



Zwei- und dreidimensionale Tabellierung

P10

Hermann Denz, Kurt Holm

Almo Statistik-System
www.almo-statistik.de
holm@almo-statistik.de
kurt.holm@jku.at

Siehe auch die beiden Almo-Dokumente Nr. 27 "Eindimensionale Häufigkeitsverteilung" und Nr. 2 "Beliebig-dimensionale Tabellierung".

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)
1. 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
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)
14. Ereignisanalyse: Sterbetafel-Methode, Kaplan-Meier-Schätzer, Cox-Regression.PDF (1,5 MB)
15. Faktorenanalyse.PDF (1,6 MB)
16. Konfirmatorische Faktorenanalyse.PDF (0,3 MB)
17. Clusteranalyse.PDF (3 MB)
18. Pisa 2012 Almo-Daten und Analyse-Programme.PDF (17 KB)
19. Guttman- und Mokken-Skalierung.PFD (0.8 MB)
20. Latent Structure Analysis.PDF (1 MB)
21. Statistische Algorithmen in C (80 KB)
22. Conjoint-Analyse (PDF 0,8 MB)
23. Ausreisser entdecken (PDF 170 KB)
24. Statistische Datenanalyse Teil I, Data Mining I
25. Statistische Datenanalyse Teil II, Data Mining II
26. Statistische Datenanalyse Teil III, Arbeiten mit Almo-Datenanalyse-System
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)

INHALTSVERZEICHNIS

P10 PROGRAMM 10: Zwei- und dreidimensionale Tabellierung	4
P10.2 Ausgabe der zweidimensionalen Tabelle.....	17
P10.3 Ausgabe der dreidimensionalen Tabelle	22
P10.4 Die statistischen Masszahlen bei Standardausgabe.....	25
P10.4.1 Chi-Quadrat-Test	25
P10.4.2 Korrelationskoeffizienten für nominale Variable	27
P10.4.3 Korrelationskoeffizienten für ordinale Variable	28
P10.4.4 Korrelationskoeffizienten für quantitative Variable	29
P10.4.5 Korrelationskoeffizienten für gemischt ordinale und quantitative Variable	30
P10.5 Programm-Maske mit Optionen	30
P10.7 Erläuterungen zu den einzelnen Optionen	37
P10.7.1 Erwartungswerte, Chi-Quadrat-Beiträge, Konfigurations-frequenzanalyse KFA	37
P10.7.2 Tests für abhängige Stichproben.....	39
P10.7.3 Ridits.....	41
P10.7.4 Exakter Fisher-Test - Exakter Freeman-Halton-Test.....	43
P10.7.5 Haldane-Dawson-Test	45
P10.7.6 Ulemans exakter Rangaufteilungs-U-Test.....	46
P10.7.7 Kolmogorov-Smirnov-Test (KS-Test) für 2 unabhängige Stichproben.....	47
P10.7.8 Konkordanz	48
P10.7.9 Polychorische Korrelation	49
P10.9 Eingabe von schon ausgezählten Tabellen.....	50
Literatur	53

P10 PROGRAMM 10: Zwei- und dreidimensionale Tabellierung

Programm 10 ermittelt zwei- oder dreidimensionale Häufigkeitstabellen und berechnet für diese eine Reihe von Koeffizienten. Außerdem können schon fertig ausgezählte Tabellen selbst eingegeben werden - mit dem Ziel, für diese verschiedene Koeffizienten zu berechnen. Das Programm wurde von **Hermann Denz** geschrieben. Mehrere Programmteile sowie die nachfolgend unter 4d, 6d, 7 bis 11,13 angegebenen Tests, wurden von Kurt Holm programmiert, der auch den vorliegende Text verfasste. Der unter 12 angegebene exakte Uleman-Test wurde von **Helmut Hoffmann** programmiert.

Folgende Typen von Tabellen werden gebildet.

1. Zweidimensionale Tabelle

Alter mit Leistung

		Leistung		Summe
		schlecht	gut	
Alter	jung	5	24	29
	alt	10	22	32
Summe		15	46	61

Die Variable vorne in der Tabelle (im Beispiel *Alter*) wird **Zeilenvariable** genannt, die Variable oberhalb der Tabelle (im Beispiel *Leistung*) **Spaltenvariable**. Die Variablen können beliebig viele Ausprägungen besitzen.

2. Dreidimensionale Tabelle

		Partialtabelle 1 Geschlecht: männlich			Partialtabelle 2 Geschlecht: weiblich		
		Leistung		Summe	Leistung		Summe
		niedrig	hoch		niedrig	hoch	
Alter	jung	4	14	18	1	10	11
	alt	8	8	16	2	14	16
Summe		12	22	34	3	24	27

Die 3. Variable (im Beispiel *Geschlecht*) wird **Kontrollvariable** genannt. Diese kann mehr als 2 Ausprägungen besitzen. Die Teiltabellen je Ausprägung der Kontrollvariablen werden **Partialtabellen** genannt. Besitzt die Spaltenvariable viele Ausprägungen oder besitzt die Kontrollvariable 3 oder mehr Ausprägungen, dann ist es nicht mehr möglich, die Partialtabellen nebeneinander zu stellen.

In Almo werden deswegen die Partialtabellen nicht, wie oben gezeigt, nebeneinander ausgegeben, sondern untereinander - in dieser Weise

Partialtabelle 1

Alter mit Leistung für Geschlecht: männlich

		Leistung		Summe
		schlecht	gut	
Alter	jung	4	14	18
	alt	8	8	16
Summe		12	22	34

Partialtabelle 2

Alter mit Leistung für Geschlecht: weiblich

		Leistung		Summe
		schlecht	gut	
Alter	jung	1	10	11
	alt	2	14	16
Summe		3	24	27

Weitere Tabellen-Typen werden im Almo-Dokument Nr 2 "Beliebig-dimensionale Tabellierung" dargestellt.

Für die Tabellen werden folgende Verfahren und Koeffizienten gerechnet.

1. Chi-Quadrat-Test für 2-dimensionale Tabellen (bzw. Partialtabellen)
2. Korrelationskoeffizienten für nominale Variable
 - a. Kontingenzkoeffizient C (auf 0-1 normiert)
 - b. Tschuprow's T
 - c. Cramer's V
 - d. Lambda
 - e. Vierfelderkorrelation Phi
3. Korrelationskoeffizienten für ordinale Variable
 - a. Gamma
 - b. Kendalls tau-b
 - c. Biseriales tau-b
 - d. Spearmans Rho
 - e. Polychorische Korrelation
4. Korrelationskoeffizienten für quantitative Variable
 - a. Produkt-Moment-Korrelation r
 - b. punktbiseriale Korrelation r (p.bis) (unabh.Variable dichotom.)
 - c. Eta (unabhängige Variable nominal, abh. Variable quantitativ)
 - d. Tetrachorische Korrelation
5. Ridits, Signifikanz der paarweisen Riditdifferenzen
6. Test für verbundene Stichproben (z.B. Messwiederholungen)
 - a. t-Test
 - b. Wilcoxon Vorzeichenrangtest
 - c. Zeichentest

- d. McNemar-Test
- 7. Kolmogorov-Smirnov 2-Stichprobentest
- 8. Exakter Fisher-Test für 2x2 Tabellen
- 9. Exakter Freeman-Halton-Test für Tabellen größer als 2*2
- 10. Haldane-Dawson-Test für große, aber schwach besetzte Tabellen
- 11. Konfigurationsfrequenzanalyse für 2-dimensionale Tabellen mit exaktem Binomialtest (für beliebig-dimensionale Tabellen in P11)
- 12. Ulemans exakter Rangaufteilungs U-Test
- 13. Kappa-Koeffizient der Urteilsübereinstimmung

P10.0 Eingabe mit Programm-Maske

Prog10m1.Msk Kurzprogramm
 2- und 3-dimensionale Tabellierung
 für Variable (auch) mit Dezimalwerten

Das Programm liefert Tabellen folgender Art:

		Beruf		
		Arbeiter	Angestell	Selbständ
Geschl.	männlich	48	18	8
	weiblich	38	11	4

Bei 3-dimensionaler Tabellierung werden für die 3. Variable Partialtabellen der ersten beiden Variablen gebildet

Für die Tabellen werden folgende Koeffizienten berechnet:
 Chi-Quadrat, Kontingenzkoeffizient, Tschuprows I, Cramers U,
 Lambda, Gamma, Kendalls tau-b, r, Spearmans Rho, punkt-
 biseriales r, Phi, Eta, tetrachorisches r.

Grafik: Balkendiagramme

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

Speicher fuer x Variable Hilfe

1 Vereinarbare Variable= 20 ;

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

Datei der Variablenamen Hilfe

3 ↔ 📁 "C:\Almo7\Testdat\Varname2.nam"
↔ ↓ zeige zeige = Namensdatei in Output zeigen
 leer = nicht

Freie Namensfelder Hilfe

4 ↔ ↔ |
[...] erzeuge zusätzliche Namensfelder

Datei aus der gelesen wird Hilfe
 bei Datei-Problemen

5 📁 "C:\Almo7\TESTDAT\Almdez.fre"
📄 frei Format der Daten Hilfe
↔ 📄 U1:10 der Datensatz enthält diese Variablen
 Bei Format DIREKT schreiben Sie: alle_U

6 ↓ Wenn Dateiformat FIX oder Nicht-Standard-FREI Hilfe

P10.0.1. Erläuterungen zu den Boxen: Box 1 bis 6

Box 1: Speicher für x Variable

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

Box 2: Optionen: Weitere Vereinbarungen

Siehe P0.2

Box 3: Datei der Variablennamen

Box 4: Freie Namensfelder

Siehe P0.3

Box 5: Datei aus der gelesen wird

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

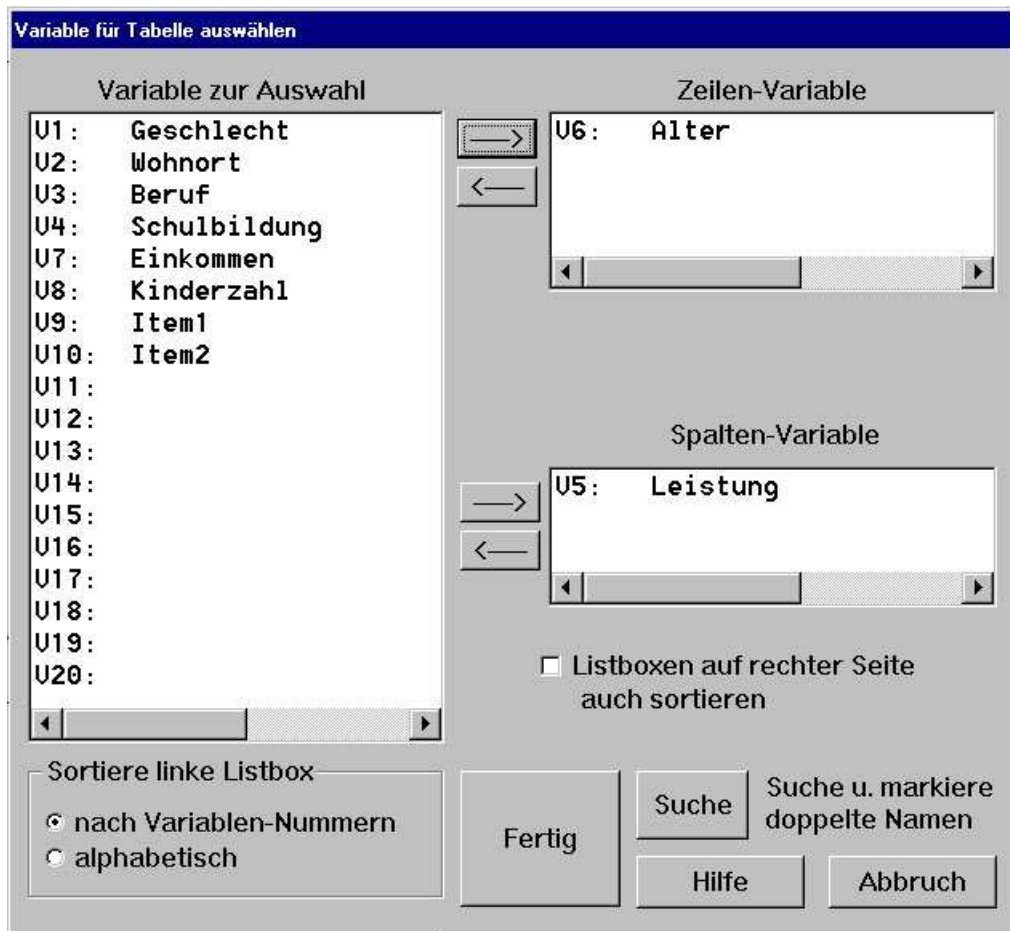
Siehe P0.4

P10.0.2 Box 7: 2-dimensionale Tabelle



Die Variablen, die in die Tabelle eingehen, dürfen ganzzahlige Werte oder Dezimalwerte besitzen. Das Maskenprogramm arbeitet im „Dezimalwert-Modus“. Siehe dazu P10.1.2 und P10.1.3.

Wenn Sie auf den Knopf mit den 2 Fenstersymbolen klicken, dann wird die Box "Variable für Tabelle auswählen" geöffnet. In Ihr geben Sie an, welche Variable als Zeilenvariable und welche als Spaltenvariable die Tabelle bilden sollen.



