dem Versuch, einen allgemeinen Korrelationskoeffizienten auf Basis des Γ -Koeffizienten zu entwickeln, Hawkes (1971), Ploch (1974), R.B. Smith (1974) und Denz (1977, 1979) beschäftigt. Im Statistikprogramm Almo (Holm, 2000) ist die Berechnung des Γ -Koeffizienten enthalten, ebenfalls die in den folgenden Abschnitten vom Verfasser entwickelte Bestimmung der Varianz aller Γ -Koeffizienten.

Die konkrete Berechnung des allgemeinen Γ -Koeffizienten werden wir hier nicht vorführen. Sie entspricht der Vorgehensweise bei der Berechnung von Kendalls τ_b unter Verwendung der oben angegebenen drei Kodierungsweisen. Siehe dazu etwa Denz (1977, S. 109 ff) oder Bortz, Lienert, Boehnke (1990, S. 429 ff).

Ein Problem blieb immer offen: das der Varianz und damit der Signifikanz der Korrelationskoeffizienten für Variable gemischten Messniveaus, beispielsweise des Koeffizienten $\Gamma_{1,3}$ für quantitative und ordinale Variable. Mit diesen gemischten Koeffizienten und ihrer Varianz werden wir uns im folgenden befassen.

1. Verallgemeinerte Korrelationskoeffizienten

Sei (X,Y) eine zweidimensionale, reelle Zufallsvariable verteilt nach F_{xy} . Um den statistischen Zusammenhang zwischen X und Y zu messen, wurden in der Literatur unterschiedliche Korrelationskoeffizienten betrachtet. Entscheidend für die Definition solcher Koeffizienten ist das Skalenniveau der Variablen X und Y . Der klassische Produktmoment-Korrelationskoeffizient von Bravais-Pearson

$$\rho(X,Y) = \frac{\text{Cov}(X,Y)}{\sqrt{V(X) \cdot V(Y)}}$$

beispielsweise ist sinnvoll, wenn beide Variablen X,Y auf metrischem Niveau gemessen sind, denn er bleibt ungeändert bei Transformationen vom Typ $x \rightarrow x' = ax + b$ und $y \rightarrow y' = cy + d$ mit $\text{sgn}(a) = \text{sgn}(c)$ – Transformationen, wie sie bei einem Wechsel der Bezugspunkte und Maßeinheiten auftreten. $\rho(X,Y)$ ist aber nicht sinnvoll, wenn eine oder beide Variablen X,Y nur auf ordinalem Niveau gemessen wurden – denn in diesem Fall müsste $\rho(X,Y)$ invariant bleiben bei Transformationen vom Typ $x \rightarrow x' = s(x)$ und $y \rightarrow y' = t(y)$, mit streng monoton wachsenden Funktionen $s(x), t(y)$ – und dies ist bekanntlich nicht der Fall.

Um zu einer Familie von Korrelationskoeffizienten mit unterschiedlichen Invarianzeigenschaften zu gelangen, schreiben wir $\rho(X,Y)$ zunächst etwas um. Es gilt, falls $(X_1, Y_1), (X_2, Y_2)$ unabhängig nach F_{xy} verteilt sind:

$$\rho(X,Y) = \frac{E[(X_2 - X_1)(Y_2 - Y_1)]}{\sqrt{E[(X_2 - X_1)^2]} \cdot \sqrt{E[(Y_2 - Y_1)^2]}} = \rho(X_2 - X_1, Y_2 - Y_1).$$

Man erhält nun eine relativ allgemeine Familie von Korrelationskoeffizienten auf folgendem Weg: - Man wählt zwei Funktionen $a(X_1, X_2)$ und $b(Y_1, Y_2)$ mit

$$a(X_1, X_2) = -a(X_2, X_1), \quad b(Y_1, Y_2) = -b(Y_2, Y_1),$$

a ist in X_1 monoton fallend und in X_2 monoton steigend,
 b ist in Y_1 monoton fallend und in Y_2 monoton steigend.

- Man bildet $\rho[a(X_1, X_2), b(Y_1, Y_2)]$,
wobei wieder $(X_1, X_2), (Y_1, Y_2)$ unabhängig nach F_{xy} verteilt sind.

Folgende Sonderfälle sind von besonderem Interesse und werden in dieser Arbeit näher untersucht:

$$a_1(X_1, X_2) = X_2 - X_1, \quad b_1(Y_1, Y_2) = Y_2 - Y_1$$

$$a_2(X_1, X_2) = F_x^*(X_2) - F_x^*(X_1), \quad b_2(Y_1, Y_2) = F_y^*(Y_2) - F_y^*(Y_1),$$

$$a_3(X_1, X_2) = \text{sgn}(X_2 - X_1), \quad b_3(Y_1, Y_2) = \text{sgn}(Y_2 - Y_1),$$

wobei F_x^* und F_y^* die symmetrisierten Verteilungsfunktionen der Randverteilungen von X bzw. Y bezeichnen: $F_x^*(x_0) = P(X < x_0) + \frac{1}{2}P(X = x_0)$.

Wir betrachten somit die verallgemeinerten Korrelationskoeffizienten

$$\rho[a_k(X_1, X_2), b_\ell(Y_1, Y_2)] \quad \text{für } k, \ell = 1, 2, 3$$

und setzen zur Abkürzung:

$$\Gamma_{k\ell}(F_{xy}) = \rho[a_k(X_1, X_2), b_\ell(Y_1, Y_2)] \quad \text{für } k, \ell = 1, 2, 3$$

um die Abhängigkeit der Größen $\Gamma_{k\ell}$ von der Verteilung F_{xy} aufzuzeigen.

Dabei ist natürlich $\Gamma_{11}(F_{xy}) = \rho(X, Y)$ der gewöhnliche Pearson-Korrelationskoeffizient. Außerdem erkennt man leicht:

- die Größen $\Gamma_{1\ell} \dots \ell=1, 2, 3$ sind invariant gegenüber Transformationen $x \rightarrow x' = ax + b$ mit $a > 0$, sind also brauchbar, wenn X ein metrisches Merkmal ist, und
- die Größen $\Gamma_{2\ell}, \Gamma_{3\ell} \dots \ell=1, 2, 3$ sind invariant gegenüber streng monoton wachsenden Transformationen $x \rightarrow x' = s(x)$, und sind daher brauchbar, falls X auf ordinalem Niveau gemessen wird. Analoges gilt für die Variable Y .

