

Kommentierte Formelsammlung multivariater statistischer Verfahren

Prof. Dr. Irene Rößler
Prof. Dr. Albrecht Ungerer

Inhaltsverzeichnis

Verfahren im Überblick	iv
Beispieldatensatz	1
1 Multiple lineare Regressionsanalyse	2
Regressionsmodell in der Stichprobe	2
Maßzahlen der Regressionsanalyse in der Stichprobe	3
Schätzmodell	4
Schätzer	4
Konfidenzintervalle	5
Testverfahren	5
Ergebnisseite der Regressionsanalyse mit WinSTAT	7
2 Varianzanalyse	9
2.1 Einfaktorielle Varianzanalyse	9
Die Stichprobe	9
Maßzahl der einfaktoriellen Varianzanalyse in der Stichprobe	10
Schätzmodell	11
Schätzer	11
Testverfahren	12
Ergebnisseite der einfaktoriellen Varianzanalyse mit WinSTAT	13
2.2 Zweifaktorielle Varianzanalyse	15
Die Stichprobe	15
Maßzahlen der zweifaktoriellen Varianzanalyse in der Stichprobe	18
Schätzmodell	18
Schätzer	19
Testverfahren	19
Ergebnisseite der zweifaktoriellen Varianzanalyse mit WinSTAT	20
3 Diskriminanzanalyse	21
Das Stichprobenmodell von Fisher	21
Maßzahlen der Diskriminanzanalyse in der Stichprobe	24
Klassifikation mit der quadrierten euklidischen Distanz	25
Das Maximum-Likelihood-Schätzmodell für Normalverteilung	25
Schätzer	25
Klassifikation nach der Maximum-Likelihood-Regel	26
Testverfahren	26
Klassifikation mit der Mahalanobis-Distanz	26
Klassifikation mit den Fisher'schen Klassifizierungsfunktionen	27
Klassifikation nach Bayes	27
Bewertung der Klassifikation	29
Ergebnisseite der Diskriminanzanalyse mit WinSTAT	30

4	Faktorenanalyse	33
	Daten- und Korrelationsmatrix	33
4.1	Hauptkomponentenanalyse	34
	Hauptachsentransformation	34
	Hauptkomponentenmethode	37
	Maßzahlen der Hauptkomponentenanalyse	39
	Ergebnisseite der Hauptkomponentenanalyse mit WinSTAT	41
4.2	Hauptachsenanalyse	42
	Kommunalitätsschätzung	42
	Hauptachsentransformation	42
	Hauptachsenmethode	43
	Ergebnisseite der Hauptachsenanalyse mit WinSTAT	43
	Kriterien zur Bestimmung der Anzahl der Faktoren	44
	Interpretation und Rotation der Faktoren	45
	Methoden zur Bestimmung der Rotationsmatrix	46
	Interpretation der rotierten Faktoren	47
	Interpretation der Faktorwerte	47
5	Clusteranalyse	49
	Datenmatrix metrischer Merkmale	49
	Distanzmaße für metrische Merkmale	50
	Ähnlichkeitsmaß für metrische Merkmale	51
	Datenmatrix nominaler binärer Merkmale	52
	Ähnlichkeitsmaße für nominale binäre Merkmale	52
	Datenmatrix nominaler binärer oder mehrstufiger Merkmale	54
	Ähnlichkeitsmaße für nominale binäre oder mehrstufige Merkmale	54
	Datenmatrix ordinaler Merkmale	55
	Ähnlichkeitsmaße für ordinale Merkmale	55
	Distanzmaße für Merkmale mit unterschiedlichem Skalenniveau	55
	Fusionierungsalgorithmen	57
	Hierarchische agglomerative Verfahren	57
	Ablauf agglomerativer Verfahren	57
	Methoden der Clusterfusionierung	58
	Rekursive Berechnung der Clusterdistanzen	59
	Dendrogramm	59
	Ergebnisseite der agglomerativen Clusteranalyse mit WinSTAT	60
6	Data Mining	62
	Der CHAID-Algorithmus von Clementine (SPSS)	62
	Der C&RT-Algorithmus von Clementine (SPSS)	63
	Data Mining des Beispieldatensatzes mit Clementine von SPSS	63
	Ergebnis einer CHAID-Analyse	63
	Ergebnis einer C&RT-Analyse	64

Anhang: Tafeln zu einigen wichtigen Verteilungen	69
A Standardnormalverteilung	69
B t -Verteilung	70
C Chi-Quadrat-Verteilung	71
D F -Verteilung	72

Multivariate statistische Verfahren im Überblick

Verfahren	Voraussetzungen	Ziele	wichtige Maßzahlen	Testverfahren
Multiple lineare Regressionsanalyse	Eine metrische zu erklärende Variable Y und mehrere metrische erklärende Variable X_1, \dots, X_k mit einem Beobachtungsvektor y und einer Beobachtungsmatrix X .	Erklärung der Varianz von Y . Prognose von Y für ein Objekt j mit $\mathbf{x}_j = (x_{1j}, \dots, x_{kj})$ durch Einsetzen von \mathbf{x}_j in die Regressionsfunktion \hat{y}_j .	zur Beurteilung der Güte der Regression in der Stichprobe: Bestimmtheitsmaß r^2 .	zur Beurteilung der Güte der Regression in der Grundgesamtheit: Overall- F -Test und t -Test für einzelne Parameter.
Varianzanalyse	Einfaktorielle Varianzanalyse: Eine metrische zu erklärende Variable Y und eine nominale erklärende Variable, sog. Faktor, X mit einer Ergebnismatrix der Beobachtungswerte von Y für die Faktorstufen des Faktors X .	Erklärung der Varianz von Y . Prognose von Y für ein Objekt j mit der Faktorstufe g durch Bestimmung des arithmetischen Mittels \bar{y}_g .	zur Beurteilung des Einflusses des Faktors X auf Y in der Stichprobe: Eta-Quadrat-Koeffizient η^2 .	zur Beurteilung des Einflusses des Faktors X auf Y in der Grundgesamtheit: F -Test.
	Zweifaktorielle Varianzanalyse: Eine metrische zu erklärende Variable Y und zwei nominale erklärende Variable, sog. Faktoren, A und B mit einer Ergebnistabelle der Beobachtungswerte von Y für die Faktorstufenkombinationen der Faktoren A und B .	Erklärung der Varianz von Y . Prognose von Y für ein Objekt j mit der Faktorstufenkombination gh durch Bestimmung des arithmetischen Mittels \bar{y}_{gh} .	zur Beurteilung des Einflusses der Faktorstufenkombinationen sowie der einzelnen Faktoren und der Interaktion auf Y in der Stichprobe: Eta-Quadrat-Koeffizienten.	zur Beurteilung des Einflusses der Faktorstufenkombinationen sowie der einzelnen Faktoren und der Interaktion auf Y in der Grundgesamtheit: F -Tests.
Diskriminanzanalyse	Zu trennende Gruppen A, B, \dots eines nominalen Merkmals und metrische erklärende Variable X_1, \dots, X_k mit nach Gruppen getrennten Beobachtungstupeln $(x_{1gi}, \dots, x_{kgi})$.	Erklärung der Trennbarkeit der Gruppen. Klassifizierung neuer Objekte, z.B. mit der quadrierten euklidischen Distanz.	zur Beurteilung der Güte der Trennbarkeit der Gruppen in der Stichprobe: kanonischer Korrelationskoeffizient c .	zur Beurteilung der Güte der Trennbarkeit der Gruppen in der Grundgesamtheit: χ^2 -Test.
Faktorenanalyse	Metrische Variable X_1, \dots, X_k mit einer Beobachtungsmatrix X .	Extrahieren von einander unabhängigen Hintergrundvariablen, sog. Faktoren, mit der Hauptkomponentenanalyse oder der Hauptachsenanalyse.	zur Interpretation der Faktoren: Faktorladungen; zur Beurteilung der erklärten Varianzen durch die Faktoren: Kommunalitäten und Eigenwerte.	
Clusteranalyse	Metrische Variable X_1, \dots, X_k mit einer Beobachtungsmatrix X . Nominale Variable X_1, \dots, X_k mit einer codierten Beobachtungsmatrix X .	Zerlegen der Gesamtheit von Objekten in disjunkte Klassen (Cluster), so dass die Klassen in sich homogen, aber deutlich voneinander getrennt sind, z.B. mit hierarchischen agglomerativen Verfahren.	Distanzmaße für metrische Variable. Ähnlichkeitsmaße für nominale Variable.	

Beispieldatensatz

Ergebnis einer statistischen Erhebung an 36 Studierenden im Vorfeld der Statistikklausur:

- Y_1 : Fachbereich (1: Betriebswirtschaft, 2: Informationstechnologien) ([excel-Datei-download](#))
- X_1 : Geschlecht (0: männlich, 1: weiblich)
- X_2 : Mathe-Note im Abitur
- Y_2 : Ausgaben für Kopien (€/Semester)
- X_3 : Nettoeinkommen (€/Semester)
- X_4 : Zeit für Nacharbeitung und Klausurvorbereitung (Std/Semester)
- X_5 : Verweildauer im Internet (Std/Semester)
- X_6 : Aufenthaltsdauer in Kinos, Discos oder Kneipen (Std/Semester)
- X_7 : Anzahl gekaufter Fachbücher im Semester
- X_8 : erwartete Leistung in der Statistikklausur (-1: unterdurchschn., 0: durchschn., +1: eher besser)

Nr.	Y_1	X_1	X_2	Y_2	X_3	X_4	X_5	X_6	X_7	X_8
1	1	1	3	24	2000	198	108	54	6	0
2	2	0	1	33	2070	108	36	18	2	0
3	1	1	2	28	2130	162	54	54	5	0
4	2	0	1	49	2840	108	18	0	2	-1
5	1	0	3	18	1820	216	105	108	6	1
6	2	1	2	24	2150	198	36	54	7	0
7	1	1	3	32	2030	180	24	36	5	-1
8	2	1	2	17	1730	180	90	90	4	1
9	1	1	3	33	2180	174	18	54	5	-1
10	1	0	2	36	2160	144	90	54	3	0
11	1	1	3	18	1880	204	90	72	7	0
12	1	0	3	45	2340	126	72	36	5	-1
13	2	0	2	34	2270	180	36	54	5	0
14	1	1	3	12	1830	270	54	90	8	1
15	2	1	2	40	2230	108	108	108	5	-1
16	1	0	3	42	2390	192	108	18	4	-1
17	2	0	2	21	1900	216	114	72	6	1
18	2	1	3	21	1900	216	90	72	5	0
19	1	0	4	44	2150	144	108	36	5	-1
20	2	0	2	15	1980	252	90	126	6	1
21	2	1	2	41	2420	252	72	18	5	-1
22	2	1	2	17	2000	216	12	54	8	0
23	1	1	3	7	1930	216	0	24	8	1
24	2	1	2	14	2040	234	78	144	5	1
25	1	0	3	27	2040	162	144	132	3	0
26	1	0	3	37	2260	162	36	54	2	0
27	1	0	3	22	1940	162	57	54	4	1
28	2	1	2	27	2290	180	18	36	9	0
29	1	0	3	30	2050	144	108	72	5	0
30	1	0	3	32	2120	126	72	36	2	0
31	2	1	1	8	1940	198	36	18	6	1
32	2	0	3	36	2150	108	30	12	1	-1
33	2	1	2	38	2150	180	72	36	5	-1
34	2	0	2	35	2240	168	60	48	4	0
35	1	1	3	25	2020	198	54	36	7	0
36	2	0	2	26	2030	198	12	0	5	1

1 Multiple lineare Regressionsanalyse

Soll die Streuung einer metrischen Variablen Y durch eine ebenso metrische Variable X erklärt werden (Einfachregression) und ergibt sich hierbei ein kleiner Wert für das Bestimmtheitsmaß, so versucht man durch Hinzunahme weiterer metrischer erklärender Variablen einen Erklärungsbeitrag zu finden (multiple lineare Regression).

Regressionsmodell in der Stichprobe

Zur Bestimmung einer Regressionsfunktion, mit der die Streuung der Zielvariablen Y erklärt werden kann, wählt man das Kleinste-Quadrate-Kriterium, denn dann ist es möglich, die zu erklärende Streuung von Y in eine nicht erklärte und eine durch die Regressionsfunktion erklärte Streuung zu zerlegen.

Y : zu erklärende metrische Variable
 X_1, \dots, X_k : k erklärende metrische, linear unabhängige Variablen; $\text{rang}(\mathbf{X}) = k + 1 \leq n$
 $(y_i, x_{1i}, x_{2i}, \dots, x_{ki})$: Beobachtungstupel des i -ten Elements, $i = 1, \dots, n$, (Stichprobenumfang n)
 $\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \dots + b_k x_{ki}$: Kleinste-Quadrate-Stichprobenregressionsfunktion
 Matrixschreibweise der Normal- bzw. Bestimmungsgleichungen aus $\min_{b_0, b_1, \dots, b_k} \sum_{i=1}^n (y_i - \hat{y}_i)^2$:

Normalgleichungen:

$$(\mathbf{X}'\mathbf{X})\mathbf{b} = \mathbf{X}'\mathbf{y}$$

$$\text{mit } \mathbf{b} := \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{pmatrix}, \mathbf{X} := \begin{pmatrix} 1 & x_{11} & x_{21} & \dots & x_{k1} \\ 1 & x_{12} & x_{22} & \dots & x_{k2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{1n} & x_{2n} & \dots & x_{kn} \end{pmatrix}, \mathbf{y} := \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

Bestimmungsgleichungen:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Zur Ermittlung von Maßzahlen für die Güte der Regressionsfunktion in der Stichprobe sowie zur Durchführung des F-Tests für die Güte der Regressionsfunktion in der Grundgesamtheit wird eine Varianzanalyse (ANalysis Of VAriance) durchgeführt. Hierzu wird eine Varianzzerlegung vorgenommen:

$$\underbrace{\frac{1}{n-1} \cdot \sum_{i=1}^n (y_i - \bar{y})^2}_{s_Y^2}_{\frac{1}{n-1}\text{SQT}} = \underbrace{\frac{1}{n-1} \cdot \sum_{i=1}^n (y_i - \hat{y}_i)^2}_{s_e^2}_{\frac{1}{n-1}\text{SQR}} + \underbrace{\frac{1}{n-1} \cdot \sum_{i=1}^n (\hat{y}_i - \bar{y})^2}_{s_{\hat{y}}^2}_{\frac{1}{n-1}\text{SQE}}$$

Die ANOVA-Tabelle beinhaltet die Quadratsummen und beschreibt die Berechnung des F_{emp} -Wertes:

	Quadratsumme SQ	Anzahl der Freiheitsgrade ν	mittlere Quadratsumme MQ	Wert der F-verteilteten Testfunktion F_{emp}
Regression	$\text{SQE} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$	$\nu_E = k$	$\text{MQE} = \frac{\text{SQE}}{k}$	$F_{\text{emp}} = \frac{\text{MQE}}{\text{MQR}}$
Residuen	$\text{SQR} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$	$\nu_R = n - k - 1$	$\text{MQR} = \frac{\text{SQR}}{n - k - 1}$	
Gesamt	$\text{SQT} = \sum_{i=1}^n (y_i - \bar{y})^2$	$\nu_T = n - 1$	$\text{MQT} = \frac{\text{SQT}}{n - 1}$	

Maßzahlen der Regressionsanalyse in der Stichprobe

Maßzahl	Symbol	Berechnung	Aussage
Regressionskoeffizienten	b_q	$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$	Geschätzte Veränderung der Zielvariablen Y bei Erhöhung der erklärenden Variablen X_q um 1 Einheit unter der Annahme, dass die restlichen erklärenden Variablen konstant bleiben, $q = 1, \dots, k$.
standardisierte Regressionskoeffizienten	b_q^*	$b_q^* = b_q \frac{s_{X_q}}{s_Y}, \quad q = 1, \dots, k$	Geschätzte Veränderung der standardisierten Zielvariablen $Y^* = \frac{Y}{s_Y}$ bei Erhöhung der standardisierten Variablen $X_q^* = \frac{X_q}{s_{X_q}}$ um 1 Einheit. Durch die Standardisierung der Regressionskoeffizienten b_q ist ein Vergleich der Stärke des Einflusses der Variablen X_1, \dots, X_k auf Y möglich.
(multiples) Bestimmtheitsmaß	r^2	$r^2 = \frac{s_{\hat{Y}}^2}{s_Y^2} = \frac{\text{SQE}}{\text{SQT}} = 1 - \frac{\text{SQR}}{\text{SQT}}$	Anteil der durch die Regressionsfunktion „erklärten“ Varianz der Zielvariablen, $0 \leq r^2 \leq 1$.
korrigiertes Bestimmtheitsmaß	r_{korr}^2	$r_{\text{korr}}^2 = 1 - \frac{\frac{s_e^2}{n-k-1}}{\frac{s_Y^2}{n-1}}$ $= 1 - \frac{n-1}{n-k-1}(1-r^2)$ $= 1 - \frac{\frac{\text{SQR}}{v_R}}{\frac{\text{SQT}}{v_T}} = 1 - \frac{\text{MQR}}{\text{MQT}}$	r^2 wird korrigiert, da es bei Hinzunahme jeder weiteren unabh. Variablen steigt, auch dann, wenn die hinzugenommene Variable keinen sinnvollen Erklärungsbeitrag liefert. Außerdem ist r_{korr}^2 erwartungstreu, während r^2 einen zu hohen Wert ausweist, d.h. $E(r_{\text{korr}}^2) = \rho^2 < E(r^2)$. r_{korr}^2 ist der (nach unten korrigierte) Anteil der durch die Regressionsfunktion „erklärten“ Varianz der Zielvariablen. Es gilt: $r_{\text{korr}}^2 \leq r^2$ und $\lim_{n \rightarrow \infty} r_{\text{korr}}^2 = r^2$.
partielles Bestimmtheitsmaß	r_{part}^2	$r_{\text{part}}^2 = \frac{r_{Y \cdot X_1, \dots, X_k}^2 - r_{Y \cdot X_1, \dots, X_{k-1}}^2}{1 - r_{Y \cdot X_1, \dots, X_{k-1}}^2}$ $= \frac{\text{SQE}_{X_1, \dots, X_k} - \text{SQE}_{X_1, \dots, X_{k-1}}}{\text{SQR}_{X_1, \dots, X_{k-1}}}$	Anteil der durch X_k zusätzlich erklärten Varianz an der durch X_1, \dots, X_{k-1} nicht erklärten Varianz.
Standardfehler des Schätzers	s_D	$s_D = \sqrt{\frac{(n-1)s_e^2}{n-k-1}} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-k-1}}$ $= \sqrt{\frac{\text{SQR}}{v_R}} = \sqrt{\text{MQR}}$	Durchschnittliche Abweichung der durch die Regressionsfunktion prognostizierten Werte von den Beobachtungswerten.

Schätzmodell

Um aus den nach der Methode der kleinsten Quadrate für die Stichprobe berechneten Regressionskoeffizienten b_0, b_1, \dots, b_k die unbekannt, wahren Regressionskoeffizienten $\beta_0, \beta_1, \dots, \beta_k$ der Regressionsfunktion der Grundgesamtheit $\hat{y} = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$ zu schätzen, wird das durch folgende Annahmen beschriebene Schätzmodell zugrundegelegt:

- Die n Beobachtungstupel $(y_i, x_{1i}, x_{2i}, \dots, x_{ki})$, $i = 1, \dots, n$, stellen eine Stichprobe aus einer übergeordneten realen oder hypothetischen $(k+1)$ -dimensionalen Grundgesamtheit dar.
- Die n Werte y_i können als Realisationen der n beobachtbaren Zufallsvariablen Y_i mit

$$Y_i = \underbrace{\beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki}}_{\text{systematische Komponente}} + \underbrace{U_i}_{\text{stochastische Komponente}}, \quad i = 1, \dots, n,$$

aufgefasst werden, wobei

x_{qi} : fest vorgegebene Werte, d.h. von Stichprobe $i = 1$ bis zu Stichprobe n beibehaltene Werte der beobachtbaren Variablen X_q , $q = 1, \dots, k$, (klassisches Modell) oder Realisationen der beobachtbaren Zufallsvariablen X_{qi} , $q = 1, \dots, k$, z.B. wenn in einer Zufallsstichprobe mit dem Umfang n simultan die n Realisationen der Variablen (Y, X_1, \dots, X_k) beobachtet werden (Modell für stochastische Regressoren)

U_i : nicht beobachtbare Zufallsvariablen mit den Realisationen $u_i = y_i - (\beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki})$

β_q : konstante, unbekannte, zu schätzende Parameter der Grundgesamtheit, $q = 0, \dots, k$

- Die Zufallsvariablen U_i , $i = 1, \dots, n$, erfüllen die Bedingungen:

$$\left. \begin{aligned} E(U_i) &= 0, \quad i = 1, \dots, n \\ \text{Var}(U_i) &= \sigma_{U_i}^2 = \sigma_U^2, \quad i = 1, \dots, n, \text{ (Homoskedastizität)} \\ \text{Cov}(U_i, U_j) &= \sigma_{U_i, U_j} = 0, \quad i, j = 1, \dots, n, \quad i \neq j, \text{ (keine Autokorrelation), d.h. die Störvariablen } U_i \text{ sind unabhängig verteilt.} \end{aligned} \right\} \text{d.h. die Störvariablen } U_i \text{ sind identisch verteilt}$$

Für stochastische Regressoren sind diese Eigenschaften unter der Bedingung $(X_{1i}, \dots, X_{ki}) = (x_{1i}, \dots, x_{ki})$ zu verstehen, womit durch $E(U_i | X_i) = 0$ gewährleistet ist, dass die Störvariablen U_i und die unabhängigen Variablen X_{qi} , $q = 1, \dots, k$, in der Grundgesamtheit nicht korrelieren.

- Zusätzliche Annahme für Tests und Konfidenzintervalle: $U_i \sim N(0, \sigma_U^2)$, $i = 1, \dots, n$.

Schätzer

Parameter	Schätzer	Berechnung	Eigenschaften
β_q	b_q	b_q , $q = 0, \dots, k$, sind die Komponenten des Vektors $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ vgl. Seite 2	Die Regressionskoeffizienten b_q der linearen KQ-Stichprobenregressionsfunktion sind erwartungstreue Schätzer für die Regressionskoeffizienten β_q der Grundgesamtheit, $q = 0, \dots, k$.

Gauß-Markov-Theorem: Von allen linearen, unverzerrten Schätzern für β_q sind die KQ-Schätzer b_q BLUE (Best Linear Unbiased Estimator), $q = 0, \dots, k$, d.h. weisen die kleinste Varianz auf. Unter der Normalverteilungsannahme sind sie überhaupt beste Schätzer und exakt normalverteilt.

Parameter	Schätzer	Berechnung	Eigenschaften
σ_U^2	$\hat{\sigma}_U^2$	$\hat{\sigma}_U^2 = s_D^2 = \frac{(n-1)s_e^2}{n-k-1}$ $= \frac{\text{SQR}}{n-k-1} = \frac{\sum (y_i - \hat{y}_i)^2}{n-k-1}$	Die aus den Stichprobendaten berechenbare Varianz $\hat{\sigma}_U^2$ ist erwartungstreuere Schätzer für die unbekannte Varianz σ_U^2 der Störvariablen U_i in der Grundgesamtheit.
σ_{B_q}	$\hat{\sigma}_{B_q}$	$\hat{\sigma}_{B_q} = \sqrt{\hat{v}_{qq}}$, wobei \hat{v}_{qq} die Elemente der Hauptdiagonalen der Matrix $\hat{V} = \hat{\sigma}_U^2 (X'X)^{-1}$ sind, $q = 0, \dots, k$.	$\hat{\sigma}_{B_q}$ sind erwartungstreue Schätzer für die Standardfehler σ_{B_q} der Regressionskoeffizienten B_q . Je geringer $\hat{\sigma}_{B_q}$ ist, d.h. je kleiner s_e^2 bzw. SQR in der Stichprobe oder je größer der Stichprobenumfang n oder je größer die Varianz der X -Werte ist, desto genauer ist die Schätzung für die Regressionskoeffizienten B_q , $q = 0, \dots, k$.

Konfidenzintervalle

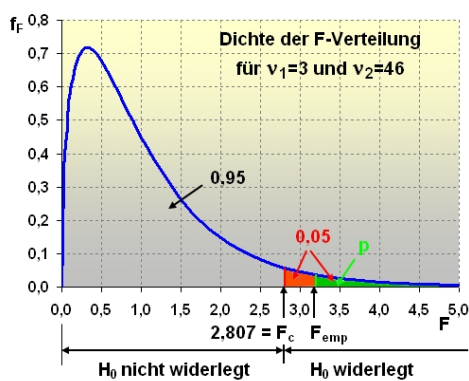
Um die Güte der aus der Stichprobe geschätzten Regressionskoeffizienten β_q , $q = 1, \dots, k$, zu beurteilen, werden Konfidenzintervalle (aus einer konkreten Stichprobe) bestimmt.

Verteilung der standardisierten Schätzer	Konfidenzintervalle für β_q , $q = 1, \dots, k$
$\frac{b_q - \beta_q}{\hat{\sigma}_{B_q}} \sim t(n-k-1)$	$b_q - t_{1-\alpha/2}(n-k-1) \cdot \hat{\sigma}_{B_q} \leq \beta_q \leq b_q + t_{1-\alpha/2}(n-k-1) \cdot \hat{\sigma}_{B_q}$
Unter der Normalverteilungsannahme sind die standardisierten Schätzer exakt t -verteilt, ansonsten approximativ für große Stichproben. D.h.: Falls die Normalverteilungsannahme verletzt ist, besitzen die Konfidenzintervalle für große Stichproben approximativ die Überdeckungswahrscheinlichkeit $1 - \alpha$.	

Testverfahren

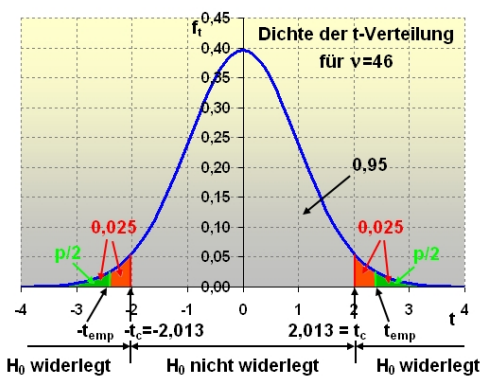
- Zur Überprüfung, ob die unabhängigen Variablen überhaupt zur Erklärung der abhängigen Variablen in der Grundgesamtheit beitragen oder anders ausgedrückt, ob der durch das Bestimmtheitsmaß der Stichprobe berechnete Erklärungsbeitrag der Regression für die Varianz der abhängigen Variablen als signifikant angesehen werden kann, wird ein Hypothesentest, der sog. **Overall - F-Test** (Goodness of fit-Test) durchgeführt.
- Zur Überprüfung, ob eine einzelne unabhängige Variable einen signifikanten Erklärungsbeitrag für die abhängige Variable liefert, wird der **t-Test für einzelne Parameter** durchgeführt.

Overall - F-Test:			Voraussetzung: $U_i \sim N(0, \sigma_U^2), i = 1, \dots, n$		
statistische Kenngröße	Nullhypothese H_0	Alternativhyp. H_1	Testfunktion F	Testverteilung F/H_0	Entscheidungsregel zur Ablehnung von H_0 bei gegebenem α , z.B. $\alpha = 0,05$
R^2	$\beta_1 = \beta_2 = \dots = \beta_k = 0$	$\beta_q \neq 0$ für mind. ein $q, q \in \{1, \dots, k\}$	$F = \frac{R^2(n-k-1)}{(1-R^2)k}$ $= \frac{MQE}{MQR}$	$f(k, n-k-1)$	$F_{emp} > \underbrace{f_{1-\alpha}(k, n-k-1)}_{=: F_c}$



In den meisten Computerprogrammen wird die zu dem aus der Stichprobe ermittelten Wert F_{emp} zugehörige Signifikanz p berechnet. Für $p < 0,05$ kann H_0 abgelehnt werden, d.h.: Man begeht nur mit $p \cdot 100\%$ Wahrscheinlichkeit einen Fehler, wenn man behauptet, dass in der Grundgesamtheit wenigstens eine der unabhängigen Variablen einen Einfluss auf die abhängige Variable hat; damit liefert mindestens eine unabhängige Variable einen signifikanten Erklärungsbeitrag für die Varianz der abhängigen Variablen.

t-Test für einzelne Parameter:			Voraussetzung: $U_i \sim N(0, \sigma_U^2), i = 1, \dots, n$		
statistische Kenngröße	Nullhypothese H_0	Alternativhyp. H_1	Testfunktion T	Testverteilung T/H_0	Entscheidungsregel zur Ablehnung von H_0 bei gegebenem α , z.B. $\alpha = 0,05$
B_q mit $\hat{\sigma}_{B_q}$	$\beta_q = 0, q \in \{1, \dots, k\}$	$\beta_q \neq 0$	$T_q = \frac{B_q}{\hat{\sigma}_{B_q}}$	$t(n-k-1)$	$ t_{emp} > \underbrace{t_{1-\alpha/2}(n-k-1)}_{=: t_c}$



Für die zu dem aus der Stichprobe ermittelten Wert t_{emp} zugehörige Signifikanz p mit $p < 0,05$ kann H_0 abgelehnt werden, d.h.: Man begeht nur mit $p \cdot 100\%$ Wahrscheinlichkeit einen Fehler, wenn man behauptet, dass in der Grundgesamtheit die unabhängige Variable X_q einen Einfluss auf die abhängige Variable hat; damit liefert X_q einen signifikanten Erklärungsbeitrag für die Varianz der abhängigen Variablen.

Ergebnisseite der Regressionsanalyse mit WinSTAT für Excel

In den gelben Feldern stehen die Symbole für die Formeln der Formelsammlung, nach denen WinSTAT die Zahlen berechnet.

Multiple Regression

X-Variable:	X_1	Mit der Methode „Direkt“ werden alle unabhängigen Variablen in die Regressionsanalyse aufgenommen. Mit den alternativen Methoden „Schrittweise“, „Vorwärts“ oder „Rückwärts“ wird eine Selektion der unabhängigen Variablen aufgrund des <i>t</i> -Tests vorgenommen; dabei werden nur solche Variablen aufgenommen, deren Signifikanzniveau kleiner als ein vorgegebener Wert (z.B. $p < 0,05$) ist.
	\vdots	
	X_k	
Y-Variable:	Y	
Methode:	Direkt	

Zusammenfassung

	N	R	R-Quadrat	Std.Fehler
normal	n	r	r^2	s_D
korrigiert		r_{korr}	r_{korr}^2	

Gleichung

95%					
	Koeffizient	Vertrauen (\pm)	Std.Fehler	T	P
Konstante	b_0	$t_{0,975}(n - k - 1)\hat{\sigma}_{b_0}$	$\hat{\sigma}_{b_0}$	$t_{\text{emp}} = \frac{b_0}{\hat{\sigma}_{b_0}}$	p_0
X_1	b_1	$t_{0,975}(n - k - 1)\hat{\sigma}_{b_1}$	$\hat{\sigma}_{b_1}$	$t_{\text{emp}} = \frac{b_1}{\hat{\sigma}_{b_1}}$	p_1
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
X_k	b_k	$t_{0,975}(n - k - 1)\hat{\sigma}_{b_k}$	$\hat{\sigma}_{b_k}$	$t_{\text{emp}} = \frac{b_k}{\hat{\sigma}_{b_k}}$	p_k

Varianzanalyse

	Quadrat-summe	Freiheits-grade	mittlere QS	F	P
Regression	SQE	k	MQE	$F_{\text{emp}} = \frac{\text{MQE}}{\text{MQR}}$	p
Residue	SQR	$n - k - 1$	MQR		
Gesamt	SQT	$n - 1$	MQT		

Auf-

gabe

1

Für den Beispieldatensatz Seite 1 mit den Semestereinkommen (X_3), der Zeit für Nacharbeit und Klausurvorbereitung (X_4), der Verweildauer im Internet (X_5), der Aufenthaltsdauer in Kinos, Discos oder Kneipen (X_6) und den Ausgaben für Kopien (Y) erhält man den unten stehenden WinSTAT-Output einer multiplen Regressionsanalyse.

- Formulieren Sie die lineare Kleinst-Quadrate-Regressionsfunktion. Ergänzen Sie den Output um die fehlenden Werte der genannten Maßzahlen und interpretieren Sie Ihr Ergebnis.
- Für eine 3-fach-Regression mit X_3 , X_5 und X_6 erhält man das Bestimmtheitsmaß $r^2 = 0,763$. Ermitteln Sie, wieviel Prozent der durch die 3-fach-Regression nicht erklärten Varianz der Ausgaben für Kopien durch die Varianz der hinzugenommenen Variablen X_4 „Zeit für Nacharbeit und Klausurvorbereitung“ erklärt werden kann.
- Formulieren Sie die Null- und Alternativhypothese des F -Tests im Sachzusammenhang. Stellen Sie den Rechengvorgang zur Bestimmung des empirischen Wertes der F -verteilten Testfunktion dar und interpretieren Sie das zugehörige Signifikanzniveau. Veranschaulichen Sie Ihr Ergebnis grafisch.
- Führen Sie die gleichen Schritte wie in c) für den t -Test einer Variablen durch. Interpretieren Sie anschließend auch die Signifikanzniveaus der restlichen Variablen.
- Interpretieren Sie die 95%-Konfidenzintervalle der Regressionskoeffizienten.

Multiple Regression

Zusammenfassung

	N	R	R-Quadrat	Std.Fehler
normal	36			
korrigiert				

Gleichung

	95%				
	Koeffizient	Vertrauen (\pm)	Std.Fehler	T	P
Konstante	-31,179	24,857	12,187	-2,558	0,016
Einkommen pro Semester	0,033	0,009	0,005	7,263	0,000
Nach- und Vorbereitungszeit	-0,078	0,044	0,022	-3,576	0,001
Verweildauer im Internet	0,101	0,055	0,027	3,747	0,001
Aufenthaltsdauer in Kinos ...	-0,066	0,062	0,030	-2,166	0,038

Varianzanalyse

	Quadrat-summe	Freiheits-grade	mittlere QS	F	P
Regression	3379,466	4	844,867	38,486	0,000
Residue	680,534	31	21,953		
Gesamt	4060	35	116		

2 Varianzanalyse

Ziel der Varianzanalyse ist es, die Streuung einer metrischen Variablen durch eine bzw. mehrere nominale (gruppierte) Variable(n) zu erklären.