Almo bildet bei dieser Eingabe die Tabelle V6 Alter mit V5 Leistung

Wie man die Dialogbox "Variable für Tabelle 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 "Zeilenvariable" oder "Spaltenvariable" 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.

Knopf **FERTIG**

Wenn Sie abschliessend auf den Knopf FERTIG klicken, dann werden die Variablenamen, die sich in den rechten Listboxen "Zeilenvariable" und "Spaltenvariable" befinden, in das Eingabefeld des Maskenprogramms eingesetzt. Wenn die hintereinander gestellten Variablenamen zu lang würden, dann verwendet Almo automatisch Variablennummern.

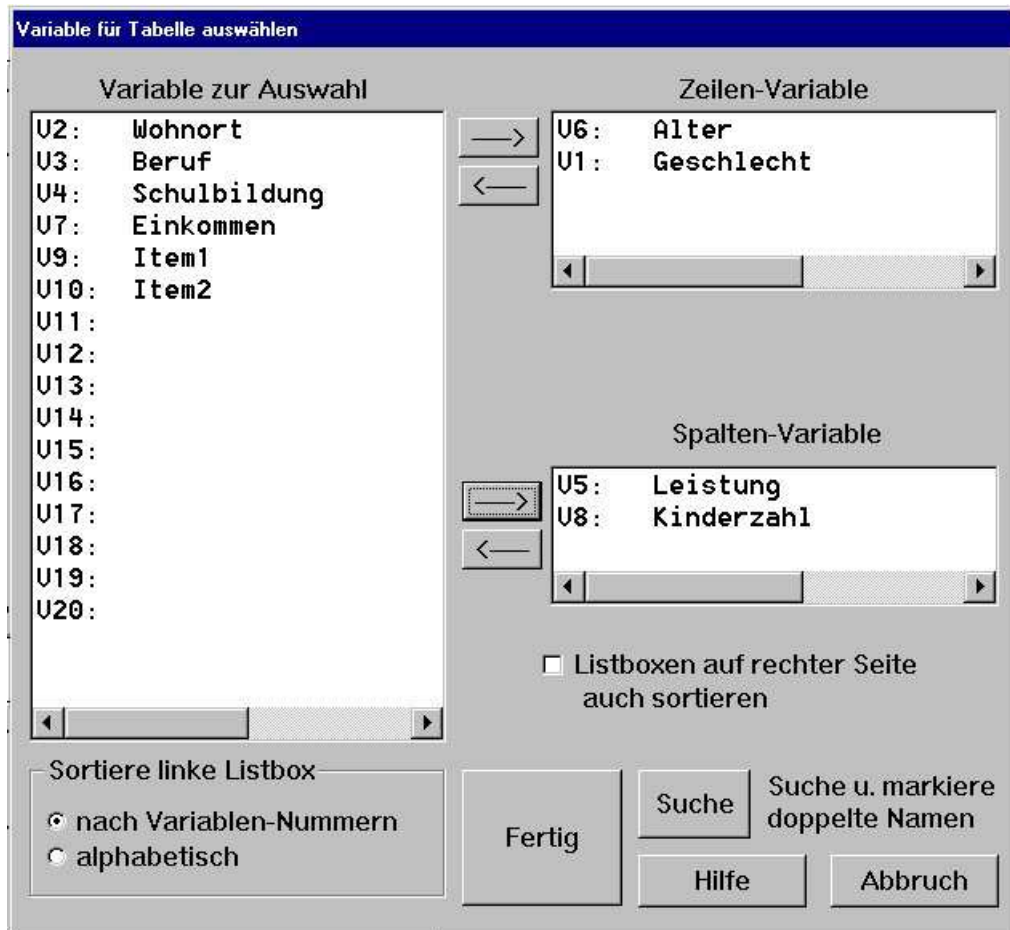
Knopf **SUCHE**

Variablenamen 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 Variablenamen werden dann durch 2 vorausgehende Unterstriche markiert, z.B. so:

V25: __Geschlecht

Diese Variablenamen dürfen dann nicht für die Analyse ausgewählt werden.

Betrachten wir ein weiteres Beispiel für eine Tabellenangabe



Nach Klick auf den Knopf FERTIG sehen Sie im Eingabefeld des Maskenprogramms folgende Eingabe:

Alter,Geschlecht mit Leistung,Kinderzahl

Alle Variable in der Box "Zeilenvariable" werden mit allen Variablen aus der Box "Spaltenvariable" gepaart. Für jedes Paar wird eine 2-dimensionale Tabelle gebildet. Also:

Alter mit Leistung
Alter mit Kinderzahl
Geschlecht mit Leistung
Geschlecht mit Kinderzahl

Direkt in das Eingabefeld der Programm-Maske hineinschreiben

Sie können auch direkt in das Eingabefeld des Maskenprogramms hineinschreiben. Wenn Sie z.B. schreiben:



dann ist V1 die Zeilen- und V2 die Spalten-Variable. V1 steht vorne in der Tabelle und V2 oben rüber in der Tabelle

Sie können in das Eingabefeld auch 2 oder mehrere Tabellenangaben schreiben, z.B. so

V1 mit V2 / V3 mit V4 / V5 mit V6

Zwischen die Tabellenangaben muss also ein Schrägstrich geschrieben werden. Zum Schluss kein Schrägstrich

Vor und hinter "mit" können mehrere (durch Beistrich getrennte) Variable stehen. Folgende Angaben für 2-dimensionale Tabellen sind beispielsweise möglich:

V1	mit	V3	tabelliert wird:	V1 mitV3
V1,2	mit	V3	tabelliert wird:	V1 mit V3, V2 mit V3
V1	mit	V3,4	tabelliert wird:	V1 mit V3, V1 mit V4
V1,2	mit	V3,4	tabelliert wird:	V1 mit V3, V1 mit V4 V2 mit V3, V2 mit V4

P10.0.3 Box 8: 3-dimensionale Tabelle mit Partialtabellen

A screenshot of a software dialog box titled "3-dimensionale Tabelle mit Partialtabellen". The dialog box has a title bar with "Hilfe" on the right. The main text reads: "Für die Ausprägungen der 3. Variable werden Partialtabellen gebildet". Below this, it says "Beispiel:". There are two tables side-by-side. The left table is titled "Männer" and the right table is titled "Frauen". Both tables have "Leistung" as a column header with sub-headers "niedrig" and "hoch". The rows are "Alter jung" and "Alter alt". The data values are: Männer (jung: 12, 5; alt: 6, 15); Frauen (jung: 10, 11; alt: 7, 9). At the bottom of the dialog box, there is an input field containing "Geschlecht mit Schulbildung mit Wohnort" and a button labeled "erzeuge zusätzliche Felder für Tabellen-Angaben".

Die Variablen, die in die Tabelle eingehen, dürfen ganzzahlige Werte oder Dezimalwerte besitzen. Das Maskenprogramm arbeitet im „Dezimalwert-Modus“. Siehe dazu P10.1.2 und P10.1.3.

Aus der Beispieltabelle, die in der Box abgebildet ist, wird ersichtlich, dass der Zusammenhang zwischen Alter und Leistung in 2 **Partialtabellen**, eine für die Männer und eine für die Frauen, dargestellt wird.

Wenn Sie auf den Knopf mit den 2 Fenstersymbolen klicken, dann wird die Dialogbox "Variablen für Tabelle auswählen" geöffnet. In Ihr geben Sie an, welche Variable als Zeilenvariable, welche als Spaltenvariable und welche als Kontrollvariable eine Tabelle bilden sollen.



Almo bildet bei dieser Eingabe die Tabelle

V6 Alter mit V5 Leistung mit Geschlecht

V6 Alter ist die Zeilen-, V5 Leistung die Spalten-Variable und V1 Geschlecht die Kontrollvariable, für deren Ausprägungen 2-dimensionale Tabellen von V6 mit V5 gebildet werden.

Beachte: Die Kontrollvariable ist jene Variable für deren Ausprägungen Partialtabellen gebildet werden, d.h. 2-dimensionale Tabellen zwischen der Zeilen- und der Spaltenvariablen.

Zur Bedienung der Dialogbox "Variablen für Tabelle auswählen" siehe die obige Darstellung zu Box 6:

In jede der 3 rechten Listboxen können mehrere Variable eingegeben werden. Es werden dann alle möglichen 3-er Kombinationen gebildet

Also: Alle Variable in der Box "Zeilenvariable" werden mit allen Variablen aus der Box "Spaltenvariable" und mit allen Variablen aus der Box "Kontrollvariable" zu einer 3-dimensionale Tabelle kombiniert

Beachte: Es können sehr viele Tabellen mit einem sehr langen unübersichtlichen Output entstehen

Sie können auch direkt in das Eingabefeld des Maskenprogramms hineinschreiben. Wenn Sie z.B. schreiben:



dann werden 2 3-dimensionale Tabellen gebildet. Zwischen die Tabellenangaben muss ein Schrägstrich geschrieben werden.

Bei der ersten ist V1 die Zeilen-, V2 die Spalten-Variable und V3 die Kontrollvariable

BEACHTE:

Sie müssen dann auf den Knopf "Variable aus Tabellenangaben" im Maskenprogramm klicken. Also registriert dann die Variablen, die zur Tabellenbildung verwendet werden.

Vor und hinter "mit" können mehrere (durch Beistrich getrennte) Variable stehen. Folgende Angabe für 3-dimensionale Tabellen sind beispielsweise möglich:

```
V1,2 mit V3,4 mit V5,6  
tabelliert wird dann: V1 mit V3 mit V5  
                     V1 mit V3 mit V6  
                     V1 mit V4 mit V5  
                     V1 mit V4 mit V6  
                     .  
                     .  
                     .  
                     V2 mit V4 mit V6
```

Das Schlüsselwort "mit"

Wir wollen die Angaben vor und nach dem Schlüsselwort MIT nochmals zusammengefasst betrachten.

V17 mit 25;

Variable 17 soll mit 25 tabelliert werden.

1. Regel: Die Variable vor MIT wird als unabhängige Variable interpretiert und in der Tabelle vorne eingetragen. Die Variable nach MIT wird als abhängige Variable interpretiert und in der Tabelle oben rüber eingetragen.

V17 mit 25/33 mit 44;

2. Regel: Als Trennzeichen gegenüber der nächsten Tabellenangabe muss ein Schrägstrich verwendet werden.

3. Regel: Hinter die letzte Tabellenangabe muss als Schlusszeichen ein Semikolon geschrieben werden.

Beachte: Das Symbol "V" bzw. das Wort "Variable" braucht nur das 1. Mal geschrieben zu werden. Bei den folgenden Variablennummern muss es nicht mehr (kann aber) geschrieben werden. Es soll eine dreidimensionale Tabelle gebildet werden.

V17 mit 25 mit 36;

4. Regel: Die Variablen vor dem 1. MIT (also V17) ist die unabhängige. Die Variable nach dem 1. MIT (also V25) ist die abhängige. Die Variable nach dem 2. MIT (also V36) ist die Kontrollvariable.

V36,43 mit 17;

Dies ist eine Kurzschreibweise für
36 MIT 17/
43 MIT 17/

V14 mit 28,36;

Dies ist eine Kurzschreibweise für
14 MIT 28/
14 MIT 36/

V14:16,25 mit 20:23;

Dies ist eine Kurzschreibweise für
14 MIT 20/ 15 MIT 20/ 16 MIT 20/ 25 MIT 20

...

14 MIT 23/ 15 MIT 23/ 16 MIT 23/ 25 MIT 23

Am besten versteht man diese Kurzschreibweise,
wenn man sich eine "Tabelle der Tabellen"
konstruiert.

		V20	21	22	23	abh.Variable
unabh. Variable	V14					
	15					
	16					
	25				X	← diese Zelle entspricht der Tabelle V25 MIT V23

V16, 17 mit 20, 21 mit 36, 43;

Dies ist eine Kurzschreibweise für dreidimensionale Tabellen. Auch hier macht man sich wieder am besten eine „Tabelle der Tabellen“.

		V36	43		Kontrollvariable	
		V20	21	20	21	abh.Variable
unabh. Variable	V16					
	17				X	← diese Zelle entspricht der Tabelle V17 MIT V21 MIT V43

Empfehlung: Man sollte von der Kurzschreibweise mit Vorsicht und Zurückhaltung Gebrauch machen. Besser ist es, alle Tabellen, die man wünscht, einzeln und für sich alleine zu schreiben.

P10.0.5 Box 10: Kein_Wert-Angabe und Umkodierung

Die Art und Weise, wie man Variable umkodiert, bzw. eine Kein-Wert-Deklaration vornimmt, haben wir in P0.5 ausführlich beschrieben.

Wie wollen hier auf folgendes Problem hinweisen:

In unseren Daten ist das Alter in 9 Stufen kodiert, ebenso die Leistung. Wird mit diesen beiden Variablen eine 2-dimensionale Tabelle gebildet, dann entsteht eine 9*9-Tabelle mit 81 Zellen. Auf diese 81 Zellen verteilen sich unsere 61 Personen. Das ist nicht sinnvoll. Außerdem ist eine so große Tabelle kaum zu überschauen. Normalerweise wird das Alter von 0 bis 100 kodiert sein. Dann entsteht eine 100*9-Tabelle mit 900 Zellen.