Ist nun $[(x_i, y_i)_n, i=1, \dots, n] = (\underline{x}, \underline{y})$ eine Stichprobe vom Umfang n , und

$\hat{F}_{xy}(x, y) = \frac{1}{n} \cdot \sum_{i=1}^n \mathbb{1}(x_i \leq x, y_i \leq y)$ die zugehörige empirische Verteilungsfunktion, dann ist

$\hat{\Gamma}_{k\ell}(\underline{x}, \underline{y}) = \Gamma_{k\ell}(\hat{F}_{xy})$, also der $\Gamma_{k\ell}$ -Koeffizient, gebildet mit der Stichprobenverteilung von $(\underline{x}, \underline{y})$, der naheliegende Schätzer für $\Gamma_{k\ell}(F_{xy})$.

Es folgt daher für $k, \ell = 1, 2, 3$:

$$\begin{aligned} \hat{\Gamma}_{k\ell}(\underline{x}, \underline{y}) &= \Gamma_{k\ell}(\hat{F}_{xy}) = \rho[a_k(X_1, X_2), b_\ell(Y_1, Y_2) \mid (X_1, Y_1), (X_2, Y_2) \text{ ua } \sim \hat{F}_{xy}] = \\ &= \frac{\frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n a_k(x_i, x_j) \cdot b_\ell(y_i, y_j)}{\sqrt{\frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n a_k^2(x_i, x_j)} \cdot \sqrt{\frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n b_\ell^2(y_i, y_j)}} = \frac{\sum_{i=1}^n \sum_{j=1}^n a_k(x_i, x_j) \cdot b_\ell(y_i, y_j)}{\sqrt{\sum_{i=1}^n \sum_{j=1}^n a_k^2(x_i, x_j)} \cdot \sqrt{\sum_{i=1}^n \sum_{j=1}^n b_\ell^2(y_i, y_j)}}. \end{aligned}$$

Stichproben-Korrelationskoeffizienten dieser Art wurden bereits von Daniels (1944) als verallgemeinerte Korrelationskoeffizienten eingeführt. Dabei ist offenbar $\hat{\Gamma}_{11}(\underline{x}, \underline{y})$ der Stichproben-Produktmoment-Korrelationskoeffizient. Um die Bedeutung von $\hat{\Gamma}_{22}(\underline{x}, \underline{y})$ zu erkennen, seien mit $R(x_i)$ und $R(y_i)$ die Mitt-Ränge von x_i in \underline{x} bzw. von y_i in \underline{y} bezeichnet:

$$R(x_i) = \sum_{j=1}^n \mathbb{1}(x_j < x_i) + \frac{1}{2} \left(1 + \sum_{j=1}^n \mathbb{1}(x_j = x_i) \right) \quad \text{für } i=1, \dots, n.$$

Analoges gilt für $R(y_i)$. Dann gilt wegen

$$n \cdot \hat{F}_x^*(x_i) = \sum_{j=1}^n \mathbb{1}(x_j < x_i) + \frac{1}{2} \sum_{j=1}^n \mathbb{1}(x_j = x_i) = R(x_i) - \frac{1}{2} :$$

$$a_2(x_i, x_j) = \hat{F}_x^*(x_j) - \hat{F}_x^*(x_i) = \frac{R(x_j) - R(x_i)}{n}.$$

Somit ist

$$\hat{\Gamma}_{22}(\underline{x}, \underline{y}) = \frac{\sum_{i=1}^n \sum_{j=1}^n (R(x_j) - R(x_i)) \cdot (R(y_j) - R(y_i))}{\sqrt{\sum \sum (R(x_j) - R(x_i))^2} \cdot \sqrt{\sum \sum (R(y_j) - R(y_i))^2}},$$

und dies ist der Rang-Korrelationskoeffizient von Spearman für die Stichprobe $(\underline{x}, \underline{y})$.

Schließlich ist

$$\hat{\Gamma}_{33}(\underline{x}, \underline{y}) = \frac{\sum_{i=1}^n \sum_{j=1}^n \text{sgn}(x_j - x_i) \cdot \text{sgn}(y_j - y_i)}{\sqrt{\sum \sum \text{sgn}^2(x_j - x_i)} \cdot \sqrt{\sum \sum \text{sgn}^2(y_j - y_i)}}$$

der Korrelationskoeffizient von Kendall, denn enthalten die Stichproben \underline{x} und \underline{y} keine Bindungen, dann gilt offenbar

$$\hat{\Gamma}_{33}(\underline{x}, \underline{y}) = \frac{C - D}{n(n-1)} = \frac{2C}{n(n-1)} - 1,$$

wobei $C = \sum_{i=1}^n \sum_{j=1}^n \mathbb{1}((x_j - x_i)(y_j - y_i) > 0)$ die Anzahl konkordanter Paare

und $D = \sum_{i=1}^n \sum_{j=1}^n \mathbb{1}((x_j - x_i)(y_j - y_i) < 0)$ die Anzahl diskordanter Paare $((x_i, y_i), (x_j, y_j))$ bezeichnen.

Die Koeffizienten $\Gamma_{k\ell}(F_{xy})$ bzw. $\hat{\Gamma}_{k\ell}(\underline{x}, \underline{y})$ für $k \neq \ell$ sind in der Literatur kaum betrachtet. Sind die Zufallsvariablen (X, Y) unabhängig, dann gilt $\Gamma_{k\ell}(F_{xy}) = 0$ für alle $k, \ell = 1, 2, 3$ – d.h. X und Y sind unkorreliert im Sinne jedes dieser Korrelationskoeffizienten.

Ziel der vorliegenden Arbeit ist:

- im Fall der Unabhängigkeit von X und Y
- die Bestimmung der Varianzen $\sigma_{k\ell}^2(n)$ der Stichprobenverteilungen der Größen $\hat{\Gamma}_{k\ell}(\underline{x}, \underline{y})$ in Abhängigkeit von n , soweit sie nicht bekannt sind,
- der Nachweis, dass $\hat{\Gamma}_{k\ell}(\underline{x}, \underline{y})$ asymptotisch normalverteilt ist.