2.1 Einfaktorielle Varianzanalyse Die Stichprobe

Y : zu erklärende metrische Variable

X : Faktor, d.h. erklärende nominale Variable

r : Anzahl der Faktorstufen (Merkmalsausprägungen) des Faktors X

n_g : Anzahl (Stichprobenumfang) der Beobachtungswerte der Faktorstufe g , $g = 1, \dots, r$

y_{gi} : i -ter Beobachtungswert der Faktorstufe g , $g = 1, \dots, r$ und $i = 1, \dots, n_g$

Ergebnismatrix:

Faktorstufen	Beobachtungswerte (Stichprobenelemente)						Stichproben- summe	Stichproben- mittel
	$i = 1$	$i = 2$	\dots	i	\dots	$i = n_g$		
$g = 1$	y_{11}	y_{12}	\dots	y_{1i}	\dots	y_{1n_1}	$y_{1\cdot}$	$\bar{y}_{1\cdot}$
$g = 2$	y_{21}	y_{22}	\dots	y_{2i}	\dots	y_{2n_2}	$y_{2\cdot}$	$\bar{y}_{2\cdot}$
\vdots	\vdots	\vdots		\vdots		\vdots	\vdots	\vdots
g	y_{g1}	y_{g2}	\dots	y_{gi}	\dots	y_{gn_g}	$y_{g\cdot}$	$\bar{y}_{g\cdot}$
\vdots	\vdots	\vdots		\vdots		\vdots	\vdots	\vdots
$g = r$	y_{r1}	y_{r2}	\dots	y_{ri}	\dots	y_{rn_r}	$y_{r\cdot}$	$\bar{y}_{r\cdot}$
Stichproben- gesamtsumme	.						$y_{\cdot\cdot}$.
Stichproben- gesamtmittel	.						.	$\bar{y}_{\cdot\cdot}$

$$y_{g\cdot} = \sum_{i=1}^{n_g} y_{gi} \quad \text{Summe der Beobachtungswerte der Faktorstufe } g$$

$$\bar{y}_{g\cdot} = \frac{1}{n_g} y_{g\cdot} = \frac{1}{n_g} \sum_{i=1}^{n_g} y_{gi} \quad \text{arithmetisches Mittel der Beobachtungswerte der Faktorstufe } g$$

$$y_{\cdot\cdot} = \sum_{g=1}^r \sum_{i=1}^{n_g} y_{gi} \quad \text{Summe aller } n \text{ Beobachtungswerte}$$

$$\bar{y}_{\cdot\cdot} = \frac{1}{n} y_{\cdot\cdot} = \frac{1}{n} \sum_{g=1}^r \sum_{i=1}^{n_g} y_{gi} \quad \text{arithmetisches Mittel aller } n \text{ Beobachtungswerte}$$

$$n = \sum_{g=1}^r n_g \quad \text{Anzahl aller Beobachtungswerte}$$

Wenn $\bar{y}_{1.} = \bar{y}_{2.} = \dots = \bar{y}_{r.}$ gelten würde, so hätte der Faktor X überhaupt keinen Einfluss auf Y . Unterscheiden sich aber die arithmetischen Mittel von mindestens zwei Faktorstufen, d.h. streuen mindestens zwei Stichprobenmittel $\bar{y}_{g.}, \bar{y}_{h.}, g, h \in \{1, \dots, r\}$, um das Stichprobengesamtmittel $\bar{y}_{..}$, so hat der Faktor X Einfluss auf die Variable Y bzw. leistet X einen Erklärungsbeitrag für die Streuung von Y . D.h.: Mit $(\bar{y}_{g.} - \bar{y}_{..})$ wird der Effekt der Faktorstufe $g, g = 1, \dots, r$, des Faktors X auf die Zielvariable Y beschrieben.

Zur Ermittlung einer Maßzahl für die Güte des Erklärungsbeitrags der Streuung der Zielvariablen Y durch den Faktor X in der Stichprobe wird eine Varianzzerlegung für die r Faktorstufen durchgeführt:

$$\underbrace{\frac{1}{n-1} \cdot \sum_{g=1}^r \sum_{i=1}^{n_g} (y_{gi} - \bar{y}_{..})^2}_{s_Y^2} = \underbrace{\frac{1}{n-1} \cdot \sum_{g=1}^r \sum_{i=1}^{n_g} (y_{gi} - \bar{y}_{g.})^2}_{s_{\text{int}}^2} + \underbrace{\frac{1}{n-1} \cdot \sum_{g=1}^r n_g (\bar{y}_{g.} - \bar{y}_{..})^2}_{s_{\text{ext}}^2}$$

$$\underbrace{\frac{1}{n-1} \text{SQT}}_{\frac{1}{n-1} \text{SQR}} \quad \text{mit } s_{\text{int}}^2 = \frac{1}{n-1} \cdot \sum_{g=1}^r (n_g - 1) \cdot s_g^2 \quad \underbrace{\frac{1}{n-1} \text{SQE}}_{\frac{1}{n-1} \text{SQR}} \quad \text{mit } s_g^2 = \frac{1}{n_g - 1} \sum_{i=1}^{n_g} (y_{gi} - \bar{y}_{g.})^2$$

Die Quadratsummen der Varianzzerlegung werden in einer ANOVA-Tabelle erfasst, die außerdem die Berechnung des Wertes der Testfunktion des F -Tests beschreibt:

	Quadratsumme SQ	Anzahl der Freiheitsgrade ν	mittlere Quadratsumme MQ	Wert der F-verteiltern Testfunktion F_{emp}
extern	$\text{SQE} = \sum_{g=1}^r n_g (\bar{y}_{g.} - \bar{y}_{..})^2$	$\nu_E = r - 1$	$\text{MQE} = \frac{\text{SQE}}{r - 1}$	$F_{\text{emp}} = \frac{\text{MQE}}{\text{MQR}}$
intern	$\text{SQR} = \sum_{g=1}^r \sum_{i=1}^{n_g} (y_{gi} - \bar{y}_{g.})^2$	$\nu_R = n - r$	$\text{MQR} = \frac{\text{SQR}}{n - r}$	
Gesamt	$\text{SQT} = \sum_{g=1}^r \sum_{i=1}^{n_g} (y_{gi} - \bar{y}_{..})^2$	$\nu_T = n - 1$	$\text{MQT} = \frac{\text{SQT}}{n - 1}$	

Aus den Quadratsummen der ANOVA-Tabelle kann der Eta-Quadrat-Koeffizient berechnet werden.

Maßzahl der einfaktoriellen Varianzanalyse in der Stichprobe

Maßzahl	Symbol	Berechnung	Aussage
Eta-Quadrat-Koeffizient	η^2	$\eta^2 = \frac{s_{\text{ext}}^2}{s_Y^2} = \frac{\text{SQE}}{\text{SQT}} = 1 - \frac{\text{SQR}}{\text{SQT}}$	Anteil der durch den Einfluss des Faktors X „erklärten“ Varianz der Zielvariablen Y , $0 \leq \eta^2 \leq 1$.

Schätzmodell

Zur Analyse, ob die in der Stichprobe ermittelten Ergebnisse: „Arithmetische Mittel der einzelnen Faktorstufen“, „globales arithmetische Mittel über alle Faktorstufen“, „Effekte der Faktorstufen des Faktors X auf die Zielvariable Y “ auch auf die Grundgesamtheit übertragen werden können, wird das durch folgende Annahmen beschriebene Schätzmodell zugrundegelegt:

- Die n Beobachtungswerte y_{gi} , $g = 1, \dots, r$ und $i = 1, \dots, n_g$, stellen eine Stichprobe aus einer übergeordneten Grundgesamtheit dar.
- Die n Werte y_{gi} können als Realisationen der n beobachtbaren Zufallsvariablen Y_{gi} mit

$$\text{Modell (I): } Y_{gi} = \mu_g + U_{gi}, \quad g = 1, \dots, r \text{ und } i = 1, \dots, n_g,$$

$$\text{Modell (II): } Y_{gi} = \mu + \alpha_g + U_{gi}, \quad g = 1, \dots, r \text{ und } i = 1, \dots, n_g,$$

aufgefasst werden, wobei

U_{gi} : nicht beobachtbare Zufallsvariablen mit den Realisationen $u_{gi} = y_{gi} - \mu_g$, (Modell (I)) bzw. $u_{gi} = y_{gi} - \mu - \alpha_g$ (Modell (II))

μ_g : unbekanntes, wahres, zu schätzendes arithmetisches Mittel der Faktorstufe g in der Grundgesamtheit

$\mu = \frac{1}{r} \sum_{g=1}^r n_g \mu_g$: unbekanntes, wahres, zu schätzendes, globales arithmetisches Mittel über alle Faktorstufen g in der Grundgesamtheit

$\alpha_g = \mu_g - \mu$: unbekannter, wahrer, zu schätzender Effekt des Faktors X auf der Faktorstufe g in der Grundgesamtheit, wobei gilt:

$$\sum_{g=1}^r n_g \alpha_g = 0.$$

- Die Zufallsvariablen Y_{gi} bzw. U_{gi} sind in jeder Faktorstufe g , $g = 1, \dots, r$, unabhängig normalverteilt mit

$$Y_{gi} \sim N(\mu_g, \sigma^2) \text{ bzw. } U_{gi} \sim N(0, \sigma^2), \quad i = 1, \dots, n_g, \quad \forall g$$

Schätzer

Modell	Parameter	Schätzer	Berechnung
Modell (I)	μ_g	$\bar{y}_{g\cdot}$	$\bar{y}_{g\cdot} = \frac{1}{n_g} \sum_{i=1}^{n_g} y_{gi}$
Modell (II)	μ	$\bar{y}_{\cdot\cdot}$	$\bar{y}_{\cdot\cdot} = \frac{1}{n} \sum_{g=1}^r \sum_{i=1}^{n_g} y_{gi}$
Modell (II)	α_g	$\bar{y}_{g\cdot} - \bar{y}_{\cdot\cdot}$	

Testverfahren

Zur Überprüfung, ob der Faktor X überhaupt eine Wirkung auf die Zielvariable Y in der Grundgesamtheit erzielt oder anders ausgedrückt, ob der durch den η^2 -Koeffizienten der Stichprobe berechnete Erklärungsbeitrag des Faktors für die Varianz der Zielvariablen als signifikant angesehen werden kann, wird ein F -Test durchgeführt.

F-Test der Varianzanalyse:			Voraussetzung: $Y_{gi} \sim N(\mu_g, \sigma^2)$, $g = 1, \dots, r$, $i = 1, \dots, n_g$		
statische Kenngröße	Nullhypothese H_0	Alternativhyp. H_1	Testfunktion F	Testverteilung F/H_0	Entscheidungsregel zur Ablehnung von H_0 bei gegebenem α , z.B. $\alpha = 0,05$
η^2	Modell (I): $\mu_1 = \mu_2 = \dots = \mu_r$ Modell (II): $\alpha_1 = \alpha_2 = \dots = \alpha_r = 0$	Modell (I): $\mu_g \neq \mu_h$ für mind. ein Paar g, h Modell (II): $\alpha_g \neq 0$ für mind. zwei α_g	$F = \frac{\eta^2(n-r)}{(1-\eta^2)(r-1)}$ $= \frac{\text{MQE}}{\text{MQR}}$	$f(r-1, n-r)$	$F_{\text{emp}} > f_{1-\alpha}(r-1, n-r)$

Zur Überprüfung der Voraussetzung der Gleichheit der Varianzen der normalverteilten Variablen Y über alle Faktorstufen in der Grundgesamtheit werden Hypothesentests durchgeführt:

F-Test zum Vergleich von Varianzen für 2 Faktorstufen:			Voraussetzung: $Y_{1i} \sim N(\mu_1, \sigma_1^2)$, $i = 1, \dots, n_1$ $Y_{2i} \sim N(\mu_2, \sigma_2^2)$, $i = 1, \dots, n_2$		
statische Kenngrößen	Nullhypothese H_0	Alternativhyp. H_1	Testfunktion F	Testverteilung F/H_0	Entscheidungsregel zur Ablehnung von H_0 bei gegebenem α , z.B. $\alpha = 0,05$
s_1^2, s_2^2	$\sigma_1^2 = \sigma_2^2$	$\sigma_1^2 \neq \sigma_2^2$	$F = \frac{s_1^2}{s_2^2}$	$f(n_1-1, n_2-1)$	$F_{\text{emp}} > f_{1-\alpha}(n_1-1, n_2-1)$
Levene-Test zum Vergleich von Varianzen für $r > 2$ Faktorstufen:			Voraussetzung: $Y_{gi} \sim N(\mu_g, \sigma_g^2)$, $i = 1, \dots, n_g$, $\forall g$ Für sehr schiefe Verteilungen besser Median \tilde{y}_g als \bar{y}_g .		
statische Kenngrößen	Nullhypothese H_0	Alternativhyp. H_1	Testfunktion F	Testverteilung F/H_0	Entscheidungsregel zur Ablehnung von H_0 bei gegebenem α , z.B. $\alpha = 0,05$
η_Z^2	$\sigma_1^2 = \sigma_2^2 = \dots = \sigma_r^2 = \sigma^2$	$\sigma_g^2 \neq \sigma_h^2$ für mind. ein Paar g, h	$F = \frac{\text{MQE}_Z}{\text{MQR}_Z}$ mit $z_{gi} = y_{gi} - \bar{y}_g $	$F(r-1, n-r)$	$F_{\text{emp}} > F_{1-\alpha}(r-1, n-r)$

Ergebnisseite der einfaktoriellen Varianzanalyse mit WinSTAT für Excel

In den gelben Feldern stehen die Symbole für die Formeln der Formelsammlung, nach denen WinSTAT die Zahlen berechnet.

Varianzanalyse

Meßvariable: Y
 gruppiert nach: X Aus der ANOVA-Tabelle kann $\eta^2 = \frac{SQE}{SQT}$ bestimmt werden.

	Quadrat-summe	Freiheits- grade	mittlere QS	F	P
Zwischen	SQE	$r - 1$	MQE	$F_{\text{emp}} = \frac{MQE}{MQR}$	p
Innerhalb	SQR	$n - r$	MQR		
Gesamt	SQT	$n - 1$	MQT		

Bartlett-Test zur Varianzgleichheit

Chi-Quadrat	Freiheits- grade	P
χ_{emp}^2	$r - 1$	p

Multiple Vergleiche:

Methode: Scheffé

Signifikanz (p): 0,05

Kritische Mittelwert-Differenzen zwischen Gruppenpaaren (rechts oben) und Signifikanzwertung (links unten):

	(Mittelwert)	Faktorstufe 1	Faktorstufe 2	...	Faktorstufe r
Faktorstufe 1	$\bar{y}_{1.}$	—	d_{12}	...	d_{1r}
Faktorstufe 2	$\bar{y}_{2.}$	z.B.: ja	—	...	d_{2r}
⋮	⋮	⋮	⋮	⋮	⋮
Faktorstufe r	$\bar{y}_{r.}$	z.B.: nein	z.B.: ja	...	—

Ist das Signifikanzniveau p des F -Tests der Varianzanalyse kleiner als eine vorgegebene Irrtumswahrscheinlichkeit α , so kann man davon ausgehen, dass die Mittelwerte der Grundgesamtheit nicht alle gleich sind. Dies bedeutet jedoch nicht, dass sich die Mittelwerte *aller* Faktorstufen voneinander unterscheiden. Daher werden Post-Hoc-Tests durchgeführt, die anhand von kritischen Differenzen d_{st} , $s, t = 1, \dots, r$, testen, ob die Unterschiede der Faktorstufen – jeweils paarweise zwischen 2 Faktorstufen s und t – signifikant sind. Mit dem Scheffé-Test werden die kritischen Differenzen nach der Formel $d_{st}(\alpha) = \sqrt{\text{MQR} \left(\frac{1}{n_s} + \frac{1}{n_t} \right) (r-1) f_{1-\alpha}(r-1, n-r)}$ berechnet. Für $|\bar{y}_{s.} - \bar{y}_{t.}| > d_{st}$ wird angenommen, dass auch in der Grundgesamtheit eine Differenz besteht. Die Ergebnisse des Tests werden unter „Multiple Vergleiche“ ausgewiesen. Ein „ja“ in der Tabelle weist bei einem vorgegebenen Signifikanzniveau (in WinSTAT mit p bezeichnet und z.B. $p=0,05$ festgelegt, d.h. in der Formel $d_{st}(\alpha = 0,05)$) auf einen signifikanten Unterschied der Mittelwerte zweier Faktorstufen hin.

Aufgabe

2

Für den Beispieldatensatz Seite 1 mit der für die Statistikklausur erwarteten Leistung (X_8) und den Ausgaben für Kopien (Y) erhält man den unten stehenden WinSTAT-Output einer einfaktoriellen Varianzanalyse.

- a) Berechnen und interpretieren Sie den η^2 -Koeffizienten.
- b) Formulieren Sie die Null- und Alternativhypothese des F -Tests im Sachzusammenhang. Stellen Sie den Rechengang zur Bestimmung des empirischen Wertes der F -verteilten Testfunktion dar und interpretieren Sie das zugehörige Signifikanzniveau.
- c) Interpretieren Sie das Ergebnis des Bartlett- und des Scheffé-Tests.

Varianzanalyse

	Quadratsumme	Freiheitsgrade	mittlere QS	F	P
Zwischen	2880	2	1440	40,271	0,000
Innerhalb	1180	33	35,758		
Gesamt	4060	35	116		

Bartlett-Test zur Varianzgleichheit

Chi-Quadrat	Freiheitsgrade	P
0,250	2	0,883

Multiple Vergleiche:

Methode: Scheffé

Signifikanz (p): 0,05

(Mittelwert)		1	0	-1
1	16	—	6,179	6,855
0	28	ja	—	6,179
-1	40	ja	ja	—

Aufgabe

3

Für den Beispieldatensatz Seite 1 mit der für die Statistikklausur erwarteten Leistung (X_8), dem Geschlecht (X_1) und den Ausgaben für Kopien (Y) erhält man den unten stehenden WinSTAT-Output einer 2-faktoriellen Varianzanalyse.

- a) Bestimmen Sie die arithmetischen Mittel aller Faktorstufen sowie der Faktorstufenkombinationen der beiden Faktoren (vgl. Seite 20). (Schlussfolgerungen?)
- b) Berechnen und interpretieren Sie die η^2 -Koeffizienten der Faktoren und Interaktion.
- c) Interpretieren Sie die Signifikanzniveaus der beiden Faktoren und der Interaktion.

2-fache Varianzanalyse

	Quadratsumme	Freiheitsgrade	mittlere QS	F	P
1. Faktor	2880	2	1440	89,256	0,000
2. Faktor	676	1	676	41,901	0,000
Interaktion	20	2	10	0,620	0,545
Residue	484	30	16,133		

2.2 Zweifaktorielle Varianzanalyse

Ergibt sich bei der einfaktoriellen Varianzanalyse ein kleiner Wert des η^2 -Koeffizienten, so versucht man durch Hinzunahme einer weiteren nominalen erklärenden Variablen, die einen Zusammenhang zu Y vermuten lässt, einen Erklärungsbeitrag zu finden (zweifaktorielle Varianzanalyse).

Die Stichprobe

- Y : zu erklärende metrische Variable
- A, B : Faktoren, d.h. erklärende nominale Variablen
- r : Anzahl der Faktorstufen des Faktors A
- q : Anzahl der Faktorstufen des Faktors B
- m : Anzahl der Beobachtungswerte der Faktorstufenkombination gh , $g = 1, \dots, r$, $h = 1, \dots, q$
- y_{ghi} : i -ter Beobachtungswert der Faktorstufenkombination gh , $i = 1, \dots, m$
- $y_{gh.} = \sum_i y_{ghi}$ Stichprobensumme der Faktorstufenkombination gh

Ergebnistabelle:

Faktor- stufen von A	Faktorstufen von B							$\sum_h \sum_i$
	$h = 1$	$h = 2$...	h	...	$h = q$		
$g = 1$	$i = 1$ $i = 2$ \vdots $i = m$	y_{111} y_{112} \vdots y_{11m}	y_{121} y_{122} \vdots y_{12m}	...	y_{1h1} y_{1h2} \vdots y_{1hm}	...	y_{1q1} y_{1q2} \vdots y_{1qm}	} $y_{1..} = \sum_{h=1}^q \sum_{i=1}^m y_{1hi}$
$g = 2$	$i = 1$ $i = 2$ \vdots $i = m$	y_{211} y_{212} \vdots y_{21m}	y_{221} y_{222} \vdots y_{22m}	...	y_{2h1} y_{2h2} \vdots y_{2hm}	...	y_{2q1} y_{2q2} \vdots y_{2qm}	
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
g	$i = 1$ $i = 2$ \vdots $i = m$	y_{g11} y_{g12} \vdots y_{g1m}	y_{g21} y_{g22} \vdots y_{g2m}	...	y_{gh1} y_{gh2} \vdots y_{ghm}	...	y_{gq1} y_{gq2} \vdots y_{gqm}	} $y_{g..} = \sum_{h=1}^q \sum_{i=1}^m y_{ghi}$
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	
$g = r$	$i = 1$ $i = 2$ \vdots $i = m$	y_{r11} y_{r12} \vdots y_{r1m}	y_{r21} y_{r22} \vdots y_{r2m}	...	y_{rh1} y_{rh2} \vdots y_{rhm}	...	y_{rq1} y_{rq2} \vdots y_{rqm}	} $y_{r..} = \sum_{h=1}^q \sum_{i=1}^m y_{rhi}$
$\sum_g \sum_i$		$y_{.1.} = \sum_{g=1}^r \sum_{i=1}^m y_{g1i}$	$y_{.2.} = \sum_{g=1}^r \sum_{i=1}^m y_{g2i}$...	$y_{.h.} = \sum_{g=1}^r \sum_{i=1}^m y_{ghi}$...	$y_{.q.} = \sum_{g=1}^r \sum_{i=1}^m y_{gqi}$	

Zur nicht-isolierten Analyse, ob die einzelnen Faktorstufenkombinationen gh der beiden Faktoren einen Einfluss auf die Zielvariable Y haben sowie zur isolierten Analyse, ob die Faktoren A, B und/oder das gemeinsame Auftreten der beiden Faktoren (Interaktion) einen Einfluss auf Y haben, werden zunächst die arithmetischen Mittel der Stichproben gebildet:

$$\begin{aligned} \bar{y}_{gh.} &= \frac{1}{m} \sum_{i=1}^m y_{ghi} : && \text{arithmetisches Mittel einer Faktorstufenkombination } gh \\ \bar{y}_{g..} &= \frac{1}{qm} \sum_{h=1}^q \sum_{i=1}^m y_{ghi} : && \text{arithmetisches Mittel der Faktorstufe } g \text{ des Faktors } A \\ \bar{y}_{.h.} &= \frac{1}{rm} \sum_{g=1}^r \sum_{i=1}^m y_{ghi} : && \text{arithmetisches Mittel der Faktorstufe } h \text{ des Faktors } B \\ \bar{y}_{...} &= \frac{1}{rqm} \sum_{g=1}^r \sum_{h=1}^q \sum_{i=1}^m y_{ghi} : && \text{arithmetisches Mittel aller } n = r \cdot q \cdot m \text{ Beobachtungswerte} \end{aligned}$$

Für $\bar{y}_{11.} = \dots = \bar{y}_{gh.} = \dots = \bar{y}_{rq.} = \bar{y}_{...}$ haben die einzelnen Faktorstufenkombinationen der Faktoren A und B keinen Einfluss auf Y .

Für $\bar{y}_{1..} = \dots = \bar{y}_{g..} = \dots = \bar{y}_{r..} = \bar{y}_{...}$ hat der Faktor A keinen Einfluss auf Y .

Für $\bar{y}_{.1.} = \dots = \bar{y}_{.h.} = \dots = \bar{y}_{.q.} = \bar{y}_{...}$ hat der Faktor B keinen Einfluss auf Y .

Für $\bar{y}_{gh.} - \bar{y}_{...} = (\bar{y}_{g..} - \bar{y}_{...}) + (\bar{y}_{.h.} - \bar{y}_{...})$ hat das gemeinsame Auftreten von A und B keinen Einfluss

bzw. für $\bar{y}_{gh.} = \bar{y}_{g..} + \bar{y}_{.h.} - \bar{y}_{...} \quad \forall g, h$ auf Y , d.h.: Es existiert keine Wechselwirkung (Interaktion) zwischen den Faktoren A und B .

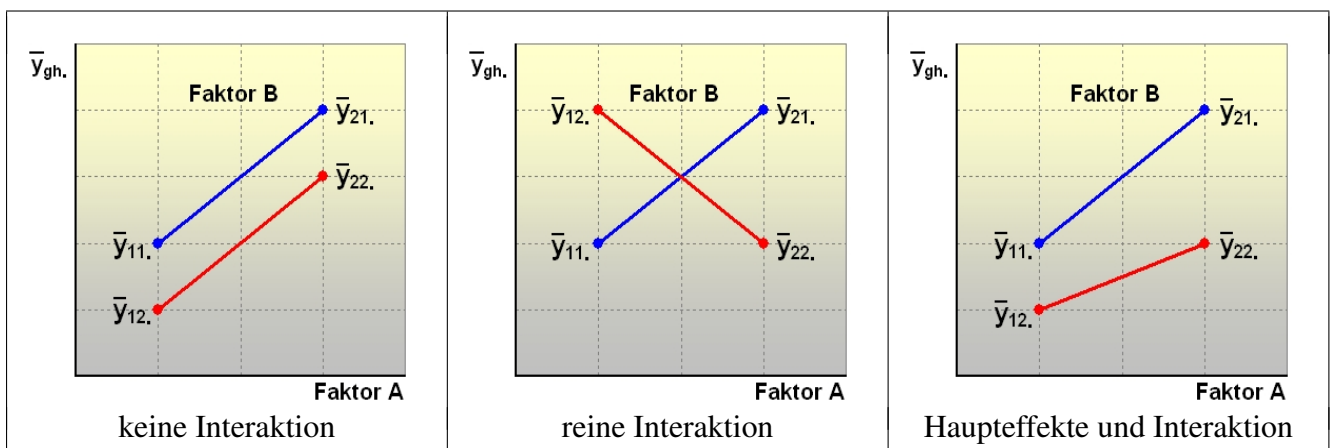
Sind die Gleichungen nicht erfüllt, so liegen entsprechende Einflüsse auf Y vor bzw. werden Erklärungsbeiträge für die Varianz von Y geliefert.

Zerlegung der Abweichung eines Beobachtungswertes y_{ghi} vom Gesamtstichprobenmittel $\bar{y}_{...}$:

$$\underbrace{y_{ghi} - \bar{y}_{...}}_{\text{zu erklärende Abweichung}} = \underbrace{(y_{ghi} - \bar{y}_{gh.})}_{\text{nicht erklärte Abweichung}} + \underbrace{(\bar{y}_{gh.} - \bar{y}_{...})}_{\text{erklärte Abweichung}}$$

Die erklärte Abweichung kann durch drei Effekte auf Y beschrieben werden:

$$\bar{y}_{gh.} - \bar{y}_{...} = \underbrace{(\bar{y}_{g..} - \bar{y}_{...})}_{\text{Einfluss von Faktor A}} + \underbrace{(\bar{y}_{.h.} - \bar{y}_{...})}_{\text{Einfluss von Faktor B}} + \underbrace{\zeta_{gh}}_{\text{Einfluss der Interaktion von A und B}}$$



Um die Stärke des Einflusses der isolierten und nicht-isolierten Betrachtung zu ermitteln, wird eine Streuungszerlegung der Abweichungsquadratsummen durchgeführt.

	$SQT = SQE_A + SQE_B + SQE_{A \times B} + SQR,$
wobei	$SQT = \sum_{g=1}^r \sum_{h=1}^q \sum_{i=1}^m (y_{ghi} - \bar{y}_{...})^2,$
SQE =	$\left\{ \begin{array}{l} SQE_A = m \cdot q \cdot \sum_{g=1}^r (\bar{y}_{g..} - \bar{y}_{...})^2, \\ + \\ SQE_B = m \cdot r \cdot \sum_{h=1}^q (\bar{y}_{.h.} - \bar{y}_{...})^2, \\ + \\ SQE_{A \times B} = m \cdot \sum_{g=1}^r \sum_{h=1}^q (\bar{y}_{gh.} - \bar{y}_{g..} - \bar{y}_{.h.} + \bar{y}_{...})^2, \end{array} \right.$
	$SQR = \sum_{g=1}^r \sum_{h=1}^q \sum_{i=1}^m (y_{ghi} - \bar{y}_{gh.})^2 = \sum_{g=1}^r \sum_{h=1}^q (m-1) s_{gh}^2.$

Die Quadratsummen der Streuungszerlegung werden in einer ANOVA-Tabelle erfasst, die außerdem die Berechnung des Wertes der Testfunktion der F -Tests beschreibt:

Streuungs- ursache	Quadrat- summe SQ	Anzahl der Freiheitsgrade ν	mittlere Quadratsumme MQ	Wert der F-verteiltern Testfunktion F_{emp}
Modell	SQE	$\nu = rq - 1$	$MQE = \frac{SQE}{rq - 1}$	$F_{\text{emp}} = \frac{MQE}{MQR}$
Faktor A	SQE_A	$\nu_A = r - 1$	$MQE_A = \frac{SQE_A}{r - 1}$	$F_{\text{emp}}^A = \frac{MQE_A}{MQR}$
Faktor B	SQE_B	$\nu_B = q - 1$	$MQE_B = \frac{SQE_B}{q - 1}$	$F_{\text{emp}}^B = \frac{MQE_B}{MQR}$
Interaktion $A \times B$	$SQE_{A \times B}$	$\nu_{A \times B} = (r-1)(q-1)$	$MQE_{A \times B} = \frac{SQE_{A \times B}}{(r-1)(q-1)}$	$F_{\text{emp}}^{A \times B} = \frac{MQE_{A \times B}}{MQR}$
Residuen	SQR	$\nu_R = rq(m-1)$	$MQR = \frac{SQR}{rq(m-1)}$	
Gesamt	SQT	$\nu_T = n - 1$		

Maßzahlen der zweifaktoriellen Varianzanalyse in der Stichprobe

Einfluss	Symbol	Berechnung	Aussage
Modell	η^2	$\eta^2 = \frac{SQE}{SQT} = 1 - \frac{SQR}{SQT}$	Anteil der durch den Einfluss der Faktoren A und B sowie der Interaktion $A \times B$ „erklärten“ Varianz der Zielvariablen Y , $0 \leq \eta^2 \leq 1$.
Faktor X , $X = A, B$	η_X^2	$\eta_X^2 = \frac{SQE_X}{SQT}$	Anteil der durch den Einfluss des Faktors X „erklärten“ Varianz der Zielvariablen Y , $0 \leq \eta_X^2 \leq 1$.
Interaktion $A \times B$	$\eta_{A \times B}^2$	$\eta_{A \times B}^2 = \frac{SQE_{A \times B}}{SQT}$	Anteil der durch den Einfluss der Wechselwirkung $A \times B$ „erklärten“ Varianz der Zielvariablen Y , $0 \leq \eta_{A \times B}^2 \leq 1$.

Schätzmodell

- Die n Beobachtungswerte y_{ghi} , $g = 1, \dots, r$, $h = 1, \dots, q$, und $i = 1, \dots, m$, stellen eine Stichprobe aus einer übergeordneten Grundgesamtheit dar.

- Die n Werte y_{ghi} können als Realisationen der n beobachtbaren Zufallsvariablen Y_{ghi} mit

$$\text{Modell (I): } Y_{ghi} = \mu_{gh} + U_{ghi}, \quad g = 1, \dots, r, \quad h = 1, \dots, q, \quad \text{und } i = 1, \dots, m,$$

$$\text{Modell (II): } Y_{ghi} = \mu + \alpha_g + \beta_h + \xi_{gh} + U_{ghi}, \quad g = 1, \dots, r, \quad h = 1, \dots, q, \quad \text{und } i = 1, \dots, m,$$

aufgefasst werden, wobei

U_{ghi} : nicht beobachtbare Zufallsvariablen mit den Realisationen $u_{ghi} = y_{ghi} - \mu_{gh}$ (Modell (I)) bzw. $u_{ghi} = y_{ghi} - \mu - \alpha_g - \beta_h - \xi_{gh}$ (Modell (II))

μ_{gh} : unbekanntes, wahres, zu schätzendes arithmetisches Mittel der Faktorstufenkombination gh in der Grundgesamtheit

$\mu = \frac{1}{r q m} \sum_{g=1}^r \sum_{h=1}^q \mu_{gh}$: unbekanntes, wahres, zu schätzendes, globales arithmetisches Mittel über alle Faktorstufenkombinationen gh der Grundgesamtheit

$\alpha_g = \mu_g - \mu$: unbekannter, wahrer, zu schätzender Effekt des Faktors A in der Faktorstufe g der Grundgesamtheit

$\beta_h = \mu_h - \mu$: unbekannter, wahrer, zu schätzender Effekt des Faktors B in der Faktorstufe h der Grundgesamtheit

$\xi_{gh} = \mu_{gh} - (\mu + \alpha_g + \beta_h)$: unbekannter, wahrer, zu schätzender Effekt der Interaktion $A \times B$ in der Faktorstufenkombination gh der Grundgesamtheit

Es gilt: $\sum_{g=1}^r \alpha_g = 0$, $\sum_{h=1}^q \beta_h = 0$, $\sum_{g=1}^r \xi_{gh} = 0$, $\sum_{h=1}^q \xi_{gh} = 0$

- Die Zufallsvariablen Y_{ghi} bzw. U_{ghi} sind in jeder Faktorstufenkombination gh , $g = 1, \dots, r$, $h = 1, \dots, q$, unabhängig normalverteilt mit

$$Y_{ghi} \sim N(\mu_{gh}, \sigma^2) \quad \text{bzw.} \quad U_{ghi} \sim N(0, \sigma^2) \quad i = 1, \dots, m, \quad \forall gh$$

Schätzer

Modell	Parameter	Schätzer	Berechnung
Modell (I)	μ_{gh}	$\bar{y}_{gh\cdot}$	vgl. Seite 16
Modell (II)	μ	\bar{y}_{\dots}	vgl. Seite 16
Modell (II)	α_g	$\bar{y}_{g\cdot\cdot} - \bar{y}_{\dots}$	vgl. Seite 16
Modell (II)	β_h	$\bar{y}_{\cdot h\cdot} - \bar{y}_{\dots}$	vgl. Seite 16
Modell (II)	ξ_{gh}	$\zeta_{gh} = \bar{y}_{gh\cdot} - \bar{y}_{g\cdot\cdot} - \bar{y}_{\cdot h\cdot} + \bar{y}_{\dots}$	vgl. Seite 16

Testverfahren

In einer nicht isolierten Betrachtung (Modell (I)) wird analysiert, ob sich in der Grundgesamtheit die Faktorstufenkombinationen unterschiedlich auf die Zielvariable auswirken. In der isolierten Betrachtung (Modell (II)) wird überprüft, ob die Faktoren A und B sowie die Interaktion $A \times B$ in der Grundgesamtheit Einfluss auf die Zielvariable Y haben oder anders ausgedrückt, ob die durch die η^2 -Koeffizienten der Stichprobe berechneten Erklärungsbeiträge für die Varianz der Zielvariablen als signifikant angesehen werden können. Die Analysen bestehen aus den folgenden F -Tests.

F-Tests der Varianzanalyse:			Voraussetzung: $Y_{ghi} \sim N(\mu_{gh}, \sigma^2), \forall gh, i = 1, \dots, m$ bzw. $U_{ghi} \sim N(0, \sigma^2), \forall gh, i = 1, \dots, m$		
Einfluss	Nullhypothese H_0	Alternativhyp. H_1	Testfunktion F	Testverteilung F/H_0	Entscheidungsregel zur Ablehnung von H_0 bei gegebenem α , z.B. $\alpha = 0,05$
Modell	$\mu_{11} = \dots = \mu_{rq}$	$\mu_{gh} \neq \mu_{st}$ für mind. ein Paar gh, st	$F = \frac{\text{MQE}}{\text{MQR}}$	$f(rq - 1, rq(m - 1))$	$F_{\text{emp}} >$ $f_{1-\alpha}(rq - 1, rq(m - 1))$
Faktor A	$\alpha_g = 0$ $\forall g = 1, \dots, r$	$\alpha_g \neq 0$ für mind. zwei α_g	$F = \frac{\text{MQE}_A}{\text{MQR}}$	$f(r - 1, rq(m - 1))$	$F_{\text{emp}}^A >$ $f_{1-\alpha}(r - 1, rq(m - 1))$
Faktor B	$\beta_h = 0$ $\forall h = 1, \dots, q$	$\beta_h \neq 0$ für mind. zwei β_h	$F = \frac{\text{MQE}_B}{\text{MQR}}$	$f(q - 1, rq(m - 1))$	$F_{\text{emp}}^B >$ $f_{1-\alpha}(q - 1, rq(m - 1))$
Interaktion $A \times B$	$\xi_{gh} = 0$ $\forall g = 1, \dots, r$ $\forall h = 1, \dots, q$	$\xi_{gh} \neq 0$ für mind. zwei Paare gh	$F = \frac{\text{MQE}_{A \times B}}{\text{MQR}}$	$f((r - 1)(q - 1),$ $rq(m - 1))$	$F_{\text{emp}}^{A \times B} >$ $f_{1-\alpha}((r - 1)(q - 1),$ $rq(m - 1))$

Ergebnisseite der zweifaktoriellen Varianzanalyse mit WinSTAT für Excel

2-fache Varianzanalyse

Meßvariable: Y
 gruppiert nach: A
 und nach: B

	Quadrat- summe	Freiheits- grade	mittlere QS	F	P
1. Faktor	SQE_A	$r - 1$	MQE_A	$F_{\text{emp}}^A = \frac{MQE_A}{MQR}$	p_A
2. Faktor	SQE_B	$q - 1$	MQE_B	$F_{\text{emp}}^B = \frac{MQE_B}{MQR}$	p_B
Interaktion	$SQE_{A \times B}$	$(r - 1)(q - 1)$	$MQE_{A \times B}$	$F_{\text{emp}}^{A \times B} = \frac{MQE}{MQR}$	$p_{A \times B}$
Residue	SQR	$rq(m - 1)$	MQR		

Die zur Ermittlung der η^2 -Koeffizienten nach den Formeln auf Seite 18 verwendete Quadratsumme SQT kann durch Addition von SQE_A , SQE_B , $SQE_{A \times B}$ und SQR bestimmt werden.