Variable mit vielen Ausprägungen müssen also umkodiert werden, wenn sie als Tabellenvariable verwendet werden sollen. Die Frage ist: Auf wieviele Ausprägungen sollen die Variable zusammengefasst werden. Wir haben, wie in der Umkodierungsbox ersichtlich ist, eine radikale Lösung für dieses Problem gewählt. Beide Variable wurden dichotomisiert, so dass eine 2*2-Tabelle entsteht. Solche Tabellen sind inhaltlich am besten zu interpretieren. Auch der Korrelationskoeffizient, in diesem Fall die Phi-Korrelation, besitzt wertvolle Eigenschaften.

Als allgemeine Regel würden wir formulieren: Wenn es sinnvoll ist, sollte man in einer 1. Analyse auf 2 Ausprägungen zusammenfassen. In einer 2. und weiteren Analyse kann man differenzierter auf 3 und mehr Ausprägungen zusammenfassen.

P10.2 Ausgabe der zweidimensionalen Tabelle

Aus dem Maskenprogramm Prog10m1.Msk und dem „selbst geschriebenen“ Programm Prog10b.Msk.

Tabelle 1

Variable 6 Alter
mit
Variable 5 Leistung

			Leistung		Summe
			schlecht 1	gut 2	
Alter	jung	1	5	24	29
	alt	2	10	22	32
Summe			15	46	61

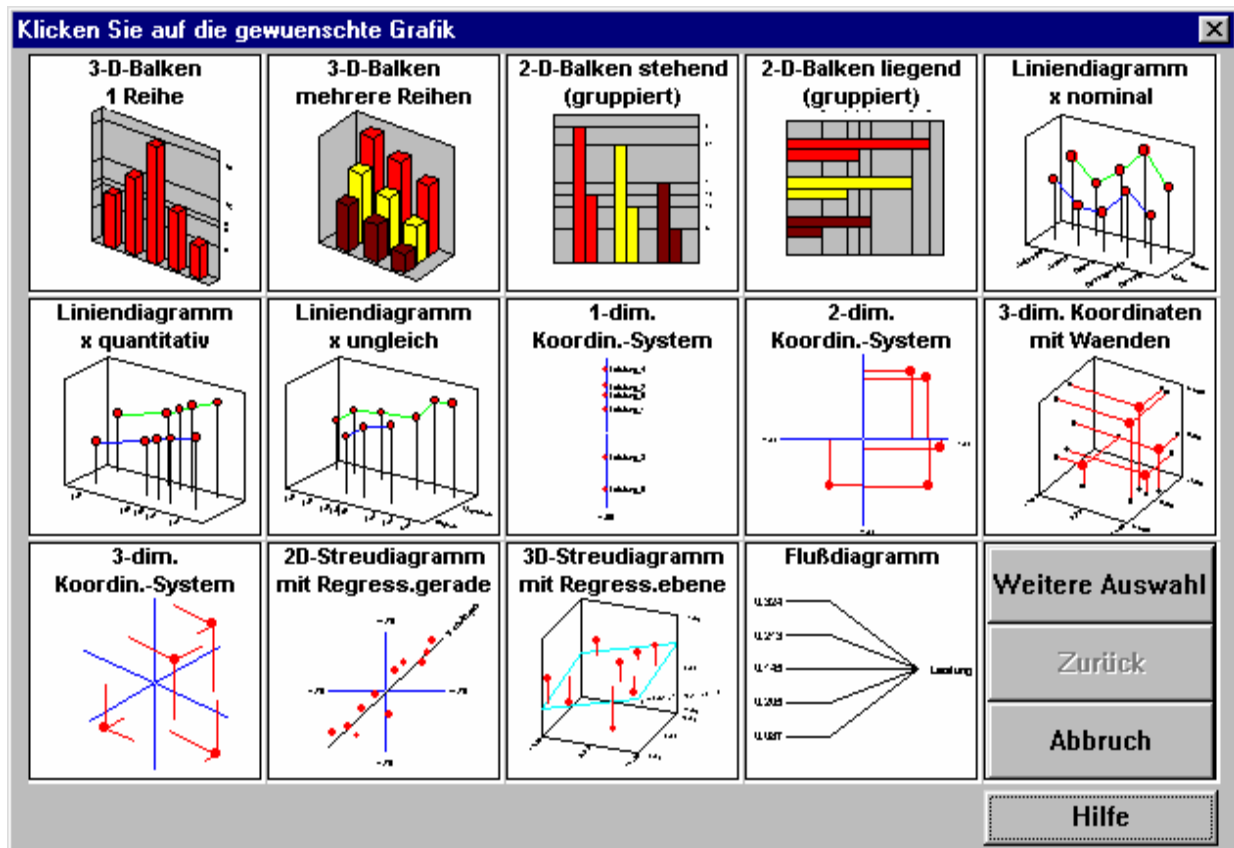
Almo liefert für diese Tabelle folgende Grafik:

2-dimensionale Verteilung
Alter
und
Leistung



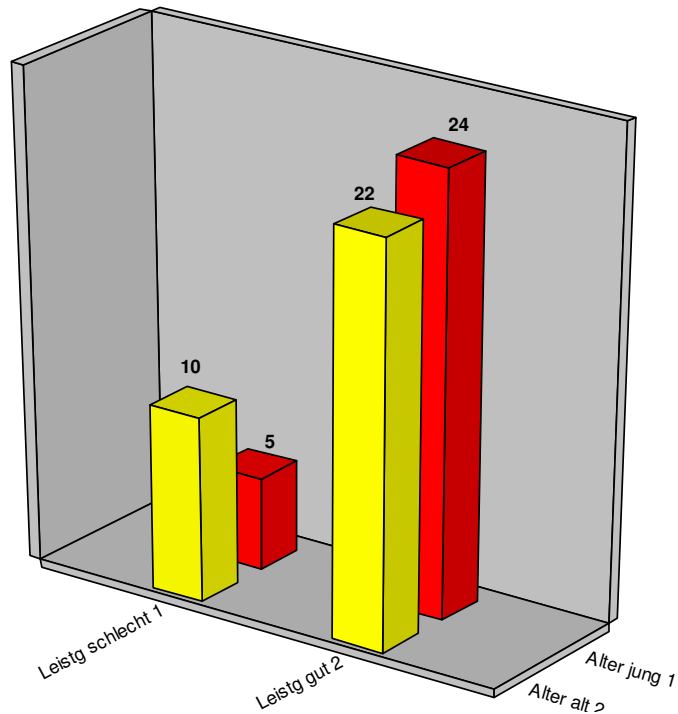
Im Grafik-Editor kann diese Grafik in vielfältiger Weise verändert werden. Siehe dazu Teil 1, Bedienungsanleitung, Abschnitt 10.2 und die Datei "AnleitungGrafik" im Wurzelverzeichnis von Almo.

Wir wollen hier nur zeigen, wie obige Grafik in ein 3D-Balkendiagramm gewandelt werden kann. Nach Klick auf "Anderer Grafiktyp" auf der linken Seite des Grafikenfensters zeigt Almo die folgende Auswahl:



Wir klicken auf "3-D -Balken, mehrere Reihen" (das 2. Bild in der 1. Reihe) Almo erzeugt dann ein Balkendiagramm, das wir noch durch einige Mausklicks auf "Perspektive", "Balkendicke" und durch Verschieben der Masszahlen verschönern.

2-dimensionale Verteilung
Alter
und
Leistung



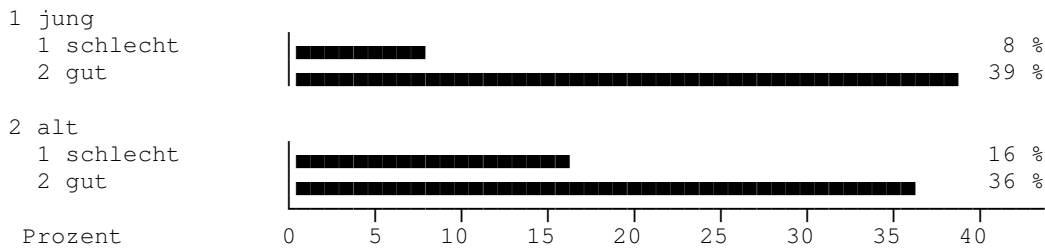
Sinnvoll wäre es auch ein Liniendiagramm zu erzeugen – vor allem dann, wenn die beiden Variablen mehrere Ausprägungen besitzen. Wir klicken zu diesem Zweck auf „Liniendiagramm, x nominal“. Das ist das letzte Bild in der ersten Reihe.

Prozentuiert nach Gesamthäufigkeit

			Leistung		Summe
			schlecht 1	gut 2	
Alter	jung	1	8.20	39.34	47.54
	alt	2	16.39	36.07	52.46
Summe			24.59	75.41	100.00

Die Werte in den Zellen sind Prozentwerte, die auf die Gesamtzahl der Untersuchungspersonen (in unserem Beispiel: 61) bezogen sind.

Almo liefert auch für diese Tabelle eine Almo-Grafik (die wir um Platz zu sparen jedoch nicht zeigen) und zusätzlich folgende einfache Säulengraphik im Textmodus.



Zeilenweise prozentuiert

			Leistung		Summe
			schlecht 1	gut 2	
Alter	jung	1	17.24	82.76	100.00
	alt	2	31.25	68.75	100.00
Summe			24.59	75.41	100.00

Almo liefert auch für diese Tabelle eine Almo-Grafik und zusätzlich eine einfache Säulengrafik im Textmodus, die wir um Platz zu sparen jedoch nicht zeigen.

Die Prozentwerte in den Zellen der Tabelle sind auf die Zeilensumme bezogen. Das zeilenweise Prozentuieren ist besonders sinnvoll. Die Gruppen der Jungen und der Alten sind jetzt gleich groß (= 100%), so dass wir sie miteinander vergleichen können. Wir sehen zum Beispiel, dass von den 100% Jungen 82,76% eine gute Leistung erbringen, von den 100% Alten jedoch nur 68,75%.

Spaltenweise prozentuiert

			Leistung		Summe
			schlecht 1	gut 2	
Alter	jung	1	33.33	52.17	47.54
	alt	2	66.67	47.83	52.46
Summe			100.00	100.00	100.00

Almo liefert auch für diese Tabelle eine Almo-Grafik und zusätzlich eine einfache Säulengrafik im Textmodus, die wir um Platz zu sparen jedoch nicht zeigen.

Koeffizienten fuer Tabelle 1

Signifikanztest

Chi-Quadrat = 1.6100 df = 1 Signifikanz (1-p)*100 = 79.562

Ein Wert ueber ca. 95 bedeutet:
Zwischen den Variablen besteht
ein signifikanter Zusammenhang

Korrelationskoeffizienten

Messniveau				zweiseitige Signifikanz	
der einen Variablen	der anderen Variablen	Korrelationskoeffizient		p	(1-p)*100
dichotom	dichotom	Phi	= 0.1625	0.20452	79.5475
dichotom	polytom	Phi'	= 0.1625	0.20452	79.5475
polytom	polytom	Kontingenzk.C(cor)	= 0.2268	0.20452	79.5475
polytom	polytom	Tschuprow's T	= 0.1625	0.20452	79.5475
polytom	polytom	Cramer's V	= 0.1625	0.20452	79.5475
polytom	polytom	Lambda (asymm.)	= 0.0000	-	-
ordinal	ordinal	Gamma	= -0.3714	0.16608	83.3921
ordinal	ordinal	tau-b	= -0.1625	0.21312	78.6880
dichotom	ordinal	biseriales tau-b	= -0.1625	0.21312	78.6880
ordinal	ordinal	Rho	= -0.1625	0.21096	78.9039
dichotom	quantit.	punktbiseriales r	= -0.1625	0.21096	78.9039
quantit.	quantit.	Produkt-Moment r	= -0.1625	0.21096	78.9039

Zeilenvariable	Spaltenvariable				
quantit.	ordinal	Gross-Gamma I	= -0.1625	0.20839	79.1608
ordinal	quantit.	Gross-Gamma II	= -0.1625	0.20839	79.1608
polytom	quantit.	Eta	= 0.1625	0.21096	78.9039
quantit.	quantit.	nichtlineares Eta	= 0.1625	0.21096	78.9039
		Signifik. der Nichtlinearitaet		1.00000	0.0000
dichotom	dichotom	tetrachorisches r	= -0.2979	0.17475	82.5247

Um den Zusammenhang zwischen 2 Variablen in einer 2-dimensionalen Tabelle beurteilen zu können, benötigt man 2 Kennwerte:

1. einen Signifikanzkoeffizienten
2. und einen Korrelationskoeffizienten

Der Signifikanzkoeffizient drückt aus, ob der Zusammenhang zwischen den beiden Variablen signifikant (oder anders formuliert: überzufällig) ist. Der Korrelationskoeffizient gibt an, wie stark der Zusammenhang mit einer Zahl zwischen 0 (= kein Zusammenhang) und 1 (= absoluter Zusammenhang) ausgedrückt, bei einigen Korrelationskoeffizienten zwischen -1 nach 0 bis +1.

In unserem Beispiel finden wir eine Signifikanz von 79.562%. D.h. die Wahrscheinlichkeit, dass die beiden Variablen zusammenhängen ist 79.562%. Anders herum betrachtet: Die Wahrscheinlichkeit, dass wir uns irren, wenn wir einen Zusammenhang von Alter und Leistung postulieren, ist $100 - 79.562 = 20.438\%$. Das ist doch sehr viel.