Daraus ergeben sich dann einfache Teststrategien zum asymptotischen Niveau α für das Testproblem

$H_0: (X, Y)$ unabhängig, $H_1: (X, Y)$ abhängig,

insbesondere für solche Fälle, wo die Variablen X und Y auf unterschiedlichem Niveau gemessen sind.

2. Die Varianz der Stichprobenverteilung von $\hat{\Gamma}_{k\ell}$

Die Stichprobenvarianzen σ_{kk}^2 der drei ungemischten Koeffizienten $\hat{\Gamma}_{kk}$ bei beliebig verteilter unabhängiger Grundgesamtheit $(\cdot, -)$ sind bekannt:

$$\begin{aligned}\sigma_{11}^2 &= \sigma^2(r) = 1/(n-1), \\ \sigma_{22}^2 &= \sigma^2(r_s) = 1/(n-1), \\ \sigma_{33}^2 &= \sigma^2(\tau) \text{ bzw. } \sigma^2(\tau^*) \text{ bzw. } \sigma^2(\tau^{**}),\end{aligned}$$

d.h. die Varianz von Kendalls τ bei keinen, ein- oder zweiseitigen Bindungen.

Nun wird $\sigma_{12}^2 = \sigma^2(\hat{\Gamma}_{12})$ hergeleitet, wobei die Beweisführung mit der von $\sigma^2(r)$ übereinstimmt, da diese für alle unabhängigen stetigen oder diskreten Zufallsvariablen gilt:

X und $R(Y)$ werden zentriert und normiert, wodurch die daraus entstehenden Lineartransformationen U und V ebenso unabhängig sind und folgende Eigenschaften aufweisen:

$$\sum u_i = \sum v_i = 0, \quad \sum u_i^2 = \sum v_i^2 = 1,$$

$$E(U_i) = E(V_i) = 0 = E(U \cdot V), \quad E(U_i^2) = E(V_i^2) = \frac{1}{n},$$

$$\hat{\Gamma}_{12}(X, Y) = r(X, R(Y)) = r(U, V) = \sum_{i=1}^n u_i \cdot v_i, \quad E(\hat{\Gamma}_{12}) = n \cdot E(U \cdot V) = 0.$$

Während $\text{Cov}(X_i, X_j) = \text{Cov}(Y_i, Y_j) = 0$ in Zufallsstichproben gilt, ist diese Beziehung für (U_i, U_j) und (V_i, V_j) wegen der Normierungen verletzt. $\text{Cov}(U_i, U_j) = \text{Cov}(V_i, V_j) = E(U_i \cdot U_j)$ läßt sich aus den zwei folgenden Gleichungen bestimmen:

$$\begin{aligned}E\left(\sum_{i=1}^n \sum_{j=1}^n U_i \cdot U_j\right) &= \sum E(U_i) \cdot \sum E(U_j) = 0 \\ &= E\left(\sum_i U_i^2 + (n^2 - n) \cdot \sum_{i \neq j} U_i \cdot U_j\right) = \sum E(U_i^2) + n(n-1) \cdot \sum_{i \neq j} E(U_i U_j) = \\ &= 1 + n(n-1) \cdot \text{Cov}(U_i, U_j).\end{aligned}$$

$$\text{Daraus folgen die Kovarianzen: } \text{Cov}(U_i, U_j) = \text{Cov}(V_i, V_j) = \frac{-1}{n(n-1)}.$$

Mit ihnen berechnet sich die Stichprobenvarianz als

$$\begin{aligned} \text{Var}(\hat{\Gamma}_{12}) &= \text{Var}\left(\sum U_i \cdot V_i\right) = n \cdot \text{Var}(U_i \cdot V_i) + (n^2 - n) \cdot \text{Cov}(U_i \cdot V_i, U_j \cdot V_j) = \\ &= n \cdot \left[E(U_i^2 \cdot V_i^2) - E^2(U_i \cdot V_i) \right] + (n^2 - n) \cdot \left[E(U_i \cdot V_i \cdot U_j \cdot V_j) - E(U_i \cdot V_i) \cdot E(U_j \cdot V_j) \right] = \\ &= n \cdot \left[E(U_i^2) \cdot E(V_i^2) - 0 \right] + n(n-1) \cdot \left[E(U_i \cdot U_j) \cdot E(V_i \cdot V_j) - 0 \cdot 0 \right] = \\ &= \frac{n}{n^2} + \frac{n(n-1)}{n^2(n-1)^2} = \frac{1}{n-1}. \end{aligned}$$

Aus Symmetriegründen gilt das Gleiche für die Varianz von $\hat{\Gamma}_{21} = r(R(X), Y)$. Somit gilt für alle $n > 1$ und $k, \ell \leq 2$: $\text{Var}(\hat{\Gamma}_{k\ell}) = \sigma_{k\ell}^2 = \frac{1}{n-1}$.

3. Die asymptotische Verteilung von $\hat{\Gamma}_{k\ell}$

In den drei ungemischten Fällen konvergieren die Stichprobenverteilungen von $r \cdot \sqrt{n-1}$, $r_S \cdot \sqrt{n-1}$ und $\tau/\sigma(\tau)$ bekanntlich schwach gegen die Standardnormalverteilung.

Sind die Zufallsvariablen X_i unabhängig und identisch verteilt mit $E(X_i) = \mu$ und $V(X_i) = \sigma^2$,

konvergiert die Folge der Verteilungen der standardisierten Summe $S_n = \frac{\sum_{i=1}^n X_i - n \cdot \mu}{\sqrt{n} \cdot \sigma}$ nach

dem zentralen Grenzwertsatz gegen eine $N(0,1)$ -Verteilung. In unserem Fall lauten die Zufallsvariablen $U_i \cdot V_i$ und deren Summe entspricht $\hat{\Gamma}_{12}$ oder $\hat{\Gamma}_{21}$. Ihr Erwartungswert $n \cdot \mu$ ist Null, ihre Standardabweichung $\sqrt{n} \cdot \sigma$ ist gleich $1/\sqrt{n-1}$.