Für die Varianzanalyse einer *nicht isolierten Betrachtung* bildet man die Quadratsumme $SQE = SQE_A + SQE_B + SQE_{A \times B}$. Der zugehörige Wert der Testfunktion F_{emp} für einen F -Test (s.u.) wird dann gemäß der Formel der ANOVA-Tabelle Seite 17 (Modell) bestimmt.

Die Mittelwerte der Zielvariablen in den Faktorstufen des Faktors A bzw. B werden mit WinSTAT nicht automatisch ausgewiesen, können aber mit excel schnell ermittelt werden: Die $m \times 3$ Datenmatrix der Beobachtungswerte wird folgendermaßen eingegeben: Spalte A – Zielvariable Y , Spalte B – Faktor A , Spalte C – Faktor B . (Beachte: In den Spalten A, B und C dürfen nur die Beobachtungswerte stehen, d.h. z.B. darf in Spalte A nicht etwa auch die Summe gebildet werden, ansonsten muss die unten stehende Formel entsprechend abgeändert werden, so dass nicht die gesamte Spalte A, sondern nur die Beobachtungswerte in der Spalte A erfasst werden.) Die r Faktorstufen des Faktors A werden in die Spalte D und die q Faktorstufen des Faktors B in die Spalte E eingegeben. Unter der Annahme, dass die Zelle D2 die Faktorstufe 1 des Faktors A enthält, kann mit folgender Formel das arithmetische Mittel $\bar{y}_{1..}$ berechnet werden:

$$=\text{SUMMEWENN}(B:B;D2; \$A:\$A)/\text{ZÄHLENWENN}(B:B;D2)$$

Analoge Formeln können für $\bar{y}_{g..}$, $g = 2, \dots, r$, und $\bar{y}_{..h}$, $h = 1, \dots, q$, angewendet werden.

Ebenso können für eine nicht isolierte Betrachtung die Mittelwerte der Faktorstufenkombinationen von A und B mit excel berechnet werden, z.B.: Zeile 1 ist Kopfzeile mit den Variablennamen, die Faktorstufen werden verschlüsselt, d.h. als Zahlen, eingegeben, in Zelle D2 wird die Faktorstufe 1 des Faktors A und in Zelle E2 die Faktorstufe 1 des Faktors B eingetragen. Dann erhält man $\bar{y}_{11.}$ (für z.B. $n = 100$) mit der Formel (in einer Zeile):

$$=\text{SUMMENPRODUKT}((B2:B101=D2)*(C2:C101=E2)*(A2:A101))/$$

$$\text{SUMMENPRODUKT}((B2:B101=D2)*(C2:C101=E2))$$

3 Diskriminanzanalyse

Die Diskriminanzanalyse prüft, ob sich 2 oder mehrere Gruppen (Merkmalsausprägungen) eines nominalen Merkmals unterscheiden. Hierzu werden vermutete metrische Einflussvariable, die für die Trennbarkeit der Gruppen verantwortlich sein könnten, auf ihre trennende Wirkung untersucht.

Das Stichprobenmodell von Fisher

A, B, C, \dots :	die Gruppen des nominalen Merkmals, die auf Trennbarkeit geprüft werden sollen
g :	Index der Gruppen mit $g = 1, \dots, G$
G :	Anzahl der Gruppen
X_1, \dots, X_k :	k für die Trennbarkeit der Gruppen verantwortlich gemachte metrische Variablen
n_g :	Anzahl (Stichprobenumfang) der Beobachtungstupel der Gruppe g
$(x_{1gi}, \dots, x_{kgi})$:	Beobachtungstupel des i -ten Elements der Gruppe g , $i = 1, \dots, n_g$
Y	Diskriminanzvariable mit
$Y = b_0 + b_1X_1 + \dots + b_kX_k$:	kanonische Diskriminanzfunktion, wobei
b_0 :	die berechnete Konstante, für die gilt, dass $\bar{y} = 0$ ist
b_1, \dots, b_k :	die nach dem Diskriminanzkriterium berechneten Diskriminanzkoeffizienten mit der Normierung
$\frac{SQ_w}{n - G} = 1$:	gepoolte Varianz innerhalb der Gruppen
$\max_{b_1, \dots, b_k} \Gamma = \max_{b_1, \dots, b_k} \frac{SQ_b}{SQ_w}$:	Diskriminanzkriterium
$SQ_b = \sum_{g=1}^G n_g (\bar{y}_g - \bar{y})^2$:	Streuung zwischen (b etween) den Gruppen
$SQ_w = \sum_{g=1}^G \sum_{i=1}^{n_g} (y_{gi} - \bar{y}_g)^2$:	Streuung innerhalb (w ithin) der Gruppen
$y_{gi} = b_0 + b_1x_{1gi} + \dots + b_kx_{kgi}$:	Diskriminanzwert des i -ten Elementes der Gruppe g , $i = 1, \dots, n_g$
$\bar{y}_g = \frac{1}{n_g} \sum_{i=1}^{n_g} y_{gi}$:	Zentroid der Gruppe g
$\bar{y} = \frac{1}{n} \sum_{g=1}^G n_g \bar{y}_g$:	arithmetisches Mittel aller Diskriminanzwerte
$n = \sum_{g=1}^G n_g$:	Anzahl der Beobachtungstupel aller Gruppen und somit Anzahl aller Diskriminanzwerte

Zur Bestimmung der Diskriminanzkoeffizienten b_1, \dots, b_k geht man zunächst von einer nicht normierten Diskriminanzfunktion

$$Y = a_1 X_1 + \dots + a_k X_k$$

aus. Unter dem Diskriminanzkriterium

$$\max_{a_1, \dots, a_k} \Gamma \quad \text{mit} \quad \Gamma = \frac{\text{SQb}}{\text{SQw}} = \frac{\sum_{g=1}^G n_g (\bar{y}_g - \bar{y})^2}{\sum_{g=1}^G \sum_{i=1}^{n_g} (y_{gi} - \bar{y}_g)^2} = \frac{\mathbf{a}' \mathbf{B} \mathbf{a}}{\mathbf{a}' \mathbf{W} \mathbf{a}}$$

wobei:

\mathbf{a} : Spaltenvektor der Diskriminanzkoeffizienten (a_1, \dots, a_k)

\mathbf{B} : $(k \times k)$ -Matrix der Streuung der Merkmalsvariablen X_k zwischen den Gruppen

$b_{qr} = \sum_{g=1}^G n_g (\bar{x}_{qg} - \bar{x}_q)(\bar{x}_{rg} - \bar{x}_r)$: Element der Matrix \mathbf{B} in q -ter Zeile und r -ter Spalte, $q = 1, \dots, k, r = 1, \dots, k$

\mathbf{W} : $(k \times k)$ -Matrix der Streuung der Merkmalsvariablen X_k innerhalb der Gruppen

$w_{qr} = \sum_{g=1}^G \sum_{i=1}^{n_g} (x_{qgi} - \bar{x}_{qg})(x_{rgi} - \bar{x}_{rg})$: Element der Matrix \mathbf{W} in q -ter Zeile und r -ter Spalte, $q = 1, \dots, k, r = 1, \dots, k$

folgt das Eigenwertproblem

$$(\mathbf{W}^{-1} \mathbf{B} - \gamma \mathbf{E}) \mathbf{a} = 0$$

wobei:

\mathbf{E} : Einheitsmatrix

γ : Eigenwerte; diese erfüllen die Bedingung $\gamma = \frac{\mathbf{a}' \mathbf{B} \mathbf{a}}{\mathbf{a}' \mathbf{W} \mathbf{a}}$.

Zur Lösung des Eigenwertproblems werden maximal $t = \text{rang}(\mathbf{B}) \leq \min\{k, G - 1\}$ positive Eigenwerte $\gamma_1 \geq \dots \geq \gamma_t > 0$, ($\gamma_{t+1} = \dots = \gamma_k = 0$) aus der Gleichung

$$\det(\mathbf{W}^{-1} \mathbf{B} - \gamma \mathbf{E}) = 0$$

berechnet, der größte ausgewählt und in die Matrixgleichung $(\mathbf{W}^{-1} \mathbf{B} - \gamma \mathbf{E}) \mathbf{a} = 0$ eingesetzt. Da $\det(\mathbf{W}^{-1} \mathbf{B} - \gamma \mathbf{E}) = 0$ ist, ergibt sich für den zugehörigen Eigenvektor \mathbf{a} keine eindeutige Lösung. Durch Hinzunahme des Normierungskriteriums

$$\frac{1}{n - G} \mathbf{a}' \mathbf{W} \mathbf{a} = \frac{\text{SQw}}{n - G} = 1, \text{ d.h. die gepoolte Innergruppen-Varianz der Diskriminanzwerte ist 1,}$$

kann eine eindeutige Lösung für \mathbf{a} ermittelt werden, die dann mit \mathbf{b} bezeichnet wird. Unter der Bedingung, dass das arithmetische Mittel \bar{y} der Diskriminanzwerte der normierten Diskriminanzfunktion gerade 0 ist, kann b_0 nach der Formel

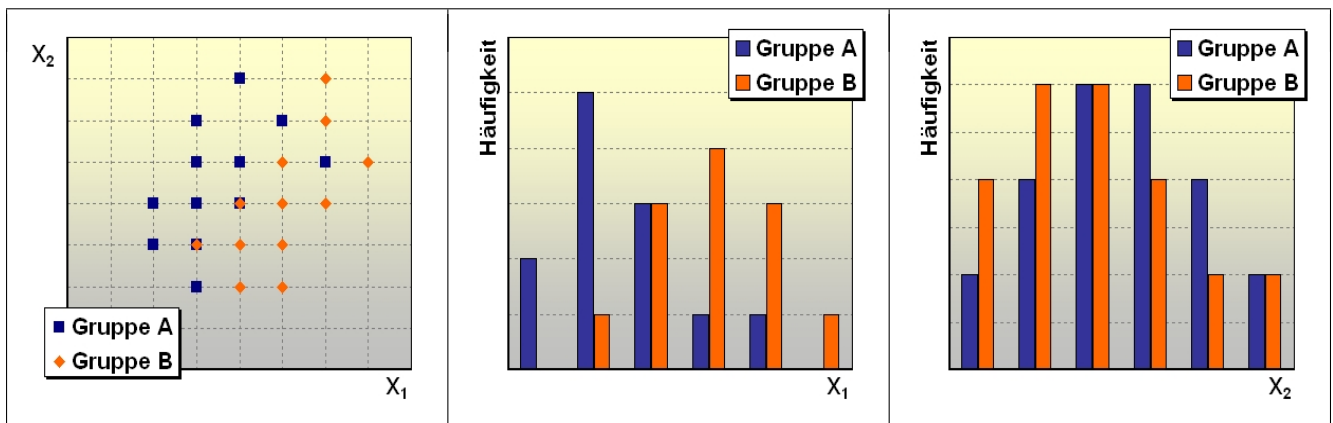
$$b_0 = - \sum_{q=1}^k b_q \bar{x}_q$$

berechnet werden. Eine so bestimmte Diskriminanzfunktion heißt kanonisch.

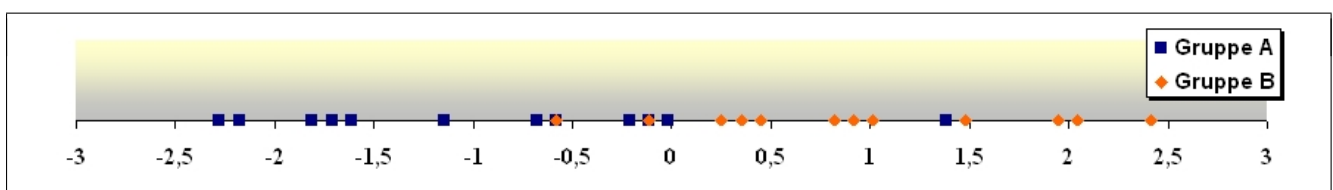
2-Gruppenfall: Wenn $\bar{y}_A = \bar{y}_B$ gelten würde, so wäre keine Trennung der beiden Gruppen möglich. Wie gut die Gruppen getrennt werden können, wenn $\bar{y}_A \neq \bar{y}_B$ gilt, hängt jedoch auch von der Streuung der Diskriminanzwerte in den Gruppen A und B ab. Die Trennbarkeit ist um so besser, je größer die Streuung der Diskriminanzwerte zwischen und je kleiner die Streuung der Diskriminanzwerte innerhalb der Gruppen ist. Daher werden die Diskriminanzkoeffizienten wie folgt bestimmt:

Diskriminanzkriterium	$\max_{b_1, \dots, b_k} \Gamma = \max_{b_1, \dots, b_k} \frac{SQb}{SQw}$, mit SQb und SQw aus der Varianzzerlegung.
Varianzzerlegung der Diskriminanzwerte	$\frac{1}{n-1} \cdot \sum_{g=1}^G \sum_{i=1}^{n_g} (y_{gi} - \bar{y})^2 = \underbrace{\frac{1}{n-1} \cdot \sum_{g=1}^G \sum_{i=1}^{n_g} (y_{gi} - \bar{y}_g)^2}_{s_Y^2} + \underbrace{\frac{1}{n-1} \cdot \sum_{g=1}^G n_g (\bar{y}_g - \bar{y})^2}_{s_{int}^2}$ $\underbrace{\frac{1}{n-1} SQT}_{\frac{1}{n-1} SQw} \quad \underbrace{\frac{1}{n-1} SQw}_{\frac{1}{n-1} SQb}$
Ergebnis des Diskriminanzkriteriums: Eigenwert γ	$\gamma = \max_{b_1, \dots, b_k} \Gamma = \max_{b_1, \dots, b_k} \frac{SQb}{SQw} = \max_{b_1, \dots, b_k} \frac{\sum_{g=1}^G n_g (\bar{y}_g - \bar{y})^2}{\sum_{g=1}^G \sum_{i=1}^{n_g} (y_{gi} - \bar{y}_g)^2}$

Im 2-Gruppenfall können folgende Grafiken einen ersten Überblick (bevor überhaupt eine Diskriminanzfunktion ermittelt wird) zur Beurteilung der Güte der Trennbarkeit der beiden Gruppen sowie für die Wichtigkeit der Variablen für die Trennbarkeit geben:



Zur Beurteilung der Güte der Trennbarkeit der Gruppen mit der Diskriminanzfunktion werden einerseits Maßzahlen gebildet und andererseits Klassifizierungstabellen erstellt, in denen die Anzahl der mit der Diskriminanzfunktion richtigen bzw. falschen Gruppe zugeordneten Elemente erfasst werden. Um einen grafischen Eindruck über die Güte der Diskriminanzfunktion zu gewinnen, können im 2-Gruppenfall die Diskriminanzwerte auf einer reellen Zahlenachse abgetragen werden:



Maßzahlen der Diskriminanzanalyse in der Stichprobe

Maßzahl	Symbol	Berechnung	Aussage
standardisierte Diskriminanzkoeffizienten	b_q^*	$b_q^* = b_q \cdot s_{X_q}$, $q = 1, \dots, k$ mit	Geschätzter Einfluss der standardisierten Variablen $X_q^* = \frac{X_q}{s_{X_q}}$ auf die Diskriminanz der Gruppen. Durch die Standardisierung der Diskriminanzkoeffizienten b_q ist ein Vergleich der diskriminatorischen Wirkung der Variablen X_1, \dots, X_k möglich. Wobei: s_{X_q} gepoolte Standardabweichung von X_q innerhalb der Gruppen; G : Anzahl der Gruppen.
mittlere Diskriminanzkoeffizienten	\bar{b}_q	$s_{X_q} = \sqrt{\frac{\sum_{g=1}^G \sum_{i=1}^{n_g} (x_{qgi} - \bar{x}_{qg})^2}{n - G}}$ $\bar{b}_q = \sum_{j=1}^t b_{qj}^* \frac{\gamma_j}{\sum_{j=1}^t \gamma_j}$ für t Diskriminanzfunktionen	
Eigenwert	γ	$\gamma = \frac{SQb}{SQw} = \frac{s_{\text{ext}}^2}{s_{\text{int}}^2}$	Die Varianz zwischen den Gruppen beträgt das γ -fache der Varianz innerhalb der Gruppen. Je größer γ ist, desto besser können die Gruppen getrennt werden.
kanonischer Korrelationskoeffizient	c	$c = \sqrt{\frac{\gamma}{1 + \gamma}}$ $= \sqrt{\frac{SQb}{SQT}} = \sqrt{\frac{s_{\text{ext}}^2}{s_Y^2}}$	c^2 gibt den Anteil der durch die Gruppenzugehörigkeit erklärten Varianz an der Gesamtvarianz der Diskriminanzwerte an. Je größer c ist, desto besser können die Gruppen getrennt werden. Es gilt: $0 \leq c \leq 1$
Wilks' Lambda	Λ	$\Lambda = \frac{1}{1 + \gamma}$ $= \frac{SQw}{SQT} = \frac{s_{\text{int}}^2}{s_Y^2}$	Λ gibt den Anteil der nicht erklärten Varianz innerhalb der Gruppen an der Gesamtvarianz der Diskriminanzwerte an. Je kleiner Λ ist, desto besser können die Gruppen getrennt werden. Es gilt: $0 \leq \Lambda \leq 1$
gepoolte Korrelationskoeffizienten	$r_{X_q Y}$	$r_{X_q Y} = \frac{1}{G} \sum_{g=1}^G r_{X_{qg} Y}$, $q = 1, \dots, k$	$r_{X_{qg} Y}$ ist der Korrelationskoeffizient von X_q und Y in der Gruppe g . Der gepoolte Korrelationskoeffizient $r_{X_q Y}$ gibt den Einfluss der Variablen X_q auf die Diskriminanz der Gruppen an. Je größer $ r_{X_q Y} $ ist, desto größer ist die diskriminatorische Wirkung der Variablen X_q . Es gilt: $-1 \leq r_{X_q Y} \leq 1$

Im Mehrgruppenfall, d.h. für $G > 2$, können sich maximal $t = \text{rang}(\mathbf{B}) \leq \min\{k, G - 1\}$ Diskriminanzfunktionen bilden lassen. Mit den obigen Maßzahlen ist es möglich, die einzelnen Diskriminanzfunktionen zu beurteilen und miteinander zu vergleichen. Möchte man aber die Trennbarkeit der Gruppen prüfen, müssen alle Diskriminanzfunktionen bzw. deren Eigenwerte gemeinsam berücksichtigt werden. Geeignete Maßzahl hierfür ist das multivariate Wilks' Lambda:

$$\Lambda = \frac{\det(\mathbf{W})}{\det(\mathbf{B} + \mathbf{W})} = \det(\mathbf{E} + \mathbf{W}^{-1}\mathbf{B}) = \prod_{j=1}^t \frac{1}{1 + \gamma_j},$$

wobei \mathbf{B} und \mathbf{W} die auf Seite 22 definierten Matrizen sind und γ_j der Eigenwert der j -ten Diskriminanzfunktion ist, $j = 1, \dots, t$.

Klassifikation mit der quadrierten euklidischen Distanz

Die **Zuordnungsvorschrift** der Klassifikation mit der quadrierten euklidischen Distanz lautet: Ein Objekt i , das den Datenvektor (x_{1i}, \dots, x_{ki}) besitzt, wird derjenigen Gruppe zugeordnet, für die der Wert $d_{ig}^2 = (y_i - \bar{y}_g)^2$ – bzw. $d_{ig}^2 = \sum_{j=1}^t (y_{ji} - \bar{y}_{jg})^2$ für t Diskriminanzfunktionen – am kleinsten ist.

Das Maximum-Likelihood-Schätzmodell für Normalverteilung

Zur Analyse, ob die in der Stichprobe ermittelten Ergebnisse der Klassifizierung der Elemente, auch auf die Grundgesamtheit übertragen werden können, werden lineare Diskriminanzfunktionen

$$Y_g \left((X_{1g}, \dots, X_{kg}), (\mu_{1g}, \dots, \mu_{kg}), \Sigma \right) = (\mu_{1g}, \dots, \mu_{kg}) \Sigma^{-1} (X_{1g}, \dots, X_{kg})' - \frac{1}{2} (\mu_{1g}, \dots, \mu_{kg}) \Sigma^{-1} (\mu_{1g}, \dots, \mu_{kg})'$$

für die Gruppen $g, g = 1, \dots, G$, in der Grundgesamtheit betrachtet. Unter den Verteilungsannahmen des Schätzmodells können die Diskriminanzfunktionen mit Kennzahlen der Stichprobe geschätzt werden.

- Die $n \cdot k$ Beobachtungswerte $x_{qgi}, q = 1, \dots, k, g = 1, \dots, G, i = 1, \dots, n_g$ und $\sum_{g=1}^G n_g = n$ stellen eine Stichprobe aus einer übergeordneten Grundgesamtheit dar.

- Für eine Gruppe g können die n_g Beobachtungstupel $(x_{1gi}, \dots, x_{kgi}), i = 1, \dots, n_g$, als Realisationen des Vektors der Zufallsvariablen $(X_{1gi}, \dots, X_{kgi})$ aufgefasst werden, $g = 1, \dots, G$.

- Der Zufallsvektor $(X_{1gi}, \dots, X_{kgi})$ einer Gruppe g ist k -dimensional normalverteilt mit gleichen Kovarianzen, d.h.

$$(X_{1gi}, \dots, X_{kgi}) \sim N_k \left((\mu_{1g}, \dots, \mu_{kg}), \Sigma \right) \quad \forall g,$$

wobei

- $(\mu_{1g}, \dots, \mu_{kg})$ der Erwartungswert(-vektor) und Σ die Kovarianzmatrix des Zufallsvektors $(X_{1gi}, \dots, X_{kgi})$ einer Gruppe g ist.

Schätzer

Parameter	Schätzer	Berechnung	Eigenschaften
μ_{qg}	\bar{x}_{qg}	$\bar{x}_{qg} = \frac{1}{n_g} \sum_{i=1}^{n_g} x_{qgi},$ $q = 1, \dots, k, g = 1, \dots, G$	Die arithmetischen Mittel der Merkmale $X_q, q = 1, \dots, k$, der Gruppe $g, g = 1, \dots, G$, sind erwartungstreue Schätzer für die Parameter μ_{qg} der Grundgesamtheit.
σ_{qr}	$\frac{w_{qr}}{n - G}$	$w_{qr} = \sum_{g=1}^G \sum_{i=1}^{n_g} (x_{qgi} - \bar{x}_{qg})(x_{rgi} - \bar{x}_{rg}),$ $q = 1, \dots, k, r = 1, \dots, k$	Die gepoolte Innergruppen-Kovarianzmatrix S der Merkmalsvariablen mit $s_{qr} = \frac{w_{qr}}{n - G}$ ist erwartungstreuer Schätzer für die Kovarianzmatrix Σ der Grundgesamtheit.

Setzt man diese Schätzer in die Diskriminanzfunktionen der Gruppen ein, so ergeben sich die geschätzten Diskriminanzfunktionen.

Klassifikation nach der Maximum-Likelihood-Regel

Die **Zuordnungsvorschrift** der Maximum-Likelihood-Regel lautet: Ein Objekt i , das den Datenvektor (x_{1i}, \dots, x_{ki}) besitzt, wird derjenigen Gruppe zugeordnet, für die der Wert der normalverteilten Dichtefunktion $f_g \left((x_{1i}, \dots, x_{ki}) | (\bar{x}_{1g}, \dots, \bar{x}_{kg}), \mathbf{S} \right)$ am größten ist, wobei $\mathbf{S} = \frac{1}{n-G} \mathbf{W}$.

Die Klassifikation nach der Maximum-Likelihood-Regel führt zu dem gleichen Ergebnis wie die Klassifikation mit der quadrierten euklidischen Distanz, d.h. die beiden Modelle sind äquivalent.

Testverfahren

- Zur Überprüfung, ob die Variablen X_q , $q = 1, \dots, k$, überhaupt zur Trennbarkeit der Gruppen in der Grundgesamtheit beitragen oder anders ausgedrückt, ob die durch Wilks' Lambda der Stichprobe berechnete diskriminatorische Wirkung der Merkmalsvariablen als signifikant angesehen werden kann, wird ein χ^2 -Test durchgeführt, der die Zentroide μ_g , $g = 1, \dots, G$, der Grundgesamtheit auf signifikante Unterschiede untersucht.
- Zur Überprüfung, ob eine einzelne Merkmalsvariable X_q , $q = 1, \dots, k$, eine signifikante diskriminatorische Wirkung für die Trennbarkeit der Gruppen liefert, kann ein χ^2 -Test einer univariaten Diskriminanzanalyse durchgeführt werden.

χ^2 -Test:			Voraussetzung: $(X_{1gi}, \dots, X_{kgi}) \sim N_k \left((\mu_{1g}, \dots, \mu_{kg}), \Sigma \right) \forall g$,		
statische Kenngröße	Nullhypothese H_0	Alternativhyp. H_1	Testfunktion (Bartlett-Approximation) χ^2	Testverteilung χ^2/H_0	Entscheidungsregel zur Ablehnung von H_0 bei gegebenem α , z.B. $\alpha = 0,01$
Wilks' Lambda Λ	$\mu_1 = \mu_2 = \dots = \mu_G$	$\mu_g \neq \mu_h$ für mind. ein Paar $g \neq h$	$\chi^2 = - \left(n - \frac{k+G}{2} - 1 \right) \ln \Lambda$	$\chi^2 (k \cdot (G-1))$	$\chi^2_{\text{emp}} > \chi^2_{1-\alpha} (k \cdot (G-1))$

Die Analyse ist nur sinnvoll, wenn der Stichprobenumfang n groß genug ist, so dass $(n - \frac{k+G}{2} - 1) > 0$ erfüllt ist. Der aus der Stichprobe berechnete χ^2_{emp} -Wert ist um so größer, je kleiner Λ ist. D.h.: Mit einem kleinen Λ besitzt die Diskriminanzfunktion sowohl in der Stichprobe als auch in der Grundgesamtheit eine größere Trennkraft.

Klassifikation mit der Mahalanobis-Distanz

Aus den Beobachtungswerten und ohne Verwendung einer Diskriminanzfunktion kann mit der Mahalanobis-Distanz:

$$d_{ig}^2 = (x_{1i} - \bar{x}_{1g}, \dots, x_{ki} - \bar{x}_{kg}) \mathbf{S}^{-1} (x_{1i} - \bar{x}_{1g}, \dots, x_{ki} - \bar{x}_{kg})', \quad \text{mit } \mathbf{S} = \frac{1}{n-G} \mathbf{W}$$

eine Klassifikation erfolgen. Die **Zuordnungsvorschrift** lautet hier: Ein Objekt i , das den Datenvektor (x_{1i}, \dots, x_{ki}) besitzt, wird derjenigen Gruppe zugeordnet, für die der Wert d_{ig}^2 am kleinsten ist. Unter der Annahme gleicher Kovarianzmatrizen der Gruppen führt die Klassifikation mit der Mahalanobis-Distanz zum gleichen Ergebnis wie die Klassifikation mit der quadrierten euklidischen Distanz (weshalb in der Notation nicht unterschieden wird).

Klassifikation mit den Fisher'schen Klassifizierungsfunktionen

Unter der Voraussetzung gleicher Streuung der Merkmalsvariablen in den Gruppen, d.h. gleicher Kovarianzmatrizen der Gruppen, kann mit den Fisher'schen Klassifizierungsfunktionen unter Berücksichtigung der Beobachtungswerte und ohne Verwendung einer Diskriminanzfunktion eine Klassifikation von Objekten i für unterschiedliche a priori-Wahrscheinlichkeiten $p(g)$ für die Gruppen $g = 1, \dots, G$, erfolgen. Die Koeffizienten $b_{0g}, b_{1g}, b_{2g}, \dots, b_{kg}$ der Fisher'schen Klassifizierungsfunktionen

$$F_g = b_{0g} + b_{1g}X_1 + b_{2g}X_2 + \dots + b_{kg}X_k, \quad g = 1, \dots, G,$$

lassen sich wie folgt berechnen:

$$b_{qg} = (n - G) \sum_{r=1}^k w_{qr}^{-1} \bar{x}_{rg}, \quad q = 1, \dots, k, \quad g = 1, \dots, G$$

$$b_{0g} = -\frac{1}{2} \sum_{q=1}^k b_{qg} \bar{x}_{qg} + \ln p(g), \quad g = 1, \dots, G.$$

Die **Zuordnungsvorschrift** lautet: Ein Objekt i , das den Datenvektor (x_{1i}, \dots, x_{ki}) besitzt, wird derjenigen Gruppe zugeordnet, für die der Wert der Fisher'schen Klassifizierungsfunktion $F_g(x_{1i}, \dots, x_{ki})$ am größten ist.

Klassifikation nach Bayes

Die Klassifikation nach Bayes erfolgt anhand eines Wahrscheinlichkeitskonzepts, mit dem auch unterschiedliche a priori-Wahrscheinlichkeiten berücksichtigt werden können. Da das Konzept auf dem Distanzkonzept aufbaut, führt es dann zu der gleichen Klassifikation wie die Klassifikation nach der quadrierten euklidischen oder Mahalanobis-Distanz, wenn keine a priori-Wahrscheinlichkeiten vorausgesetzt werden. Das Konzept kann sowohl mit den Beobachtungswerten als auch unter Verwendung der Fisher'schen Diskriminanzfunktion formuliert werden.

A priori-Wahrscheinlichkeit: $p(g)$ für die Gruppe g , $g = 1, \dots, G$.

A posteriori-Wahrscheinlichkeit: $P(g|y_i)$, d.h. die Wahrscheinlichkeit für die Zugehörigkeit von Objekt i mit dem Diskriminanzwert y_i zur Gruppe g , $g = 1, \dots, G$.

Nach dem Satz von Bayes gilt:

$$P(g|y_i) = \frac{P(y_i|g)p(g)}{\sum_{h=1}^G P(y_i|h)p(h)}.$$

Für eine stetige Verteilung der Diskriminanzwerte wird die diskrete Formulierung des Satzes von Bayes modifiziert, indem die bedingten Wahrscheinlichkeiten $P(y_i|g)$, – d.h. die Wahrscheinlichkeiten, dass sich für das Objekt i , das zur Gruppe g gehört, der Diskriminanzwert y_i ergibt, – durch die Dichten $f(y_i|g)$ ersetzt und nach der Formel

$$f(y_i|g) = \frac{1}{\sqrt{2\pi}s_g} \exp\left(-\frac{d_{ig}^2}{2s_g^2}\right)$$

berechnet werden. Unter der Annahme, dass alle Gruppen gleiche Streuung haben, d.h. (normiert) die Standardabweichung $s_g = 1$ ist für alle $g = 1, \dots, G$, vereinfacht sich die Formel entsprechend, so dass die a posteriori-Wahrscheinlichkeiten $P(g|y_i)$ nach der Formel

$$P(g|y_i) = \frac{\exp\left(-\frac{d_{ig}^2}{2}\right) p(g)}{\sum_{h=1}^G \exp\left(-\frac{d_{ih}^2}{2}\right) p(h)}$$

bestimmt werden können.

Die **Zuordnungsvorschrift** von Bayes lautet: Ein Objekt i , das den Datenvektor (x_{1i}, \dots, x_{ki}) mit dem zugehörigen Diskriminanzwert y_i besitzt, wird der für den y_i -Wert erwarteten Gruppe $\hat{g} = E(g|y_i)$ zugeordnet, für die die a posteriori-Wahrscheinlichkeit $P(\hat{g}|y_i)$ am größten ist, also der Gruppe \hat{g} , für die gilt: $P(\hat{g}|y_i) \geq P(h|y_i)$ für $h = 1, \dots, G$.

Die Bayes'schen Zuordnungsvorschrift kann auch durch Einführung einer Kostenfunktion C , welche die Kosten der Fehlklassifikation mitberücksichtigt, beschrieben werden: $C(g, \hat{g})$ seien die Kosten, wenn g die wahre Gruppe des Objekts ist und die Entscheidung \hat{g} getroffen wird. Für $g = \hat{g}$ gelte $C(g, \hat{g}) = 0$.

Beispiele für Kostenfunktionen:

- 1) Einfache symmetrische Kostenfunktion, d.h.: Bewertung aller Fehlklassifikationen mit gleichen Kosten:

$$C(g, \hat{g}) = \begin{cases} 0 & \text{für } g = \hat{g} \\ c > 0 & \text{für } g \neq \hat{g}. \end{cases}$$

- 2) Umgekehrt proportionale Kostenfunktion, d.h.: Bewertung von Fehlklassifikationen für Objekte aus Gruppen mit geringerer a priori-Wahrscheinlichkeit mit höheren Kosten:

$$C(g, \hat{g}) = \begin{cases} 0 & \text{für } g = \hat{g} \\ \frac{c}{p(g)} & \text{für } g \neq \hat{g}. \end{cases}$$

Diese Kostenfunktion wird z.B. bei der Kreditvergabe verwendet, da ein Fehler bei der Kreditvergabe eines schlechten Kredits durch Zuordnung zur Gruppe der Kreditwürdigen höhere Kosten nach sich zieht als die Opportunitätskosten einer falschen Zuordnung eines guten Kredits zur Gruppe der Kreditunwürdigen.

Wenn y_i der zugehörige Diskriminanzwert des Beobachtungstupels (x_{1i}, \dots, x_{ki}) und \hat{g} die für y_i erwartete Gruppe ist, so entstehen die bedingten erwarteten Kosten $C(\hat{g}|y_i) = \sum_{g=1}^G C(g, \hat{g}) \cdot P(g|y_i)$.

Die Bayes'sche **Zuordnungsvorschrift** lautet: Ein Objekt i , das den Datenvektor (x_{1i}, \dots, x_{ki}) mit dem zugehörigen Diskriminanzwert y_i besitzt, wird derjenigen für y_i erwarteten Gruppe \hat{g} zugeordnet, für die die bedingten Kosten $C(\hat{g}|y_i)$ mit der einfachen symmetrischen Kostenfunktion minimal werden.

Die Minimierung der bedingten Kosten mit der umgekehrt proportionalen Kostenfunktion entspricht der Maximum-Likelihood-Regel.