Üblicherweise wird für die Signifikanz ein niedrigster noch zu akzeptierender Wert von 95% angenommen. Unser Wert von 78.562% liegt deutlich darunter. Also müssen wir folgern: Zwischen Alter und Leistung besteht kein Zusammenhang. Im Prinzip brauchen wir jetzt gar nicht mehr nachschauen, wie groß der Korrelationskoeffizient ist. Wir wollen dies aber trotzdem tun.

Da Almo das Messniveau der beiden tabellierten Variablen nicht kennt, gibt es für bestimmte Kombinationen von Messniveaus Korrelationskoeffizienten aus. Der Benutzer muss selbst entscheiden, welcher Korrelationskoeffizient für ihn der richtige ist. In unserem Beispiel sind die beiden Variable dichotom. Dann ist der zutreffende Koeffizient der Phi-Koeffizient. Wir finden einen Phi-Koeffizienten von 0.1625.

Beachte: Mit Ausnahme des Lambda-Koeffizienten, sind alle Korrelationskoeffizienten symmetrisch. Die Zeilen- und die Spalten-Variablen können also vertauscht werden. Wir werden später, in Abschnitt P10.4.2 ausführlich auf die Korrelationskoeffizienten eingehen.

P10.3 Ausgabe der dreidimensionalen Tabelle

Das Charakteristikum von P10 ist, dass es 3-dimensionale Tabellen in Partialtabellen auflöst und für jede Partialtabelle die verschiedenen Koeffizienten berechnet. Damit wird es z.B. möglich die Lazarsfeldsche Methode der 3-Variablen-Analyse durchzuführen (siehe dazu Schmierer, 1975). Im Unterschied dazu wird in P11 für die 3- und mehrdimensionale Tabelle ein pauschaler Chi-Quadrat-Wert und eine mehrdimensionale Konfigurationsfrequenzanalyse gerechnet. Siehe P11.3.

Tabelle
Partialtabelle 1

Kontrollvariable: V1 - mit dem Wert 1 Geschlecht: männlich

Variable 6 Alter
mit
Variable 5 Leistung

			Leistung		Summe
			schlecht 1	gut 2	
Alter	jung	1	4	14	18
	alt	2	8	8	16
Summe			12	22	34

Dies ist also die Partialtabelle, die den Zusammenhang zwischen Alter und Leistung für die Gruppe der Männer darstellt.

Almo liefert auch für diese und alle nachfolgenden Tabellen eine Almo-Grafik, die wir um Platz zu sparen jedoch nicht zeigen

Prozentuiert nach Gesamthäufigkeit

			Leistung		Summe
			schlecht 1	gut 2	
Alter	jung	1	11.76	41.18	52.94
	alt	2	23.53	23.53	47.06
Summe			35.29	64.71	100.00

Zeilenweise prozentuiert

			Leistung		Summe
			schlecht 1	gut 2	
Alter	jung	1	22.22	77.78	100.00
	alt	2	50.00	50.00	100.00
Summe			35.29	64.71	100.00

Spaltenweise prozentuiert

			Leistung		Summe
			schlecht 1	gut 2	
Alter	jung	1	33.33	63.64	52.94
	alt	2	66.67	36.36	47.06
Summe			100.00	100.00	100.00

Koeffizienten fuer Tabelle
Partialtabelle 1

Signifikanztest

Chi-Quadrat = 2.8620

df = 1

Signifikanz (1-p)*100 = 90.939

Ein Wert ueber ca. 95 bedeutet:
Zwischen den Variablen besteht
ein signifikanter Zusammenhang

Korrelationskoeffizienten

Messniveau				zweiseitige Signifikanz		
der einen	der anderen	Korrelationskoeffizient		p	(1-p)*100	
Variablen	Variablen					
dichotom	dichotom	Phi	=	0.2901	0.09073	90.9272
dichotom	polytom	Phi'	=	0.2901	0.09073	90.9272
polytom	polytom	Kontingenzk.C(cor)	=	0.3941	0.09073	90.9272
polytom	polytom	Tschuprow's T	=	0.2901	0.09073	90.9272
polytom	polytom	Cramer's V	=	0.2901	0.09073	90.9272
polytom	polytom	Lambda (asymm.)	=	0.0000	1.00000	0.0000
ordinal	ordinal	Gamma	=	-0.5556	0.03352	96.6476
ordinal	ordinal	tau-b	=	-0.2901	0.10505	89.4954
dichotom	ordinal	biserials tau-b	=	-0.2901	0.10505	89.4954
ordinal	ordinal	Rho	=	-0.2901	0.09602	90.3982
dichotom	quantit.	punktbiserials r	=	-0.2901	0.09602	90.3982
quantit.	quantit.	Produkt-Moment r	=	-0.2901	0.09602	90.3982

Zeilenvariable	Spaltenvariable					
quantit.	ordinal	Gross-Gamma I	=	-0.2901	0.09574	90.4263
ordinal	quantit.	Gross-Gamma II	=	-0.2901	0.09574	90.4263
polytom	quantit.	Eta	=	0.2901	0.09602	90.3982
quantit.	quantit.	nichtlineares Eta	=	0.2901	0.09602	90.3982
		Signifik. der Nichtlinearitaet			1.00000	0.0000
dichotom	dichotom	tetrachorisches r	=	-0.4587	0.09749	90.2505

Tabelle
Partialtabelle 2

Kontrollvariable: V1 - mit dem Wert 2 Geschlecht: weiblich

Variable 6 Alter
mit
Variable 5 Leistung

		Leistung		Summe
		schlecht 1	gut 2	
Alter	jung	1	10	11
	alt	2	14	16
Summe		3	24	27

Almo liefert auch für Partialtabelle 2 eine Tabelle

1. prozentuiert nach Gesamthäufigkeit
2. zeilenweise prozentuiert
3. spaltenweise prozentuiert

sowie die verschiedenen Koeffizienten.

Almo liefert auch wieder für diese Tabellen eine Almo-Grafik. Um Platz zu sparen werden wir jedoch diese Tabellen und Grafiken nicht zeigen

P10.4 Die statistischen Masszahlen bei Standardausgabe

P10.4.1 Chi-Quadrat-Test

Durch den Chi-Quadrat-Test wird die Signifikanz des Zusammenhangs zweier Variablen ermittelt. Genauer formuliert: Durch den Chi-Quadrat-Test wird die Nullhypothese der Unabhängigkeit der beiden Variablen getestet. Also gibt aus: Testwert, Freiheitsgrade, die Sicherheitswahrscheinlichkeit für die Ablehnung der Nullhypothese, der Anteil an Erwartungswerten (e_{ij}), die kleiner 1 und die kleiner 5 sind (der Test sollte nur verwendet werden, wenn kein Erwartungswert kleiner 1 und nicht mehr als 10 % kleiner 5 sind).

Beim Maskenprogramm mit Optionen (Prog10m2) und beim „selbst geschriebenen“ Programm mit Optionen (Prog10c), die beide anschließend dargestellt werden, besteht die Möglichkeit, die Erwartungswerte und die Chi-Quadrat-Beiträge je Zelle zu ermitteln.

Der Benutzer hat dann noch die Möglichkeit auf vermutete Typen zu prüfen, wie dies im Rahmen der von Krauth/Lienert entwickelten Konfigurationsfrequenzanalyse geschieht.

Die Berechnung von Chi-Quadrat wollen wir an einem Beispiel vorführen: Der Einfachheit halber betrachten wir folgende 2*2-Tabelle, die ALMO ermittelt hat

		Leistung		Summe
		niedrig	hoch	
Alter	jung	5	24	29
	alt	10	22	32
Summe		15	46	61

Zuerst werden die Zellenhäufigkeiten berechnet, die sich ergeben würden, wenn der Zusammenhang zwischen Alter und Leistung zufällig wäre. Die so ermittelten Zellenhäufigkeiten werden "Erwartungswerte" genannt. Zu beachten ist, dass bei der Berechnung der Erwartungswerte die Randhäufigkeiten der beiden Variablen erhalten bleiben müssen.

Die Formel ist sehr einfach folgende:

$$E(i,k) = R(i) * R(k) / N$$

- i = Index für Zeile
- k = Index für Spalte
- $E(i,k)$ = Erwartungswert für die Zelle ik
- $R(i)$ = Randhäufigkeit in Zeile i
- $R(k)$ = Randhäufigkeit in Spalte k
- N = Gesamthäufigkeit

Beispiel für die Zelle Alter:jung / Leistung:hoch
i ist also 1 und k ist 2

$$E(1,2) = 29 * 46 / 61 = 21.87$$

So entsteht folgende Tabelle der Erwartungswerte

		Leistung		Summe
		niedrig	hoch	
Alter	jung	7.13	21.87	29.00
	alt	7.87	24.13	32.00
Summe		15.00	46.00	61.00

Die Zelle (1,2) hat also einen Erwartungswert von 21.87. Die tatsächliche Zellenhäufigkeit ist jedoch 24. Die Differenz der beiden Werte ist klein. D.h. die empirische Häufigkeit weicht von der zufälligen nur wenig ab.

Den Beitrag zum Chi-Quadrat-Wert, den diese Zelle (1,2) leistet erhalten wir gemäß folgender Formel:

$$B(i,k) = (F(i,k) - E(i,k)) * (F(i,k) - E(i,k)) / E(i,k)$$

i = Index für Zeile
k = Index für Spalte
B(i,k) = Beitrag der Zelle ik zum Chi-Quadrat
F(i,k) = Tatsächliche Häufigkeit in Zelle ik
E(i,k) = Erwartungswert für die Zelle ik

Für die Zelle (1,2) erhalten wir

$$B(1,2) = (24 - 21.87) * (24 - 21.87) / 21.87 = 0.208$$

So entsteht folgende Tabelle der Beiträge zum Chi-Quadrat

		Leistung		Summe
		niedrig	hoch	
Alter	jung	0.637	0.208	0.845
	alt	0.577	0.188	0.765
Summe		1.214	0.396	1.610

Der Gesamt-Chi-Quadrat-Wert ist dann die Summe der Beiträge, in unserem Falle also

$$\text{Chi-Quadrat} = 1.610$$

Nun wäre es noch interessant zu überprüfen, ob beispielsweise in Zelle (1,2) die Differenz zwischen tatsächlicher Häufigkeit (=24) und Erwartungswert (=21.87) überzufällig ist. Anders formuliert, ob der Chi-Quadrat-Beitrag von 0.208 in dieser Zelle signifikant ist.

Diese Fragestellung wird im Rahmen der "Konfigurationsfrequenzanalyse" (kurz: KFA) beantwortet. Wir kommen später darauf zurück.

Wir müssen folgende Anmerkungen zum Chi-Quadrat-Test machen.

- (a) Der Chi-Quadrat-Test unterstellt, dass die beiden Tabellenvariablen nominal sind. Ist eine von ihnen oder sind beide ordinal oder sogar quantitativ, dann wird die Signifikanz falsch eingeschätzt, in der Regel unterschätzt.
- (b) Die Signifikanz hängt sehr stark von der Zahl der Untersuchungseinheiten ab.

Betrachten wir ein Beispiel:

		Leistung	
		niedrig	hoch
Alter	jung	5	24
	alt	10	22

Chi-Quadrat: 1.61
Signifikanz: 79.562

		Leistung	
		niedrig	hoch
Alter	jung	50	240
	alt	100	460

Chi-Quadrat: 16.1
Signifikanz: 99.999...

In der 2. Tabelle haben wir die Zellenhäufigkeit um den Faktor 10 erhöht. Der Chi-Quadrat-Wert ist dann mit 16.1 auch 10 mal größer. Da die Zahl der Freiheitsgrade mit $df = 1$ gleich bleibt, haben wir dann eine Signifikanz von 99.999...%.

P10.4.2 Korrelationskoeffizienten für nominale Variable

Korrelationskoeffizienten sind symmetrisch (mit der Ausnahme von Lambda und Eta). D.h. es spielt keine Rolle welche Variable in der Tabelle die Zeilenvariable und welche die Spaltenvariable ist.

Lambda ist ein asymmetrischer Koeffizient für Nominaldaten. Dabei wird die Zeilenvariable als unabhängige und die Spaltenvariable als abhängige Variable betrachtet. Lambda ist als PRE-Koeffizient definiert (siehe P20.6.3) als Reduktion des (Prognose-) Fehlers für die abhängige Variable durch Einführung einer unabhängigen Variablen. Für diesen Koeffizienten wird auch die Sicherheitswahrscheinlichkeit berechnet.

Es wird zwischen den folgenden 3 Konstellationen unterschieden

- Beide variable sind dichotom
- Beide Variable sind polytom (d.h. haben mehr als 2 Ausprägungen)
- Die eine Variable ist dichotom, die andere ist polytom

Beide Variable sind dichotom

Almo berechnet den Phi-Koeffizienten und die tetrachorische Korrelation

$$\text{Phi} = \sqrt{\text{Chi} / N}$$

Chi = Chi-Quadrat-Wert

N = Gesamtzahl der Untersuchungseinheiten

Phi ist bei genügend großem N identisch mit der Produkt-Moment-Korrelation r . Siehe dazu Bortz, Lienert, Boehnke, 1990, S. 330 f.

Auf die *tetrachorische Korrelation* werden wir in Abschnitt P17.7.9 eingehen.