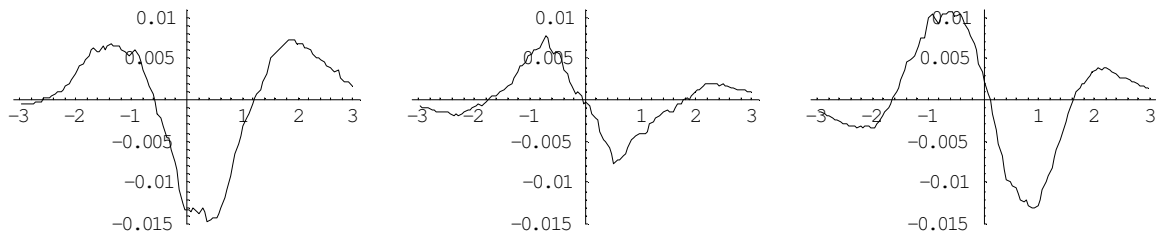
Nach dem ZGWS müsste nun die Folge der Verteilungen von $\hat{\Gamma}_{12} \cdot \sqrt{n-1}$ mit wachsendem n gegen eine $N(0,1)$ -Verteilung konvergieren. Dies trifft, wie folgende Simulationen zeigen, zu - allerdings mit anderem Konvergenzverhalten, da die Variablen $U_i \cdot V_i$ zwar identisch verteilt, aber nicht unabhängig sind. Ihre Kovarianz lautet:

$$\text{Cov}(U_i \cdot V_i, U_j \cdot V_j) = E(U_i \cdot V_i \cdot U_j \cdot V_j) - E(U_i \cdot V_i) \cdot E(U_j \cdot V_j) = E(U_i \cdot U_j) \cdot E(V_i \cdot V_j) - 0 = \frac{1}{n^2 \cdot (n-1)^2};$$

ihre Korrelation ist wegen $\text{Var}(U_i^2 \cdot V_i^2) = \frac{1}{n^2} \cdot \frac{1}{n^2}$ gleich $\frac{1}{(n-1)^2}$ und verschwindet mit zunehmender Stichprobengröße.

Somit sind unter $H_0 : \Gamma_{k\ell} = 0$ für $k, \ell \leq 2$ die Prüfgrößen $z_{k\ell} = \hat{\Gamma}_{k\ell} \cdot \sqrt{n-1}$ asymptotisch $N(0,1)$ -verteilt und geeignet, die Unabhängigkeit von X und Y zu testen.

Um die Konvergenzgeschwindigkeit von $z_{12} = \hat{\Gamma}_{12} \cdot \sqrt{n-1}$ zu veranschaulichen, wurden 50000 Stichproben des Umfangs $n=20$ simuliert. Als Verteilungsannahmen von X wurden – in der Reihenfolge der drei Abbildungen – die Normal-, Gleich- und Exponentialverteilung gewählt. Gegen die z -Achse aufgetragen sind die Differenzen $\hat{F}(z) - \Phi(z)$, die nach dem ZGWS mit zunehmendem Stichprobenumfang verschwinden. Im Bereich $|z| > 1.28$, in dem üblicherweise Testentscheidungen getroffen werden, ist als maximale absolute Differenz ca. 0.005 abzulesen. Dies bedeutet, dass sich bei dieser Stichprobengröße und einseitigen Entscheidungen das angenommene α -Niveau höchstens um einen halben Prozentpunkt vom tatsächlichen Fehler unterscheidet.



Die Differenzen $\hat{F}(z_F) - \Phi(z_F)$ der Fisher-Transformierten $z_F = \frac{1}{2} \ln \frac{1+\hat{F}}{1-\hat{F}} \sqrt{n-3}$ veränderten sich je nach Verteilungsannahme verschiedenartig, ihre Maximalwerte verringerten sich aber nur unwesentlich. Eine Verdoppelung des Stichprobenumfangs $n=40$ verbesserte die maximalen Differenzen der Verteilungsfunktionen um ca. ein Drittel.

4. Die übrigen gemischten Koeffizienten $\hat{\Gamma}_{k\ell}$

Es folgen algebraische Zusammenhänge, aus denen hervorgeht, wie sich die vier gemischten Koeffizienten $\hat{\Gamma}_{13}$, $\hat{\Gamma}_{23}$, $\hat{\Gamma}_{31}$ und $\hat{\Gamma}_{32}$ als konstante Vielfache der Produktmomentkoeffizienten $\hat{\Gamma}_{12}$, $\hat{\Gamma}_{22}$, $\hat{\Gamma}_{21}$ und $\hat{\Gamma}_{22}$ berechnen lassen.

Die drei möglichen Quadratsummen Q_{xk} , deren Wurzeln gemeinsam mit denen von $Q_{y\ell}$ bei

der Berechnung von $\hat{\Gamma}_{k\ell} = \frac{\sum_{i=1}^n \sum_{j=1}^n a_k(x_i, x_j) \cdot b_\ell(y_i, y_j)}{\sqrt{Q_{xk} \cdot Q_{y\ell}}}$ in den Nenner eingesetzt werden,

vereinfachen sich zu:

$$Q_{x1} = \sum_{i=1}^n \sum_{j=1}^n (x_i - x_j)^2 = 2n \sum x_i^2 - 2(\sum x_i)^2 = 2n^2 \sigma_x^2.$$

$$\text{Ebenso ergibt sich } Q_{y1} = 2n^2 \sigma_y^2.$$

$$Q_{x2} = \sum_{i=1}^n \sum_{j=1}^n (R(x_i) - R(x_j))^2 = 2n \sum R_i^2 - 2(\sum R_i)^2.$$

Bei der Methode, verbundenen ordinalen Beobachtungen den Mittelwert der betroffenen Rangplätze zuzuordnen, ist die Rangplatzsumme verbundener Ränge gleich der von ungebundenen und lautet $n(n+1)/2$.