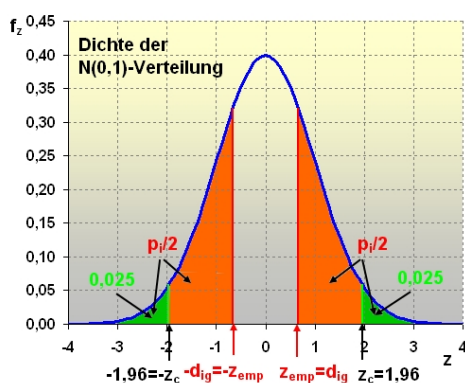
Bewertung der Klassifikation

Zur Überprüfung der Güte der Klassifikation aller Elemente $i, i = 1, \dots, n$, werden Klassifikationstabellen erstellt, aus denen der Prozentsatz der richtig klassifizierten Fälle berechnet werden kann:

tatsächliche Gruppe	Anzahl der zugeordneten Fälle zur				Anzahl der tatsächlichen Fälle
Gruppe	Gruppe 1	Gruppe 2	...	Gruppe G	Σ
Gruppe 1	n_{11}	n_{12}	...	n_{1G}	$n_1 = \sum_{g=1}^G n_{1g}$
Gruppe 2	n_{21}	n_{22}	...	n_{2G}	$n_2 = \sum_{g=1}^G n_{2g}$
⋮	⋮	⋮	⋮	⋮	⋮
Gruppe G	n_{G1}	n_{G2}	...	n_{GG}	$n_G = \sum_{g=1}^G n_{Gg}$

Wird ein Objekt i aufgrund der Zuordnungsvorschrift von Bayes einer Gruppe g zugeordnet, so ist damit nicht gesagt, dass die Klassifikation gut ist, da es nicht möglich ist, ein Element keiner Gruppe zuzuordnen. Eine Klassifikation ist sicher dann gut, wenn das Objekt in der Nähe des Zentroiden der zugeordneten Gruppe liegt, während die Güte der Klassifikation mit zunehmender Distanz zum Zentroiden abnimmt.

Test eines Objekts i auf richtige Klassifikation:		Voraussetzung: $(X_{1gi}, \dots, X_{kgi}) \sim N_k((\bar{x}_{1g}, \dots, \bar{x}_{kg}), \mathbf{S}) \quad \forall g,$			
statistische Kenngröße	Nullhypothese H_0	Alternativhyp. H_1	Testfunktion Z	Testverteilung Z/H_0	Entscheidungsregel zur Ablehnung von H_0 bei gegebenem α , z.B. $\alpha = 0,05$
d_{ig}	$d_{ig} = 0$	$d_{ig} \neq 0$	$Z = d_{ig}$	$N(0, 1)$	$ z_{emp} > z_{1-\alpha/2}$



Je größer d_{ig} ist, desto geringer ist die Wahrscheinlichkeit, dass Objekt i der Gruppe g angehört. Die zu dem aus der Stichprobe ermittelten Wert $z_{emp} = d_{ig}$ zugehörige Signifikanz $p_i \cdot 100\%$ kann als Prozentsatz der Fälle der Gruppe g , die weiter vom Zentroiden der Gruppe g entfernt sind als das Objekt i , interpretiert werden, also $p_i = P(D^2 > d_{ig}^2 | g)$. Somit weist ein großes Signifikanzniveau p_i auf ein typisches Objekt der Gruppe g hin, während ein kleines Signifikanzniveau eher auf ein untypisches Objekt der Gruppe g und daher auf eine nicht signifikante Klassifikation deutet.

Zur Berechnung des Signifikanzniveaus $p_i = P(D^2 > d_{ig}^2 | g)$ für ein Objekt i , das den Datenvektor (x_{1i}, \dots, x_{ki}) mit dem zugehörigen Diskriminanzwert y_i besitzt, berechnet man $d_{ig} = \sqrt{d_{ig}^2} = \sqrt{(y_i - \bar{y}_g)^2}$. Dann ermittelt man aus der [Tafel zur Standardnormalverteilung](#) für $z = d_{ig}$ den zugehörigen $F(z)$ -Wert. Aus $1 - p_i = 2 \cdot F(z) - 1$ folgt dann das Signifikanzniveau $p_i = 2 - 2 \cdot F(z)$.

Zur Überprüfung, ob sich jeweils zwei Gruppen signifikant unterscheiden, wird ein F -Test durchgeführt:

F-Test:			Voraussetzung: $(X_{1gi}, \dots, X_{kgi}) \sim N_k((\mu_{1g}, \dots, \mu_{kg}), \Sigma) \forall g,$		
statische Kenngröße	Nullhypothese H_0	Alternativhyp. H_1	Testfunktion F	Testverteilung F/H_0	Entscheidungsregel zur Ablehnung von H_0 bei gegebenem α , z.B. $\alpha = 0,05$
d_{gh}^2	$\mu_g = \mu_h$	$\mu_g \neq \mu_h$	$F = \frac{n_1 n_2 (n_1 + n_2 - k - 1)}{k(n_1 + n_2 - 2)(n_1 + n_2)} d_{gh}^2$	$f(k, n_1 + n_2 - k - 1)$	$F_{\text{emp}} > f_{1-\alpha}(k, n_1 + n_2 - k - 1)$
Die Mahalanobis-Distanz zwischen zwei Gruppen g und h ist: $d_{gh}^2 = (\bar{x}_{1g} - \bar{x}_{1h}, \dots, \bar{x}_{kg} - \bar{x}_{kh}) \mathbf{S}^{-1} (\bar{x}_{1g} - \bar{x}_{1h}, \dots, \bar{x}_{kg} - \bar{x}_{kh})', \text{ mit } \mathbf{S} = \frac{1}{n-G} \mathbf{W}.$					

Ergebnisseite der Diskriminanzanalyse mit WinSTAT für Excel

In den gelben Feldern stehen die Symbole für die Formeln der Formelsammlung, nach denen WinSTAT die Zahlen berechnet.

Diskriminanzanalyse

X-Variable: X_1
 X_2
 \vdots
 X_k

Y-Variable: nominales Merkmal, dessen Gruppen auf Trennbarkeit analysiert werden sollen

	Funktion	
	1	2
Eigenwert	γ_1	γ_2
Varianz Prozent	$s_1^2\% := \frac{\text{SQb}_{\text{Funktion 1}}}{\sum_q \text{SQb}_{\text{Funktion } q}} \cdot 100\%$	$s_2^2\% := \frac{\text{SQb}_{\text{Funktion 2}}}{\sum_q \text{SQb}_{\text{Funktion } q}} \cdot 100\%$
Prozent kumuliert	$s_1^2\%$	$s_1^2\% + s_2^2\%$
Kanonische Korrelation	c_1	c_2
Wilk's Lambda	Λ_1	Λ_2
Chi-Quadrat	χ_1^2	χ_2^2
Freiheitsgrade	$k \cdot (G - 1)$	$k \cdot (G - 1)$
P	p_1	p_2

Varianz Prozent gibt den Anteil der auf die q -te Funktion entfallende Streuung an der Gesamtstreuung an, $q = 1, \dots, t$, mit $t \leq \min\{k, G - 1\}$. Bei nur 2 Gruppen hat diese Maßzahl keinen Aussagegehalt, da hier die maximale Anzahl der Diskriminanzfunktionen $t \leq \min\{k, G - 1\} = 1$ ist.

Standardisierte Koeffizienten der Diskriminanzfunktionen

	Funktion		
	1	2	...
X_1	$b_{1(1)}^*$	$b_{1(2)}^*$	
X_2	$b_{2(1)}^*$	$b_{2(2)}^*$	
\vdots	\vdots	\vdots	
X_k	$b_{k(1)}^*$	$b_{k(2)}^*$	

Nicht-standardisierte Koeffizienten der Diskriminanzfunktionen

	Funktion		
	1	2	...
X_1	$b_{1(1)}$	$b_{1(2)}$	
X_2	$b_{2(1)}$	$b_{2(2)}$	
\vdots	\vdots	\vdots	
X_k	$b_{k(1)}$	$b_{k(2)}$	
(Konstante)	$b_{0(1)}$	$b_{0(2)}$	

Werte der Diskriminanzfunktionen bei Gruppenzentroiden

	Funktion		
	1	2	...
1	$\bar{y}_{1(1)}$	$\bar{y}_{1(2)}$	
2	$\bar{y}_{2(1)}$	$\bar{y}_{2(2)}$	
\vdots	\vdots	\vdots	
G	$\bar{y}_{G(1)}$	$\bar{y}_{G(2)}$	

Mahalanobis-Distanzen zwischen den Gruppen (rechts oben) und p-Werte der damit verbundenen F-Tests (links unten)

	1	2	3	...	g	...	G
1	–	d_{12}	d_{13}	...	d_{1g}	...	d_{1G}
2	p_{12}	–	d_{23}	...	d_{2g}	...	d_{2G}
3	p_{13}	p_{23}	–	...	d_{3g}	...	d_{3G}
\vdots	\vdots	\vdots	\vdots		\vdots		\vdots
g	p_{1g}	p_{2g}	p_{3g}	...	–	...	d_{gG}
\vdots	\vdots	\vdots	\vdots		\vdots		\vdots
G	p_{1G}	p_{2G}	p_{3G}	...	p_{gG}	...	–

Klassifizierungsergebnisse (priore Wahrscheinlichkeiten gleich):

	tatsächliche Anzahl	berechnet			
		1	2	...	G
1	$n_1 = \sum_{g=1}^G n_{1g}$	n_{11}	n_{12}	...	n_{1G}
2	$n_2 = \sum_{g=1}^G n_{2g}$	n_{21}	n_{22}	...	n_{2G}
\vdots	\vdots	\vdots	\vdots		\vdots
G	$n_G = \sum_{g=1}^G n_{Gg}$	n_{G1}	n_{G2}	...	n_{GG}

Mit WinSTAT ist die Klassifizierung neuer Fälle möglich, indem man im Datenblatt in *kursiver Schrift* Werte für die X-Variablen eingibt und „Kursive Zeilen neu berechnen und Y-Werte überschreiben“ aktiviert. Die zugeordnete Gruppe wird dann im Datenblatt angegeben.

$$\left(\frac{n_{11} + n_{22} + \dots + n_{GG}}{n_1 + n_2 + \dots + n_G} \right) \cdot 100\% \text{ der Fälle wurden richtig klassifiziert.}$$

Aufgabe

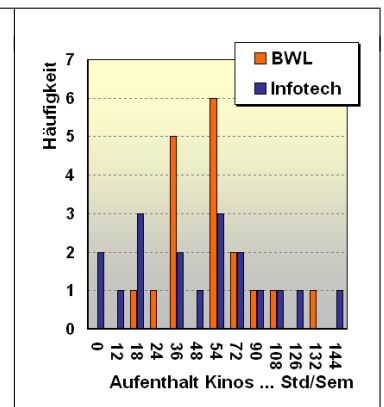
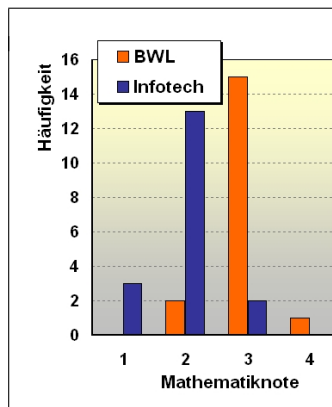
4

Für den Beispieldatensatz Seite 1 mit der Mathematiknote (X_2), der Aufenthaltsdauer in Kinos, Discos oder Kneipen (X_6) und dem Fachbereich (Y) erhält man den unten stehenden WinSTAT-Output einer Diskriminanzanalyse.

- Ergänzen Sie den Output um die fehlenden Werte der Maßzahlen c und Λ . Interpretieren Sie Ihr Ergebnis.
- Formulieren Sie die Null- und Alternativhypothese des χ^2 -Tests im Sachzusammenhang und interpretieren Sie das dem empirisch ermittelten Wert der Testfunktion zugehörige Signifikanzniveau.
- Begründen Sie mit einer geeigneten Maßzahl, welche Variable die größere diskriminatorische Wirkung hat. Beurteilen Sie die Diskriminanz auch anhand der Grafiken.
- Überprüfen Sie die Güte der Klassifikation anhand der Klassifizierungstabelle (wieviel Prozent werden richtig klassifiziert?) und dem Signifikanzniveau der Mahalanobis-Distanz zwischen den Gruppen.
- Ermitteln Sie den Fachbereich eines Studierenden i mit den Angaben: Mathematiknote 2 und Aufenthaltsdauer in Kinos, Discos, Kneipen 48 Std/Semester mit Hilfe der quadrierten euklidischen Distanz.

Diskriminanzanalyse

	Funktion 1
Eigenwert	1,153
Varianz Prozent	100
Prozent kumuliert	100
Kanonische Korrelation	
Wilk's Lambda	
Chi-Quadrat	25,300
Freiheitsgrade	2
P	0,000



Standardisierte Koeffizienten der Diskriminanzfunktionen

	Funktion 1
Mathematiknote	1,010
Zeit in Kinos ...	-0,102

Nicht-standardisierte Koeffizienten der Diskriminanzfunktionen

	Funktion 1
Mathematiknote	2,096
Zeit in Kinos ...	-0,003
(Konstante)	-4,969

Werte der Diskriminanzfunktionen bei Gruppenzentroiden

	Funktion 1
1	1,043
2	-1,043

Mahalanobis-Distanzen zwischen den Gruppen (rechts oben) und p-Werte der damit verbundenen F-Tests (links unten)

	1	2
1	–	2,087
2	0,000	–

Klassifizierungs-Ergebnisse (priore Wahrscheinlichkeiten gleich):

	tatsächliche Anzahl	berechnet	1	2
1	18	16	2	
2	18	2		16

4 Faktorenanalyse

Wurden in einer empirischen Analyse Beobachtungswerte für eine Vielzahl von Variablen erhoben, so werden einige Variablen miteinander mehr oder weniger korrelieren. Die explorative Faktorenanalyse fasst viele miteinander korrelierende Variablen zu wenigen voneinander unabhängigen, hypothetischen Variablen, sog. Faktoren, zusammen. Dabei gilt es, eine sinnvolle Interpretation der Faktoren zu finden.

Daten- und Korrelationsmatrix

X_1, \dots, X_k :

k metrische, annähernd normalverteilte Variablen

$$\mathbf{X} := \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix}$$

Matrix der Beobachtungswerte x_{iq} , des i -ten Objekts der Variablen X_q , $i = 1, \dots, n$, $q = 1, \dots, k$; es wird vorausgesetzt, dass $n > k$ ist, und die Beobachtungsvektoren der k Variablen (d.h. die Spaltenvektoren der Matrix \mathbf{X}) linear unabhängig sind, d.h. $\text{rang}(\mathbf{X}) = k$.

$$z_{iq} = \frac{x_{iq} - \bar{x}_q}{s_{x_q}}$$

standardisierter Beobachtungswert des i -ten Objekts der Variablen X_q , $i = 1, \dots, n$, $q = 1, \dots, k$, mit

$$\bar{x}_q = \frac{1}{n} \sum_{i=1}^n x_{iq}$$

arithmetisches Mittel der Variablen X_q , $q = 1, \dots, k$

$$s_{x_q} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_{iq} - \bar{x}_q)^2}$$

Standardabweichung der Variablen X_q , $q = 1, \dots, k$

$$r_{pq} := r_{X_p X_q} = r_{Z_p Z_q}$$

Korrelationskoeffizient der Variablen X_p mit der Variablen X_q bzw. der standardisierten Variablen Z_p mit Z_q , $p = 1, \dots, k$, $q = 1, \dots, k$, d.h.

$$r_{pq} := r_{X_p X_q} = \frac{\sum_{i=1}^n (x_{ip} - \bar{x}_p)(x_{iq} - \bar{x}_q)}{\sqrt{\sum_{i=1}^n (x_{ip} - \bar{x}_p)^2 \cdot \sum_{i=1}^n (x_{iq} - \bar{x}_q)^2}} = \frac{1}{n-1} \sum_{i=1}^n z_{ip} z_{iq} = r_{Z_p Z_q}$$

$$\mathbf{R} = \frac{1}{n-1} \mathbf{Z}' \mathbf{Z}$$

Matrizengleichung zur Berechnung der Korrelationskoeffizienten r_{pq} , $p = 1, \dots, k$, $q = 1, \dots, k$, wobei

$$\mathbf{R} := \begin{pmatrix} r_{11} & r_{12} & \dots & r_{1k} \\ r_{21} & r_{22} & \dots & r_{2k} \\ \vdots & \vdots & & \vdots \\ r_{k1} & r_{k2} & \dots & r_{kk} \end{pmatrix}, \quad \mathbf{Z} := \begin{pmatrix} z_{11} & z_{12} & \dots & z_{1k} \\ z_{21} & z_{22} & \dots & z_{2k} \\ \vdots & \vdots & & \vdots \\ z_{n1} & z_{n2} & \dots & z_{nk} \end{pmatrix}$$

Mit der Voraussetzung, dass die Variablen X_1, \dots, X_k annähernd normalverteilt sind, wird der Einfluss unterschiedlicher Verteilungen der Variablen auf die Korrelationskoeffizienten ausgeschlossen.

4.1 Hauptkomponentenanalyse

Ziel der Hauptkomponentenanalyse ist es, untereinander korrelierende, standardisierte Beobachtungsvariablen so zu wenigen Faktoren zusammenzufassen, dass die Faktoren einen möglichst großen Anteil der Gesamtvarianz der standardisierten Beobachtungsvariablen erklären. Dazu wird eine lineare Transformation der standardisierten Beobachtungsvariablen Z_1, \dots, Z_k zu Faktoren (sog. Hauptkomponenten) F_1, \dots, F_r vorgenommen, wobei die Faktoren untereinander unkorreliert und nach fallender Varianz geordnet sind. Wenn die ersten $r < k$ Faktoren den größten Prozentsatz der Gesamtvarianz erfassen, so können die restlichen Faktoren für die Erklärung der Streuung vernachlässigt werden.

Hauptachsentransformation

Hier werden die normierten Hauptachsen F_1, \dots, F_k eines Koordinatensystems, als Faktoren ausgewählt. Die Hauptkomponentenanalyse transformiert (dreht) das k -dimensionale Koordinatensystem, in dem die standardisierten Beobachtungstupel als k -dimensionale Punktwolke vorliegen, so, dass die erste Achse, die durch die Punktwolke geht, die Varianz der Beobachtungswerte in dieser Richtung maximiert. Die zweite senkrecht auf der ersten stehende Achse wird so bestimmt, dass die Varianz der Beobachtungswerte in dieser Richtung am zweitgrößten ist ... Für die k -dimensionalen Beobachtungstupel existieren k senkrecht aufeinanderstehende (d.h. orthogonale) Achsen.

Beispielsweise wird für $k = 2$ das (z_1, z_2) -Koordinatensystem der Variablen Z_1, Z_2 , in dem die standardisierten Beobachtungspaare (z_{i1}, z_{i2}) , $i = 1, \dots, n$, als Punktwolke um den Nullpunkt ($\bar{z}_1 = \bar{z}_2 = 0$), vorliegen, so gedreht, dass die Hauptachsen einer Ellipse mit Mittelpunkt $\mathbf{0}$ ein Koordinatensystem aufspannen. Bei einer Drehung des (z_1, z_2) -Koordinatensystems um α in ein (h_1, h_2) -Koordinatensystem können die Koordinaten eines Punktes $P = (z_{i1}, z_{i2})$ in die Koordinaten (h_{i1}, h_{i2}) des neuen Koordinatensystems mit den Formeln

$$h_{i1} = z_{i1} \cos \alpha + z_{i2} \sin \alpha \quad \text{und} \quad h_{i2} = -z_{i1} \sin \alpha + z_{i2} \cos \alpha$$

umgerechnet werden. Mit $t_{11} = \cos \alpha$, $t_{21} = \sin \alpha$, $t_{12} = -\sin \alpha$, $t_{22} = \cos \alpha$ kann die lineare Transformation der standardisierten Beobachtungswertpaare (z_{i1}, z_{i2}) , $i = 1, \dots, n$, in das durch die Hauptachsen der Ellipse aufgespannte Koordinatensystem mit

$$\begin{pmatrix} h_{i1} \\ h_{i2} \end{pmatrix} = \underbrace{\begin{pmatrix} t_{11} & t_{21} \\ t_{12} & t_{22} \end{pmatrix}}_{\text{Rotationsmatrix}} \begin{pmatrix} z_{i1} \\ z_{i2} \end{pmatrix}$$

beschrieben werden. Gesucht wird dann diejenige Ellipse, deren erste Hauptachse in Richtung der größten Streuung aller h_{i1} -Werte, $i = 1, \dots, n$, weist und deren zweite Hauptachse, die senkrecht zur ersten steht, in Richtung der größten Streuung aller h_{i2} -Werte, $i = 1, \dots, n$, weist. D.h.: Gesucht werden die Vektoren $\mathbf{t}'_1 = (t_{11}, t_{21})$, $\mathbf{t}'_2 = (t_{12}, t_{22})$ der Rotationsmatrix für die gilt:

$$\underbrace{\begin{pmatrix} h_{11} \\ h_{21} \\ \vdots \\ h_{n1} \end{pmatrix}}_{=\mathbf{H}_1} = \underbrace{\begin{pmatrix} z_{11} & z_{12} \\ z_{21} & z_{22} \\ \vdots & \vdots \\ z_{n1} & z_{n2} \end{pmatrix}}_{=\mathbf{Z}} \underbrace{\begin{pmatrix} t_{11} \\ t_{21} \end{pmatrix}}_{=\mathbf{t}'_1} \quad \text{und} \quad \underbrace{\begin{pmatrix} h_{12} \\ h_{22} \\ \vdots \\ h_{n2} \end{pmatrix}}_{=\mathbf{H}_2} = \underbrace{\begin{pmatrix} z_{11} & z_{12} \\ z_{21} & z_{22} \\ \vdots & \vdots \\ z_{n1} & z_{n2} \end{pmatrix}}_{=\mathbf{Z}} \underbrace{\begin{pmatrix} t_{12} \\ t_{22} \end{pmatrix}}_{=\mathbf{t}'_2}$$

und die Lösungen der folgenden Optimierungsprobleme darstellen:

1) Optimierungsproblem der 1. Hauptachse:

$$\begin{aligned} \max_{t_{11}, t_{21}} s_{\mathbf{H}_1}^2 & \quad \text{mit } s_{\mathbf{H}_1}^2 = \frac{1}{n-1} \sum_{i=1}^n (h_{i1} - \bar{h}_1)^2 = \frac{1}{n-1} \sum_{i=1}^n h_{i1}^2 = \frac{1}{n-1} \mathbf{H}'_1 \mathbf{H}_1 \\ \text{u.d.N.: } \|\mathbf{t}_1\| = 1 & \quad (\text{da } t_{11}^2 + t_{21}^2 = \cos^2 \alpha + \sin^2 \alpha = 1). \end{aligned}$$

Unter Berücksichtigung, dass $\|\mathbf{t}_1\| = 1$ ist, kann zu dem größten Eigenwert λ_1 des Eigenwertproblems (allg. Darstellung der Lösung Seite 36) ein zugehöriger eindeutiger Eigenvektor \mathbf{t}_1 bestimmt werden. Multiplikation der Matrix \mathbf{Z} mit dem Eigenvektor \mathbf{t}_1 liefert die Hauptachsenwerte $\mathbf{H}_1 = \mathbf{Z} \cdot \mathbf{t}_1$.

2) Optimierungsproblem der 2. Hauptachse:

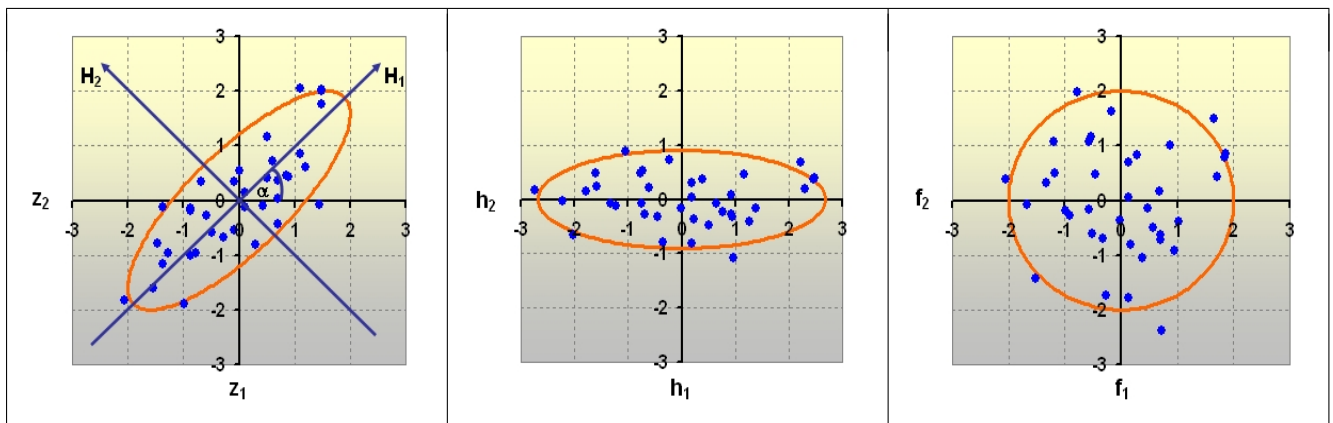
$$\begin{aligned} \max_{t_{12}, t_{22}} s_{\mathbf{H}_2}^2 & \quad \text{mit } s_{\mathbf{H}_2}^2 = \frac{1}{n-1} \mathbf{H}'_2 \mathbf{H}_2 \\ \text{u.d.N.: } \mathbf{H}'_2 \mathbf{H}_1 = 0 & \quad (\text{orthogonale Hauptachsen}) \\ \|\mathbf{t}_2\| = 1 & \quad (\text{da } t_{12}^2 + t_{22}^2 = \sin^2 \alpha + \cos^2 \alpha = 1). \end{aligned}$$

Unter Beachtung, dass die Bedingungen $\|\mathbf{t}_2\| = 1$ und $\mathbf{H}'_1 \mathbf{H}_2 = 0$ erfüllt sind, ergibt sich als Lösung für den zweitgrößten Eigenwert λ_2 des Eigenwertproblems der Eigenvektor \mathbf{t}_2 mit $t_{12} = -t_{21} = -\sin \alpha$ und $t_{22} = t_{11} = \cos \alpha$ und somit der Winkel $\alpha = 45^\circ$, $\alpha = 135^\circ$, $\alpha = 225^\circ$ oder $\alpha = 315^\circ$. Da $\|\mathbf{t}_q\| = 1$, $q = 1, 2$, und $\mathbf{t}'_1 \mathbf{t}_2 = 0$ ist, sind \mathbf{t}_1 und \mathbf{t}_2 orthonormierte Vektoren. Multiplikation der Matrix \mathbf{Z} mit dem Eigenvektor \mathbf{t}_2 liefert die Hauptachsenwerte $\mathbf{H}_2 = \mathbf{Z} \cdot \mathbf{t}_2$.

Die Längen der Hauptachsen der Ellipse sind proportional zu den Wurzeln der Eigenwerte λ_1, λ_2 . Durch Umskalieren der Hauptachsen \mathbf{H}_1 und \mathbf{H}_2 erhält man die Faktoren $\mathbf{F}_1, \mathbf{F}_2$ als orthonormierte Hauptachsen. Die Koordinaten eines Punktes (f_{i1}, f_{i2}) eines Objekts i , $i = 1, \dots, n$, in dem durch die normierten, orthogonalen Hauptachsen $\mathbf{F}_1, \mathbf{F}_2$ aufgespannten Koordinatensystem lassen sich durch

$$f_{iq} = \lambda_q^{-1/2} \cdot h_{iq} = \lambda_q^{-1/2} \cdot (z_{i1} t_{1q} + z_{i2} t_{2q}), \quad q = 1, 2$$

berechnen.



Liegen die Punkte (h_{i1}, h_{i2}) , $i = 1, \dots, n$, nahe an der ersten Hauptachse, d.h. ist der Eigenwert $\lambda_2 = s_{\mathbf{H}_2}^2$ klein, so kann auf die zweite Hauptachse als Informationsträger der Varianz der Beobachtungswerte in dieser Richtung verzichtet werden, d.h. dann können die beiden Variablen X_1 und X_2 auf einen Faktor reduziert werden.

Das formale Vorgehen zur Bestimmung der Hauptachsen für k standardisierte Beobachtungsvariablen wird auf der folgenden Seite beschrieben.

Die Hauptachsen $\mathbf{H}_1, \dots, \mathbf{H}_k \in \mathbb{R}^n$ können als lineare Transformation der standardisierten Beobachtungsvektoren $\mathbf{Z}_1, \dots, \mathbf{Z}_k \in \mathbb{R}^n$ dargestellt werden:

$$\mathbf{H}_q = (\mathbf{Z}_1, \dots, \mathbf{Z}_k) \cdot \mathbf{t}_q \quad \text{mit } \mathbf{t}_q \in \mathbb{R}^k,$$

$$\text{bzw. } \mathbf{H}_q = \mathbf{Z} \cdot \mathbf{t}_q \quad \text{mit } \mathbf{Z} \in \mathbb{R}^{n,k},$$

wobei:

$\mathbf{Z}'_q = (z_{1q}, \dots, z_{nq})$, $q = 1, \dots, k$: standardisierte Beobachtungsvektoren

$\mathbf{H}'_q = (h_{1q}, \dots, h_{nq})$, $q = 1, \dots, k$: mit $\mathbf{H}'_p \mathbf{H}_q = 0$ für $p \neq q$ (Orthogonalität der Hauptachsen)

$\mathbf{t}'_q = (t_{1q}, \dots, t_{kq})$, $q = 1, \dots, k$: Normierungsvektoren mit $\|\mathbf{t}_q\| = 1$.

Aus dem Optimierungsproblem zur Bestimmung der 1. Hauptachse \mathbf{H}_1 :

$$\max_{t_{11}, \dots, t_{k1}} s_{\mathbf{H}_1}^2 \quad \text{mit } s_{\mathbf{H}_1}^2 = \frac{1}{n-1} \mathbf{H}'_1 \mathbf{H}_1 = \frac{1}{n-1} \mathbf{t}'_1 \mathbf{Z}' \mathbf{Z} \mathbf{t}_1 = \mathbf{t}'_1 \mathbf{R} \mathbf{t}_1$$

$$\text{u.d.N.: } \mathbf{t}'_1 \mathbf{t}_1 = 1 \quad (\text{Normierung})$$

folgt das Eigenwertproblem

$$(\mathbf{R} - \lambda \mathbf{E}) \mathbf{t}_1 = 0,$$

wobei:

\mathbf{E} : Einheitsmatrix

λ : Eigenwerte; diese erfüllen die Bedingung $\lambda = s_{\mathbf{H}_1}^2$.

Zur Lösung des Eigenwertproblems werden maximal $k = \text{rang}(\mathbf{R}) = \text{rang}(\mathbf{X})$ positive Eigenwerte $\lambda_1 \geq \dots \geq \lambda_k > 0$, aus der Gleichung

$$\det(\mathbf{R} - \lambda \mathbf{E}) = 0$$

berechnet, der größte ausgewählt, d.h. $\lambda_1 = s_{\mathbf{H}_1}^2$, und in die Matrizengleichung $(\mathbf{R} - \lambda \mathbf{E}) \mathbf{t}_1 = 0$ eingesetzt. Unter Beachtung der Bedingung $\|\mathbf{t}_1\| = 1$ ergibt sich der zugehörige Eigenvektor \mathbf{t}_1 .

Aus dem Optimierungsproblem zur Bestimmung der 2. Hauptachse \mathbf{H}_2 :

$$\max_{t_{12}, \dots, t_{k2}} s_{\mathbf{H}_2}^2 \quad \text{mit } s_{\mathbf{H}_2}^2 = \frac{1}{n-1} \mathbf{H}'_2 \mathbf{H}_2 = \frac{1}{n-1} \mathbf{t}'_2 \mathbf{Z}' \mathbf{Z} \mathbf{t}_2 = \mathbf{t}'_2 \mathbf{R} \mathbf{t}_2$$

$$\text{u.d.N.: } \mathbf{H}'_2 \mathbf{H}_1 = 0 \quad (\text{Orthogonalität der Hauptachsen})$$

$$\mathbf{t}'_2 \mathbf{t}_2 = 1 \quad (\text{Normierung})$$

folgt das Eigenwertproblem $(\mathbf{R} - \lambda \mathbf{E}) \mathbf{t}_2 = 0$. Zur Lösung des Eigenwertproblems wird aus den mit der Gleichung $\det(\mathbf{R} - \lambda \mathbf{E}) = 0$ berechneten Eigenwerten der zweitgrößte ausgewählt, d.h. $\lambda_2 = s_{\mathbf{H}_2}^2$, in die Matrizengleichung $(\mathbf{R} - \lambda \mathbf{E}) \mathbf{t}_2 = 0$ eingesetzt und unter Berücksichtigung von $\|\mathbf{t}_2\| = 1$ hieraus der Eigenvektor \mathbf{t}_2 bestimmt.

Analog werden die restlichen Eigenvektoren ermittelt und die Hauptachsen $\mathbf{H}_q = \mathbf{Z} \cdot \mathbf{t}_q$ bestimmt, d.h.: $\mathbf{H} = (\mathbf{H}_1, \dots, \mathbf{H}_k) = \mathbf{Z} \mathbf{T}$ mit $\mathbf{H} \in \mathbb{R}^{n,k}$, $\mathbf{T} = (\mathbf{t}_1, \dots, \mathbf{t}_k) \in \mathbb{R}^{k,k}$. Danach werden die Hauptachsen in orthonormierte Faktoren, sog. Hauptkomponenten, $\mathbf{F}_q \in \mathbb{R}^n$, $q = 1, \dots, k$, umskaliert, so dass gilt:

$$\mathbf{Z} = \mathbf{F} \mathbf{A}' \quad \text{mit } \mathbf{F} = \mathbf{H} \mathbf{L}^{-1/2}, \quad \mathbf{A} = \mathbf{T} \mathbf{L}^{1/2} \quad \text{und } \mathbf{R} = \mathbf{T} \mathbf{L} \mathbf{T}' \implies \mathbf{F} = \mathbf{Z} \mathbf{A} (\mathbf{A}' \mathbf{A})^{-1} \quad \text{und } \mathbf{R} = \mathbf{A} \mathbf{A}', \quad \text{wobei}$$

$$\mathbf{F} = (\mathbf{F}_1, \dots, \mathbf{F}_k) \in \mathbb{R}^{n,k}, \quad \mathbf{L} = \text{diag}\{\lambda_1, \dots, \lambda_k\} \in \mathbb{R}^{k,k}, \quad \frac{1}{n-1} \mathbf{F}' \mathbf{F} = \mathbf{E}, \quad \text{d.h. } s_{\mathbf{F}_j}^2 = 1, \quad j = 1, \dots, k.$$

Hauptkomponentenmethode

Die Hauptkomponentenmethode besteht darin, die ersten r Hauptkomponenten F_1, \dots, F_r , die sich bei einer Hauptachsentransformation der standardisierten Daten Z anhand der Korrelationsmatrix R ergeben, als Faktoren auszuwählen unter Vernachlässigung der restlichen $k - r$ Faktoren.