Beide Variable sind polytom

Almo berechnet:

- (1) Tschuprows T nach der Formel

$$T = \sqrt{\text{Chi} / (N \cdot (k-1)(m-1))}$$

- (2) Cramers V

$$V = \sqrt{\text{Chi} / (N \cdot t)}$$

dabei ist $t = \min(k-1, m-1)$, d.h. der kleinere Wert von $k-1$ bzw. $m-1$.

(3) den korrigierten Kontingenzkoeffizienten

$$C_{\text{cor}} = \sqrt{\text{Chi}/(N + \text{Chi})/C_{\text{max}}}$$

dabei ist $C_{\text{max}} = \sqrt{(t-1)/t}$ und $t = \min(k, m)$

Zeichenerklärung:

Chi = Chi-Quadrat-Wert

N = Gesamtzahl aller Untersuchungseinheiten

k, m = Zahl der Ausprägungen der einen und der anderen Variable

$\min(k, m)$ = die kleinere Zahl der beiden k und m

Diese 3 Koeffizienten beruhen auf dem Chi-Quadrat-Wert. Für $k = m = 2$ sind T und V identisch mit dem Phi-Koeffizienten.

C_{cor} ist in aller Regel größer als T und V.

T ist nicht normiert. Es kann nur den Maximalwert 1.0 erreichen, wenn mindestens eine der beiden Variablen dichotom ist. In diesem Fall ist es identisch mit V. T ist der am wenigsten brauchbare Koeffizient.

Der *empfehlenswerte* Korrelationskoeffizient ist Cramers V. Er kann abgeleitet werden aus dem Allgemeinen Linearen Modell. Siehe dazu P20.9.5.1 und Bortz, Lienert, Boehnke, 1990, S. 357.

Die eine Variable ist dichotom, die andere polytom

Almo berechnet Phi' nach derselben Formel wie Phi. D.h. die Phi-Formel kann ausgedehnt werden, auf den Fall einer $2 \times k$ -Tabelle. Phi' entspricht dann dem multiplen Korrelationskoeffizienten R aus einer Regression der $k-1$ unabhängige Dummies der polytomen Variablen auf die dichotome abhängige Variable (siehe Bortz, Lienert, Boehnke, 1990, S.342)

P10.4.3 Korrelationskoeffizienten für ordinale Variable

Für Ordinaldaten werden berechnet

- (1) der Spearman'sche Rangkorrelationskoeffizient rho,
- (2) Gamma (Schmieder, 1975, S. 99)
- (3) tau-b (Denz, 1977, S. 109 ff): tau-b kann auch berechnet werden, wenn eine Variable ordinal und die andere dichotom ist (biserial tau-b)
- (4) Optional: Polychorische Korrelation. Siehe dazu Abschnitt 10.7.9.

Wir wollen die Berechnung der ordinalen Korrelationskoeffizienten Gamma und tau-b an einem kleinen Beispiel erläutern:

Person	x	y
A	1	1
B	3	2
C	4	1
D	3	2

x und y seien 2 ordinale Variable. Die Zahlenwerte drücken also nur die Relationen größer, kleiner, gleich aus.

Die 4 Personen werden nun zu Paaren kombiniert.

Im Paar A – B hat die Person B sowohl in x als auch in y einen höheren Wert als die Person A.

Im Paar B – C hat Person C einen höheren Wert in x aber einen niedrigeren Wert in y als Person B. Dieses Paar weist also eine „Inversion“ der Rangwerte auf – während das Paar A – B eine „Proversion“ der Rangwerte aufweist.
 Beim Paar A – C haben beide in y denselben Wert. Sie besitzen eine „Bindung in y“.

Die Formel für Kendalls tau, für Gamma und für Kendalls tau-b lauten

$$\text{tau} = \frac{N_s - N_d}{N(N-1)/2}$$

$$\text{Gamma} = \frac{N_s - N_d}{N_s + N_d}$$

$$\text{tau-b} = \frac{N_s - N_d}{\sqrt{(N_s + N_d + T_y)(N_s + N_d + T_x)}}$$

N_s = Zahl der Paare, die eine „Proversion“ der Rangwerte besitzen

N_d = Zahl der Paare mit einer „Inversion“ der Rangwerte

N = Zahl der Untersuchungseinheiten

$N(N-1)/2$ = Gesamtzahl der Paare

T_y = Zahl der Paare mit einer Bindung in y

T_x = Zahl der Paare mit Bindung in x

Der genaue Rechengang für die Größen dieser Formeln ist beschrieben bei Bortz, Lienert, Boehnke (1990, S. 429ff).

Der „beste“ ordinale Korrelationskoeffizient ist tau-b, da er die Bindungen in x und y berücksichtigt und da er als „PRE-Koeffizient“ interpretiert werden kann (siehe dazu P20.6.3).

P10.4.4 Korrelationskoeffizienten für quantitative Variable

- (1) Produkt-Moment-Korrelationskoeffizient, wenn beide Variablen quantitativ sind,
- (2) Eta, wenn die abhängige quantitativ und die unabhängige Variable nominal ist, oder wenn beide Variable quantitativ, aber nicht-linear verbunden sind.
- (3) Der punktbiseriale Koeffizient $r_{p.bis}$, wenn die eine Variable quantitativ und die andere nominal-dichotomisch ist. Sind beide Variable quantitativ, dann kann aus der Differenz von r und Eta die Linearitätsannahme überprüft werden (Schmierer 1975, S. 103).

P10.4.5 Korrelationskoeffizienten für gemischt ordinale und quantitative Variable

Ist die eine Variable ordinal und die andere quantitativ, dann errechnet Almo einen Korrelationskoeffizienten nach dem Groß-Gamma-Kalkül. Dabei ist es nicht belanglos, ob die ordinale Variable die Zeilen- oder die Spalten-Variable ist. Wir sprechen von "Gross-Gamma I", wenn sie die Spaltenvariable ist und von "Gross-Gamma II" wenn sie die Zeilenvariable ist. Die beiden Koeffizienten können verschieden groß sein. Der Groß-Gamma-Kalkül wird ausführlich in Abschnitt P19.0.3 und vor allem im Anhang dargestellt.

P10.5 Programm-Maske mit Optionen

Almo bietet eine Fülle von Optionen an, die der Benutzer beim Maskenprogramm oder beim selbst geschriebenen Programm verwenden kann.

Prog10m2.Msk Kurzprogramm
 2- und 3-dimensionale Tabellierung
 für Variable (auch) mit Dezimalwerten
 mit vielen Optionen

Das Programm liefert Tabellen folgender Art:

		Beruf		
		Arbeiter	Angestell	Selbständ
Geschl.	männlich	48	18	8
	weiblich	38	11	4

Bei 3-dimensionaler Tabellierung werden für die 3. Variable Partialtabellen der ersten beiden Variablen gebildet

Für die Tabellen werden folgende Koeffizienten berechnet: Chi-Quadrat, Kontingenzkoeffizient, Tschuprows T, Cramers U, Lambda, Gamma, Kendalls tau-b, r, Spearmans Rho, punkt-biseriales r, Phi, Eta.

Folgende Koeffizienten könne zusätzlich gerechnet werden: Ridits, Kolmogorov-Smirnov 2-Stichprobentest, exakter Fisher-Test, exakter Uleman Rangaufteilung U-Test, exakter Freeman-Halton-Test, Haldane-Dawson-Test, Konfigurationsfrequenz-analyse mit exaktem Binomialtest, kappa-Koeffizient der Urteilskonkordanz, polychorische Korrelation
 Test für verbundene Stichproben: t-Test, Wilcoxon Vorzeichenrangtest, McNemar-Test, Zeichentest.

Grafik: Balkendiagramme

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

1

Speicher fuer x Variable Hilfe

Vereinbare Variable= 20 ;

2

↓ Option: Weitere Vereinbarungen - nur wenn Almo dazu auffordert

3

Datei der Variablennamen Hilfe

↔ 📁 "C:\Almo7\Testdat\Varname2.nam"

↔ ↓ zeige zeige = Namensdatei in Output zeigen
 leer = nicht

4

Freie Namensfelder Hilfe

↔ Name 5=Leistung:niedrig, hoch;

↔ Name 6=Alter:jung, alt;

⋮ erzeuge zusätzliche Namensfelder

5 **Datei der Variablennamen** Hilfe

zeige = Namensdatei in Output zeigen
leer = nicht

6 **Freie Namensfelder** Hilfe

7 **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

8 **Wenn Dateiformat FIX oder Nicht-Standard-FREI** Hilfe

9 **2-dimensionale Tabelle** Hilfe

Beispiel:

		Leistung	
		niedrig	hoch
Alter	jung	22	16
	alt	13	24

10 **3-dimensionale Tabelle mit Partialtabellen** Hilfe

Für die Ausprägungen der 3. Variable werden Partialtabellen gebildet

Beispiel:

		Männer		Frauen	
		Leistung		Leistung	
		niedrig	hoch	niedrig	hoch
Alter	jung	12	5	10	11
	alt	6	15	7	9

11 **Variable aus Tabellenangabe**
 U1,5,6 <----- keine Benutzereingabe
 Also ermittelt die Variablen, die in obige Tabellen eingegangen sind selbst. Sie können aber auch auf diesen Knopf klicken

12 **Option: Ein- und Ausschliessen von Untersuchungseinheiten**

13 **Loesche wieder diese Box**
Umkodierungen und Kein-Wert-Angaben
 Umkodierungen
 Kein_Wert-Angabe
 Leistung <1:3=1; 3:10=2>
 Alter <1:4=1; 4:10=2>

 erzeuge zusätzliche Felder für Umkodierungen / Kein_Wert-Angaben

Kontrollieren, ob Umkodierung so erfolgt wie gewünscht
 diese Variablen ...
 Leistung, Alter
 1:20 ... aus diesen Datensätzen vor und nach der Umkodierung zur Kontrolle anzeigen

14 **Option: Untersuchungseinheiten gewichten**

15 **Option: Verschiedene Tests und Koeffizienten**

16 **Option: Verzichte auf Teile der Ausgabe**

17 **Option: "Aussehen" der auszugebenden Tabelle bzw. Matrix**

18 **Grafik-Optionen**

19

Erläuterungen zu den Boxen:

Box 1 bis Box 6: Siehe die Erläuterungen zu vorausgehendem Maskenprogramm Prog10m1 in P10.0.1

Box 7, 8: die zu bildenden Tabellen

Siehe vorausgehendes Maskenprogramm in P10.0.2 und P10.0.3

Box 9: Variable für Tabellen

Siehe vorausgehendes Maskenprogramm Prog10m1.Msk in P10.0.4

Box 10: Ein- und Ausschliessen von Untersuchungseinheiten

Siehe Almo-Dokument Nr. 0 "Arbeiten mit Almo", P0.7.

Box 11: Kein_Wert-Angabe und Umkodierung

Siehe Almo-Dokument Nr. 0 "Arbeiten mit Almo", P0.5 und P10.0.5

Box 12: Kontrollieren, ob Umkodierung so erfolgt, wie gewünscht

Siehe P0.6.

Box 13: Gewichtung

Siehe P0.8.

Box 14: Option: Verschiedene Tests und Koeffizienten

↓ Loesche wieder diese Box		Optionen	Hilfe
a	↔ ↓	Erwartungswert	Erwartungswerte aus Chi-Quadrat ausgeben
b	↔ ↓	Chi_Quadrat_Beitrag	Chi-Quadrat-Beitrag je Zelle ausgeben
c	↑↓	1	1= Konfigurationsfrequenzanalyse rechnen nur bei 2-dimensionalen Tabellen 0= nicht
d	↔ ↓	abh_Stichproben	Tests für abhängige Stichproben rechnen: t-Test, Wilcoxon Vorzeichenrangtest, Zeichentest, McNemar-Test, Bowker-Test
e	↔ ↓	Ridit	Ridits. Nur sinnvoll für ordinale Variablen
f	↔ ↓	Kolmogorov_Smirnov	Kolmogorov-Smirnov-Test für 2 unabh. Stichproben
g	↔ ↓	exakt	exakter Fisher-Test für 2*2-Tabellen bzw. exakter Freeman-Halton für größere Tab.
h	↔	33	Rechne exakten Fisher- bzw. exakten Freeman-Halton-Test bis $n \leq xx$. Wird diese Option nicht geschrieben, dann gilt die Voreinstellung. Entnehmen Sie diese aus dem Handbuch
i	↔ ↓	Haldane_Dawson	Haldane-Dawson-Test für sehr große, aber schwach besetzte Tabellen
j	↔ ↓	■	Ulemans exakter Rangaufteilungs U-Test ACHTUNG: Ist sehr rechenintensiv !!
k	↔ ↓	Konkordanz	kappa-Koeffizient der Urteils Konkordanz nur für k*k-Tabellen
l	↑↓	0	1 = polychorische Korrelation 2 = polychorische Korrelation und Schwellenwerte 0 = nicht berechnen

In dieser Box werden Ihnen weitere Tests und Koeffizienten angeboten. Diese werden in den folgenden Abschnitten P10.7.1 ff dargestellt. Löschen Sie einen Eintrag dadurch, dass Sie auf den Raus-Rein-Knopf klicken. Setzen Sie einen Eintrag ein, indem Sie auf den Knopf mit dem Pfeil nach unten klicken.

Box 15: Option: Verzichte auf Teile der Ausgabe



Optionsbox geöffnet:



Löschen Sie einen Eintrag dadurch, dass Sie auf den Raus-Knopf klicken. Setzen Sie einen Eintrag ein, indem Sie auf den Knopf mit dem Pfeil nach unten klicken.