Die Quadratsumme ungebundener Ränge lautet $n(n+1)(2n+1)/6$. Sind t geordnete Rangplätze von der Stelle i bis $i+t-1$ gebunden, vermindert sich die Quadratsumme bei

ungeradem Bindungsumfang um $2 \sum_{j=1}^{(t-1)/2} j^2$, und bei gerader Anzahl um $2 \sum_{j=1}^{t/2} \left(\frac{2j-1}{2}\right)^2$ - in beiden Fällen somit um $t(t^2-1)/12$, und dies unabhängig von der Stelle i . Sind weitere t_2 Rangplätze gebunden, beträgt die Verminderung zusätzlich $t_2(t_2^2-1)/12$, da deren Stellen sich mit den vorigen nicht überschneiden können. Existieren m_x Bindungen mit den Umfängen $\underline{t}_x = \{t_1, \dots, t_{m_x}\}$, lautet die gesamte Verminderung $\sum_{j=1}^{m_x} t_j(t_j^2-1)/12$, woraus folgt:

$$Q_{x2} = n \cdot [n(n^2 - 1) - \sum t_x (t_x^2 - 1)] / 6.$$

Ist \underline{t}_y die in beliebiger Reihenfolge angegebene Liste der Bindungsumfänge von Y , folgt:

$$Q_{y2} = n \cdot [n(n^2 - 1) - \sum t_y (t_y^2 - 1)] / 6.$$

$Q_{x3} = \sum_{i=1}^n \sum_{j=1}^n \text{sgn}^2(x_i - x_j)$. Die Quadratsumme dieser n^2 Vorzeichen ist $n(n-1)$. Sind t

geordnete Rangplätze von der Stelle i bis $i+t-1$ gebunden, vermindert sich die Quadratsumme um $t(t-1)$, unabhängig von der Stelle i . Existieren m_x Bindungen mit den Umfängen \underline{t}_x , lautet die gesamte Verminderung $\sum_{j=1}^{m_x} t_j(t_j - 1)$, woraus folgt:

$$Q_{x3} = n(n-1) - \sum t_x (t_x - 1).$$

Für die Vorzeichenquadrate von Y gilt $Q_{y3} = n(n-1) - \sum t_y (t_y - 1)$.

Die drei möglichen Produktsummen, die zur Berechnung von $\hat{\Gamma}_{k\ell}$ in den Fällen $k < \ell$ als Zähler

$P_{k\ell} = \sum_{i=1}^n \sum_{j=1}^n a_k(x_i, x_j) \cdot b_\ell(y_i, y_j)$ gebraucht werden, vereinfachen sich zu:

$$P_{12} = \sum_{i=1}^n \sum_{j=1}^n (x_i - x_j) \cdot (R(y_i) - R(y_j)) = 2n \cdot \sum x_i \cdot R(y_i) - 2 \cdot \sum x_i \cdot \sum R_i = 2n(\sum x_i R_i - n \cdot \bar{x} \cdot \bar{R}).$$

$$P_{13} = \sum_{i=1}^n \sum_{j=1}^n (x_i - x_j) \cdot \text{sign}(y_i - y_j) = 2 \cdot \sum_{i=1}^n \left(x_i \cdot \sum_{j=1}^n \text{sign}(y_i - y_j) \right).$$

Bei festgehaltenem Index i und $t = \sum_{j=1}^n \mathbb{1}(y_j = y_i)$ Beobachtungen mit Rangplatz $R(y_i)$ folgt

aus der Eigenschaft von Mitt-Rängen, dass $R(y_i) - (t+1)/2$ Beobachtungen kleiner, und $n - R(y_i) - (t-1)/2$ größer als y_i sind. Daraus folgt die Vorzeichensumme $2 \cdot R(y_i) - n - 1 = 2 \cdot (R(y_i) - \bar{R})$, und ergibt

$$P_{13} = 4 \cdot \sum_{i=1}^n x_i \cdot (R(y_i) - \bar{R}) = 4(\sum x_i \cdot R(y_i) - n \cdot \bar{x} \cdot \bar{R}) = \frac{2}{n} \cdot P_{12}.$$

P_{23} : Ersetzt man in P_{13} die Beobachtungen x_i durch ihre Rangplätze, ergibt sich

$$P_{23} = 4(\sum R(x_i) \cdot R(y_i) - n \cdot \bar{R}^2).$$

Ersetzt man in P_{12} ebenso, ergibt sich $P_{22} = 2n(\sum R(x_i) \cdot R(y_i) - n \cdot \bar{R}^2)$. Daraus folgt

$$P_{23} = \frac{2}{n} \cdot P_{22}.$$

Vertauscht man die Variablen X und Y, um die Zähler von $\hat{\Gamma}_{k\ell}$ für $k > \ell$ zu vereinfachen, folgt $P_{31} = \frac{2}{n} \cdot P_{21}$ und $P_{32} = \frac{2}{n} \cdot P_{22} = P_{23}$.

Aus den obigen Zusammenhängen folgt

$$\hat{\Gamma}_{13} = \frac{P_{13}}{\sqrt{Q_{x1} \cdot Q_{y3}}} = \frac{2}{n} \cdot \frac{P_{12}}{\sqrt{Q_{x1} \cdot Q_{y3}}} \quad \text{und} \quad \hat{\Gamma}_{23} = \frac{P_{23}}{\sqrt{Q_{x2} \cdot Q_{y3}}} = \frac{2}{n} \cdot \frac{P_{22}}{\sqrt{Q_{x2} \cdot Q_{y3}}}.$$

$\hat{\Gamma}_{13}$ läßt sich als Vielfaches von $\hat{\Gamma}_{12}$ darstellen. Bildet man den Quotienten $\hat{\Gamma}_{13}/\hat{\Gamma}_{12}$, entsteht eine Konstante C_{ty} , die allein von n und den Bindungsumfängen t_y abhängt:

$$C_{ty} = \frac{2}{n} \cdot \sqrt{\frac{Q_{y2}}{Q_{y3}}} = \sqrt{\frac{2}{3n} \cdot \frac{n(n^2 - 1) - \sum t_y(t_y^2 - 1)}{n(n-1) - \sum t_y(t_y - 1)}}.$$

Weist Y keine Bindungen auf, lautet sie $C = \sqrt{\frac{2}{3} \cdot \frac{n+1}{n}}$.