Anzahl der Faktoren gleich der Anzahl der Variablen, d.h. $r = k$:	Anzahl der Faktoren kleiner als die Anzahl der Variablen, d.h. $r < k$:
Datenmatrix $Z = FA'$	Zerlegung der Datenmatrix in $Z = FA' + U$
Vollständige Erklärung der Gesamtvarianz der k standardisierten Beobachtungsvariablen: $s_{Z_q}^2 = \sum_{j=1}^k a_{qj}^2 = 1, \quad q = 1, \dots, k$	Zerlegung der zu erklärenden Gesamtvarianz der k standardisierten Beobachtungsvariablen in eine durch die r Faktoren erklärte und eine nicht erklärte Varianz: (vgl. Seite 38) $s_{Z_q}^2 = \sum_{j=1}^r a_{qj}^2 + s_{qe}^2 = 1, \quad q = 1, \dots, k.$
Korrelationsmatrix: $R = AA'$ Reproduzierte Korrelationsmatrix: $\hat{R} = AA' = R$	Korrelationsmatrix: $R = AA' + \frac{1}{n-1}U'U$ Reproduzierte Korrelationsmatrix: $\hat{R} = AA'$
<p>mit $A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1r} \\ a_{21} & a_{22} & \dots & a_{2r} \\ \vdots & \vdots & & \vdots \\ a_{k1} & a_{k2} & \dots & a_{kr} \end{pmatrix}$, $F = (F_1, \dots, F_r) = \begin{pmatrix} f_{11} & f_{12} & \dots & f_{1r} \\ f_{21} & f_{22} & \dots & f_{2r} \\ \vdots & \vdots & & \vdots \\ f_{n1} & f_{n2} & \dots & f_{nr} \end{pmatrix}$, $1 \leq r \leq k$,</p> <p>$U = \begin{pmatrix} u_{11} & u_{12} & \dots & u_{1k} \\ u_{21} & u_{22} & \dots & u_{2k} \\ \vdots & \vdots & & \vdots \\ u_{n1} & u_{n2} & \dots & u_{nk} \end{pmatrix}$</p> <p>Die Vektoren F_1, \dots, F_r der Matrix F sind die orthonormierten Hauptachsen. Die Bestimmung der Matrizen A und F ist für $r = k$ auf Seite 36 beschrieben. Für $r < k$ ergeben sich die entsprechenden um $k - r$ Spalten reduzierten Matrizen.</p>	
<p>Die Faktorwerte f_{ij} des Faktors j des Objekts i, $i = 1, \dots, n$, $j = 1, \dots, r$, können mit der Regressionsanalyse geschätzt oder durch</p> $F = ZA(A'A)^{-1}$ <p>berechnet werden.</p>	

Ziel der Hauptkomponentenanalyse ist es, eine Zerlegung $Z = FA' + U$ zu finden, so dass der Rest U bzw. der Anteil der nicht erklärten Varianz s_{qe}^2 der standardisierten Beobachtungsvariablen Z_q , $q = 1, \dots, k$, „möglichst klein“ wird. Approximativ entspricht die Faktorisierung der standardisierten Beobachtungswerte einer Regressionsanalyse für k lineare Regressionfunktionen $\hat{Z}_q = a_{q1}F_1 + a_{q2}F_2 + \dots + a_{qr}F_r$, wobei $\min_{a_{q1}, \dots, a_{qr}} \sum_{i=1}^n (z_{iq} - \hat{z}_{iq})^2$ mit $\hat{z}_{iq} = a_{q1}f_{i1} + a_{q2}f_{i2} + \dots + a_{qr}f_{ir}$, $i = 1, \dots, n$, $q = 1, \dots, k$.

Erläuterung zur Berechnung der durch die r Faktoren erklärten Varianz einer standardisierten Variablen Z_q , der sog. Kommunalität der Variablen Z_q : $h_q^2 = \sum_{j=1}^r a_{qj}^2$, $q = 1, \dots, k$:

$$\hat{z}_{iq} = a_{q1}f_{i1} + a_{q2}f_{i2} + \dots + a_{qr}f_{ir}, \quad \text{wobei für } r = k \text{ gilt: } \hat{z}_{iq} = z_{iq}$$

⇒ Varianzzerlegung für Z_q :

$$\begin{aligned} s_{Z_q}^2 &= \frac{1}{n-1} \sum_{i=1}^n (z_{iq} - \bar{z}_q)^2 = \frac{1}{n-1} \sum_{i=1}^n (z_{iq} - \hat{z}_{iq})^2 + \frac{1}{n-1} \sum_{i=1}^n \hat{z}_{iq}^2 = 1 \\ &= \underbrace{\frac{1}{n-1} \sum_{i=1}^n (z_{iq} - \hat{z}_{iq})^2}_{\text{nicht erklärte Varianz } s_{qe}^2} + \underbrace{\frac{1}{n-1} \sum_{i=1}^n (a_{q1}f_{i1} + a_{q2}f_{i2} + \dots + a_{qr}f_{ir})^2}_{\text{durch die Faktoren } F_1, \dots, F_r \text{ erklärte Varianz } h_q^2}, \quad \text{wobei für } r = k \text{ gilt: } s_{qe}^2 = 0 \end{aligned}$$

$$\begin{aligned} \Rightarrow h_q^2 &= \frac{1}{n-1} (a_{q1}f_{11} + a_{q2}f_{12} + \dots + a_{qr}f_{1r})^2 \\ &\quad + \frac{1}{n-1} (a_{q1}f_{21} + a_{q2}f_{22} + \dots + a_{qr}f_{2r})^2 \\ &\quad \vdots \\ &\quad + \frac{1}{n-1} (a_{q1}f_{n1} + a_{q2}f_{n2} + \dots + a_{qr}f_{nr})^2 \\ &= \frac{1}{n-1} \left[\left(\sum_{j=1}^r a_{qj}f_{1j} \right)^2 + \left(\sum_{j=1}^r a_{qj}f_{2j} \right)^2 + \dots + \left(\sum_{j=1}^r a_{qj}f_{nj} \right)^2 \right] \\ &= \frac{1}{n-1} \left(\sum_{j=1}^r a_{qj}f_{1j}, \sum_{j=1}^r a_{qj}f_{2j}, \dots, \sum_{j=1}^r a_{qj}f_{nj} \right) \begin{pmatrix} \sum_{j=1}^r a_{qj}f_{1j} \\ \sum_{j=1}^r a_{qj}f_{2j} \\ \vdots \\ \sum_{j=1}^r a_{qj}f_{nj} \end{pmatrix} \\ &= \frac{1}{n-1} (a_{q1}, a_{q2}, \dots, a_{qr}) \underbrace{\begin{pmatrix} f_{11} & f_{21} & \dots & f_{n1} \\ f_{12} & f_{22} & \dots & f_{n2} \\ \vdots & \vdots & & \vdots \\ f_{1r} & f_{2r} & \dots & f_{nr} \end{pmatrix}}_{= F'} \underbrace{\begin{pmatrix} f_{11} & f_{12} & \dots & f_{1r} \\ f_{21} & f_{22} & \dots & f_{2r} \\ \vdots & \vdots & & \vdots \\ f_{n1} & f_{n2} & \dots & f_{nr} \end{pmatrix}}_{= F} \begin{pmatrix} a_{q1} \\ a_{q2} \\ \vdots \\ a_{qr} \end{pmatrix} \\ &= \sum_{j=1}^r a_{qj}^2, \quad \text{da } \frac{1}{n-1} F'F = E. \end{aligned}$$

Maßzahlen der Hauptkomponentenanalyse

Maßzahl	Symbol	Berechnung	Aussage
Faktorladung	a_{qj}	$a_{qj} = t_{qj} \sqrt{\lambda_j}$, vgl. Seite 36: $\mathbf{A} = \mathbf{TL}^{1/2}$	Einfluss (Korrelationskoeffizient) des j -ten Faktors auf die standardisierte Variable Z_q . Es gilt: $-1 \leq a_{qj} \leq 1$, $j = 1, \dots, r$, $q = 1, \dots, k$. Je größer $ a_{qj} $ ist, desto größer ist der Einfluss. Der Zusammenhang zwischen einer Variablen Z_q und den Faktoren F_1, \dots, F_r kann durch $Z_q = a_{q1}F_1 + a_{q2}F_2 + \dots + a_{qr}F_r, \quad q = 1, \dots, k,$ dargestellt werden.
Varianz der standardisierten Variablen	$s_{Z_q}^2$	$s_{Z_q}^2 = \sum_{j=1}^r a_{qj}^2 + s_{qe}^2 = 1$	Die zu erklärende Gesamtvarianz $s_{Z_q}^2$ einer standardisierten Variablen Z_q , $q = 1, \dots, k$, lässt sich zerlegen in eine durch alle r Faktoren gemeinsam erklärte und eine nicht erklärte Varianz.
Kommunalität	h_q^2	$h_q^2 = \sum_{j=1}^r a_{qj}^2$	Die Kommunalität h_q^2 gibt den Anteil der durch alle r Faktoren gemeinsam erklärten Varianz der standardisierten Variablen Z_q an. Es gilt: $0 \leq h_q^2 \leq s_{Z_q}^2 = 1$. Für $h_q^2 = 1$ wird die Varianz von Z_q vollständig durch die r Faktoren erklärt. Dies ist dann der Fall, wenn $r = k$ ist.
Eigenwert	λ_j	$\lambda_j = \sum_{q=1}^k a_{qj}^2$, $\lambda_1 \geq \dots \geq \lambda_r$, $r = k \implies \sum_{j=1}^r \lambda_j = k$	Der Eigenwert λ_j (bzw. $\frac{\lambda_j}{k}$) gibt die (den Anteil der) durch den j -ten Faktor erklärte(n) Varianz aller k standardisierten Variablen an. Es gilt: $0 < \lambda_j \leq k$, $j = 1, \dots, r$. Die zu erklärende Gesamtvarianz aller k standardisierten Variablen beträgt $\sum_{q=1}^k s_{Z_q}^2 = k$.
reproduzierte Korrelationskoeffizienten	\hat{r}_{pq}	$\hat{r}_{pq} = \sum_{j=1}^r a_{pj}a_{qj}$, $p \neq q$	\hat{r}_{pq} misst die geschätzte Stärke des linearen Zusammenhangs zwischen der p -ten und der q -ten standardisierten Variablen. Es gilt: $-1 \leq \hat{r}_{pq} \leq 1$, $p, q = 1, \dots, k$. Für $r = k$ ist $\hat{r}_{pq} = r_{pq}$.
Faktorwert	f_{ij}	$f_{ij} = \sum_{q=1}^k \alpha_{qj}z_{iq}$, mit $(\alpha_{qj})_{k,k} = \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}$	Erklärung des j -ten Faktors durch alle k Variablen, $j = 1, \dots, r$, $i = 1, \dots, n$. Der Zusammenhang zwischen einem Faktor F_j und den Variablen Z_1, \dots, Z_k kann durch $F_j = \alpha_{1j}Z_1 + \alpha_{2j}Z_2 + \dots + \alpha_{kj}Z_k, \quad j = 1, \dots, r$ dargestellt werden. Ziel der Faktorenanalyse ist es, dass ein Faktor nur von einem Teil der Variablen erklärt wird, d.h. für einige q die Koeffizienten α_{qj} klein sind.

Erläuterung zur Berechnung der durch den Faktor F_j erklärten Varianz aller standardisierten Variablen, des sog. Eigenwerts: $\lambda_j = \sum_{q=1}^k a_{qj}^2$, $j = 1, \dots, r$:

$$\begin{aligned}
 z_{iq(j)} &= a_{qj} f_{ij} \implies \\
 \lambda_j &= \sum_{q=1}^k s_{Z_{q(j)}}^2 = \underbrace{\frac{1}{n-1} \sum_{i=1}^n (a_{1j} f_{ij})^2}_{\text{durch den Faktor } F_j \text{ erklärte Varianz von } Z_1} + \underbrace{\frac{1}{n-1} \sum_{i=1}^n (a_{2j} f_{ij})^2}_{\text{durch den Faktor } F_j \text{ erklärte Varianz von } Z_2} + \dots + \underbrace{\frac{1}{n-1} \sum_{i=1}^n (a_{kj} f_{ij})^2}_{\text{durch den Faktor } F_j \text{ erklärte Varianz von } Z_k} \\
 &= \frac{1}{n-1} \left[(a_{1j} f_{1j})^2 + (a_{2j} f_{1j})^2 + \dots + (a_{kj} f_{1j})^2 \right] \\
 &\quad + \frac{1}{n-1} \left[(a_{1j} f_{2j})^2 + (a_{2j} f_{2j})^2 + \dots + (a_{kj} f_{2j})^2 \right] \\
 &\quad \vdots \\
 &\quad + \frac{1}{n-1} \left[(a_{1j} f_{nj})^2 + (a_{2j} f_{nj})^2 + \dots + (a_{kj} f_{nj})^2 \right] \\
 &= \frac{1}{n-1} \left[\sum_{q=1}^k (a_{qj} f_{1j})^2 + \sum_{q=1}^k (a_{qj} f_{2j})^2 + \dots + \sum_{q=1}^k (a_{qj} f_{nj})^2 \right] \\
 &= \frac{1}{n-1} \left[f_{1j}^2 \sum_{q=1}^k a_{qj}^2 + f_{2j}^2 \sum_{q=1}^k a_{qj}^2 + \dots + f_{nj}^2 \sum_{q=1}^k a_{qj}^2 \right] \\
 &= \left(\sum_{q=1}^k a_{qj}^2 \right) \underbrace{\frac{1}{n-1} (f_{1j}^2 + f_{2j}^2 + \dots + f_{nj}^2)}_{= 1, \text{ da } \frac{1}{n-1} F'F = E} \\
 &= \sum_{q=1}^k a_{qj}^2.
 \end{aligned}$$

Außerdem vgl. Seite 36:

$$\begin{aligned}
 \mathbf{A} = \mathbf{T}\mathbf{L}^{1/2} \implies \mathbf{L}^{1/2} = \mathbf{T}^{-1}\mathbf{A} \implies \mathbf{L} &= (\mathbf{T}^{-1}\mathbf{A})' \mathbf{T}^{-1}\mathbf{A} = \mathbf{A}' \underbrace{(\mathbf{T}^{-1})' \mathbf{T}^{-1}}_{=\mathbf{E}} \mathbf{A} = \mathbf{A}'\mathbf{A} \implies \\
 \lambda_j &= (a_{1j}, a_{2j}, \dots, a_{kj}) \begin{pmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{kj} \end{pmatrix} = \sum_{q=1}^k a_{qj}^2, \quad j = 1, \dots, r.
 \end{aligned}$$

Ziel der Hauptkomponentenanalyse ist es, dass r möglichst klein und gleichzeitig $\frac{\sum_{j=1}^r \lambda_j}{k}$ möglichst groß ist, wobei die Faktoren nach der Größe von λ_j geordnet sind, d.h. $\lambda_1 \geq \dots \geq \lambda_r$.

Ergebnisseite der Hauptkomponentenanalyse mit WinSTAT für Excel

Faktorenanalyse

Gültige Fälle: n

Kommunalitäten

	Analyse	
	geschätzt	1
X_1	1	h_1^2
X_2	1	h_2^2
\vdots	\vdots	\vdots
X_k	1	h_k^2

In den gelben Feldern stehen die Symbole für die Formeln der Formelsammlung, nach denen WinSTAT die Zahlen berechnet. Um mit WinSTAT die Ergebnisse der Hauptkomponentenanalyse zu erhalten, muss unter der Faktorenanalyse die Einstellung „geschätzte Kommunalitäten: 1,0“ gewählt werden, womit vorausgesetzt wird, dass die Varianzen $s_{Z_q}^2$, $q = 1, \dots, k$, vollständig durch die gemeinsamen Faktoren erklärt werden können. Zur Berechnung ohne Iteration muss „Analyse der Kommunalitäten wiederholen“ deaktiviert werden.

Eigenwerte

Faktor	Eigenwert	Varianz Prozent	Prozent kumuliert
1	λ_1	$\frac{\lambda_1}{k} \cdot 100$	$\frac{\lambda_1}{k} \cdot 100$
2	λ_2	$\frac{\lambda_2}{k} \cdot 100$	$\frac{1}{k}(\lambda_1 + \lambda_2) \cdot 100$
\vdots	\vdots	\vdots	\vdots
k	λ_k	$\frac{\lambda_k}{k} \cdot 100$	$\frac{1}{k} \sum_{j=1}^k \lambda_j \cdot 100$

In dieser Tabelle werden die Eigenwerte stets als Quadratsummen der Faktorladungen vor einer Rotation ausgewiesen. Im Falle einer Rotation (vgl. Seite 45) unterscheiden sich die Eigenwerte von der Quadratsumme der rotierten Faktorladungen. Die Summe der Eigenwerte der extrahierten Faktoren stimmt aber mit der Summe der Quadratsummen der rotierten Faktorladungen überein. Die folgende Tabelle gilt für den Fall unrotierter Faktorladungen, aus deren Quadratsummen sich die Eigenwerte berechnen lassen.

Unrotierte Faktorladungen Hierbei werden die X_q , $q = 1, \dots, k$, nach der Größe der Faktorladungen des ersten Faktors geordnet.

	Faktor 1	Faktor 2	...	Faktor r	Kommunalität
Z.B.: X_1	a_{11}	a_{12}	...	a_{1r}	$h_1^2 = \sum_{j=1}^r a_{1j}^2$
X_3	a_{31}	a_{32}	...	a_{3r}	$h_3^2 = \sum_{j=1}^r a_{3j}^2$
X_2	a_{21}	a_{22}	...	a_{2r}	$h_2^2 = \sum_{j=1}^r a_{2j}^2$
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
X_k	a_{k1}	a_{k2}	...	a_{kr}	$h_k^2 = \sum_{j=1}^r a_{kj}^2$
Quadratsumme	$\lambda_1 = \sum_{q=1}^k a_{q1}^2$	$\lambda_2 = \sum_{q=1}^k a_{q2}^2$...	$\lambda_r = \sum_{q=1}^k a_{qr}^2$	$\sum_{j=1}^r \lambda_j$
Prozent der Varianz	$\frac{\lambda_1}{k} \cdot 100$	$\frac{\lambda_2}{k} \cdot 100$...	$\frac{\lambda_r}{k} \cdot 100$	$\frac{1}{k} \sum_{j=1}^r \lambda_j \cdot 100$

4.2 Hauptachsenanalyse

Im Unterschied zur Hauptkomponentenanalyse geht man hier davon aus, dass ein Teil der Varianz $s_{Z_q}^2$, $q = 1, \dots, k$, nicht durch gemeinsame Faktoren erklärt werden kann. Ziel der Hauptachsenanalyse ist es, lediglich die Varianz in Höhe der Kommunalitäten zu erklären.

Kommunalitätenschätzung

Zunächst stellt sich die Frage, wie groß der Anteil der durch die gemeinsamen Faktoren erklärten Varianz h_q^2 an der zu erklärenden Varianz $s_{Z_q}^2$, $q = 1, \dots, k$, sein soll. Ein mögliches Kriterium für einen Schätzwert der Kommunalitäten, das auch WinSTAT verwendet, ist die „höchste Korrelation“, d.h.: die Kommunalität der Variablen Z_q wird so geschätzt, dass $\hat{h}_q^2 = \max_p |r_{pq}|$, $p \neq q$, $p = 1, \dots, k$, gilt. Damit liefern die gemeinsamen Faktoren den gleichen Erklärungsbeitrag für die zu erklärende Varianz $s_{Z_q}^2$ wie die höchste Korrelation der Variablen Z_q mit den restlichen Variablen Z_p , $p \neq q$, $p = 1, \dots, k$, ausmacht.

Hauptachsentransformation

Die Hauptachsenanalyse transformiert das k -dimensionale Koordinatensystem der standardisierten Beobachtungsvariablen so, dass die erste Achse, die durch die Punktwolke geht, die Varianz in Höhe der geschätzten Kommunalität der Beobachtungswerte in dieser Richtung maximiert. Die zweite senkrecht auf der ersten stehende Achse wird so bestimmt, dass die Varianz in Höhe der geschätzten Kommunalität in dieser Richtung am zweitgrößten ist ...

Analog zu Seite 36 werden die Hauptachsen $\mathbf{H}_q = \mathbf{Z}\mathbf{t}_q$, $q = 1, \dots, k$, bestimmt:

Aus dem Optimierungsproblem zur Bestimmung der 1. Hauptachse \mathbf{H}_1 :

$$\max_{t_{11}, \dots, t_{k1}} \hat{s}_{\mathbf{H}_1}^2 \quad \text{mit } \hat{s}_{\mathbf{H}_1}^2 = \hat{h}_{\mathbf{H}_1}^2 = \frac{1}{n-1} \mathbf{H}'_1 \mathbf{H}_1 - \mathbf{V} = \mathbf{t}'_1 \left(\frac{1}{n-1} \mathbf{Z}'\mathbf{Z} - \mathbf{V} \right) \mathbf{t}_1 = \mathbf{t}'_1 \mathbf{R}_h \mathbf{t}_1$$

und der reduzierten Korrelationsmatrix $\mathbf{R}_h = \mathbf{R} - \mathbf{V}$, wobei

$$\mathbf{V} = \begin{pmatrix} v_1^2 & 0 & \dots & 0 \\ 0 & v_2^2 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & v_k^2 \end{pmatrix} = \begin{pmatrix} s_{1e}^2 & 0 & \dots & 0 \\ 0 & s_{2e}^2 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & s_{ke}^2 \end{pmatrix},$$

$s_{qe}^2 = 1 - \hat{h}_q^2$, $q = 1, \dots, k$: Restvarianzen der Kommunalitätenschätzung

u.d.N.: $\mathbf{t}'_1 \mathbf{t}_1 = 1$ (Normierung)

folgt das Eigenwertproblem $(\mathbf{R}_h - \lambda \mathbf{E}) \mathbf{t}_1 = 0$. Aus der Gleichung $\det(\mathbf{R}_h - \lambda \mathbf{E}) = 0$ wird der größte Eigenwert $\lambda_1 = \hat{s}_{\mathbf{H}_1}^2$ ermittelt.

Aus den Lösungen der restlichen Optimierungsprobleme werden die Eigenwerte $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_k$, die zugehörigen Eigenvektoren $\mathbf{t}_1, \dots, \mathbf{t}_k$ und Hauptachsen $\mathbf{H}_q = \mathbf{Z}\mathbf{t}_q$, $q = 1, \dots, k$, bestimmt. Durch Umskalieren ergeben sich dann die normierten Hauptachsen \mathbf{F}_q , $q = 1, \dots, k$, so dass gilt:

$$\mathbf{Z} = \mathbf{F}\mathbf{A}', \mathbf{F} = \mathbf{Z}\mathbf{T}\mathbf{L}^{-1/2}, \mathbf{A} = \mathbf{T}\mathbf{L}^{1/2} \text{ und } \mathbf{R}_h = \mathbf{T}\mathbf{L}\mathbf{T}' \implies \mathbf{F} = \mathbf{Z}\mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}, \mathbf{R}_h = \mathbf{A}\mathbf{A}' \text{ und } \mathbf{R} = \mathbf{A}\mathbf{A}' + \mathbf{V},$$

mit $\mathbf{F} = (\mathbf{F}_1, \dots, \mathbf{F}_k) \in \mathbb{R}^{n,k}$, $\mathbf{T} = (\mathbf{t}_1, \dots, \mathbf{t}_k) \in \mathbb{R}^{k,k}$, $\mathbf{L} = \text{diag}\{\lambda_1, \dots, \lambda_k\} \in \mathbb{R}^{k,k}$, $\frac{1}{n-1} \mathbf{F}'\mathbf{F} = \mathbf{E}$.

Hauptachsenmethode

Die Hauptachsenmethode besteht darin, die ersten r Hauptachsen F_1, \dots, F_r , die sich bei einer Hauptachsentransformation der standardisierten Daten Z anhand der reduzierten Korrelationsmatrix R_h ergeben, als Faktoren auszuwählen unter Vernachlässigung der restlichen $k - r$ Faktoren.

Anzahl der Faktoren	$r < k$
Zerlegung der Datenmatrix	$Z = FA' + U$
Zerlegung der Varianz in Höhe der geschätzten Kommunalität der k standardisierten Beobachtungsvariablen in eine durch die r gemeinsamen Faktoren erklärte Varianz und eine nicht erklärte Varianz	$\hat{s}_{Z_q}^2 = \hat{h}_q^2 = \sum_{j=1}^r a_{qj}^2 + \tilde{s}_{qe}^2, \quad q = 1, \dots, k$ $\implies s_{Z_q}^2 = \sum_{j=1}^r a_{qj}^2 + \tilde{s}_{qe}^2 + s_{qe}^2$
Korrelationsmatrix	$R = AA' + V + \text{Rest}$
Reproduzierte Korrelationsmatrix	$\hat{R} = AA'$
<p>Die Bestimmung der Matrizen A und F ist für $r = k$ auf Seite 42 beschrieben. Für $r < k$ ergeben sich die entsprechenden um $k - r$ Spalten reduzierten Matrizen. WinSTAT ermittelt die Matrizen A und F durch einen Iterationsprozess, in dem die Kommunalitätsschätzung und die Hauptachsentransformation iterativ wiederholt werden. Anschließend können die Faktorwerte f_{ij} des Faktors j des Objekts i, $i = 1, \dots, n$, $j = 1, \dots, r$, durch</p> $F = ZA(A'A)^{-1}$ <p>geschätzt werden, wenn die Iteration konvergiert bzw. nicht vorher abgebrochen wurde.</p>	

Ziel der Hauptachsenanalyse ist es, eine Zerlegung $Z = FA' + U$ zu finden, so dass der Anteil der nicht erklärten Varianz \tilde{s}_{qe}^2 „möglichst klein“ wird (bei dem Iterationsprozess null ergibt) und damit die zu erklärende Varianz – in Höhe der geschätzten Kommunalität – der k standardisierten Variablen „möglichst vollständig“ (bei dem Iterationsprozess vollständig) durch r gemeinsame Faktoren erklärt wird. Am Ende der Iteration verbleibt als nicht erklärte Restvarianz von $s_{Z_q}^2$ nur s_{qe}^2 . D.h.: Für die Diagonalelemente der Matrix R gilt: $r_{qq} = \sum_{j=1}^r a_{qj}a'_{jq} + s_{qe}^2 = \sum_{j=1}^r a_{qj}a'_{jq} + v_q^2 = 1$. I.d.R. sind am Ende des Iterationsprozesses einige Kommunalitäten größer als die am Anfang des Prozesses geschätzten Kommunalitäten.

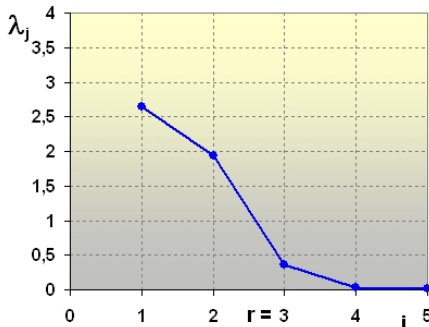
Maßzahlen und ihre Interpretation können aus der Hauptkomponentenanalyse Seite 39 übernommen werden. Kommunalitäten und Eigenwerte können wieder als Anteile der zu erklärenden Varianzen bzw. der zu erklärenden Gesamtvarianz der standardisierten Variablen aufgefasst werden, wobei jedoch die auf Seite 39 beschriebenen oberen Grenzen nicht angenommen werden, da für die geschätzten oder durch Iteration erhaltenen Kommunalitäten gilt: $h_q^2 < s_{Z_q}^2 = 1$ bzw. für $r = k$: $\sum_{j=1}^k \lambda_j = \sum_{q=1}^k h_q^2 < \sum_{q=1}^k s_{Z_q}^2 = k$.

Ergebnisseite der Hauptachsenanalyse mit WinSTAT für Excel

Um mit WinSTAT die Ergebnisse der Hauptachsenanalyse zu erhalten, muss unter der Faktorenanalyse die Einstellung „geschätzte Kommunalitäten: Höchste Korrelation“ gewählt werden, womit vorausgesetzt wird, dass die Varianzen $s_{Z_q}^2$, $q = 1, \dots, k$, nicht vollständig durch gemeinsame Faktoren erklärt werden können. Dadurch sind die geschätzten Kommunalitäten im Unterschied zur Hauptkomponentenanalyse nicht mehr gleich sondern kleiner 1. Zur Iteration muss „Analyse der Kommunalitäten wiederholen“ aktiviert werden. Der Aufbau der Seite entspricht der Seite 41 für die Hauptkomponentenanalyse.

Kriterien zur Bestimmung der Anzahl der Faktoren

Ziel der Faktorenanalyse ist es, dass eine Vielzahl von Variablen nur durch wenige Faktoren repräsentiert wird, gleichzeitig aber durch die Faktoren ein großer Teil der Gesamtvarianz $\sum_{q=1}^k s_{Z_q}^2$ aller standardisierten Beobachtungsvariablen erklärt wird. Ausgehend von einer Anzahl r von Faktoren, die kleiner als die Anzahl k der Variablen ist, kann stets durch Hinzunahme eines weiteren Faktors zusätzliche Varianz erklärt werden. Die folgenden Kriterien, die alle als Maßzahl die Eigenwerte verwenden, geben an, wie viele Faktoren verwendet werden sollten bzw. wann eine Erhöhung der Faktoren sinnvoll ist.

Kriterium	Aussage/Begründung
$r = \max_{1, \dots, k} \{j \lambda_j \geq 1\}$	Wähle die Anzahl r der Faktoren so, dass die Eigenwerte der Faktoren größer gleich 1 sind, d.h.: Die durch einen Faktor F_j erklärte Varianz λ_j aller k standardisierten Variablen soll mindestens so groß sein wie diejenige, die durchschnittlich auf eine Variable entfällt, $\frac{1}{k} \sum_{q=1}^k s_{Z_q}^2 = 1$.
$r = \min_{1, \dots, k} \{j \sum_{i=1}^j \lambda_i \geq \frac{c}{100} k\}$	Wähle die Anzahl r der Faktoren so, dass ein best. Anteil $c\%$ der zu erklärenden Gesamtvarianz $\sum_{q=1}^k s_{Z_q}^2 = k$ durch die Faktoren erklärt wird.
Knickpunkt des Screeplots	<div style="display: flex; align-items: center;">  <div style="margin-left: 20px;"> <p>Wähle die Anzahl r der Faktoren so, dass die Eigenwerte λ_j relativ große Werte aufweisen – dies ist gerade bis zum Knickpunkt der Fall –, da Faktoren mit kleineren Eigenwerten – nach dem Knickpunkt – nur noch zufällig sind.</p> </div> </div>

Prüfung auf Signifikanz der Hauptkomponenten: Es wird getestet, ob sich die $k - r$ kleinsten Eigenwerte $\lambda_{r+1}, \dots, \lambda_k$ signifikant unterscheiden und daher noch weitere Hauptkomponenten aufgenommen werden sollten.

χ^2-Test für $0 < r < k - 1$:			Voraussetzung: $X_q \sim N(\mu_q, \sigma_q^2), q = 1, \dots, k$		
statische Kenngröße	Nullhypothese H_0	Alternativhypothese H_1	Testfunktion (Bartlett-Approximation) χ_r^2	Testverteilung χ^2/H_0	Entscheidungsregel zur Ablehnung von H_0 bei gegebenem α , z.B. $\alpha = 0,01$
$\lambda_1, \dots, \lambda_r$	$\lambda_{r+1} = \dots = \lambda_k$	$\lambda_i \neq \lambda_j$ für mind. ein Paar $i \neq j, i, j \in \{r+1, \dots, k\}$	$\chi_r^2 = (N - 1)[-\ln(\det(\mathbf{R})) + \ln(\lambda_1 \dots \lambda_r) + (k - r) \ln \lambda]$ mit $\lambda = (k - \lambda_1 - \dots - \lambda_r)(k - r)$	$\chi^2(\frac{1}{2}(k - r + 2) \cdot (k - r - 1))$	$\chi_{r, \text{emp}}^2 > \chi_{1-\alpha}^2(\cdot)$

Schrittweises Vorgehen beginnend mit $r = 1$: Solange H_0 abgelehnt werden kann, nimmt man eine weitere Hauptkomponente hinzu. Sobald für ein r die Nullhypothese nicht mehr abgelehnt werden kann, hat man die Anzahl der Faktoren gefunden. Für kleine Stichprobengrößen N bringt $N' = N - r - \frac{1}{6}(2(k - r) + 1 + \frac{2}{k-r})$ statt N im Ausdruck der Testfunktion eine bessere Anpassung an die χ^2 -Verteilung.

Interpretation und Rotation der Faktoren

Mit dem linearen Zusammenhang zwischen Variablen und Faktoren, der aus der Ladungsmatrix A einer Hauptkomponenten- oder Hauptachsenanalyse folgt:

$$Z_q = a_{q1}F_1 + a_{q2}F_2 + \dots + a_{qr}F_r, \quad q = 1, \dots, k,$$

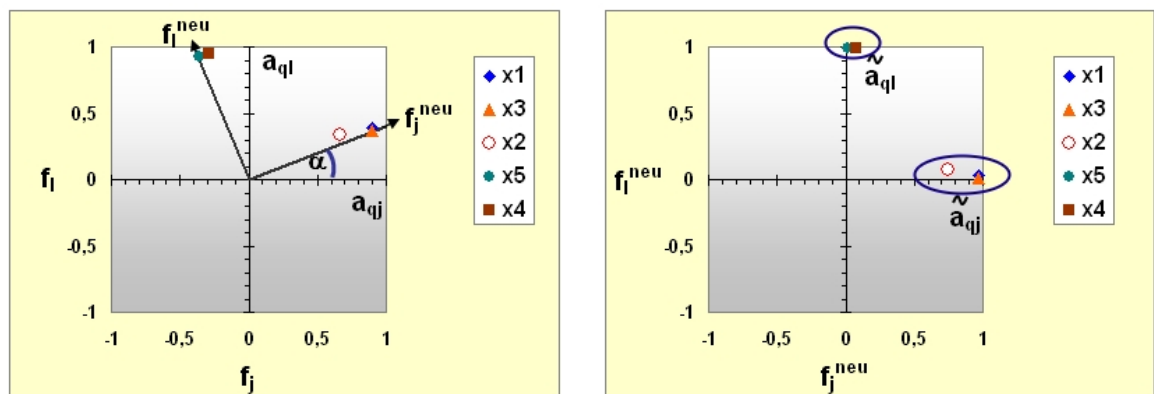
ist noch nicht der Sinn der Faktoren geklärt. Man weiß nur, dass die Faktoren der Reihe nach abnehmende Varianzen der Gesamtvarianz $\sum_{q=1}^k s_{Z_q}^2 = k$ erklären (vgl. Tabelle der Eigenwerte auf der Ergebnisseite WinSTAT Seite 41), d.h.:

$$\sum_{q=1}^k a_{q1}^2 \geq \sum_{q=1}^k a_{q2}^2 \geq \dots \geq \sum_{q=1}^k a_{qr}^2.$$

Entscheidend für die Interpretation der Faktoren bzw. für die Suche nach einem „Sammelbegriff“, zu dem mehrere Variablen zusammengefasst werden können, sind die Koeffizienten der Ladungsmatrix. Liegt eine Einfachstruktur der Ladungsmatrix vor, d.h. laden die Variablen nur auf einen Faktor hoch und auf alle anderen Faktoren niedrig, so lassen sich Variablen (unter Sachkenntnis des Anwenders) leicht durch eine gemeinsame Hintergrundvariable interpretieren. Ist jedoch eine solche Struktur der Ladungsmatrix nicht gegeben, so kann für eine bessere Interpretierbarkeit eine Rotation (lineare Transformation) der Faktoren vorgenommen werden. Damit die Beiträge der Faktoren an der Gesamtvarianz nach der Rotation unverändert bleiben, muss gleichzeitig die Ladungsmatrix A von rechts mit der inversen Rotationsmatrix multipliziert werden. Ziel der Rotation ist es, eine Einfachstruktur der neuen Ladungsmatrix zu erreichen, so dass die Faktorladungen entweder sehr große oder sehr kleine Werte annehmen.