Dabei sind folgende Einträge möglich:

CHI	(=kein Chi-Quadrat)
KORRELATION	(=keine Korrelationskoeffizienten)
KOEFFIZIENTEN	(=überhaupt keine Koeffizienten)
ABSOLUT_TABELLE	(=keine Tabellen der absoluten Häufigkeiten)
PROZENT_TABELLE	(=keine prozentuierten Tabellen)
GESAMTPROZENT	(=keine Gesamtprozenttabelle)
ZEILENPROZENT	(=keine zeilenweis prozentuierte Tabelle)
SPALTENPROZENT	(=keine spaltenweis prozentuierte Tabelle)

Box 16: Optionen, die das „Aussehen“ der Tabellen steuern

Sie können das Erscheinungsbild der Tabellen in der Almo-Ausgabe beeinflussen. Siehe P0.9.

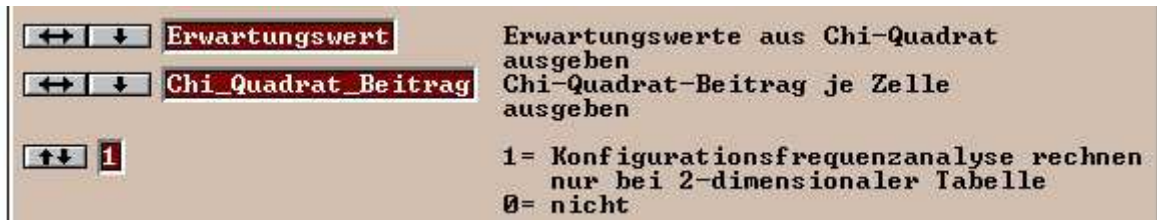
Box 17: Grafik-Optionen

Siehe P0.10.

P10.7 Erläuterungen zu den einzelnen Optionen

P10.7.1 Erwartungswerte, Chi-Quadrat-Beiträge, Konfigurationsfrequenzanalyse KFA

Beim Maskenprogramm, Box 14, Eingabefeld a, b, c:



Beim „selbst geschriebenen“ Almo-Programm:

Mit "**Zeige=Erwartungswert, Chi-Quadrat-Beitrag;**" wird die Tabelle der Erwartungswerte und der Chi-Quadrat-Beiträge je Zelle der Tabelle ausgegeben. Es kann auch nur eine der beiden Angaben gemacht werden.

Mit "**Konfig_Freq_Analyse=1 ;**" wird eine Konfigurations-Frequenz-Analyse gerechnet.

Die KFA untersucht, ob einzelne Chi-Quadrat-Beiträge signifikant sind, bzw. auf einen bestimmten Typus hinweisen. Siehe die Darstellung bei Lienert (1978, S.534) und Krauth/Lienert (1973). Wir rechnen ein Beispiel aus Lienert (1978, S.535) mit folgender Tabelle:

	Einstellung zu x	
	positiv	negativ
Studenten der Naturw.	135	108
Technik	31	40
Geistw.	57	95
Jura	39	65

Die Frage lautet: Ist ein spezifischer Typ feststellbar? Almo liefert folgendes Ergebnis

Chi-Quadrat = 16.5363 df = 3 Signifikanz (1-p)*100 = 99.877

Beitraege zum Chi-Quadrat

	Einstellung		Summe
	positiv	negativ	
Studienr Naturw	4.863	4.136	8.999
Technik	0.081	0.069	0.151
Geistw	2.370	2.016	4.385
Jura	1.621	1.379	3.000
Summe	8.935	7.601	16.536

durchschnittlicher Beitrag = 2.067

Schranke fuer	bei Signifikanz
Chi-Quadrat-Beitrag	(1-p) * 100
4.324543	85.0
5.027256	90.0
6.240185	95.0
7.502315	97.5
9.266205	99.0
13.447144	99.9

Der höchste Chi-Quadrat-Beitrag ist mit 4.863 in Zelle Naturw/positiv. Er ist allerdings nur mit nicht ganz 90 % signifikant. Ein Chi-Quadrat-Beitrag müsste mindestens 6.240185 sein, damit er mit 95% Signifikanz auf einen „Typ“ hinweist.

Die KFA kann verwendet werden, um

1. zu überprüfen, ob ein Chi-Quadrat-Beitrag signifikant ist
2. und ob ein Typ identifizierbar ist.

Ein Chi-Quadrat-Beitrag ist signifikant, wenn er den Schrankenwert (je gewählte Signifikanz) überschreitet.

Ein Typ ist existent, wenn

- a. der Chi-Quadrat-Beitrag signifikant ist
- b. und wenn die tatsächliche Häufigkeit grösser ist als der Erwartungswert. Ist sie (bei signifikantem Chi-Quadrat-Beitrag) kleiner, dann kann von einem "Antityp" gesprochen werden. Dieser ist inhaltlich nicht immer interpretierbar und sollte dann negiert werden. Siehe dazu: Krauth: Einführung in die Konfigurationsfrequenzanalyse (KFA), Beltz-Verlag, Weinheim, 1993, S. 23 ff.

Beim exakteren Binomialtest ergeben sich folgende p-Werte

Matrix der p-Werte aus KFA-Binomialtest

	Einstellung		Summe
	positiv	negativ	
Studienr Naturw	0.0091	0.0104	-
Technik	0.4290	0.4156	-
Geistw	0.0541	0.0721	-
Jura	0.1022	0.1228	-
Summe	-	-	-

Schranke fuer p-WERT in Tabelle	bei Signifikanz (1-p)*100
0.018750	85.0
0.012500	90.0
0.006250	95.0
0.003125	97.5
0.001250	99.0
0.000125	99.9

(Nur ein p-Wert kleiner .00625 ist mit 95% signifikant. In obiger Tabelle trifft das auf keinen zu).

Die Zelle Naturw/positiv. ist mit etwas über 90 % signifikant.

Schlussfolgerung:

D.h. Studenten der Naturwissenschaft, die eine positive Einstellung zu x haben bilden einen schwachen spezifischen Typ.

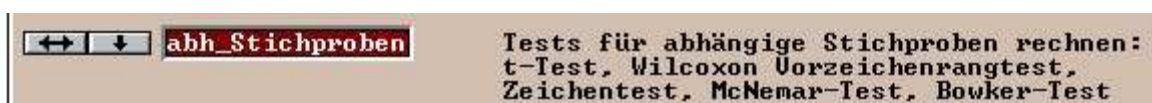
Anmerkung: Der p-Wert für Naturw/positiv von 0.0091 wird bei Lienert mit 0.007 angegeben. Der Binomialtest wird in ALMO exakt (über die F- Transformation) gerechnet. Bei großen Binomialkoeffizienten können jedoch Rechner- bzw. Compiler-bedingte Ungenauigkeiten auftreten. In diesem Falle glauben wir jedoch, dass Lienert (er möge uns verzeihen) mit dem Taschenrechner noch ungenauer gerechnet hat.

Zur Erläuterung der Ergebnisse aus der KFA siehe auch P11.3.

P10.7.2 Tests für abhängige Stichproben

- t-Test
- Wilcoxon Vorzeichenrangtest
- Zeichentest
- McNemar-Test bzw. Bowker-Test

Beim Maskenprogramm, Box 14, Eingabefeld d:



Beim „selbst geschriebenen“ Almo-Programm:

Mit "**Test = abh_Stichproben;**" werden oben angegebene Tests gerechnet.

Hinter dem Wort Test = ... können mehrere durch Komma getrennte Anweisungen

stehen. Siehe dazu die nachfolgenden bzw. die vorausgegangenen Optionen. Wird die Test-Anweisung wiederholt,

z.B.: Test=abh_Stichproben;
 Test=Ridit;

dann wird von ALMO immer nur die letzte ausgewertet. Die vorausgehenden Test-Anweisungen gehen verloren.

Die Logik der Tests für abhängige Stichproben ist nun etwas anders: Es gibt nicht mehr eine unabhängige und eine abhängige Variable, sondern es sollen Wertepaare (z.B. Test- und Retest-Werte derselben Variablen), die je an einer Untersuchungseinheit erhoben wurden, miteinander verglichen werden (Claus/Ebner 1970, S.217 ff).

Die *Nullhypothese* ist in allen Fällen, dass zwischen den Werten eines Paares kein Unterschied besteht. Für jeden dieser drei Tests wird die Sicherheitswahrscheinlichkeit angegeben. Diese Tests für verbundene Stichproben werden von ALMO nur dann gerechnet, wenn beide Variablen dieselbe Anzahl von Ausprägungen haben.

Der **t-Test** erfordert eine quantitative Skala, da die Differenzen der beiden Messungen ermittelt werden; er testet dann, ob die Mittelwerte der beiden Messungen signifikant voneinander verschieden sind. Der t-Test für abhängige Stichproben ist auch in Prog 18 (siehe P18.3) enthalten.

Der **Wilcoxon-Vorzeichenrangtest** erfordert eine "ordered metric" Skala, d.h. die Differenzen müssen Ordinaleigenschaft besitzen. Auch in Prog 8 ist der Wilcoxon-Test enthalten. Der Unterschied ist folgender: Der Test in Prog 10 ist vor allem für gruppierte Daten mit nicht zu vielen Ausprägungen der ordinalen Variablen geeignet. Bindungen und Nulldifferenzen werden nicht berücksichtigt. Der Wilcoxon-Test in Prog 8 ist genauer, aber auch rechenintensiver. Er berücksichtigt Bindungen und Nulldifferenzen und rechnet auch den sogenannten exakten Test.

Der **Zeichentest** benötigt ordinale Skalen. Dieser Test ist auch in Programm 18 und in Programm 8 als besondere Version des Friedman-Tests enthalten. Siehe P8.11. Im Unterschied zu Prog 10 werden in Prog 8 Bindungen berücksichtigt.

McNemar-Test und Bowker-Test

Betrachten wir ein Beispiel:

		Retest	
		+	-
Test	+	A	B
	-	C	D

Die Untersuchungsobjekte, die von Test nach Retest sich geändert haben, befinden sich in den Zellen C und B. Beachte: wird die Tabelle so gebildet, dass die Wechsler, sich in A und D befinden, dann entsteht ein falsches Ergebnis.

Der McNemar-Test wird nur angewendet, wenn beide Variable dichotom sind. ALMO führt bei der Berechnung des McNemar-Tests die Kontinuitätskorrektur nach Yates aus (indem 0.5 subtrahiert wird).

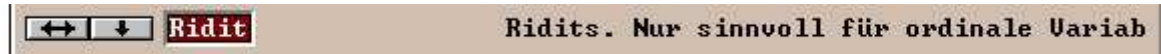
Die ausgegebene Signifikanz ist zweiseitig. Ist die Besetzung der beiden Felder B + C zusammen kleiner 10, dann wird anstelle des Chi-Quadrat der Binomialtest gerechnet. Zum McNemar-Test siehe Siegel, 1987, S. 60, Botz, Lienert, Boehnke 1990, S.160 u. S. 165.

Der **Bowker-Test** ist eine Verallgemeinerung des McNemar-Test auf k*k-Tabellen.

Almo führt ihn automatisch anstelle des McNemar-Tests durch, wenn es erkennt, dass die beiden Variablen nicht dichotom sind.

P10.7.3 Ridits

Beim Maskenprogramm, Box 14, Eingabefeld e:



Beim „selbst geschriebenen“ Almo-Programm:

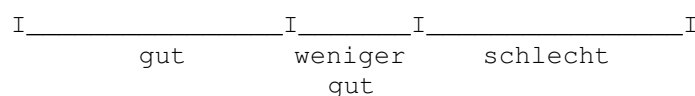
Mit "**Test=Ridit**," werden Ridits gerechnet. Siehe Regeln zur Testanweisung in P10.7.2.

Da die RIDITS relativ wenig bekannt sind, sollen die Anwendung und Berechnung dieses Verfahrens etwas ausführlicher beschrieben werden (siehe dazu Fleiss 1973, S.102 und Agresti 1984, S.167)

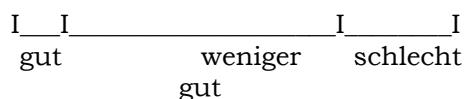
Es gibt oft Variablen, denen zwar eine quantitative Messdimension zugrunde liegt, die jedoch nur **ordinal** erhoben werden können. Bei einer Umfrage wird beispielsweise gefragt:

"Finden Sie den Politiker A: gut, weniger gut oder schlecht?" "Glauben Sie, dass unsere Umwelt sehr bedroht, etwas bedroht oder nicht bedroht ist?" "Stimmen Sie der Aussage: 'Eine Ohrfeige hat noch niemandem geschadet' ganz, teilweise oder gar nicht zu?"

Man erkennt intuitiv, dass die Messdimensionen, welche diese Fragen messen sollen, Quantitäten wären: Das Ausmaß der Zustimmung zu einem Politiker, das (subjektive) Ausmaß der Bedrohtheit der Umwelt, die Einstellung zu körperlichen Strafen. Aber für alle diese Messdimensionen konnten keine Antwortalternativen gefunden werden, welche diese Quantitäten abbilden. Die Antwortalternativen sind Kategorien, welche Abschnitte der quantitativen Messdimension repräsentieren, aber es kann nicht festgestellt werden, wie groß diese Abschnitte sind. Beim Beispiel des Politikers (bei einem Politiker, der eher polarisierend wirkt) könnte man sich vorstellen, dass die Kategorien die Messdimension in folgender Weise repräsentieren:



Hingegen könnten diese Kategorien bei einem anderen Politiker (der eher ausgleichend wirken möchte) die Messdimensionen ganz anders repräsentieren:



Über die Größe und Anordnung der Kategorien entlang der quantitativen Messdimension können zwar Hypothesen aufgestellt werden, eine Quantifizierung ist aber nicht möglich.