Da sich bei der Division $\hat{\Gamma}_{23}/\hat{\Gamma}_{22}$ die Quadratsumme Q_{x2} kürzt, gilt ebenso $\hat{\Gamma}_{23} = \hat{\Gamma}_{22} \cdot C_{ty}$.

Für die Zusammenhänge $\hat{\Gamma}_{31} = \hat{\Gamma}_{21} \cdot C_{tx}$ und $\hat{\Gamma}_{32} = \hat{\Gamma}_{22} \cdot C_{tx}$ sind zur Berechnung der Konstanten die Bindungsumfänge t_x einzusetzen.

Für die Stichprobenvarianzen der vier gemischten Koeffizienten gilt

$$\sigma_{13}^2 = \sigma_{23}^2 = \frac{C_{ty}^2}{n-1} \quad \text{und} \quad \sigma_{31}^2 = \sigma_{32}^2 = \frac{C_{tx}^2}{n-1}.$$

Daraus folgt für die Prüfgrößen: $z_{13} = z_{12}$, $z_{31} = z_{21}$ und $z_{23} = z_{32} = z_{22}$.

Daran erkennt man, dass sich die Berechnung der vier gemischten Koeffizienten erübrigt, falls nur ihre Signifikanz von Interesse ist. Von den neun möglichen Γ -Koeffizienten verbleibt allein Kendalls τ bzw. $\hat{\Gamma}_{33}$, das nicht als PM-Koeffizient berechenbar ist.

5. Ein Demonstrationsbeispiel zur Berechnung verschiedener Koeffizienten $\hat{\Gamma}(X, Y)$ samt Prüfgrößen $z = \hat{\Gamma} / \sigma(\hat{\Gamma})$

Stichprobenumfang: $n = 6$

Stichprobe: $\underline{x} = \{3, 3, 3, 5, 10, 10\}$, $\bar{x} = 5.6$, $\sigma_x^2 = 9.8$
 $\underline{y} = \{1, 2, 3, 2, 4, 3\}$, $\bar{y} = 2.5$, $\sigma_y^2 = 0.916$

Rangplätze: $R(\underline{x}) = \{2, 2, 2, 4, 5.5, 5.5\}$
 $R(\underline{y}) = \{1, 2.5, 4.5, 2.5, 6, 4.5\}$

Anzahl und Umfänge der Bindungen: $m_x = 2$, $t_x = \{3, 2\}$
 $m_y = 2$, $t_y = \{2, 2\}$

Je Variable können deren n^2 Differenzen auf eine von drei Weisen definiert werden:

- als Differenzen der Ausprägungen, z.B. $A1_{15} = x_1 - x_5 = 3 - 10 = -7$,
- als Differenzen der Rangplätze, z.B. $B2_{15} = R(y_1) - R(y_5) = 1 - 6 = -5$,
- nur als Größer-Kleiner-Gleich-Beziehungen, d.h. als Signum paarweiser Differenzen, z.B. $B3_{15} = \text{sgn}(y_1 - y_5) = \text{sgn}(1 - 4) = -1$.

Die drei möglichen Differenzmatrizen der Variablen X werden A1 bis A3 genannt, die von Y lauten B1 bis B3:

$$A1 = \begin{pmatrix} 0 & 0 & 0 & -2 & -7 & -7 \\ 0 & 0 & 0 & -2 & -7 & -7 \\ 0 & 0 & 0 & -2 & -7 & -7 \\ 2 & 2 & 2 & 0 & -5 & -5 \\ 7 & 7 & 7 & 5 & 0 & 0 \\ 7 & 7 & 7 & 5 & 0 & 0 \end{pmatrix}, \quad A2 = \begin{pmatrix} 0 & 0 & 0 & -2 & -3.5 & -3.5 \\ 0 & 0 & 0 & -2 & -3.5 & -3.5 \\ 0 & 0 & 0 & -2 & -3.5 & -3.5 \\ 2 & 2 & 2 & 0 & -1.5 & -1.5 \\ 3.5 & 3.5 & 3.5 & 1.5 & 0 & 0 \\ 3.5 & 3.5 & 3.5 & 1.5 & 0 & 0 \end{pmatrix}, \quad A3 = \begin{pmatrix} 0 & 0 & 0 & -1 & -1 & -1 \\ 0 & 0 & 0 & -1 & -1 & -1 \\ 0 & 0 & 0 & -1 & -1 & -1 \\ 1 & 1 & 1 & 0 & -1 & -1 \\ 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 \end{pmatrix}$$

$$B1 = \begin{pmatrix} 0 & -1 & -2 & -1 & -3 & -2 \\ 1 & 0 & -1 & 0 & -2 & -1 \\ 2 & 1 & 0 & 1 & -1 & 0 \\ 1 & 0 & -1 & 0 & -2 & -1 \\ 3 & 2 & 1 & 2 & 0 & 1 \\ 2 & 1 & 0 & 1 & -1 & 0 \end{pmatrix}, \quad B2 = \begin{pmatrix} 0 & -1.5 & -3.5 & -1.5 & -5 & -3.5 \\ 1.5 & 0 & -2 & 0 & -3.5 & -2 \\ 3.5 & 2 & 0 & 2 & -1.5 & 0 \\ 1.5 & 0 & -2 & 0 & -3.5 & -2 \\ 5 & 3.5 & 1.5 & 3.5 & 0 & 1.5 \\ 3.5 & 2 & 0 & 2 & -1.5 & 0 \end{pmatrix}, \quad B3 = \begin{pmatrix} 0 & -1 & -1 & -1 & -1 & -1 \\ 1 & 0 & -1 & 0 & -1 & -1 \\ 1 & 1 & 0 & 1 & -1 & 0 \\ 1 & 0 & -1 & 0 & -1 & -1 \\ 1 & 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & -1 & 0 \end{pmatrix}$$