Bsp. 1: Orthogonale Transformation für zwei Faktoren. Drehung des Koordinatensystems um α mit

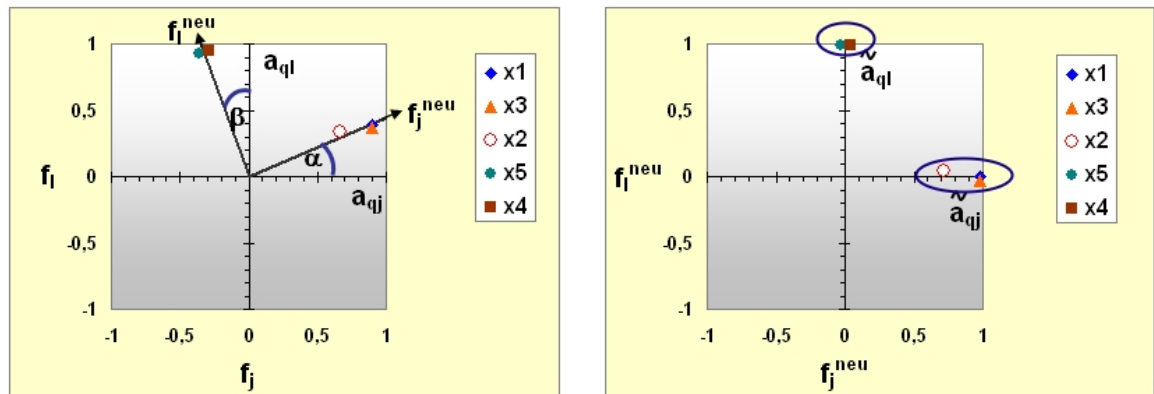
$$\begin{pmatrix} f_j^{\text{neu}} \\ f_l^{\text{neu}} \end{pmatrix} = \underbrace{\begin{pmatrix} \cos \alpha & \sin \alpha \\ -\sin \alpha & \cos \alpha \end{pmatrix}}_{\text{Rotationsmatrix } C} \begin{pmatrix} f_j \\ f_l \end{pmatrix}, \quad \tilde{A} = AC^{-1} = A \begin{pmatrix} \cos \alpha & -\sin \alpha \\ \sin \alpha & \cos \alpha \end{pmatrix}.$$



Es wird versucht, das Koordinatensystem eines Ladungsdiagramms so zu drehen, dass die neuen Achsen mitten durch die Gruppen verwandter Variablen gehen und somit die rotierten Faktoren die Gruppenzugehörigkeit von Variablen kennzeichnen. Dies ist in diesem Beispiel – wie man sieht – nur ungefähr möglich im Unterschied zu der Rotation des Beispiels 2. Die Drehung reicht jedoch auf jeden Fall aus, um die Aussage treffen zu können, dass Faktor j die Variablen x_1 , x_2 und x_3 zusammenfasst, während Faktor l die Variablen x_4 und x_5 gruppiert.

Bsp. 2: Oblique (nicht orthogonale) Transformation für zwei Faktoren – diese hat zur Folge, dass die Faktoren nach der Rotation nicht mehr unabhängig (orthogonal) sind. Eine getrennte Drehung der Faktoren um die Winkel α und β mit

$$\begin{pmatrix} f_j^{\text{neu}} \\ f_l^{\text{neu}} \end{pmatrix} = \underbrace{\begin{pmatrix} \cos \alpha & \sin \alpha \\ -\sin \beta & \cos \beta \end{pmatrix}}_{\text{Rotationsmatrix } C} \begin{pmatrix} f_j \\ f_l \end{pmatrix}, \quad \tilde{A} = AC^{-1} = A \begin{pmatrix} \cos \alpha & -\sin \beta \\ \sin \alpha & \cos \beta \end{pmatrix}.$$



Methoden zur Bestimmung der Rotationsmatrix

Ziel der Rotationsmethoden ist es, die Einfachstruktur der Ladungsmatrix zu verbessern. Zuvor kann eine Kaiser-Normalisierung vorgenommen werden:

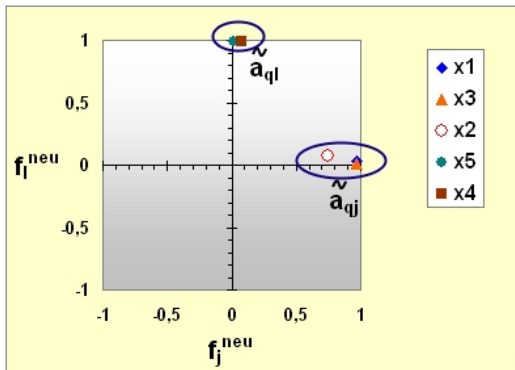
$$\tilde{a}_{qj} = \frac{a_{qj}}{\sqrt{h_q^2}} \text{ für Varimax und Promax bzw. } \tilde{a}_{qj} = \frac{a_{qj}}{\sqrt{h_j^2}} \text{ für Quartimax und Oblimax.}$$

Hierdurch werden Variablen mit höherer Kommunalität stärker berücksichtigt.

Methode	Transformation	Kriterium
Varimax	orthogonal	Maximierung der Quadrate der Faktorladungen innerhalb der Spalten der Ladungsmatrix, so dass eine deutlichere Ausprägung der Ladungen innerhalb der Spalten auftritt.
Quartimax	orthogonal	Maximierung der Quadrate der Faktorladungen innerhalb der Zeilen der Ladungsmatrix, so dass eine deutlichere Ausprägung der Ladungen innerhalb der Zeilen auftritt.
Promax	oblique	Verbesserung des Ergebnisses einer Varimax-Rotation durch eine oblique Transformation, so dass eine noch deutlichere Ausprägung der Ladungen innerhalb der Spalten auftritt.
Oblimax	oblique	Verbesserung des Ergebnisses einer Quartimax-Rotation durch eine oblique Transformation, so dass eine noch deutlichere Ausprägung der Ladungen innerhalb der Zeilen auftritt.

Interpretation der rotierten Faktoren

Faktorladungsdiagramm nach Rotation



Inhaltliche Interpretation der rotierten Faktoren

Die Faktoren werden am besten durch die auf sie hochladenden Variablen beschrieben. Alle Variablen X_q mit $|\tilde{a}_{qj}| \geq 0,5$ werden dem Faktor j zugeordnet, alle Variablen X_q mit $|\tilde{a}_{ql}| \geq 0,5$ werden dem Faktor l zugewiesen. Hierbei sind die Koeffizienten \tilde{a}_{qj} bzw. \tilde{a}_{ql} Faktorladungen der rotierten Faktorladungsmatrix. Beachte: Lädt eine Variable auf mehrere Faktoren hoch, so muss sie bei allen hochladenden (General-)Faktoren als inhaltlich bedeutend miteinbezogen werden (s.u.).

Schematische Darstellung der rotierten Faktorladungsmatrix, z.B.:

	Faktor 1	Faktor 2	Faktor 3
X_1	+		
X_2	+	+	
X_3	+		
X_4		-	
X_5		-	
X_6	+	+	
X_7			+
X_8			-

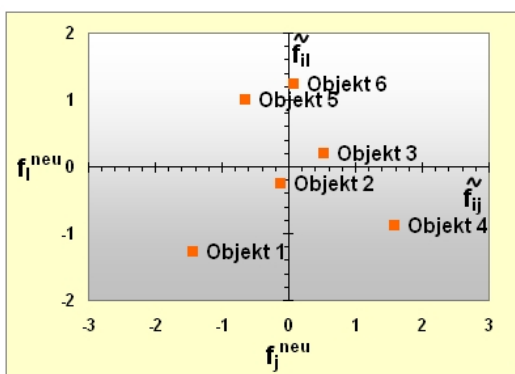
Die auf einen Faktor hochladenden Variablen werden mit einem „+“ oder „-“ versehen, d.h.:

$$\begin{aligned} \text{„+“: } & \tilde{a}_{qj} \geq 0,5, \\ \text{„-“: } & \tilde{a}_{qj} \leq -0,5. \end{aligned}$$

Diejenigen Variablen, die in derselben Spalte ein „+“ oder „-“ aufweisen, beschreiben inhaltlich den Faktor dieser Spalte. Dabei deuten Variablen, die nur in einer Spalte ein „+“ oder „-“ haben, auf einen Gruppenfaktor hin, der nur einen Teil der Variablen beeinflusst. Variablen, die in mehreren Spalten ein „+“ oder „-“ besitzen, geben Generalfaktoren wieder, die mehrere Variable mitbestimmen.

Interpretation der Faktorwerte

Faktorwertediagramm nach Rotation



Beurteilung der Objekte

Da die Faktorwerte aus einer linearen Transformation der standardisierten Daten Z hervorgehen, sind sie ebenso standardisiert, d.h. das arithmetische Mittel der Faktorwerte aller n Objekte eines Faktors ist null und die Standardabweichung der Faktorwerte aller n Objekte eines Faktors beträgt eins.

Damit lässt sich jedes einzelne Objekt in Bezug auf einen Faktor im Vergleich zu allen anderen Objekten beurteilen.

Der Faktorwert eines Objekts ist größer (kleiner, gleich) null

Das Objekt ist in Bezug auf den betrachteten Faktor im Vergleich zu allen anderen Objekten überdurchschnittlich (unterdurchschnittlich, dem Durchschnitt entsprechend) ausgeprägt.

Aufgabe

5

Für den Beispieldatensatz Seite 1 mit der Zeit für Nacharbeitung und Klausurvorbereitung (X_4), der Verweildauer im Internet (X_5), der Aufenthaltsdauer in Kinos, Discos oder Kneipen (X_6) und der Anzahl gekaufter Fachbücher (X_7) erhält man den unten stehenden WinSTAT-Output einer Hauptachsenanalyse.

- a) Interpretieren Sie die Kommunalitäten, die Eigenwerte sowie die Faktorladungen.
- b) Interpretieren Sie die rotierten Faktoren inhaltlich.
- c) Interpretieren Sie die Faktorwerte der Studierenden Nr. 4, 20, 23 und 25 anhand des Faktorwertediagramms.

Faktorenanalyse

Kommunalitäten

	Analyse	
	geschätzt	1
Nach- und Vorbereitungszeit	0,666	0,679
Verweildauer im Internet	0,575	0,603
Aufenthaltsdauer in Kinos ...	0,575	0,609
Anzahl gekaufter Fachbücher	0,666	0,676

Eigenwerte

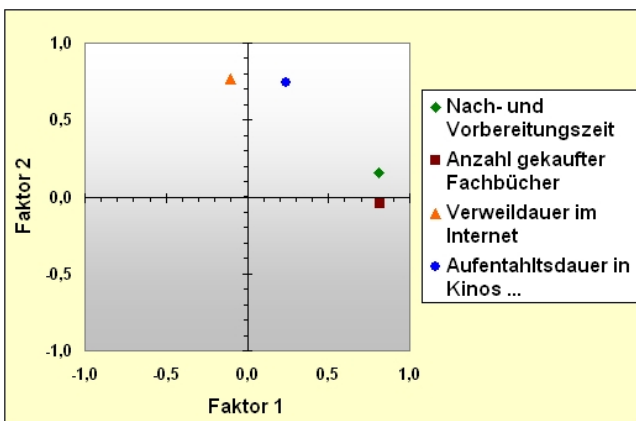
Faktor	Eigenwert	Varianz Prozent	Prozent kumuliert
1	1,504	37,596	37,596
2	1,064	26,590	64,186
3	0,065	1,621	65,806
4	0,020	0,503	66,310

Varimax Faktorladungen

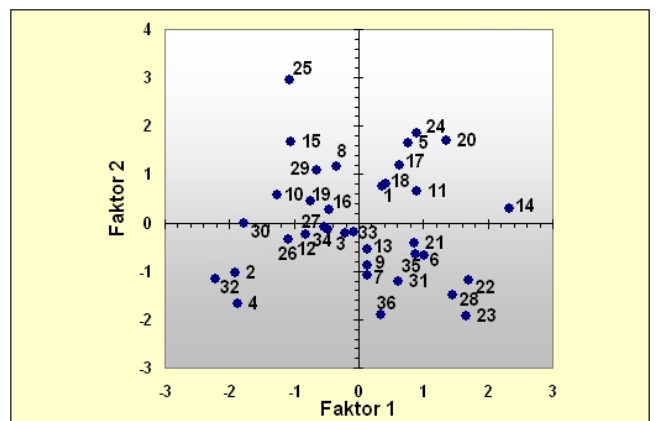
	Faktor 1	Faktor 2	Kommunalität
Anzahl gekaufter Fachbücher	0,821	-0,045	0,676
Nach- und Vorbereitungszeit	0,809	0,157	0,679
Verweildauer im Internet	-0,106	0,769	0,603
Aufenthaltsdauer in Kinos ...	0,242	0,742	0,609

Quadratsumme	1,398	1,169	2,567
Prozent der Varianz	34,951	29,235	64,186

Varimax Faktorladungen:



Faktorwerte nach Rotation:



5 Clusteranalyse

Die Clusteranalyse beinhaltet Klassifikationsverfahren, die – auf der Basis erhobener (Klassifizierungs-) Merkmale – die Gesamtheit von Objekten in disjunkte Teilmengen (Partitionen, Klassen, Cluster) zerlegt,¹ mit den Zielen:

- Die Beobachtungstupel der Objekte innerhalb einer Teilmenge sind möglichst ähnlich.
- Die Beobachtungstupel der Objekte unterschiedlicher Teilmengen sind möglichst verschieden.
- Die Teilmengen sind sachlich interpretierbar (idealtypischerweise durch Variable, die nicht zur Klassenbildung benutzt wurden) bzw. für die Zielsetzung der statistischen Analyse anwendbar.

Zur Beurteilung der Ähnlichkeit bzw. Verschiedenartigkeit von Beobachtungstupeln zwischen je zwei Objekten werden in Abhängigkeit der Skalierbarkeitseigenschaft der Merkmale Distanz- und Ähnlichkeitsmaße eingesetzt. Durch Anwenden eines Fusionierungsalgorithmus werden dann auf der Basis berechneter Distanz- bzw. Ähnlichkeitswerte Objekte zu homogenen Klassen zusammengefasst.

Datenmatrix metrischer Merkmale

Bei den im folgenden Abschnitt dargestellten skaleninvarianten Distanzmaßen, bei denen die Distanzen der Daten von den Maßeinheiten der Merkmale abhängen, müssen insbesondere bei verschiedenen Maßeinheiten die Daten vor Berechnung der Distanzen normiert werden. Im Falle der euklidischen Metrik kann als Normierung eine z -Transformation durchgeführt werden.

X_1, \dots, X_k :

k metrische Variablen

$$X := \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix}$$

Matrix der Beobachtungswerte x_{iq} , des i -ten Objekts der Variablen X_q , $i = 1, \dots, n$, $q = 1, \dots, k$; $n > k$; d.h.: $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ik})$ ist Beobachtungstupel des i -ten Objekts, und X kann als Punktwolke der Tupel \mathbf{x}_i , $i = 1, \dots, n$, im k -dimensionalen euklidischen Raum aufgefasst werden.

$$\tilde{x}_{iq} = \frac{x_{iq} - \bar{x}_q}{s_q^{(p)}}$$

normierter Beobachtungswert des i -ten Objekts der Variablen X_q , $i = 1, \dots, n$, $q = 1, \dots, k$, mit

$$\bar{x}_q = \frac{1}{n} \sum_{i=1}^n x_{iq}$$

arithmetisches Mittel der Variablen X_q , $q = 1, \dots, k$

$$s_q^{(p)} = \sqrt[p]{\frac{1}{n-1} \sum_{i=1}^n (x_{iq} - \bar{x}_q)^p}$$

für $p = 2$: Standardabweichung der Variablen X_q , $q = 1, \dots, k$

$$\tilde{X} := \begin{pmatrix} \tilde{x}_{11} & \tilde{x}_{12} & \dots & \tilde{x}_{1k} \\ \tilde{x}_{21} & \tilde{x}_{22} & \dots & \tilde{x}_{2k} \\ \vdots & \vdots & & \vdots \\ \tilde{x}_{n1} & \tilde{x}_{n2} & \dots & \tilde{x}_{nk} \end{pmatrix}$$

normierte Datenmatrix; für $p = 2$ ist \tilde{X} die z -transformierte Beobachtungsmatrix.

¹ Die sog. Fuzzy Clusteranalysen, d.h. Klassifikationsverfahren mit möglichen nicht disjunkten, d.h. überlappenden Teilmengen, werden hier nicht behandelt. Außerdem beschränkt sich die Darstellung nur auf deskriptive Klassifikationsverfahren, weshalb keine Verteilungsannahme innerhalb einer Klasse getroffen wird. Bei der deskriptiven Analyse wird aber die erwartungstreue Standardabweichung zugrundegelegt, da WinSTAT mit dieser Standardabweichung rechnet.

Distanzmaße für metrische Merkmale

Für metrische Merkmale können als Distanzmaße Metriken eines k -dimensionalen Punktraums $M^k \subseteq \mathbb{R}^k$ verwendet werden.

Def.: Sei M^k ein k -dimensionaler Punktraum mit n Tupeln $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ik})$, $i = 1, \dots, n$.

Eine Funktion $d: M^k \times M^k \rightarrow \mathbb{R}^+$ mit den Eigenschaften

- $$\left. \begin{array}{l} (1) \quad d(\mathbf{x}_i, \mathbf{x}_j) = 0 \text{ dann und nur dann, wenn } \mathbf{x}_i = \mathbf{x}_j \\ (2) \quad d(\mathbf{x}_i, \mathbf{x}_j) = d(\mathbf{x}_j, \mathbf{x}_i) \end{array} \right\} \text{ für jedes Paar von Tupeln } \mathbf{x}_i \text{ und } \mathbf{x}_j \text{ des Punktraums } M^k$$

heißt **Distanzfunktion** und die Zahl $d(\mathbf{x}_i, \mathbf{x}_j)$ heißt **Distanz** (Abstand) der Punkte $\mathbf{x}_i, \mathbf{x}_j$. Die symmetrische $n \times n$ -Matrix $\mathbf{D} = (d(\mathbf{x}_i, \mathbf{x}_j))$ der Distanzwerte heißt **Distanzmatrix**.

Erfüllt die Distanzfunktion d zusätzlich die Dreiecksungleichung

- $$(3) \quad d(\mathbf{x}_i, \mathbf{x}_j) \leq d(\mathbf{x}_i, \mathbf{x}_l) + d(\mathbf{x}_l, \mathbf{x}_j) \quad \text{für alle } \mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_l \text{ des Punktraums } M^k,$$

dann heißt d metrische Distanzfunktion oder kurz **Metrik** des Punktraums M^k .

Schreibweise: $d(i, j) := d(\mathbf{x}_i, \mathbf{x}_j)$ oder $d_{ij} := d(\mathbf{x}_i, \mathbf{x}_j)$. Je kleiner die Distanz $d(i, j)$ zwischen den Beobachtungstupeln von zwei Objekten i und j ist, desto ähnlicher sind die beiden Objekte.

Distanzmaß	Symbol	Berechnung	Eigenschaften
Minkowski-Metrik (L_p -Distanz)	$d_p(i, j)$	$d_p(i, j) = \sqrt[p]{\sum_{q=1}^k (x_{iq} - x_{jq})^p},$ $p > 1$	<ul style="list-style-type: none"> • nicht skaleninvariant • translationsinvariant (unabhängig vom Koordinatenursprung)
City-Block-Metrik (Manhattan-Metrik, L_1 -Distanz)	$d_1(i, j)$	$d_1(i, j) = \sum_{q=1}^k x_{iq} - x_{jq} ,$	<ul style="list-style-type: none"> • nicht skaleninvariant • translationsinvariant
euklidische Metrik (L_2 -Distanz)	$d_2(i, j)$	$\begin{aligned} d_2(i, j) &= \sqrt{\sum_{q=1}^k (x_{iq} - x_{jq})^2} \\ &= \sqrt{(\mathbf{x}_i - \mathbf{x}_j)'(\mathbf{x}_i - \mathbf{x}_j)} \\ &= \ \mathbf{x}_i - \mathbf{x}_j\ \end{aligned}$	<ul style="list-style-type: none"> • anschaulich: für $k = 2$ Berechnung der Distanz nach dem Satz von Pythagoras und für $k > 2$ analoge Def. der Distanz • nicht skaleninvariant • translationsinvariant • invariant gegenüber orthogonalen Transformationen (unabhängig von einer Drehung oder Spiegelung des Koordinatensystems)
Supremum-Metrik (Tschebyscheff-Metrik, L_∞ -Distanz)	$d_\infty(i, j)$	$d_\infty(i, j) = \max_{q=1, \dots, k} x_{iq} - x_{jq} $	<ul style="list-style-type: none"> • nicht skaleninvariant • translationsinvariant

Distanzmaß	Symbol	Berechnung	Eigenschaften
quadrierte euklidische Distanz	$d_2^2(i, j)$	$d_2^2(i, j) = \sum_{q=1}^k (x_{iq} - x_{jq})^2$ $= (\mathbf{x}_i - \mathbf{x}_j)'(\mathbf{x}_i - \mathbf{x}_j)$ $= \ \mathbf{x}_i - \mathbf{x}_j\ ^2$	<ul style="list-style-type: none"> • keine Metrik • einfache Berechnung • nicht skaleninvariant • translationsinvariant
(quadrierte) Mahalanobis-Distanz	$d_M(i, j)$	$d_M(i, j) = (\mathbf{x}_i - \mathbf{x}_j)' \mathbf{S}^{-1} (\mathbf{x}_i - \mathbf{x}_j),$ <p>mit $\mathbf{S} = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$</p> <p>und $\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$</p> <p>d.h. $\mathbf{S} = (s_{qt})$ ist die $k \times k$-Kovarianzmatrix mit</p> $s_{qt} = \frac{1}{n-1} \sum_{i=1}^n (x_{iq} - \bar{x}_q)(x_{it} - \bar{x}_t),$ <p style="text-align: center;">$q, t = 1, \dots, k$</p> <p>und $\bar{\mathbf{x}}' = (\bar{x}_1, \dots, \bar{x}_q, \dots, \bar{x}_k)$</p> <p>mit $\bar{x}_q = \frac{1}{n} \sum_{i=1}^n x_{iq}, q = 1, \dots, k.$</p>	<ul style="list-style-type: none"> • keine Metrik • skaleninvariant • translationsinvariant • dekorreliert korrelierte Merkmale, d.h. die Mahalanobis-Distanzen werden unter Verwendung von k unkorrelierten Merkmalen berechnet, auch wenn die ursprünglichen k Merkmale korreliert sind, denn: d_2^2 ist für die normierten unkorrelierten Beobachtungstupel \mathbf{y}_i mit $\mathbf{y}_i = \mathbf{S}^{-1/2}(\mathbf{x}_i - \bar{\mathbf{x}})$ gleich d_M der nicht normierten Beobachtungstupel \mathbf{x}_i, d.h.: $\ \mathbf{y}_i - \mathbf{y}_j\ ^2 = (\mathbf{x}_i - \mathbf{x}_j)' \mathbf{S}^{-1} (\mathbf{x}_i - \mathbf{x}_j)$

Ähnlichkeitsmaß für metrische Merkmale

Sollen in einer statistischen Analyse Daten von metrischen Merkmalen für verschiedene Objekte auf ein ähnliches Profil, z.B. Zeitreihenwerte einer metrischen Variablen auf eine ähnliche Entwicklung, untersucht werden, so ist für die Ähnlichkeit von Objekten nicht die Distanz zwischen den Beobachtungstupeln entscheidend sondern ein Maß, das einen ähnlichen Verlauf der (Zeitreihen-)Tupel von je zwei Objekten wiedergibt. Ein geeignetes Ähnlichkeitsmaß ist z.B. der Korrelationskoeffizient von Pearson:

$$r_{ij} := r_{\mathbf{x}_i \mathbf{x}_j} = \frac{\sum_{q=1}^k (x_{iq} - \bar{x}_i)(x_{jq} - \bar{x}_j)}{\sqrt{\sum_{q=1}^k (x_{iq} - \bar{x}_i)^2 \cdot \sum_{q=1}^k (x_{jq} - \bar{x}_j)^2}}, \quad -1 \leq r_{ij} \leq 1,$$

wobei: $\bar{x}_i = \frac{1}{k} \sum_{q=1}^k x_{iq}$: arithmetisches Mittel aller Variablen (Zeitreihenwerte) X_1, \dots, X_k des Objekts i (analog für j).

Zwei Objekte i und j sind um so ähnlicher, je größer der Betrag des Korrelationskoeffizienten der Beobachtungstupel der beiden Objekte ist. Die Ähnlichkeitsmatrix entspricht der symmetrischen Korrelationsmatrix.

Datenmatrix nominaler binärer Merkmale

Es wird nur der Fall nominaler binärer (dichotomer) Merkmale betrachtet, d.h. es wird vorausgesetzt, dass alle nominalen Merkmale genau zwei Ausprägungen haben.

X_1, \dots, X_k :

k nominale binäre Variablen mit den numerisch kodierten Merkmalsausprägungen 1 und 0

$$X := \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix}$$

Matrix der kodierten Beobachtungswerte x_{iq} , des i -ten Objekts der Variablen X_q , $i = 1, \dots, n$, $q = 1, \dots, k$; $n > k$;

d.h.: $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ik})$ ist Beobachtungstupel des i -ten Objekts mit $x_{iq} \in \{0, 1\}$, $i = 1, \dots, n$, $q = 1, \dots, k$.

Objekt i	Objekt j		Σ
	1	0	
1	a_{ij}	c_{ij}	$a_{ij} + c_{ij}$
0	b_{ij}	e_{ij}	$b_{ij} + e_{ij}$
Σ	$a_{ij} + b_{ij}$	$c_{ij} + e_{ij}$	k

Kontingenztafel für zwei Objekte i, j , d.h.: a_{ij} , b_{ij} , c_{ij} und e_{ij} sind die Häufigkeiten, mit denen die kodierten Merkmalsausprägungspaare bei den beiden Objekten auftreten. Die Randhäufigkeiten beschreiben dann die Häufigkeiten, mit denen das Objekt i bzw. j die Merkmalsausprägungen 1 oder 0 annimmt.

Ähnlichkeitsmaße für nominale binäre Merkmale

Ein **Ähnlichkeitsmaß** $t(\mathbf{x}_i, \mathbf{x}_j)$ zwischen zwei Objekten i und j mit den Tupeln \mathbf{x}_i und \mathbf{x}_j nominaler Merkmale ist eine aus den Häufigkeiten von übereinstimmenden und nichtübereinstimmenden Komponenten von \mathbf{x}_i und \mathbf{x}_j gebildete Maßzahl. Da die verwendeten Häufigkeiten von Beobachtungswertpaaren der Objekte i und j unabhängig von der Reihenfolge des Zählens an den Objekten i und j sind, wird mit der Bildung der Ähnlichkeitsmaße aus den Häufigkeiten die Symmetrieeigenschaft

$$(1) \quad t(\mathbf{x}_i, \mathbf{x}_j) = t(\mathbf{x}_j, \mathbf{x}_i)$$

erfüllt. Die hier dargestellten Ähnlichkeitsmaße sind auf das Intervall $[0, 1]$ normiert, d.h. es gilt: $0 \leq t(\mathbf{x}_i, \mathbf{x}_j) \leq 1$. Damit erfüllen alle dargestellten Ähnlichkeitsmaße die Eigenschaft:²

$$(2) \quad t(\mathbf{x}_i, \mathbf{x}_j) = 1 \text{ dann und nur dann, wenn } \mathbf{x}_i = \mathbf{x}_j.$$

Die symmetrische $n \times n$ -Matrix $T = (t(\mathbf{x}_i, \mathbf{x}_j))$ der Ähnlichkeitswerte heißt **Ähnlichkeitsmatrix**.

Schreibweise: $t(i, j) := t(\mathbf{x}_i, \mathbf{x}_j)$ oder $t_{ij} := t(\mathbf{x}_i, \mathbf{x}_j)$. Je größer die Ähnlichkeit $t(i, j)$ zwischen den Beobachtungstupeln von zwei Objekten i und j ist, desto ähnlicher sind die beiden Objekte.

Es werden zwei Typen von Ähnlichkeitsmaßen vorgestellt: M-verwandte und S-verwandte Koeffizienten. Die Rangordnungen der M-verwandten Koeffizienten unterscheiden sich i.d.R. von denjenigen der S-verwandten Koeffizienten.

² Für die S-verwandten Koeffizienten wird der Fall $\mathbf{x}_i = \mathbf{x}_j = \mathbf{0}$ nicht gezählt.

Ähnlichkeitsmaß	Berechnung	Eigenschaften
M-verwandte Koeffizienten mit Gewicht u für Überein- und Gewicht $(1-u)$ für Nichtübereinstimmung	$t_{ij} = \frac{u \cdot (a_{ij} + e_{ij})}{u \cdot (a_{ij} + e_{ij}) + (1-u) \cdot (b_{ij} + c_{ij})},$ $0 < u < 1$	<ul style="list-style-type: none"> • gleiche Ähnlichkeitsrangordnung für alle u • invariant gegenüber eineindeutiger Transformation eines oder mehrerer Merkmale, d.h.: Die Ähnlichkeit ist unabhängig davon, was als 0 und was als 1 bezeichnet wird.
M-Koeffizient (matching coefficient, Koeffizient der einfachen Übereinstimmung) ($u = 1/2$)	$t(i, j) = \frac{a_{ij} + e_{ij}}{k},$ $= \frac{\sum_{q=1}^k I(x_{iq}, x_{jq})}{k},$ mit $I(x_{iq}, x_{jq}) = \begin{cases} 1 & \text{für } x_{iq} = x_{jq} \\ 0 & \text{sonst} \end{cases}$	<ul style="list-style-type: none"> • Anteil der übereinstimmenden Beobachtungswertpaare (1,1) oder (0,0) an allen Wertepaaren der Objekte i, j • Übereinstimmungen und Nichtübereinstimmungen werden gleichgewichtet
Sokal/Sneath-Koeffizient 1 ($u = 2/3$)	$t(i, j) = \frac{2(a_{ij} + e_{ij})}{2(a_{ij} + e_{ij}) + b_{ij} + c_{ij}},$	<ul style="list-style-type: none"> • Übereinstimmungen (1,1), (0,0) erhalten doppeltes Gewicht
Rogers/Tanimoto-Koeffizient ($u = 1/3$)	$t(i, j) = \frac{a_{ij} + e_{ij}}{a_{ij} + e_{ij} + 2(b_{ij} + c_{ij})},$	<ul style="list-style-type: none"> • Nichtübereinstimmungen (0,1), (1,0) erhalten doppeltes Gewicht
S-verwandte Koeffizienten mit Gewicht u für Überein- und Gewicht $(1-u)$ für Nichtübereinstimmung	$t(i, j) = \frac{u \cdot a_{ij}}{u \cdot a_{ij} + (1-u) \cdot (b_{ij} + c_{ij})},$ $0 < u < 1$	<ul style="list-style-type: none"> • übereinstimmende Beobachtungswertpaare (0,0) werden nicht gezählt • gleiche Ähnlichkeitsrangordnung für alle u • nicht invariant gegenüber eineindeutiger Transformation eines oder mehrerer Merkmale, d.h.: Die Ähnlichkeit ist i.d.R. abhängig davon, was als 0 und was als 1 bezeichnet wird.
S-Koeffizient (similarity coefficient, Jaccard-Koeff.) ($u = 1/2$)	$t(i, j) = \frac{a_{ij}}{a_{ij} + b_{ij} + c_{ij}},$	<ul style="list-style-type: none"> • Anteil der übereinstimmenden Beobachtungswertpaare (1,1) an allen Wertepaaren außer (0,0) der Objekte i, j • Übereinstimmungen und Nichtübereinstimmungen werden gleichgewichtet
Dice- (Würfel-) Koeffizient ($u = 2/3$)	$t(i, j) = \frac{2a_{ij}}{2a_{ij} + b_{ij} + c_{ij}},$	<ul style="list-style-type: none"> • Übereinstimmungen (1,1) erhalten doppeltes Gewicht
Sokal/Sneath-Koeffizient 2 ($u = 1/3$)	$t(i, j) = \frac{a_{ij}}{a_{ij} + 2(b_{ij} + c_{ij})},$	<ul style="list-style-type: none"> • Nichtübereinstimmungen (0,1), (1,0) erhalten doppeltes Gewicht

Datenmatrix nominaler binärer oder mehrstufiger Merkmale

Nominale Merkmale mit mehr als zwei Merkmalsausprägungen werden als mehrstufig bezeichnet.

$X_1, \dots, X_k:$	k nominale binäre oder mehrstufige Variablen
$X := \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix}$	Matrix der geeignet numerisch kodierten Beobachtungswerte x_{iq} , des i -ten Objekts der Variablen X_q , $i = 1, \dots, n$, $q = 1, \dots, k$; $n > k$; d.h.: $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ik})$ ist Beobachtungstupel des i -ten Objekts.
$v_{ij}:$	Anzahl übereinstimmender Komponenten von \mathbf{x}_i und \mathbf{x}_j
$w_{ij}:$	Anzahl nichtübereinstimmender Komponenten von \mathbf{x}_i und \mathbf{x}_j
$m_q:$	Anzahl der Merkmalsausprägungen des Merkmals X_q , $q = 1, \dots, k$
$m^* = \sum_{q=1}^k m_q:$	Summe aller Merkmalsausprägungen der Merkmale X_1, \dots, X_k

Ähnlichkeitsmaße für nominale binäre oder mehrstufige Merkmale

Ähnlichkeitsmaß	Berechnung	Eigenschaften
verallgemeinerte M-verwandte Koeffizienten mit Gewicht u für Übereinstimmung und Gewicht $(1-u)$ für Nichtübereinstimmung	$t(i, j) = \frac{u \cdot v_{ij}}{u \cdot v_{ij} + (1-u) \cdot w_{ij}},$ $0 < u < 1$	<ul style="list-style-type: none"> • gleiche Ähnlichkeitsrangordnung für alle u • invariant gegenüber eindeutiger Transformation eines oder mehrerer Merkmale, d.h.: Die Ähnlichkeit ist unabhängig von der Bezeichnung der Merkmale. • die Anzahl der Merkmalsausprägungen, die die Merkmale besitzen, wird nicht berücksichtigt.
verallgemeinerter M-Koeffizient ($u = 1/2$)	$t(i, j) = \frac{v_{ij}}{k},$ $= \frac{\sum_{q=1}^k I(x_{iq}, x_{jq})}{k},$ $I(x_{iq}, x_{jq}) = \begin{cases} 1 & \text{für } x_{iq} = x_{jq} \\ 0 & \text{sonst} \end{cases}$	<ul style="list-style-type: none"> • Anteil der übereinstimmenden Beobachtungswertpaare an allen Wertepaaren der Objekte i, j • Übereinstimmung und Nichtübereinstimmung werden gleich gewichtet
modifizierter verallgemeinerter M-Koeffizient	$t(i, j) = \frac{1}{m^*} \sum_{q=1}^k m_q I(x_{iq}, x_{jq}),$ $I(x_{iq}, x_{jq}) = \begin{cases} 1 & \text{für } x_{iq} = x_{jq} \\ 0 & \text{sonst} \end{cases}$	<ul style="list-style-type: none"> • den Objekten i und j, die in einem Merkmal mit vielen Ausprägungen übereinstimmen, wird eine höhere Ähnlichkeit zugeordnet als solchen Objekten, die in einem Merkmal mit wenigen Ausprägungen übereinstimmen • invariant gegenüber eindeutiger Transformation eines oder mehrerer Merkmale

Datenmatrix ordinaler Merkmale

Für jedes ordinale Merkmal kann eine Rangordnung der Merkmalsausprägungen aufgestellt werden. Zwei Objekte i und j werden bzgl. eines Merkmals X_q als umso ähnlicher betrachtet, je näher die Beobachtungswerte x_{iq} und x_{jq} des Merkmals X_q hinsichtlich der Rangordnung beieinanderliegen. Zur Berücksichtigung der Rangordnung werden Hilfsvariablen eingeführt. Aus den Tupeln der Zeilen i und j der Beobachtungsmatrix der Hilfsvariablen aller ordinalen Merkmale X_1, \dots, X_k wird dann die Ähnlichkeit zwischen zwei Objekten i und j bestimmt.