Solche Variablen können nun wie nominale oder wie quantitative behandelt werden: Im einen Fall würde man auf Information (nämlich die Ordinalität der Kategorien) verzichten, im anderen Fall einen Fehler machen, weil man Information verwendet, die gar nicht enthalten ist (nämlich die Abstände). So bietet sich die

RIDIT-Analyse als ordinales Verfahren an.

RIDIT ist die Abkürzung von: "relative to an identified distribution". Es werden also mehrere Verteilungen mit einer bestimmten Verteilung verglichen (der Referenzgruppe). Diese Gruppe erhält den RIDIT-Wert 0.5. Die anderen Gruppen werden mit dieser verglichen und die Abweichungen festgestellt: ein Wert kleiner 0.5 bedeutet, dass mehr Personen in niedrigen Kategorien sind als in der Referenzgruppe, ein Wert von über 0.5, dass diese Personen eher in den oberen Kategorien sind.

Ein Beispiel: "Finden Sie den Politiker A: gut, weniger gut oder schlecht?"

	gut	weniger gut	schlecht	Summe
Arbeiter	72	30	18	120
Angestellte	34	40	26	100
Beamte	12	22	26	60

Ridits zeilenweise – Bezugsgruppe = 2 (Angestellte)

1	2	3
0.368	0.5	0.609

Die Berechnung selbst ist sehr einfach. Allerdings muss noch eine zusätzliche Annahme getroffen werden, nämlich: Die Fälle innerhalb einer Kategorie sind (bezogen auf die quantitative Messdimension) gleichverteilt. Man stellt nun für jede Kategorie fest, wie groß die Wahrscheinlichkeit ist, dass eine Person einen Wert aufweist, der kleiner ist als die Mitte dieser Kategorie.

Ein einfaches Rechenschema ist:

(1)	(2)	(3)	(4)	(5)	(6)
gut	34	0	17	17	0.17
weniger gut	40	34	20	54	0.54
schlecht	26	74	13	87	0.87

- (1) die einzelnen ordinalen Kategorien
- (2) die absoluten Häufigkeiten der Referenzgruppe (Angestellte)
- (3) die kumulierten Häufigkeiten
- (4) die Hälfte der absoluten Häufigkeiten
- (5) Summe aus (3) und (4)
- (6) (5) dividiert durch die Anzahl der Fälle (n) = RIDITS

Nun müssen die RIDITS für die Vergleichsgruppen berechnet werden. Sie sind die Summe der RIDITS der Referenzgruppe multipliziert mit den absoluten Häufigkeiten der Vergleichsgruppe, diese Summe wird durch die Anzahl der Fälle der Vergleichsgruppe dividiert:

(1)	(2)	(3)	(4)	(5)	(6)
gut	0.17	72	12.24	12	2.04
weniger gut	0.54	30	16.20	22	11.88
schlecht	0.87	18	15.66	26	22.62
Summe		120	44.10	60	36.54
Ridits			0.3675		0.6090

- (1) die ordinalen Kategorien
- (2) die RIDITS der Referenzgruppe
- (3) die absoluten Häufigkeiten der Gruppe 1 (Arbeiter)

- (4) (2) mal (3)
- (5) die absoluten Häufigkeiten der Gruppe 3 (Beamte)
- (6) (2) mal (5)

RIDIT für die Gruppe 1 ist Summe (4) dividiert durch Summe (3)
 RIDIT für die Gruppe 3 ist Summe (6) dividiert durch Summe (5)

Durch diese Berechnungen erhält man Masszahlen zum Vergleich von Gruppen hinsichtlich einer ordinalen Variablen. Dazu ist auch die Information wichtig, ob der Unterschied zwischen zwei Gruppen nur zufällig oder systematisch (signifikant) ist. Die Differenzen zwischen zwei RIDITS bilden eine Normalverteilung mit dem Mittelwert Null und der Varianz:

$$\text{Varianz der RIDIT-Differenzen} = \frac{n_1 + n_2}{12 * n_1 * n_2}$$

Für die obige Tabelle:

Test Gruppe 1 gegen 2: Varianz = (100+120)/12*100*120) = 0.00153
 Standardabweichung=Wurzel(0.00153) = 0.039
 z = (0.3675-0.5)/0.039=3.390
 Sicherheitswahrscheinlichkeit = 99.9 %

Dieser Test kann auch für die beiden anderen möglichen Vergleiche durchgeführt werden:

Test Gruppe 1 gegen 3: Sicherheitswahrscheinlichkeit = 99.99 %
 Test Gruppe 2 gegen 3: Sicherheitswahrscheinlichkeit = 98 %

Dieses Verfahren wurde in die Tabellenanalyse integriert. Das Programm geht davon aus, dass die unabhängige Variable die (nominale) Gruppenzugehörigkeit ist, die abhängige (ordinale) Variable die Variable ist, hinsichtlich derer die Gruppen verglichen werden sollen: die RIDITS werden also zeilenweise berechnet.

Die Referenzgruppe sollte entweder nach theoretischen Gesichtspunkten gewählt werden oder man wählt eine Gruppe, die eher in der Mitte liegt (wie im obigen Beispiel). Das Programm geht davon aus, dass die erste Zeile der Tabelle die Referenzgruppe darstellt. Sind die Gruppen anders codiert, müsste man für diese Berechnung die Gruppenvariable entsprechend umcodieren.

P10.7.4 Exakter Fisher-Test - Exakter Freeman-Halton-Test

Beim Maskenprogramm, Box 14, Eingabefeld g:



Beim „selbst geschriebenen“ Almo-Programm:

Mit "**Test=Exakt;**" wird der exakte Fisher Test - bei 2*2-Tabellen bzw. der exakte Freeman-Halton-Test - bei größeren Tabellen, gerechnet (siehe die Regeln zur "Test"-Anweisung in P10.5.1).

Dem Chi-Quadrat-Test sollte nicht mehr vertraut werden, wenn mehr als 10 % der Erwartungswerte kleiner als 5 sind. Manche Autoren lassen auch noch 20 % zu. Tritt dieser Fall nun ein, dann gibt es 3 Tests, die anstelle des Chi-Quadrat-Tests

gerechnet werden können.

1. Der **exakte Fisher-Test** ist anwendbar auf 2*2 Tabellen. Wir geben folgende Tabelle ein:

	dafür	dagegen
Männer	8	3
Frauen	6	14

ALMO liefert folgende Ausgabe:

```
Chi-Quadrat          = 5.2314
                    df = 1
Signifikanz (1-p)*100 = 97.783
```

Exakter Fisher-Test

Teil-Wahrscheinlichkeiten

```
0.0000
0.0002
0.0032
0.0241
-----
p = 0.0275
```

```
Signifikanz (1-p)*100 = 97.246%
```

Zur Bedeutung der Teilwahrscheinlichkeiten siehe Siegel (1987, S.96) oder Büning/Trenkler (1978, S.249).

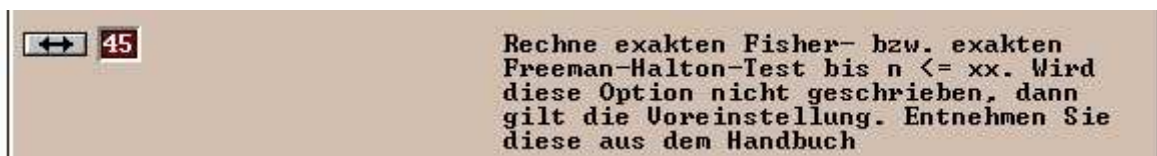
Unser obiges Beispiel zeigt, dass Chi-Quadrat- und Fisher-Test praktisch übereinstimmen.

ALMO rechnet den Fisher-Test nicht mehr, wenn $n > 33$.

Diese Sperre kann beim „selbst geschriebenen“ Almo-Programm über **Option 7** aufgehoben werden. Beispiel: **Option 7=45**; der exakte Fisher-Test wird bis $n \leq 45$ gerechnet. Die Rechenzeit steigt etwas an, die Rechengenauigkeit nimmt ab.

Beim Maskenprogramm, Box 14 wird n in folgendes Eingabefeld eingeschrieben:

Eingabefeld j:



2. Der **exakte Freeman-Halton-Test** wird von ALMO automatisch anstelle des exakten Fisher-Tests gerechnet, wenn die Tabelle größer als 2*2 ist. Wir geben folgende Tabelle ein (Beispiel aus Lienert, 1978. S.406)

	Leistung		
	gut	mittel	schlecht
Unterrichtsmethode 1	0	3	2
Unterrichtsmethode 2	6	5	1

ALMO liefert folgendes Ergebnis (wir geben zuerst die Ergebnisse aus dem normalen Chi-Quadrat-Test und dann die aus dem Freeman- Halton-Test an)

```
Chi-Quadrat = 1.6593      df = 2      Signifikanz (1-p)*100 = 56.028  
-----  
Exakter Freeman-Halton-Test  
-----  
p=0.4805      Signifikanz (1-p)*100 = 51.948%
```

Der Freeman-Halton-Test ist sehr rechenaufwendig. Die Rechenzeit wächst mit der Zahl n der Untersuchungseinheiten und der Zahl der Tabellenzellen.

Tabellen, die größer 4×4 sind und einem $n > 40$ sollten nicht mehr gerechnet werden. ALMO verwendet als Sperre nur die Zahl n der Untersuchungseinheiten. Ab $n > 33$ wird der Test nicht mehr gerechnet. Mit **Option 7** beim „selbst geschriebenen“ Almo-Programm kann diese Sperre hinaufgesetzt werden.

Beispiel: Option 7=35; der Test wird bis $n \leq 35$ gerechnet.

Beim Maskenprogramm ist in Box 14, wie schon oben gezeigt, das gewünschte n einzutragen.

P10.7.5 Haldane-Dawson-Test

Beim Maskenprogramm, Box 14, Eingabefeld h:



Beim „selbst geschriebenen“ Almo-Programm:

Mit "**Test=Haldane_Dawson;**" wird der Haldane-Dawson-Test gerechnet. Siehe die Regel zur "Test"-Anweisung in P10.7.2.

Der Test kann angewendet werden, wenn die Zahl der Zellen ca. größer als 25 ist. Die Erwartungswerte dürfen kleiner 5 bzw. 1 sein, siehe dazu Lienert (1978, S.399). Wir rechnen ein Beispiel mit einer 16×4 - Tabelle mit vielen unbesetzten Zellen und vielen Zellen mit nur wenigen Besetzungen.

ALMO liefert folgendes Ergebnis (wir geben zuerst die Ergebnisse aus dem normalen Chi-Quadrat-Test und dann die aus dem Haldane-Dawson-Test an)

		Alter			
		unter 20 1	20-30 2	30-40 3	über 40 4
Leistungspunkte	1.1	0	3	0	0
	1.2	1	0	0	1
	2.1	1	1	2	1
	2.2	2	0	2	1
	3.1	2	1	0	0
	3.2	4	4	1	1
	4.1	0	2	2	0
	4.2	5	3	0	2
	5.1	0	2	0	2
	5.2	0	2	2	0
	6.2	0	2	0	0
	7.1	2	1	0	0
	7.2	1	0	0	0
	8.1	1	0	0	1
8.2	0	0	0	1	
9.1	0	0	0	1	
Summe		19	21	9	11

Chi-Quadrat = 65.8876 df =60 Signifikanz (1-p)*100 = 71.904

Erwartungswerte kleiner 1 = 74%
kleiner 5 =100% der Zellen

Haldane-Dawson-Test (einseitig)

Chi-Quadrat= 65.8876
Erwartungswert von Chi-Q.= 61.0000
Varianz von Chi-Quadrat =204.6125
z-Wert= 0.3417 p/2 = 0.3666 Signifikanz (1-p/2)*100 = 63.340%

P10.7.6 Ulemans exakter Rangaufteilungs-U-Test

Beim Maskenprogramm, Box 14, Eingabefeld i:



Beim „selbst geschriebenen“ Almo-Programm:

Mit; "**Test=Uleman;**" wird der Uleman-Test gerechnet. Siehe die Regel zur "Test"-Anweisung in P10.7.2.

Der U-Test ist anwendbar, wenn wir eine dichotome, nominale Variable (z.B. männlich-weiblich, oder Versuchs-Kontrollgruppe) und eine mindestens ordinale Variable haben. Siehe die ausführliche Darstellung in P8.

Beim U-Test mit gruppierten Daten, wie wir ihn im Rahmen von Prog 10 rechnen, sind sehr viele Rangteilungen vorzunehmen. Der Test nach Uleman ist nun ein exakter U-Test auch für den Fall vieler Rangteilungen.