Die Quadratsummen Q_{xk} und Q_{yk} , die in die Nenner der Γ -Koeffizienten eingesetzt werden, lauten $Q_x = \{712, 180, 22\}$ und $Q_y = \{66, 198, 26\}$. Für sie gelten folgende Zusammenhänge, wobei sinngemäß für X, t_x und A das gleiche gilt wie für Y, t_y und B:

$$\sum_{i=1}^n \sum_{j=1}^n A1_{ij}^2 = 2n^2 \sigma_x^2 = 72 \cdot 9.8 = 712, \quad \sum_{i=1}^n \sum_{j=1}^n B1_{ij}^2 = 72 \cdot 0.916 = 66$$

$$\sum_{i=1}^n \sum_{j=1}^n A2_{ij}^2 = \frac{n}{6} [n(n^2 - 1) - \sum_{i=1}^{m_x} t_{x_i} (t_{x_i}^2 - 1)] = 210 - 24 - 6 = 180, \quad \sum_{i=1}^n \sum_{j=1}^n B2_{ij}^2 = 210 - 6 - 6 = 198$$

$$\sum_{i=1}^n \sum_{j=1}^n A_{ij}^2 = n(n-1) - \sum_{i=1}^{\text{mx}} t_{x_i}(t_{x_i} - 1) = 30 - 6 - 2 = 22,$$

$$\sum_{i=1}^n \sum_{j=1}^n B_{ij}^2 = 30 - 2 - 2 = 26.$$

Die neun möglichen Produktsummen $P_{k\ell}$, die in die Zähler der Γ -Koeffizienten eingesetzt

werden, lauten $\underline{P} = \sum_{i=1}^n \sum_{j=1}^n A_{ij} \cdot B_{ij} = 2 \cdot \begin{pmatrix} 78 & 135 & 45 \\ 36 & 61.5 & 20.5 \\ 12 & 20.5 & 7 \end{pmatrix}$.

Für die Produktsummen P_{13} , P_{23} , P_{21} und P_{31} gelten folgende Zusammenhänge: sie sind die $2/n$ -fachen Werte von P_{12} , P_{22} , P_{22} und P_{21} .

Es ergeben sich folgende neun möglichen Koeffizienten, gerundet auf 4 Dezimalen:

$$\hat{\Gamma}(X, Y) = \frac{\sum_{i=1}^n \sum_{j=1}^n A_{ij} \cdot B_{ij}}{\sqrt{\sum_{i=1}^n \sum_{j=1}^n A_{ij}^2 \cdot \sum_{i=1}^n \sum_{j=1}^n B_{ij}^2}} = \begin{pmatrix} 0.7196 & 0.7191 & 0.6615 \\ 0.6606 & 0.6515 & 0.5993 \\ 0.6298 & 0.6212 & 0.5854 \end{pmatrix}.$$

Die Produktmomentkorrelation $r(X, Y)$, Spearmans $\rho_s(X, Y)$ und Kendalls $\tau^{**}(X, Y)$ berechnen sich entweder nach den bekannten Formeln als $\{r, \rho_s, \tau^{**}\}$ oder durch Einsetzen von $(A1, B1)$, $(A2, B2)$ bzw. $(A3, B3)$ in $\hat{\Gamma}_{kk}(X, Y)$.

$\hat{\Gamma}_{12}(X, Y)$ berechnet sich entweder als PM-Korrelation $\rho[X, R(Y)]$ oder durch Einsetzen von $(A1, B2)$ in den Γ -Kalkül. $\hat{\Gamma}_{21}(X, Y)$ berechnet sich entweder als PM-Korrelation $\rho[R(X), Y]$ oder durch Einsetzen von $(A2, B1)$ in den Γ -Kalkül.

Die übrigen vier gemischten Koeffizienten $\hat{\Gamma}_{13}(X, Y)$, $\hat{\Gamma}_{31}$, $\hat{\Gamma}_{23}$, und $\hat{\Gamma}_{32}$ berechnen sich entweder durch Einsetzen von $(A1, B3)$, $(A3, B1)$, $(A2, B3)$ bzw. $(A3, B2)$ in den Γ -Kalkül oder durch folgende Zusammenhänge:

$$\hat{\Gamma}_{13} = \hat{\Gamma}_{12} \cdot C_{ty} = \hat{\Gamma}_{12} \cdot \sqrt{\frac{2}{3n} \cdot \frac{n(n^2 - 1) - \sum t_y(t_y^2 - 1)}{n(n-1) - \sum t_y(t_y - 1)}} = 0.7191 \cdot \sqrt{\frac{2}{18} \cdot \frac{210 - 12}{30 - 4}} = 0.6615,$$

$$\hat{\Gamma}_{31} = \hat{\Gamma}_{21} \cdot C_{tx} = \hat{\Gamma}_{21} \cdot \sqrt{\frac{2}{3n} \cdot \frac{n(n^2 - 1) - \sum t_x(t_x^2 - 1)}{n(n-1) - \sum t_x(t_x - 1)}} = 0.6606 \cdot \sqrt{\frac{2}{18} \cdot \frac{210 - 30}{30 - 8}} = 0.6298,$$

$$\hat{\Gamma}_{23} = \hat{\Gamma}_{22} \cdot C_{ty} = 0.5993, \text{ sowie } \hat{\Gamma}_{32} = \hat{\Gamma}_{22} \cdot C_{tx} = 0.6212.$$

Die Umrechnungsfaktoren $C_{ty} = \sqrt{\frac{11}{13}}$ und $C_{tx} = \sqrt{\frac{10}{11}}$ sind Konstante, die nur von n und den Bindungsumfängen von \underline{y} bzw. \underline{x} abhängen. Würde die Stichprobe keine Bindungen aufweisen, d.h. $\underline{t}_x = \underline{t}_y = \emptyset$, wäre der Umrechnungsfaktor in allen vier Fällen der gleiche:

$$C = \sqrt{\frac{2}{3} \cdot \frac{n+1}{n}} = 0.8819$$

Es besteht folgende Invarianzeigenschaft: die Koeffizienten $\hat{\Gamma}_{12}(X, Y)$ und $\hat{\Gamma}_{13}(X, Y)$ bleiben

unverändert bei linearen Transformationen von X (z.B. $X' = 2X-3$) und bei streng monoton wachsenden Transformationen von Y (z.B. $Y' = \ln Y$).

Die Prüfvarianzen lauten $\underline{\sigma}^2 = \begin{pmatrix} 0.2 & 0.2 & 0.1692 \\ 0.2 & 0.2 & 0.1692 \\ 0.1818 & 0.1818 & 0.1552 \end{pmatrix}$.