X_1, \dots, X_k : k ordinale Variablen
 m_q : Anzahl der Merkmalsausprägungen des Merkmals X_q , $q = 1, \dots, k$
 $x_{q,1} \prec x_{q,2} \prec \dots \prec x_{q,m_q}$: Rangordnung der m_q Merkmalsausprägungen des Merkmals X_q

Für jedes ordinale Merkmal werden so viele binäre Hilfsvariablen eingeführt, wie das Merkmal Ausprägungen aufweist, d.h.: Für Merkmal X_q gibt es m_q Hilfsvariablen. Nimmt der Beobachtungswert x_{iq} des Objekts i an der Variablen X_q die Position $x_{q,h}$ mit $1 \leq h \leq m_q$ der Rangordnung ein, dann wird den ersten h Hilfsvariablen der Wert 1 und den verbleibenden Hilfsvariablen der Wert 0 zugewiesen. Die Beobachtungswerte (zugewiesene Werte 1 oder 0) der Hilfsvariablen aller Objekte und aller ordinalen Merkmale werden in einer $n \times m^*$ -Matrix mit $m^* = \sum m_q$ erfasst.

Ähnlichkeitsmaße für ordinale Merkmale

Als Ähnlichkeitsmaße können alle vorgestellten Koeffizienten für nominale binäre Merkmale auf die Hilfsvariablen angewendet werden.

Bsp.: Beispieldatensatz Seite 1 mit dem Merkmal X_8 : erwartete Leistung in der Statistiklausur

Merkmalsausprägungen	Beobachtungswerte der Studierenden $i = 3, 4, 5$	Beobachtungstupel der Hilfsvariablen von X_8
$x_{8,1}$: unterdurchschnittlich	x_{38} : durchschnittlich	$\mathbf{x}'_3 = (1, 1, 0)$
$x_{8,2}$: durchschnittlich	x_{48} : unterdurchschnittlich	$\mathbf{x}'_4 = (1, 0, 0)$
$x_{8,3}$: eher besser	x_{58} : eher besser	$\mathbf{x}'_5 = (1, 1, 1)$
Ähnlichkeiten für X_8 mit dem M-Koeffizienten: $t(3, 4) = 2/3$, $t(3, 5) = 2/3$ und $t(4, 5) = 1/3$.		

Aus der Beobachtungsmatrix der Hilfsvariablen von X_2 (Mathenote) und X_8 ergeben sich die Ähnlichkeiten mit dem M-Koeffizienten: $t(3, 4) = 5/7$, $t(3, 5) = 5/7$ und $t(4, 5) = 3/7$.

Zur Vereinfachung werden auch Distanzmaße für ordinale Merkmale verwendet, indem die Merkmalsausprägungen der Merkmale entsprechend der Rangfolge numerisch kodiert und die Distanzen für die Tupel \mathbf{x}_i und \mathbf{x}_j der kodierten Beobachtungswerte mit einem metrischen Distanzmaß berechnet werden.

Distanzmaße für Merkmale mit unterschiedlichem Skalenniveau

Bei vielen multivariaten Datenerhebungen werden die Merkmale X_1, \dots, X_k unterschiedliche Skalenniveaus aufweisen. Für die Analyse der Ähnlichkeit zwischen je zwei Objekten i und j mit gemischt-skalierten Beobachtungstupeln \mathbf{x}_i und \mathbf{x}_j werden zwei mögliche Vorgehen, die jeweils zwei Schritte beinhalten, vorgestellt:

- (I) Im ersten Schritt werden Ähnlichkeiten für Beobachtungstupel mit ausschließlich nominalen bzw. ordinalen Komponenten von zwei Objekten i und j mit Ähnlichkeitsmaßen für nominale Merkmale (z.B. dem M-Koeffizienten) bzw. für ordinale Merkmale bestimmt und anschließend in Distanzen $d(i, j) = 1 - t(i, j)$ überführt. Ebenso werden Distanzen für (normierte) Beobachtungstupel mit ausschließlich metrischen Komponenten von zwei Objekten i und j mit einem Distanzmaß für metrische Merkmale (z.B. der euklidischen Metrik) berechnet.

Im zweiten Schritt wird das arithmetische Mittel – oder das nach der Anzahl der nominalen, ordinalen bzw. metrischen Variablen gewogene arithmetische Mittel – aller berechneten Distanzen gebildet und als Distanz für die Objekte i und j mit den Beobachtungstupeln \mathbf{x}_i und \mathbf{x}_j verwendet.

- (II) Im ersten Schritt werden Distanzmaße für *eine Variable* X_q , $q = 1, \dots, k$, in Abhängigkeit der Skalierbarkeitseigenschaft gebildet. Im zweiten Schritt wird zur Vergleichbarkeit der Distanzmaße eine Normierung vorgenommen und anschließend durch Summieren der normierten Distanzmaße über alle Variablen ein Distanzmaß für die Beobachtungstupel \mathbf{x}_i und \mathbf{x}_j bestimmt.

Skalenniveau	Distanzmaß für die Variable X_q															
nominal	$d^{(q)}(i, j) = \begin{cases} 0 & \text{für } x_{iq} = x_{jq} \\ 1 & \text{sonst} \end{cases}$															
ordinal	<p>$d^{(q)}(i, j) = x_{iq} - x_{jq}$ mit x_{iq}: Beobachtungswert des Objekts i der – entsprechend der Rangfolge – numerisch kodierten Merkmalsausprägungen der Variablen X_q (analog j) Bsp.: Beispieldatensatz Seite 1 mit dem Merkmal X_2: Mathenote</p> <table border="1" style="margin-left: auto; margin-right: auto;"> <thead> <tr> <th>kodierte Merkmalsausprägungen</th> <th>Beobachtungswerte der Studierenden $i = 3, 4, 5$</th> <th>Distanzen der Studierenden $i = 3, 4, 5$</th> </tr> </thead> <tbody> <tr> <td>$x_{2,1} = 1$ (Note 4)</td> <td>$x_{32} = 3$</td> <td>$d^{(2)}(3, 4) = 3 - 4 = 1$</td> </tr> <tr> <td>$x_{2,2} = 2$ (Note 3)</td> <td>$x_{42} = 4$</td> <td>$d^{(2)}(3, 5) = 3 - 2 = 1$</td> </tr> <tr> <td>$x_{2,3} = 3$ (Note 2)</td> <td>$x_{52} = 2$</td> <td>$d^{(2)}(4, 5) = 4 - 2 = 2$</td> </tr> <tr> <td>$x_{2,4} = 4$ (Note 1)</td> <td></td> <td></td> </tr> </tbody> </table>	kodierte Merkmalsausprägungen	Beobachtungswerte der Studierenden $i = 3, 4, 5$	Distanzen der Studierenden $i = 3, 4, 5$	$x_{2,1} = 1$ (Note 4)	$x_{32} = 3$	$d^{(2)}(3, 4) = 3 - 4 = 1$	$x_{2,2} = 2$ (Note 3)	$x_{42} = 4$	$d^{(2)}(3, 5) = 3 - 2 = 1$	$x_{2,3} = 3$ (Note 2)	$x_{52} = 2$	$d^{(2)}(4, 5) = 4 - 2 = 2$	$x_{2,4} = 4$ (Note 1)		
kodierte Merkmalsausprägungen	Beobachtungswerte der Studierenden $i = 3, 4, 5$	Distanzen der Studierenden $i = 3, 4, 5$														
$x_{2,1} = 1$ (Note 4)	$x_{32} = 3$	$d^{(2)}(3, 4) = 3 - 4 = 1$														
$x_{2,2} = 2$ (Note 3)	$x_{42} = 4$	$d^{(2)}(3, 5) = 3 - 2 = 1$														
$x_{2,3} = 3$ (Note 2)	$x_{52} = 2$	$d^{(2)}(4, 5) = 4 - 2 = 2$														
$x_{2,4} = 4$ (Note 1)																
metrisch	$d^{(q)}(i, j) = x_{iq} - x_{jq} ^p$, i.d.R. $p = 1$															

Normierung:
$$\tilde{d}^{(q)}(i, j) = \frac{d^{(q)}(i, j)}{\sum_{\alpha=1}^n \sum_{\beta=1}^n d^{(q)}(\alpha, \beta)}, \quad 0 \leq \tilde{d}^{(q)}(i, j) \leq 1 \quad \text{und} \quad \sum_{\alpha=1}^n \sum_{\beta=1}^n \tilde{d}^{(q)}(\alpha, \beta) = 1.$$

$\tilde{d}^{(q)}(i, j)$ gibt den Heterogenitätsanteil eines Paares (i, j) an der Heterogenität aller Paare (α, β) mit $\alpha, \beta = 1, \dots, n$ für ein Merkmal X_q an.

Distanzmaß für zwei Objekte i und j mit den gemischt-skalierten Beobachtungstupeln \mathbf{x}_i und \mathbf{x}_j :

$$d(i, j) = \sum_{q=1}^k \tilde{d}^{(q)}(i, j).$$

Fusionierungsalgorithmen

Nachdem alle Distanzen bzw. Ähnlichkeiten zwischen je zwei Objekten berechnet worden sind, werden die Objekte nach ihrer Distanz bzw. Ähnlichkeit zu Clustern kombiniert.

Hierarchische agglomerative Verfahren

Hierarchische Verfahren werden immer dann angewendet, wenn es interessiert, welche Verbindungen zwischen den Klassen (Clustern) bestehen bzw. wenn in der Objektmenge hierarchische Strukturen zu vermuten sind. Bei den agglomerativen Verfahren wird durch schrittweises Senken der Homogenität ein Dendrogramm (Stammbaum) von unten nach oben konstruiert, d.h.: Zunächst bildet jedes Objekt ein Cluster, dann werden die beiden Cluster mit der kleinsten Distanz (bzw. größten Ähnlichkeit) ermittelt und zu einem Cluster fusioniert. Aus der so erhaltenen neuen Menge von Clustern werden wieder die beiden Cluster mit der kleinsten Distanz (bzw. größten Ähnlichkeit) ermittelt und zu einem Cluster zusammengefasst. Dieses schrittweise Vorgehen der Reduktion der Anzahl der Cluster wird solange wiederholt, bis alle Objekte in einem einzigen großen Cluster fusioniert sind. Die einzelnen Schritte liefern eine Folge von Zerlegungen, aus der – z.B. nach inhaltlichen Überlegungen – eine Partition ausgewählt und damit gleichzeitig die Anzahl der Cluster festgelegt wird.

Ablauf agglomerativer Verfahren

$A = \{A_1, A_2, \dots, A_n\}$:	zu klassifizierende Objektmenge
$\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ik})$:	Beobachtungstupel des i -ten Objekts A_i , $i = 1, \dots, n$
$\mathcal{C} = \{C_1, \dots, C_g\}$:	Partition von A in g disjunkte Teilmengen C_1, \dots, C_g , d.h.
$C_1 \cup \dots \cup C_g = A$:	vollständige Zerlegung von A und
$C_\mu \cap C_\nu = \emptyset$ für alle $\mu \neq \nu$:	disjunkte Teilmengen

Iteratives Verfahren der Clusterfusionierung:

Schritt 1, $\tau = 0$: $\mathcal{C}_0 = \{\{A_1\}, \dots, \{A_n\}\}$ sei die feinste Partition der Objektmenge A .

Schritt für $\tau \geq 1$: Bilde aus der Partition $\mathcal{C}_{\tau-1}$ eine neue Partition \mathcal{C}_τ durch Fusionierung derjenigen zwei Cluster C_u und C_w aus $\mathcal{C}_{\tau-1}$, für die ein vorgegebenes Distanzmaß D (vgl. Seite 58) minimal ist, d.h.: $D(C_u, C_w) = \min_{\substack{C_\mu, C_\nu \in \mathcal{C}_{\tau-1} \\ \mu \neq \nu}} D(C_\mu, C_\nu)$.

Schrittfolge: Man iteriert Schritt 2 bis n , d.h. von $\tau = 1$ bis $\tau = n - 1$, so dass man im letzten Schritt, also für $\tau = n - 1$, die Partition $\mathcal{C}_{n-1} = \{A\}$ erhält.

Heterogenitätsindex: Dem im Schritt für $\tau \geq 1$ durch Fusion entstandenen Cluster $C_u \cup C_w$ wird der Heterogenitätsindex $h_\tau := D(C_u, C_w)$ zugeordnet. Definiert man für $\tau = 0$ einfach $h_0 := 0$, so erhält man von $\tau = 0$ bis $\tau = n - 1$ eine Folge von Heterogenitätsindizes, für die gelten sollte: $0 =: h_0 \leq h_1 \leq \dots \leq h_{n-1}$.

Existiert bei einer Partition kein eindeutiges Minimum von Distanzen, so kann man alle Cluster fusionieren, welche die gleiche minimale Distanz aufweisen.

Methoden der Clusterfusionierung

Die bisher dargestellten Distanzmaße wurden für die Beobachtungstupel von *zwei Objekten* formuliert. Sobald jedoch im ersten Schritt Objekte zu Clustern zusammengefasst worden sind, werden in den folgenden Schritten Distanzmaße für zwei Cluster mit Beobachtungstupeln von mehr als zwei Objekten benötigt. Die Clusterbildung aufgrund der Minimierung der auf dieser Seite definierten Distanzmaße für Cluster mit mehr als zwei Objekten führen zu den entsprechenden Fusionierungsmethoden. Da die für nominale oder ordinale Beobachtungstupel vorgestellten Ähnlichkeitsmaße $t(i, j)$ zwischen zwei Objekten in Distanzmaße $d(i, j) = 1 - t(i, j)$ überführt werden können, werden alle Methoden anhand von Distanzmaßen beschrieben. Alternativ könnten Ähnlichkeitsmaße für zwei Cluster definiert werden, wobei einer Minimierung der Distanzen eine Maximierung der Ähnlichkeiten entspricht.

Distanzmaß	Berechnung	Skalenniveau	Aussage
Single-Linkage (Nächstgelegener Nachbar)	$D(C_\mu, C_\nu) = \min_{\substack{A_i \in C_\mu \\ A_j \in C_\nu}} d(i, j)$	beliebig	Die Distanz D zweier Cluster C_μ und C_ν wird durch die Distanz d derjenigen Objekte aus C_μ und C_ν bestimmt, die die kleinste Distanz liefern.
Complete-Linkage (Entferntester Nachbar)	$D(C_\mu, C_\nu) = \max_{\substack{A_i \in C_\mu \\ A_j \in C_\nu}} d(i, j)$	beliebig	Die Distanz D zweier Cluster C_μ und C_ν wird durch die Distanz d derjenigen Objekte aus C_μ und C_ν bestimmt, die die größte Distanz liefern.
Average-Linkage (Linkage zwischen den Gruppen)	$D(C_\mu, C_\nu) = \frac{1}{n_\mu n_\nu} \sum_{A_i \in C_\mu} \sum_{A_j \in C_\nu} d(i, j)$ mit $n_\mu = C_\mu $: Anzahl der Objekte im Cluster C_μ und $n_\nu = C_\nu $: Anzahl der Objekte im Cluster C_ν	beliebig	Die Distanz D zweier Cluster C_μ und C_ν wird durch den Durchschnitt aller Distanzen zwischen den Objekten aus C_μ und C_ν bestimmt, d.h.: Im Mittel sind die Objekte der beiden (kompakten) Cluster ähnlich.
Zentroid	$D(C_\mu, C_\nu) = d_2^2(\bar{x}_\mu, \bar{x}_\nu) = \ \bar{x}_\mu - \bar{x}_\nu\ ^2,$ $\bar{x}_\mu = \frac{1}{n_\mu} \sum_{A_i \in C_\mu} x_i: \text{Zentroid der Klasse } C_\mu$ $\bar{x}_\nu = \frac{1}{n_\nu} \sum_{A_j \in C_\nu} x_j: \text{Zentroid der Klasse } C_\nu$	metrisch	Die Distanz D zweier Cluster C_μ und C_ν wird durch die quadrierte euklidische Distanz der Zentroide \bar{x}_μ und \bar{x}_ν der beiden Cluster bestimmt, d.h.: Im Mittel sind die Objekte der beiden Cluster ähnlich.
Ward	$D(C_\mu, C_\nu) = \frac{n_\mu n_\nu}{n_\mu + n_\nu} d_2^2(\bar{x}_\mu, \bar{x}_\nu)$ $= \frac{n_\mu n_\nu}{n_\mu + n_\nu} \ \bar{x}_\mu - \bar{x}_\nu\ ^2$	metrisch	Die Distanz D zweier Cluster C_μ und C_ν wird so bestimmt, dass sie den Homogenitätsverlust, der bei Fusionierung der beiden Cluster auftritt, beschreibt.

Rekursive Berechnung der Clusterdistanzen

In jedem Schritt entsteht durch die Fusion von zwei Clustern ein neues Cluster. Die Distanzen zwischen dem neuen Cluster und den verbleibenden Clustern lassen sich rekursiv berechnen. Die Rekursionsformel zur Berechnung einer Distanz zwischen einem fusionierten Cluster $C_u \cup C_w$ und einem verbleibenden Cluster C_λ ist von der Fusionierungsmethode, d.h. von dem verwendeten Distanzmaß für zwei Cluster, abhängig.

Methode	Rekursionsformel
Single-Linkage (Nächstgelegener Nachbar)	$D(C_u \cup C_w, C_\lambda) = \min\{D(C_u, C_\lambda), D(C_w, C_\lambda)\}$ $= \frac{1}{2}D(C_u, C_\lambda) + \frac{1}{2}D(C_w, C_\lambda) - \frac{1}{2} D(C_u, C_\lambda) - D(C_w, C_\lambda) $
Complete-Linkage (Entferntester Nachbar)	$D(C_u \cup C_w, C_\lambda) = \max\{D(C_u, C_\lambda), D(C_w, C_\lambda)\}$ $= \frac{1}{2}D(C_u, C_\lambda) + \frac{1}{2}D(C_w, C_\lambda) + \frac{1}{2} D(C_u, C_\lambda) - D(C_w, C_\lambda) $
Average-Linkage (Linkage zwischen den Gruppen)	$D(C_u \cup C_w, C_\lambda) = \frac{n_u}{n_u + n_w} D(C_u, C_\lambda) + \frac{n_w}{n_u + n_w} D(C_w, C_\lambda)$
Zentroid	$D(C_u \cup C_w, C_\lambda) = \frac{n_u}{n_u + n_w} D(C_u, C_\lambda) + \frac{n_w}{n_u + n_w} D(C_w, C_\lambda)$ $- \frac{n_u n_w}{(n_u + n_w)^2} D(C_u, C_w)$
Ward	$D(C_u \cup C_w, C_\lambda) = \frac{n_u + n_\lambda}{n_u + n_w + n_\lambda} D(C_u, C_\lambda) + \frac{n_w + n_\lambda}{n_u + n_w + n_\lambda} D(C_w, C_\lambda)$ $- \frac{n_\lambda}{n_u + n_w + n_\lambda} D(C_u, C_w)$

In Stufe $\tau = 1$ des iterativen Verfahrens der Clusterfusionierung werden die Distanzen zwischen den Objekten mit einem vorgegebenen Distanzmaß für zwei Objekte berechnet und diejenigen beiden Objekte fusioniert, welche die kleinste Distanz aufweisen. In den Stufen $\tau \geq 2$ werden die Distanzen zwischen dem in der Stufe $\tau - 1$ fusionierten Cluster und den verbleibenden Clustern bei vorgegebener Methode nach der Rekursionsformel berechnet, mit allen (in Stufe $\tau - 1$ berechneten) Distanzen zwischen den verbleibenden Clustern verglichen und diejenigen Cluster mit der kleinsten Distanz fusioniert.

Dendrogramm

In einem Dendrogramm (Stammbaum) wird der Ablauf der Clusterbildung von der ersten bis zur letzten Stufe des iterativen Verfahrens grafisch veranschaulicht. Da die Darstellung in einem Koordinatensystem erfolgt, in dem auf der einen Achse die Objekte und auf der anderen Achse die Heterogenitätsindizes abgetragen werden, wird nicht nur abgebildet, welche Cluster auf einer Stufe fusioniert werden, sondern auch wie groß die Distanz zwischen den zusammengefassten Clustern ist.

Ergebnisseite der agglomerativen Clusteranalyse mit WinSTAT für Excel

Mit WinSTAT kann eine Clusteranalyse mit hierarchischen agglomerativen Verfahren für metrische Variable durchgeführt werden. Die Beobachtungswerte werden zur Vergleichbarkeit zunächst z-transformiert und als Distanzmaß für die Objekte wird die **quadrierte euklidische Distanz** verwendet. Zur Fusionierung der Cluster kann eine der auf Seite 58 beschriebenen Methoden gewählt werden.

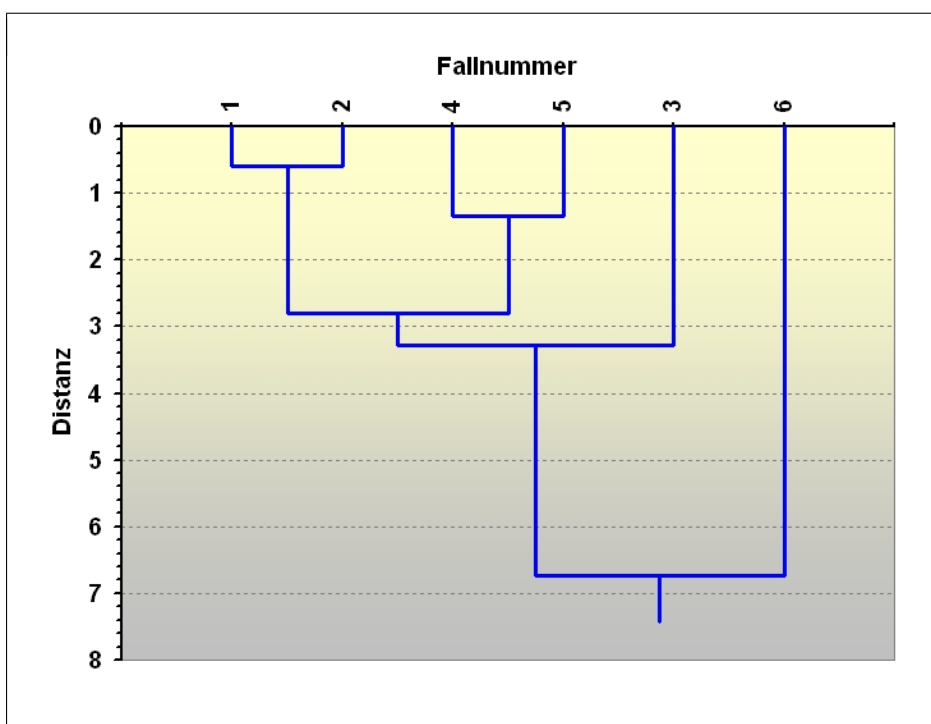
Clusteranalyse

Messvariable: X_1
 \vdots
 X_k

Agglomerationsmethode: z.B. Durchschnittsverbindung (Average-Linkage)

z.B.:

Schritt τ	verbinde		mit		Distanz	
	Cluster 1	Größe	Cluster 2	Größe	Heterogenitätsindex h_τ	
$\tau = 1$	1	1	2	1	$h_1 = D(\{A_1\}, \{A_2\}) = d_2^2(12)$	= 0,586
$\tau = 2$	4	1	5	1	$h_2 = D(\{A_4\}, \{A_5\}) = d_2^2(45)$	= 1,344
$\tau = 3$	1	2	4	2	$h_3 = D(\{A_1\} \cup \{A_2\}, \{A_4\} \cup \{A_5\})$	= 2,801
$\tau = 4$	1	4	3	1	$h_4 = D(\{A_1, A_2\} \cup \{A_4, A_5\}, \{A_3\})$	= 3,288
$\tau = 5$	1	5	6	1	$h_5 = D(\{A_1, A_2, A_4, A_5\} \cup \{A_3\}, \{A_6\})$	= 6,742



WinSTAT bietet die Möglichkeit, eine Clustertrennung vorzunehmen, d.h. eine Anzahl der Cluster vorzugeben und die hierbei zusammengefassten Objekte durch Benennung mit derselben Clusternummer in einer Spalte des Tabellenblattes auszuweisen. Der Spalte wird zuvor ein Variablenname, z.B. „Cluster“, gegeben, so dass mit der neuen Variablen anschließend eine Diskriminanzanalyse durchgeführt werden kann.

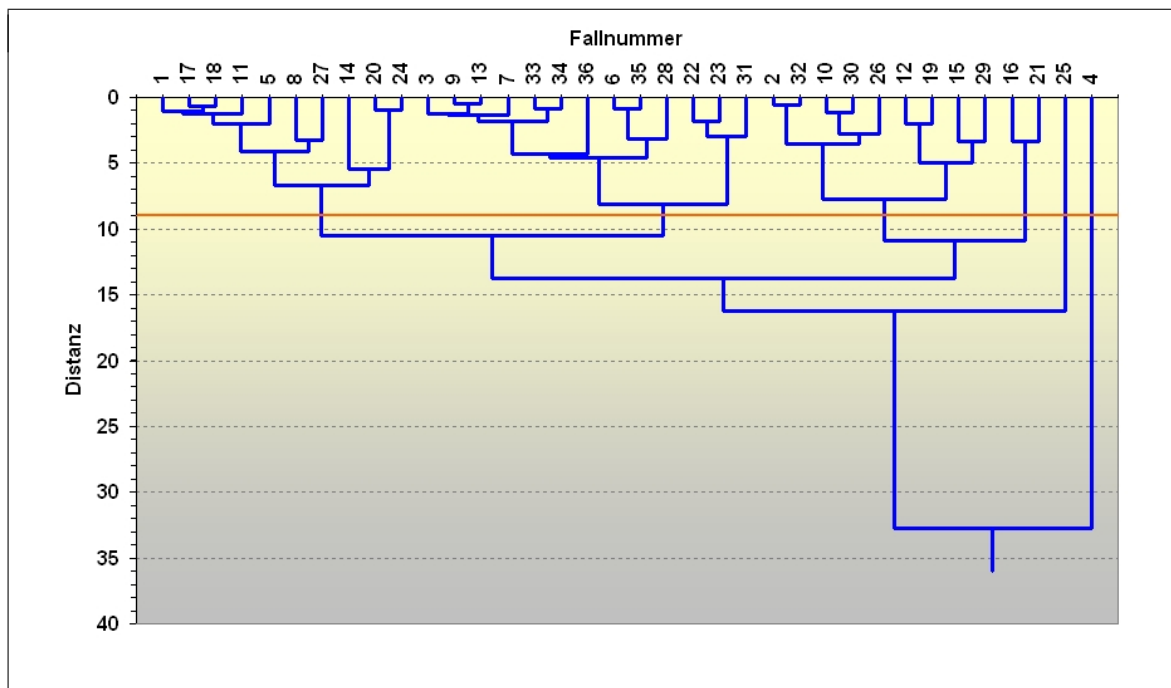
**Auf-
gabe**

6

Für den Beispieldatensatz Seite 1 mit den sechs Merkmalsvariablen: Ausgaben für Kopien, Nettoeinkommen, Zeit für Nacharbeitung und Klausurvorbereitung, Verweildauer im Internet, Aufenthaltsdauer in Kinos, Discos oder Kneipen und der Anzahl gekaufter Fachbücher erhält man den unten stehenden WinSTAT-Output einer Clusteranalyse.

- Bilden Sie anhand des Dendrogramms sechs Cluster.
- Charakterisieren Sie die Studierenden der Cluster.
- Interpretieren Sie die mittleren Diskriminanzkoeffizienten einer Diskriminanzanalyse für die sechs Cluster.

Clusteranalyse Agglomerationsmethode: Durchschnittsverbindung (Average-Linkage)



Diskriminanzanalyse

	mittlerer Diskriminanz- koeffizient
Ausgaben für Kopien	0,339
Nettoeinkommen	0,810
Vorbereitungszeit	0,426
Verweildauer im Internet	0,373
Aufenthaltsdauer in Kinos ...	0,526
Anzahl gekaufter Fachbücher	0,268

Die mittleren Diskriminanzkoeffizienten wurden gemäß der Formel

$$\bar{b}_q = \sum_{j=1}^t |b_{qj}^*| \frac{\gamma_j}{\sum_{j=1}^t \gamma_j}$$

Seite 24 mit Excel aus dem WinSTAT-Output der Eigenwerte γ_j und standardisierten Diskriminanzkoeffizienten b_{qj}^* der Diskriminanzfunktionen berechnet.

Mit WinSTAT kann durch eine Diskriminanzanalyse eine Person mit den Angaben: Ausgaben für Kopien: 30 €/Sem, Semesternettoeinkommen: 2 000€, Vor- und Nacharbeitungszeit: 144 Std/Sem, Verweildauer im Internet: 90 Std/Sem, Aufenthaltsdauer in Kinos, ... : 72 Std/Sem, Anzahl gekaufter Fachbücher: 4 einem der 6 Cluster zugeordnet werden (vgl. Seite 25 und 31). Welche Zuordnung vermuten Sie?

6 Data Mining

Die Anwendung geeigneter Verfahren zur Extraktion von Informationen durch Identifizieren von bedeutsamen und aussagekräftigen Mustern eines Datenbestandes wird als Data Mining bezeichnet. Somit zählen die in Kapitel 4 und 5 dargestellten strukturentdeckenden Verfahren der Faktoren- und Clusteranalyse zu Data Mining. In diesem Kapitel werden die Entscheidungsbaumverfahren herausgegriffen. Als Beispiel eines Klassifikationsbaums wird das von Kaas (1980) stammende (klassische) bzw. von Biggs et al. (1991) vorgeschlagene (exhaustive) CHAID-Verfahren und als Beispiel eines Klassifikations- und Regressionsbaums das von Breimann et al. (1984) entwickelte C&RT-Verfahren vorgestellt. Mit den ermittelten homogenen Teilgruppen der Entscheidungsbäume kann die Streuung einer metrischen abhängigen Variablen bzw. die Trennbarkeit der Gruppen einer nominalen abhängigen Variablen begründet werden. Die extrahierten Informationen können verwendet werden, um für ein neues Objekt Schätzungen der metrischen Zielvariablen vorzunehmen oder eine Prognose zu erstellen, welcher Kategorie der abhängigen nominalen Zielvariablen das neue Objekt zuzuordnen ist. Vorteil gegenüber den strukturprüfenden Methoden der Regressions-, Varianz- und Diskriminanzanalyse ist, dass die unabhängigen Variablen beliebig skalierbar sein können. Allerdings werden beim CHAID-Algorithmus metrische unabhängige Variable ordinal kategorisiert – im Unterschied zum C&RT-Algorithmus, weshalb dieser dem CHAID-Algorithmus überlegen ist, wenn auch metrische unabhängige Variable vorliegen. Der CHAID-Algorithmus kann jedoch im Unterschied zum C&RT-Algorithmus nicht binäre Entscheidungsbäume bilden. Weitere Vorteile von Entscheidungsbäumen bestehen darin, dass kein linearer Zusammenhang zwischen den Variablen oder eine bestimmte Verteilung der Variablen gefordert werden.

Der CHAID-Algorithmus von Clementine (SPSS)

Ziel des CHAID (**C**hi-squared **a**utomatic **i**nteractive **d**etector)-Algorithmus ist es, eine Menge von Objekten so in Gruppen aufzuteilen, dass sich die Gruppen bezüglich eines vorgegebenen Kriteriums möglichst deutlich voneinander unterscheiden. Das Kriterium wird durch eine abhängige Variable (Zielvariable) vorgegeben. Als unabhängige (erklärende) Variablen werden Merkmale gewählt, die für die Bildung von Gruppen geeignet erscheinen. Der CHAID-Algorithmus besteht im wesentlichen aus zwei Schritten:

- 1) **Zusammenfassung von Kategorien der unabhängigen Variablen:** Mit Hilfe des χ^2 -Tests bei einer nominalen abhängigen Variablen und des ANOVA- F -Tests bei einer metrischen abhängigen Variablen wird für jede unabhängige (kategorisierte) Variable mit mehr als 2 Merkmalsausprägungen getestet, ob (bei ordinalen oder metrischen Merkmalen benachbarte) Kategorien zusammengefasst werden können. Dies ist dann der Fall, wenn für zwei getestete Kategorien bzgl. der abhängigen Variablen kein signifikanter Unterschied besteht. Dieser Test wird solange wiederholt bis keine Zusammenfassungen zu Kategorienpaaren einer unabhängigen Variablen mehr möglich sind.
- 2) **Auswahl der Variablen zur Unterteilung des CHAID-Baumes:** Mit Hilfe des χ^2 -Tests bei einer nominalen abhängigen Variablen und des ANOVA- F -Tests bei einer metrischen abhängigen Variablen wird für jede unabhängige (kategorisierte) Variable der Zusammenhang zwischen der abhängigen und unabhängigen Variablen überprüft. Für die Unterteilung der Stichprobe wird diejenige unabhängige Variable ausgewählt, die das kleinste (unter einem vorgegebenen) Signifikanzniveau aufweist, d.h. für die ein Zusammenhang mit der abhängigen Variablen am wahrscheinlichsten ist. Die ursprünglichen Kategorien bzw. die unter Punkt 1) zusammengefassten Kategorien dieser unabhängigen Variablen bilden dann die Knoten der ersten Ebene des Entscheidungsbaums.

Wurde eine Stichprobe in zwei oder mehr Kategorien (Knoten) unterteilt, wird für jede dieser Kategorien geprüft, ob sie anhand einer der unabhängigen Variablen weiter unterteilt werden soll. Hierzu werden für jede Kategorie, d.h. nur für die Daten der Stichprobe, die zu der Kategorie gehören, die Schritte 1) und 2) wiederholt. Die CHAID-Analyse ist beendet, wenn sich für keine der in Frage kommenden unabhängigen Variablen ein signifikanter Zusammenhang mit der abhängigen Variablen für die betrachtete Kategorie ergibt. Im Gegensatz zur C&RT-Analyse beendet der CHAID-Algorithmus das Wachsen des Baumes, bevor der Baum zu groß geworden ist, so dass hinterher kein Stutzen mit einer Pruning-Methode notwendig ist. Exhaustive CHAID ist eine Variante von CHAID, die *alle* von einer unabhängigen Variablen möglichen Zerlegungen (und nicht nur Kategorienpaare) auf signifikante Unterschiede analysiert.

Der C&RT-Algorithmus von Clementine (SPSS)

Ziel des C&RT (Classification And Regression Trees, auch CART genannt)- Algorithmus ist es, mit Hilfe einer unabhängigen Variablen die Daten in zwei disjunkte Teilmengen (Unter-knoten) mit möglichst hoher Homogenität zu zerlegen. Hierzu werden für jede unabhängige Variable alle möglichen (bei ordinalen oder metrischen Variablen nach Ordnen der Daten aneinandergrenzende) Dichotomisierungen und ihre zugehörigen Inhomogenitäten (impurity measures) bestimmt. Diejenige Variable, bei der die Inhomogenitäten zweier Teilmengen gegenüber der Inhomogenität der gesamten Datenmenge (Wurzelknoten) am stärksten sinken, wird für die Unterteilung der Stichprobe ausgewählt, d.h. die zwei Teilmengen dieser unabhängigen Variablen bilden dann die Knoten der ersten Ebene des binären Entscheidungsbaums. Mit demselben Verfahren wird für die beiden Teilmengen (Knoten) eine Zerlegung in zwei homogene Teilmengen gesucht. Die C&RT-Analyse ist beendet, wenn bei keiner Zerlegung eines Knotens eine Mindeständerung der Inhomogenität erreicht wird.