Wir rechnen ein Beispiel aus Lienert (1973, S.220) mit folgender Tabelle:

	Leistung			
	sehr gut	gut	mittel	schlecht
Unterrichtsmethode A	0	1	2	1
Unterrichtsmethode B	2	3	1	0

ALMO liefert folgendes Ergebnis:

Chi-Quadrat = 4.0972 df = 3 Signifikanz (1-p)*100 = 75.002

Erwartungswerte kleiner 1 = 38%
kleiner 5 =100% der Zellen

Ulemans exakter Rangaufteilungs U-Test (einseitig)

U1= 20.5 U2= 3.5 p=0.0619 Signifikanz (1-p)*100 = 93.810%

Wir erkennen, dass der Uleman-Test einen (halbwegs) signifikanten Unterschied zwischen den beiden Unterrichtsmethoden nachweist, während der Chi-Quadrat-Test diese Signifikanz nicht identifizieren kann.

Der Uleman-Test ist außerordentlich rechenintensiv. Die Rechenzeit hängt ab von der Zahl der Ränge der ordinalen Variablen und der Zahl der Untersuchungseinheiten. Wir empfehlen nicht über 4 Ränge und ca. 30 Untersuchungseinheiten hinauszugehen. Eine Sperre ist in ALMO nicht eingebaut.

P10.7.7 Kolmogorov-Smirnov-Test (KS-Test) für 2 unabhängige Stichproben

Beim Maskenprogramm, Box 14, Eingabefeld f:



Beim „selbst geschriebenen“ Almo-Programm:

Mit "**Test=Kolmogorov_Smirnov;**" wird der KS-Test angewendet.

Der Kolmogorov_Smirnov-Test ist zulässig, wenn die unabhängige Variable nominal-dichotomisch und die abhängige Variable zumindest ordinal ist.

Wir rechnen ein Beispiel aus Siegel (1987, S.129) mit folgender Eingabe-Tabelle:

		Variable 2						
		1	2	3	4	5	6	7
Variable 1	1	11	7	8	3	5	5	5
	2	1	3	6	12	12	14	6

Chi-Quadrat = 22.0648 df = 6 Signifikanz (1-p)*100 = 99.848

Erwartungswerte kleiner 1 = 0%
kleiner 5 = 14% der Zellen

Kolmogorov-Smirnov 2-Stichproben-Test (einseitig)

		kumulierte relative Haeufigkeiten		Differenz
		-----		-----
		V1	V1	
		Auspraeg.1	Auspraeg.2	
		niedrig	hoch	
Auspraegungen	von V2			
1	0-2	0.2500	0.0185	0.2315
2	3-5	0.4091	0.0741	0.3350
3	6-8	0.5909	0.1852	0.4057
4	9-11	0.6591	0.4074	0.2517
5	12-14	0.7727	0.6296	0.1431
6	15-17	0.8864	0.8889	-0.0025
7	18-20	1.0000	1.0000	-0.0000

max. Diff. = 0.4057 asymptotischer Test: Signifikanz (1-p)*100 = 99.966%
Chi-Quadrat = 15.9640 df = 2 Signifikanz (1-p)*100 = 99.938%

Zum Vergleich geben wir zunächst die Ergebnisse aus dem normalen Chi- Quadrat-Test an.

ALMO ermittelt für die 2 zu vergleichenden Gruppen die kumulierten relativen Häufigkeiten und gibt die größte Differenz D aus. Das Vorzeichen von D ist in folgender Weise zu interpretieren: An dem Punkt, an dem die beiden relativen, kumulierten Häufigkeiten maximal differieren, ist das Vorzeichen von D positiv, wenn die 1. Gruppe über der 2. liegt und negativ, wenn umgekehrt.

ALMO errechnet die Irrtumswahrscheinlichkeit p nach einer von Smirnov hergeleiteten asymptotischen Entwicklung (siehe van der Waerden 1971, S. 72). ALMO errechnet für den KS-Test weiterhin eine Chi-Quadrat-Approximation.

Bei kleineren Häufigkeiten ist der asymptotische Test der bessere. Er liefert auch bei sehr kleinen n1 bzw. n2 einen p-Wert, der den tabellierten Werten (siehe etwa Siegel, Tabelle E) sehr nahe kommt. Die Chi-Quadrat-Approximation unterschätzt dann die Signifikanz des Zusammenhangs. Zum KS-Test siehe van der Waerden (1971, S. 72), Siegel (1987, S. 123ff).

P10.7.8 Konkordanz

Beim Maskenprogramm: Box 14, Eingabefeld k

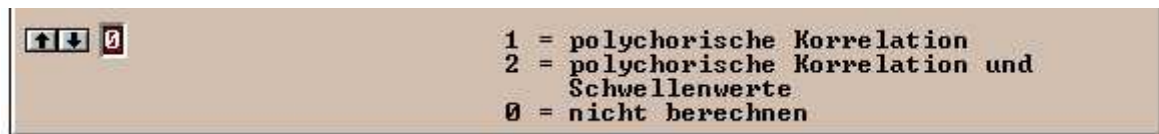


Betrachten wir ein Beispiel: 2 Lehrer beurteilen die Deutsch-Aufsätze von m Schülern. Wie stark stimmen sie in ihren Urteilen überein? Der Grad der

Übereinstimmung, die Konkordanz kann durch den "kappa-Koeffizient der Urteilskonkordanz" ausgedrückt werden. Der Koeffizient bewegt sich, wie ein Korrelationskoeffizient, zwischen -1.0 (totale Gegenläufigkeit der Urteile) über .0 (keinerlei Übereinstimmung) bis zu +1.0 (totale Übereinstimmung): Siehe dazu Bortz,Lienert,Boehnke, 1990, Kap. 9.

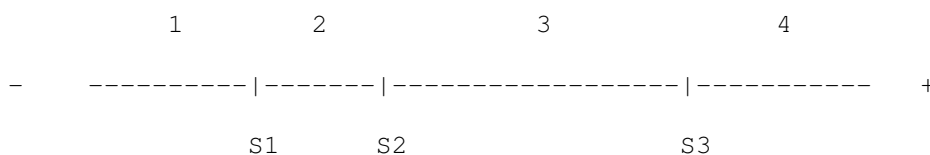
P10.7.9 Polychorische Korrelation

Beim Maskenprogramm: Box 14, Eingabefeld 1



Die beiden Variablen A und B, die gegeneinander tabelliert werden, müssen ordinal sein. Dabei wird unterstellt, dass sie in Wirklichkeit quantitativ und 2-dimensional normalverteilt sind.

Wir wollen das für die Variable A grafisch veranschaulichen:



Die Variable A besitzt 4 Ausprägungen. Es wird unterstellt, dass der Variablen eine Intervallskala zugrunde liegt, die in (ungleich große) Abstände unterteilt ist. Die Intervallskala reicht links nach minus unendlich und rechts nach plus unendlich. Die sogenannten "Schwellenwerte" S1, S2 und S3 zerschneiden die Intervallskala in (ungleich große) Abschnitte. Diesen Abschnitten werden die ordinalen Zahlen 1,2,3,4 zugeordnet (in denen die Variable gemessen ist). Eine weitere Unterstellung ist, dass Normalverteilung besteht.

Im Kalkül der polychorischen Korrelation, werden für beide Variable A und B die "Schwellenwerte" und ihre Korrelation ermittelt.

Die "tetrachorische Korrelation", die wir in Abschnitt P10.4.2 erwähnt haben, ist der Sonderfall der polychorischen Korrelation, bei dem beide Variable nur 2 Ausprägungen besitzen. Die tetrachorische Korrelation wird in Almo immer gerechnet, wenn eine 2*2-Tabelle vorliegt. Almo verwendet dafür einen kurzen Näherungskalkül (siehe dazu Bortz,Lienert,Boehnke, 1990, S.337). Wird die polychorische Korrelation zusätzlich in Box 14 aktiviert, dann entsteht für diese in der Regel ein geringfügig anderes Ergebnis, da dann ein anderer Kalkül gerechnet wird (siehe dazu Ulf Olsen: Maximum Likelihood estimation of the polychoric correlation coefficient, in: Psychometrica, Vol.44, No.4, Dec. 1979, S. 443-460).

P10.9 Eingabe von schon ausgezählten Tabellen (mit Programm-Maske Prog10m3)

Gelegentlich kommt es vor, dass man über eine 2-dimensionale Tabelle schon verfügt - für die man nun die Koeffizienten berechnen möchte, die im Rahmen von Programm 10 ermittelt werden.

Prog10m3.Msk
 2-dimensionale Tabellierung
 mit Eingabe fertiger Tabellen

Für die fertigen Tabellen sollen die Koeffizienten und Tests gerechnet werden, die im Rahmen des ALMO-Tabellierungsprogramms Prog10m1 bzw. Prog10m2 errechenbar sind

Die Tabelle, die eingegeben werden soll, ist etwa folgende

		U2 Schulabschluss		
		1	2	3
U1 Geschlecht	1	25	29	14
	2	38	27	2

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

1 **Speicher fuer Tabelle**

TabelleA = , # Zeilen der Tabelle plus 1 #
 ; # Spalten der Tabelle plus 1 #

2 **Namen für die zu tabellierenden Variablen**

3 Zahl der Ausprägungen der Zeilenvariablen
 Zahl der Ausprägungen der Spaltenvariablen

4 Option: Verschiedene Tests und Koeffizienten

5 Option: Verzichte auf Teile der Ausgabe

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

7 Grafik-Optionen

8 **Schreiben**

Schreiben Sie hier dahinter die Tabelle
 In der 1. Spalte stehen die Ausprägungen von U1
 In der 1. Zeile stehen die Ausprägungen von U2
 Im Eck oben links muss eine 0 geschrieben werden

Schalten Sie dazu die Schreibsperre aus

<--- EIN : rot
 AUS : grau

0	1	2	3
1	25	29	14
2	38	27	2

Erläuterungen zu den Boxen:

Box 1: Speicher für x Variable

Speicher fuer Tabelle

```
Tabelle# =  , # Zeilen der Tabelle plus 1 #  
 ; # Spalten der Tabelle plus 1 #
```

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

Sie müssen auch die Zahl der Zeilen und Spalten der Tabelle angeben, die eingelesen werden soll. Dabei muss die 1. Spalte, in der die Ausprägungsnummern von V1 steht, mitgezählt werden. Entsprechend auch die 1. Zeile.

Box 2: Name für die Variablen V1 und V2, die die Tabelle bilden.

Namen für die zu tabellierenden Variablen

Sie können einen Variablennamen und (hinter dem Doppelpunkt) Ausprägungsnamen angeben. Wenn Sie dies nicht wollen, dann löschen Sie das Editfeld durch Klick auf den Knopf mit dem doppelköpfigen Pfeil. Siehe dazu auch P0.3.

Box 3: Zahl der Ausprägungen der Variablen

Zahl der Ausprägungen der Zeilenvariablen
 Zahl der Ausprägungen der Spaltenvariablen

Geben Sie an wieviel Ausprägungen die beiden Variablen besitzen.

Box 4: Optionen

Siehe dazu die Erläuterungen zu Prog10m2.Msk in P10.7.

Box 5: Optionen, die das „Aussehen“ der Tabelle steuern

Siehe dazu Almo-Dokument Nr. 0 "Arbeiten mit Almo", P0.9.

Box 6: Grafik-Optionen

Siehe Almo-Dokument Nr. 0 "Arbeiten mit Almo", P0.10.

Box 7: Schreiben

Schreiben

Schreiben Sie hier dahinter die Tabelle
In der 1. Spalte stehen die Ausprägungen von U1
In der 1. Zeile stehen die Ausprägungen von U2
Im Eck oben links muss eine 0 geschrieben werden

Schalten Sie dazu die Schreibsperr aus

<--- EIN : rot
AUS : grau

```
0   1   2   3  
1  25  29  14  
2  38  27   2
```

Hinter Box 7 wird die Tabelle in folgender Weise geschrieben. (Wir betrachten das Beispiel einer 2*3 Tabelle):

0	1	2	3
1	25	29	14
2	38	27	2

Beachte: In der 1. Spalte stehen die Ausprägungsnummern von V1 und in der 1. Zeile die von V2. Sie werden ebenfalls geschrieben. Im Eck oben links muss eine 0 geschrieben werden. Die Striche werden nicht geschrieben. Wir haben sie nur zur Verdeutlichung eingesetzt.

Um hinter Box 7 überhaupt schreiben, löschen oder ändern zu können, müssen Sie durch Klick auf den Knopf mit Pfeil nach unten die Schreibsperre ausschalten.

Literatur

Agresti A.: Analysis of ordinal categorical data, Wiley, 1984

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

Clauss, G., Ebner, H.: Grundlagen der Statistik für Psychologen, Pädagogen und Soziologen, Berlin 1970

Denz, H.: Einführung in die empirische Sozialforschung, Springer, Wien, 1989

Denz, H.: Regressionsanalyse für ordinale Variable, in Holm (Hrsg.) Befragung 5, Francke, München, 1977

Fleiss, J.: Statistical methods for rates and proportions, New York-London-Sidney-Toronto 1973

in Holm (Hg.): Befragung 5, UTB 435, München, 1977

Krauth, J./ Lienert, G.A.: KFA - Die Konfigurationsfrequenzanalyse, Freiburg, München, 1973

Krauth, J.: Einführung in die Konfigurationsfrequenzanalyse, Beltz Verlag, Weinheim, 1993

Lienert, G.A.: Verteilungsfreie Methoden in der Biostatistik, Meisenheim am Glan, Band 1 (1973), Band 2 (1978)

Schmierer, Chr.: Tabellenanalyse in Holm (Hrsg.): Befragung 2, Francke, UTB 373, München 1975

van der Waerden, B.L.: Mathematische Statistik, Springer Verlag, 1971