Es ergeben sich $\sigma_{11}^2 = \sigma_{12}^2 = \sigma_{21}^2 = \sigma_{22}^2 = \frac{1}{n-1} = 0.2$ und $\sigma_{33}^2 = 0.1552$ als Varianz von τ^{**} .

Die übrigen vier Prüfvarianzen berechnen sich mit den beiden quadrierten konstanten Vielfachheiten von 0.2, mit welchen die Koeffizienten zusammenhängen.

Die 9 Prüfgrößen $z_{k\ell} = \frac{\hat{\Gamma}_{k\ell}}{\sigma_{k\ell}}$ sowie deren Fisher-Transformierte $\frac{1}{2} \ln \frac{1+\hat{\Gamma}}{1-\hat{\Gamma}} \sqrt{n-3}$ (in den Fällen,

wo $\hat{\Gamma}$ als PM-Korrelation berechenbar ist) sind asymptotisch $N(0,1)$ -verteilt und lauten

$$\underline{z} = \begin{pmatrix} 1.6092 & 1.6080 & 1.6080 \\ 1.4771 & 1.4569 & 1.4569 \\ 1.4771 & 1.4569 & 1.4857 \end{pmatrix}, \quad \underline{z}_F = \begin{pmatrix} 1.5708 & 1.5689 & - \\ 1.3750 & 1.3475 & - \\ - & - & - \end{pmatrix}.$$

Die Prüfgrößen $z_{12} = z_{13}$, $z_{21} = z_{31}$ und $z_{22} = z_{23} = z_{32}$ sind identisch, da sich bei der Division $\hat{\Gamma}_{k3} = C_{ty} \cdot \hat{\Gamma}_{k2}$ durch $\sigma_{k3} = C_{ty} \cdot \sigma_{k2}$ oder bei $\hat{\Gamma}_{3k} = C_{tx} \cdot \hat{\Gamma}_{2k}$ durch $\sigma_{3k} = C_{tx} \cdot \sigma_{2k}$ die Konstante kürzt. Aus diesem Grund erübrigt sich die Berechnung der Signifikanz der gemischten Koeffizienten $\hat{\Gamma}_{13}$, $\hat{\Gamma}_{23}$, $\hat{\Gamma}_{31}$ und $\hat{\Gamma}_{32}$. Da $\hat{\Gamma}_{12}$, $\hat{\Gamma}_{21}$ und die drei ungemischten Koeffizienten als PM-Korrelationen berechenbar sind, muss nur im Fall $k=\ell=3$ der Γ -Kalkül verwendet werden – ist man an den numerischen Werten der vier gemischten Koeffizienten interessiert, können sie wie oben angeführt als Vielfache von $\hat{\Gamma}_{12}$, $\hat{\Gamma}_{22}$ und $\hat{\Gamma}_{21}$ berechnet werden.

Berechnet man (trotz des kleinen Stichprobenumfangs) Überschreitungswahrscheinlichkeiten unter der Nullhypothese, (X,Y) entstamme einer unkorrelierten Verteilung, ergeben sich am Niveau $\alpha = 0.05$ die einseitigen p-values

$$P(\underline{z} > 1.645) = \begin{pmatrix} 0.0538 & 0.0539 & 0.0539 \\ 0.0698 & 0.080 & 0.0726 \\ 0.0698 & 0.0726 & 0.0687 \end{pmatrix}.$$

Relativierte Γ - Koeffizienten:

Ordnet man \underline{x} und \underline{y} jeweils aufsteigend (bzw. \underline{y} absteigend, wenn $\hat{\Gamma}_{kl} < 0$) und berechnet man $\hat{\Gamma}^0$, kann man diese Koeffizienten als „maximal bei fester Bindungsstruktur“ bezeichnen, und $\hat{\Gamma}/|\hat{\Gamma}^0|$ als relativierte Koeffizienten. Ihre Prüfgrößen bleiben unverändert, ihre möglichen Maxima sind nur in Ausnahmefällen gleich ± 1 . Berechnet man alle neun möglichen $\hat{\Gamma}/|\hat{\Gamma}^0|$, ergeben sich die selben Identitäten wie bei den neun Prüfgrößen; dies kann ebenso algebraisch bewiesen werden. Im Demonstrationsbeispiel ist nur $\underline{y} = \{1,2,2,3,3,4\}$ zu ordnen und man berechnet:

$$\hat{\Gamma}^0 = \begin{pmatrix} 0.8303 & 0.8469 & 0.7791 \\ 0.8808 & 0.9058 & 0.8332 \\ 0.8398 & 0.8636 & 0.8362 \end{pmatrix} \quad \text{und} \quad \hat{\Gamma}/|\hat{\Gamma}^0| = \begin{pmatrix} 0.8667 & 0.8491 & 0.8491 \\ 0.75 & 0.7193 & 0.7193 \\ 0.75 & 0.7193 & 0.7 \end{pmatrix}.$$

Literatur zum Groß-Gamma-Koeffizienten

- Bortz, Lienert, Boehnke:** Verteilungsfreie Methoden in der Biostatistik, Berlin, Heidelberg, 1990
- Daniels, H. E. (1944):** The relation between measures of correlation in the universe of sample permutations. *Biometrika* 35, 129-135.
- Denz, H.:** Das Groß-Gamma- Modell, in: Holm: Befragung 6, Francke-Verlag, UTB 436
- Denz, H.:** Die Einbeziehung ordinaler Variablen in das ALM, in Holm: Almo-Statistiksystem, Handbuch zu Allgemeines Lineares Modell
- Hawkes, R. K.:** The multivariate analysis of ordinal measures, *American Journal of Sociology*, Vol. 76
- Kendall, M. G. (1962):** Rank correlation methods, 3rd ed., London: C. Griffin and Co.
- Ploch, D. R.:** Ordinal measures of association and the general linear model, in H. M. Blalock (ed). *Measurement in the social sciences*, Chicago, 1974
- Smith, R. B.:** Continuities in ordinal path-analysis, *Social Forces*, Vol. 53, 1974