Data Mining des Beispieldatensatzes mit Clementine von SPSS

Die Daten des fiktiven Beispieldatensatzes Seite 1 wurden mit Data Mining Verfahren des Programmpakets Clementine von SPSS analysiert, um zu klären, mit welchen Variablen sich die unterschiedliche Höhe der Kopierausgaben von Studierenden einer Statistikvorlesung erklären lässt.

Ergebnis einer CHAID-Analyse

Unter Einbeziehung aller Daten des Datensatzes resultiert unter Verwendung des (klassischen oder exhaustiven) CHAID-Algorithmus der auf den Seiten 65/66 dargestellte Entscheidungsbaum. Hieraus ist zu entnehmen:

- Die erwartete Leistung in der Statistik Klausur beeinflusst die Kopierausgaben am stärksten, was sich daran ablesen lässt, dass dieses Merkmal auf der obersten Ebene in die Analyse einbezogen wird. Die Beziehung gestaltet sich folgendermaßen: Je schlechter die erwartete Leistung in der Statistik Klausur, desto höher die Kopierausgaben. Studierende, die in der Statistik Klausur ein unterdurchschnittliches Ergebnis erwarten, geben am meisten für Kopien aus, da sie wohl meinen, dadurch ihre Chancen bei der Klausur verbessern zu können. Dieses Ergebnis ist unabhängig von Geschlecht, Einkommen, Fachbereich etc.
- Konsequenterweise weist Knoten 1, dem sämtliche 10 Studierende angehören, die eine unterdurchschnittliche Leistung in der Statistik Klausur erwarten, mit durchschnittlich 40€ die höchsten Kopierausgaben auf.

- Studierender Nr. 23 (= Knoten 28) weist mit 7€ die geringsten Ausgaben für Kopien aus. Ausschlaggebend hierfür sind seine Merkmale:
 - Erwartete überdurchschnittliche Leistung in der Statistiklausur
 - Geschlecht: weiblich
 - Verweildauer im Internet < 12 Stunden.
- Die Mathematiknote und die Fachrichtung beeinflussen die Kopierausgaben nicht.
- Insgesamt lässt sich am vorgestellten CHAID-Baum folgendes für den Datensatz ablesen:
 - Männliche Studierende kopieren grundsätzlich mehr als ihre Kommilitoninnen.
 - Studenten machen ihre Entscheidung, wie viel sie für Kopien ausgeben, vom Einkommen abhängig: Studenten mit einem höheren Einkommen haben höhere Ausgaben.
 - Die Höhe der Kopierausgaben wird von der Vorbereitungs- bzw. Nacharbeitungszeit beeinflusst: Nehmen sich die Studenten weniger Zeit fürs Studium, so kopieren sie mehr.
 - Die Ausgaben für Kopien hängen von der Anzahl gekaufter Fachbücher ab: Kaufen die Studenten mehr Fachbücher, so müssen sie weniger kopieren.
 - Studentinnen entscheiden über die Höhe ihrer Ausgaben für Kopien unabhängig vom Einkommen, sondern in erster Linie aufgrund ihrer Zeit zur Nacharbeitung und Klausurvorbereitung und danach bei höherem Zeitengagement für das Studium in Abhängigkeit der Anzahl gekaufter Fachbücher. Ebenso wie bei den Studenten gilt hierbei: Weniger Zeit für das Studium oder weniger Fachbücher führen zu höheren Ausgaben für Kopien.
 - Bei Studierenden, die ein überdurchschnittliches Ergebnis in der Statistiklausur erwarten, folgt: Die Ausgaben für Kopien sind bei männlichen Studierenden von ihrer Zeit, die sie in Kinos etc. verbringen, abhängig: Studenten mit hoher Aufenthaltsdauer in Kinos etc. geben hierfür viel Geld dafür wenig für Kopien aus. Studentinnen kopieren dann mehr, wenn sie viel Zeit im Internet verbringen.

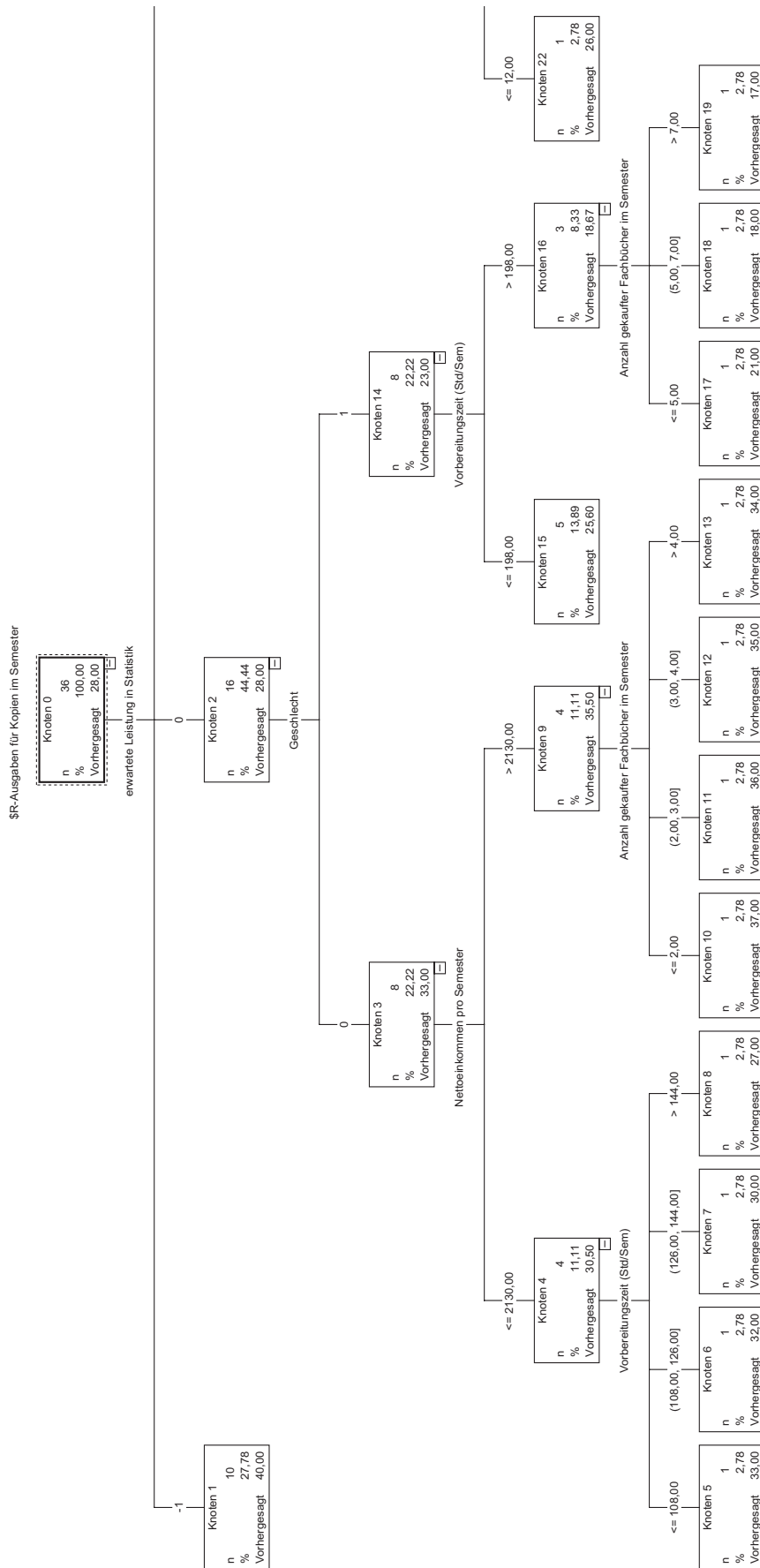
Ergebnis einer C&RT-Analyse

Der mit dem CHAID-Algorithmus entwickelte Entscheidungsbaum eignet sich sehr gut für die Klassifikation, die gut erklärbar und leicht nachvollziehbar ist. Da unabhängige metrische Variable einbezogen werden, ist für die Prognose jedoch die C&RT-Analyse besser geeignet. Bei der Prognose der Kopierausgaben weist Clementine für die CHAID-Analyse einen mittleren absoluten Fehler von 1,389 und eine Standardabweichung von 2,797 für die Abweichungen der tatsächlichen von den geschätzten Werten für Kopierausgaben aus. Bei Anwendung des C&RT-Algorithmus ergibt sich ein mittlerer absoluter Fehler von 0,25 und eine Standardabweichung von 0,487.

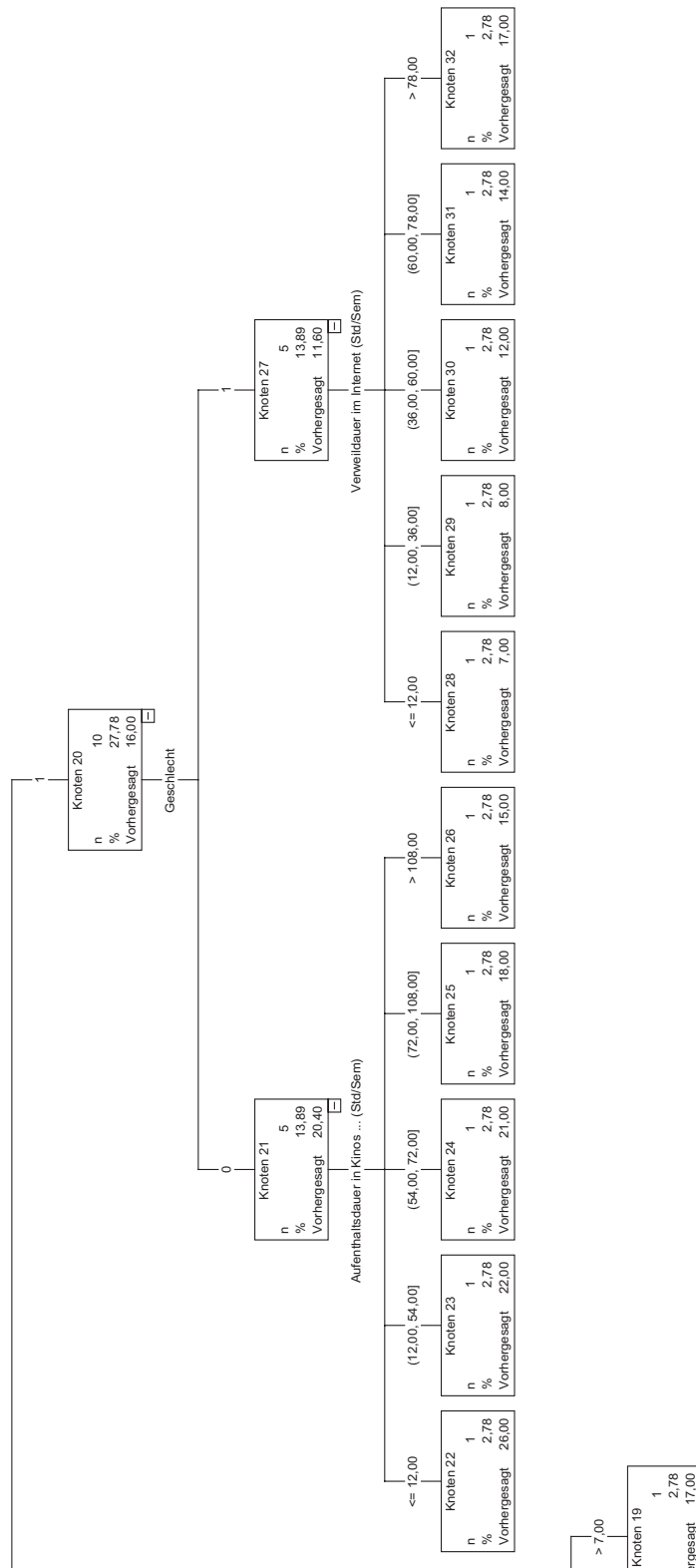
Aufgabe

7

Interpretieren Sie den Entscheidungsbaum des C&RT-Algorithmus, Seite 67. Bei einer Befragung im folgenden Jahr der Erhebung des Beispieldatensatzes machte ein Student der Betriebswirtschaftslehre folgende Angaben: Semesternettoeinkommen: 2 000€, Vorbereitungs- und Nacharbeitungszeit: 144 Std/Sem, Verweildauer im Internet: 90 Std/Sem, Aufenthaltsdauer in Kinos, Discos, Kneipen: 60 Std/Sem, erwartete Leistung in der Statistiklausur: 0. Schätzen Sie die Höhe der Ausgaben für Kopien des Studenten mit dem Entscheidungsbaum des CHAID-Algorithmus, Seite 65/66, sowie des C&RT-Algorithmus, Seite 67. Vergleichen Sie Ihre Ergebnisse mit der Schätzung einer Regressionsanalyse in Aufgabe 1 und einer 2-faktoriellen Varianzanalyse in Aufgabe 3.

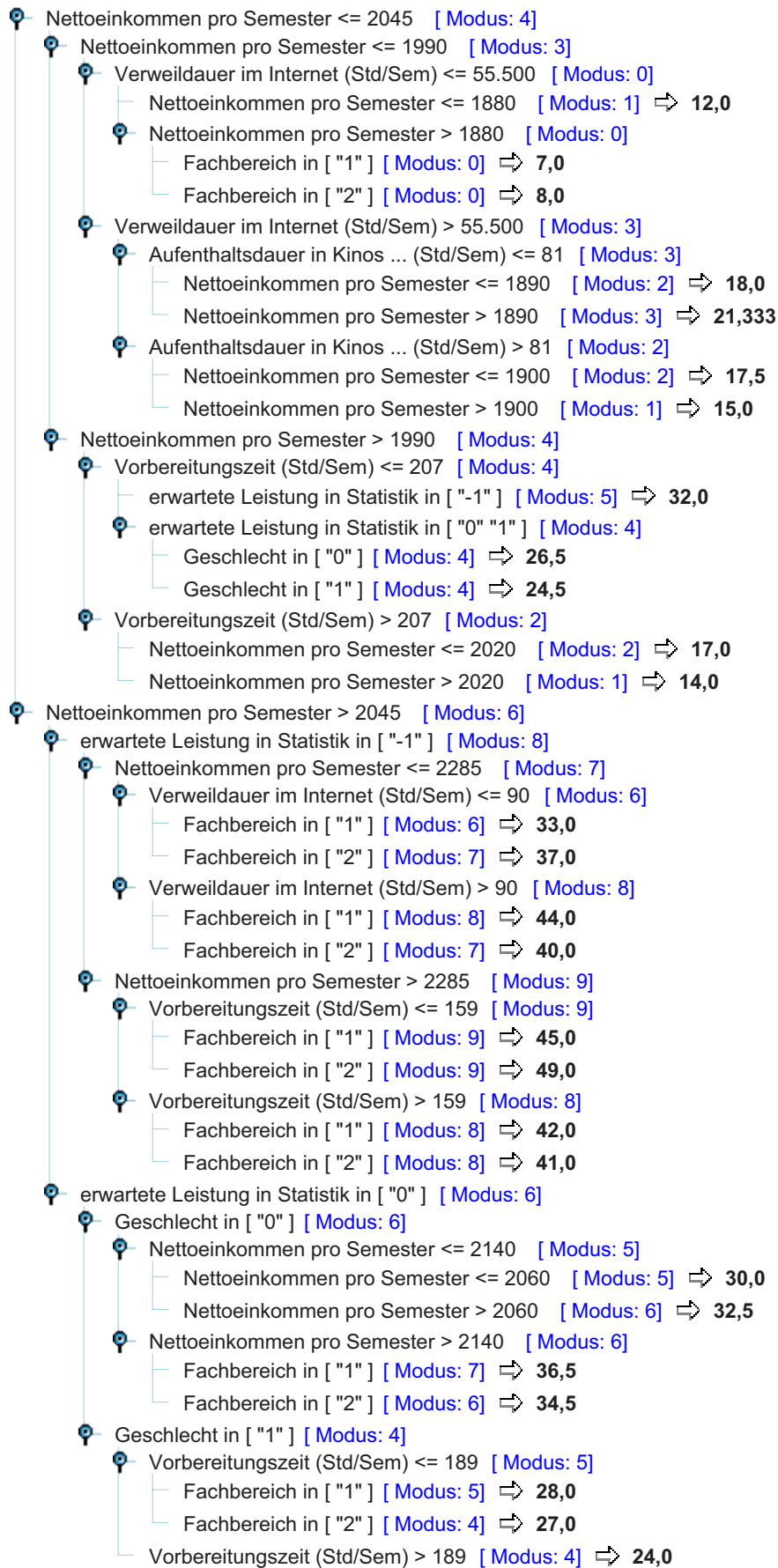


- Knoten 1: Stud.-Nr.: 4, 7, 9, 12, 15, 16, 19, 21, 32, 33
- Knoten 5: Stud.-Nr.: 2
- Knoten 6: Stud.-Nr.: 30
- Knoten 7: Stud.-Nr.: 29
- Knoten 8: Stud.-Nr.: 25
- Knoten 10: Stud.-Nr.: 26
- Knoten 11: Stud.-Nr.: 10
- Knoten 12: Stud.-Nr.: 34
- Knoten 13: Stud.-Nr.: 13
- Knoten 15: Stud.-Nr.: 1, 3, 6, 28, 35
- Knoten 17: Stud.-Nr.: 18
- Knoten 18: Stud.-Nr.: 11
- Knoten 19: Stud.-Nr.: 22

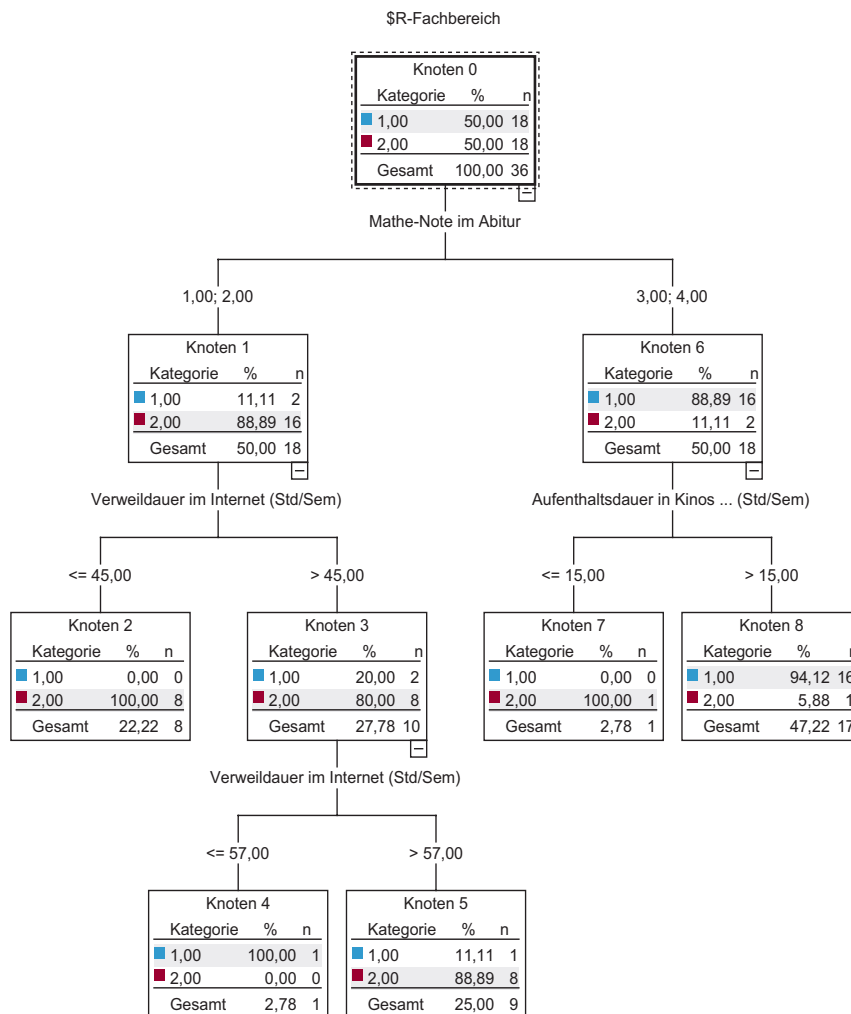


Knoten 28: Stud.-Nr.: 23
 Knoten 29: Stud.-Nr.: 31
 Knoten 30: Stud.-Nr.: 14
 Knoten 31: Stud.-Nr.: 24
 Knoten 32: Stud.-Nr.: 8

Knoten 22: Stud.-Nr.: 36
 Knoten 23: Stud.-Nr.: 27
 Knoten 24: Stud.-Nr.: 17
 Knoten 25: Stud.-Nr.: 5
 Knoten 26: Stud.-Nr.: 20



Unter Verwendung des C&RT-Algorithmus ergibt sich folgender Entscheidungsbaum für die Analyse, mit welchen Variablen sich die Fachbereichsentscheidung eines Studierenden erklären lässt. Als mögliche Einflussvariable wurden das Geschlecht, die Mathematiknote, die Aufenthaltsdauer in Kinos etc. und die Verweildauer im Internet gewählt. Ergebnis (hier werden alte Klischees bedient!): Studierende der Betriebswirtschaftslehre verbringen mehr Zeit in Kneipen etc., Studierende der Informationstechnologien verbringen mehr Zeit im Internet.



- Knoten 2: Fachbereich 1: 0
 Fachbereich 2: Stud.-Nr.: 2, 4, 6, 13, 22, 28, 31, 36
- Knoten 4: Fachbereich 1: Stud.-Nr.: 3
 Fachbereich 2: 0
- Knoten 5: Fachbereich 1: Stud.-Nr.: 10
 Fachbereich 2: Stud.-Nr.: 8, 15, 17, 20, 21, 24, 33, 34
- Knoten 7: Fachbereich 1: 0
 Fachbereich 2: Stud.-Nr.: 32
- Knoten 8: Fachbereich 1: Stud.-Nr.: 1, 5, 7, 9, 11, 12, 14, 16, 19, 23, 25, 26, 27, 29, 30, 35
 Fachbereich 2: Stud.-Nr.: 18

Anhang: Tafeln zu einigen wichtigen Verteilungen

A Standardnormalverteilung

Vertafelt sind die Werte der Verteilungsfunktion $F(z) = P(Z \leq z)$ für $z \geq 0$.

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998
3,5	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998
3,6	0,9998	0,9998	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999

B t -Verteilung

Vertafelt sind die Werte von t zu gegebenen Werten der Verteilungsfunktion für ν Freiheitsgrade. Für $t_{1-\alpha}(\nu)$ gilt $F(t_{1-\alpha}(\nu)) = 1 - \alpha$.

ν	$1 - \alpha$									
	0,600	0,700	0,750	0,800	0,900	0,950	0,975	0,990	0,995	0,999
1	0,325	0,727	1,000	1,376	3,078	6,314	12,706	31,821	63,656	318,289
2	0,289	0,617	0,816	1,061	1,886	2,920	4,303	6,965	9,925	22,328
3	0,277	0,584	0,765	0,978	1,638	2,353	3,182	4,541	5,841	10,214
4	0,271	0,569	0,741	0,941	1,533	2,132	2,776	3,747	4,604	7,173
5	0,267	0,559	0,727	0,920	1,476	2,015	2,571	3,365	4,032	5,894
6	0,265	0,553	0,718	0,906	1,440	1,943	2,447	3,143	3,707	5,208
7	0,263	0,549	0,711	0,896	1,415	1,895	2,365	2,998	3,499	4,785
8	0,262	0,546	0,706	0,889	1,397	1,860	2,306	2,896	3,355	4,501
9	0,261	0,543	0,703	0,883	1,383	1,833	2,262	2,821	3,250	4,297
10	0,260	0,542	0,700	0,879	1,372	1,812	2,228	2,764	3,169	4,144
11	0,260	0,540	0,697	0,876	1,363	1,796	2,201	2,718	3,106	4,025
12	0,259	0,539	0,695	0,873	1,356	1,782	2,179	2,681	3,055	3,930
13	0,259	0,538	0,694	0,870	1,350	1,771	2,160	2,650	3,012	3,852
14	0,258	0,537	0,692	0,868	1,345	1,761	2,145	2,624	2,977	3,787
15	0,258	0,536	0,691	0,866	1,341	1,753	2,131	2,602	2,947	3,733
16	0,258	0,535	0,690	0,865	1,337	1,746	2,120	2,583	2,921	3,686
17	0,257	0,534	0,689	0,863	1,333	1,740	2,110	2,567	2,898	3,646
18	0,257	0,534	0,688	0,862	1,330	1,734	2,101	2,552	2,878	3,610
19	0,257	0,533	0,688	0,861	1,328	1,729	2,093	2,539	2,861	3,579
20	0,257	0,533	0,687	0,860	1,325	1,725	2,086	2,528	2,845	3,552
21	0,257	0,532	0,686	0,859	1,323	1,721	2,080	2,518	2,831	3,527
22	0,256	0,532	0,686	0,858	1,321	1,717	2,074	2,508	2,819	3,505
23	0,256	0,532	0,685	0,858	1,319	1,714	2,069	2,500	2,807	3,485
24	0,256	0,531	0,685	0,857	1,318	1,711	2,064	2,492	2,797	3,467
25	0,256	0,531	0,684	0,856	1,316	1,708	2,060	2,485	2,787	3,450
26	0,256	0,531	0,684	0,856	1,315	1,706	2,056	2,479	2,779	3,435
27	0,256	0,531	0,684	0,855	1,314	1,703	2,052	2,473	2,771	3,421
28	0,256	0,530	0,683	0,855	1,313	1,701	2,048	2,467	2,763	3,408
29	0,256	0,530	0,683	0,854	1,311	1,699	2,045	2,462	2,756	3,396
30	0,256	0,530	0,683	0,854	1,310	1,697	2,042	2,457	2,750	3,385
40	0,255	0,529	0,681	0,851	1,303	1,684	2,021	2,423	2,704	3,307
50	0,255	0,528	0,679	0,849	1,299	1,676	2,009	2,403	2,678	3,261
100	0,254	0,526	0,677	0,845	1,290	1,660	1,984	2,364	2,626	3,174
150	0,254	0,526	0,676	0,844	1,287	1,655	1,976	2,351	2,609	3,145
∞	0,253	0,524	0,674	0,842	1,282	1,645	1,960	2,326	2,576	3,090

C Chi-Quadrat-Verteilung

Vertafelt sind die Werte von χ^2 zu gegebenen Werten der Verteilungsfunktion für ν Freiheitsgrade. Für $\chi^2_{1-\alpha}(\nu)$ gilt $F(\chi^2_{1-\alpha}(\nu)) = 1 - \alpha$. Approximation für $\nu > 35$: $\chi^2_{1-\alpha}(\nu) \approx \frac{1}{2}(z_{1-\alpha} + \sqrt{2\nu - 1})^2$.

ν	$1 - \alpha$									
	0,600	0,700	0,800	0,900	0,950	0,975	0,980	0,990	0,995	0,999
1	0,708	1,074	1,642	2,706	3,841	5,024	5,412	6,635	7,879	10,827
2	1,833	2,408	3,219	4,605	5,991	7,378	7,824	9,210	10,597	13,815
3	2,946	3,665	4,642	6,251	7,815	9,348	9,837	11,345	12,838	16,266
4	4,045	4,878	5,989	7,779	9,488	11,143	11,668	13,277	14,860	18,466
5	5,132	6,064	7,289	9,236	11,070	12,832	13,388	15,086	16,750	20,515
6	6,211	7,231	8,558	10,645	12,592	14,449	15,033	16,812	18,548	22,457
7	7,283	8,383	9,803	12,017	14,067	16,013	16,622	18,475	20,278	24,321
8	8,351	9,524	11,030	13,362	15,507	17,535	18,168	20,090	21,955	26,124
9	9,414	10,656	12,242	14,684	16,919	19,023	19,679	21,666	23,589	27,877
10	10,473	11,781	13,442	15,987	18,307	20,483	21,161	23,209	25,188	29,588
11	11,530	12,899	14,631	17,275	19,675	21,920	22,618	24,725	26,757	31,264
12	12,584	14,011	15,812	18,549	21,026	23,337	24,054	26,217	28,300	32,909
13	13,636	15,119	16,985	19,812	22,362	24,736	25,471	27,688	29,819	34,527
14	14,685	16,222	18,151	21,064	23,685	26,119	26,873	29,141	31,319	36,124
15	15,733	17,322	19,311	22,307	24,996	27,488	28,259	30,578	32,801	37,698
16	16,780	18,418	20,465	23,542	26,296	28,845	29,633	32,000	34,267	39,252
17	17,824	19,511	21,615	24,769	27,587	30,191	30,995	33,409	35,718	40,791
18	18,868	20,601	22,760	25,989	28,869	31,526	32,346	34,805	37,156	42,312
19	19,910	21,689	23,900	27,204	30,144	32,852	33,687	36,191	38,582	43,819
20	20,951	22,775	25,038	28,412	31,410	34,170	35,020	37,566	39,997	45,314
21	21,992	23,858	26,171	29,615	32,671	35,479	36,343	38,932	41,401	46,796
22	23,031	24,939	27,301	30,813	33,924	36,781	37,659	40,289	42,796	48,268
23	24,069	26,018	28,429	32,007	35,172	38,076	38,968	41,638	44,181	49,728
24	25,106	27,096	29,553	33,196	36,415	39,364	40,270	42,980	45,558	51,179
25	26,143	28,172	30,675	34,382	37,652	40,646	41,566	44,314	46,928	52,619
26	27,179	29,246	31,795	35,563	38,885	41,923	42,856	45,642	48,290	54,051
27	28,214	30,319	32,912	36,741	40,113	43,195	44,140	46,963	49,645	55,475
28	29,249	31,391	34,027	37,916	41,337	44,461	45,419	48,278	50,994	56,892
29	30,283	32,461	35,139	39,087	42,557	45,722	46,693	49,588	52,335	58,301
30	31,316	33,530	36,250	40,256	43,773	46,979	47,962	50,892	53,672	59,702
31	32,349	34,598	37,359	41,422	44,985	48,232	49,226	52,191	55,002	61,098
32	33,381	35,665	38,466	42,585	46,194	49,480	50,487	53,486	56,328	62,487
33	34,413	36,731	39,572	43,745	47,400	50,725	51,743	54,775	57,648	63,869
34	35,444	37,795	40,676	44,903	48,602	51,966	52,995	56,061	58,964	65,247
35	36,475	38,859	41,778	46,059	49,802	53,203	54,244	57,342	60,275	66,619

D F-Verteilung

Vertafelt sind die Werte von f zu gegebenen Werten der Verteilungsfunktion für (v_1, v_2) Freiheitsgrade. Für $f_{1-\alpha}(v_1, v_2)$ gilt $F(f_{1-\alpha}(v_1, v_2)) = 1 - \alpha$.

v_1	$1 - \alpha$	v_2								
		1	2	3	4	5	6	7	8	9
1	0,900	39,864	8,526	5,538	4,545	4,060	3,776	3,589	3,458	3,360
1	0,950	161,446	18,513	10,128	7,709	6,608	5,987	5,591	5,318	5,117
1	0,975	647,793	38,506	17,443	12,218	10,007	8,813	8,073	7,571	7,209
1	0,990	4052,185	98,502	34,116	21,198	16,258	13,745	12,246	11,259	10,562
2	0,900	49,500	9,000	5,462	4,325	3,780	3,463	3,257	3,113	3,006
2	0,950	199,499	19,000	9,552	6,944	5,786	5,143	4,737	4,459	4,256
2	0,975	799,482	39,000	16,044	10,649	8,434	7,260	6,542	6,059	5,715
2	0,990	4999,340	99,000	30,816	18,000	13,274	10,925	9,547	8,649	8,022
3	0,900	53,593	9,162	5,391	4,191	3,619	3,289	3,074	2,924	2,813
3	0,950	215,707	19,164	9,277	6,591	5,409	4,757	4,347	4,066	3,863
3	0,975	864,151	39,166	15,439	9,979	7,764	6,599	5,890	5,416	5,078
3	0,990	5403,534	99,164	29,457	16,694	12,060	9,780	8,451	7,591	6,992
4	0,900	55,833	9,243	5,343	4,107	3,520	3,181	2,961	2,806	2,693
4	0,950	224,583	19,247	9,117	6,388	5,192	4,534	4,120	3,838	3,633
4	0,975	899,599	39,248	15,101	9,604	7,388	6,227	5,523	5,053	4,718
4	0,990	5624,257	99,251	28,710	15,977	11,392	9,148	7,847	7,006	6,422
5	0,900	57,240	9,293	5,309	4,051	3,453	3,108	2,883	2,726	2,611
5	0,950	230,160	19,296	9,013	6,256	5,050	4,387	3,972	3,688	3,482
5	0,975	921,835	39,298	14,885	9,364	7,146	5,988	5,285	4,817	4,484
5	0,990	5763,955	99,302	28,237	15,522	10,967	8,746	7,460	6,632	6,057
6	0,900	58,204	9,326	5,285	4,010	3,405	3,055	2,827	2,668	2,551
6	0,950	233,988	19,329	8,941	6,163	4,950	4,284	3,866	3,581	3,374
6	0,975	937,114	39,331	14,735	9,197	6,978	5,820	5,119	4,652	4,320
6	0,990	5858,950	99,331	27,911	15,207	10,672	8,466	7,191	6,371	5,802
7	0,900	58,906	9,349	5,266	3,979	3,368	3,014	2,785	2,624	2,505
7	0,950	236,767	19,353	8,887	6,094	4,876	4,207	3,787	3,500	3,293
7	0,975	948,203	39,356	14,624	9,074	6,853	5,695	4,995	4,529	4,197
7	0,990	5928,334	99,357	27,671	14,976	10,456	8,260	6,993	6,178	5,613
8	0,900	59,439	9,367	5,252	3,955	3,339	2,983	2,752	2,589	2,469
8	0,950	238,884	19,371	8,845	6,041	4,818	4,147	3,726	3,438	3,230
8	0,975	956,643	39,373	14,540	8,980	6,757	5,600	4,899	4,433	4,102
8	0,990	5980,954	99,375	27,489	14,799	10,289	8,102	6,840	6,029	5,467
9	0,900	59,857	9,381	5,240	3,936	3,316	2,958	2,725	2,561	2,440
9	0,950	240,543	19,385	8,812	5,999	4,772	4,099	3,677	3,388	3,179
9	0,975	963,279	39,387	14,473	8,905	6,681	5,523	4,823	4,357	4,026
9	0,990	6022,397	99,390	27,345	14,659	10,158	7,976	6,719	5,911	5,351

v_1	$1 - \alpha$	v_2									
		10	11	12	13	14	15	16	17	18	19
1	0,900	3,285	3,225	3,177	3,136	3,102	3,073	3,048	3,026	3,007	2,990
1	0,950	4,965	4,844	4,747	4,667	4,600	4,543	4,494	4,451	4,414	4,381
1	0,975	6,937	6,724	6,554	6,414	6,298	6,200	6,115	6,042	5,978	5,922
1	0,990	10,044	9,646	9,330	9,074	8,862	8,683	8,531	8,400	8,285	8,185
2	0,900	2,924	2,860	2,807	2,763	2,726	2,695	2,668	2,645	2,624	2,606
2	0,950	4,103	3,982	3,885	3,806	3,739	3,682	3,634	3,592	3,555	3,522
2	0,975	5,456	5,256	5,096	4,965	4,857	4,765	4,687	4,619	4,560	4,508
2	0,990	7,559	7,206	6,927	6,701	6,515	6,359	6,226	6,112	6,013	5,926
3	0,900	2,728	2,660	2,606	2,560	2,522	2,490	2,462	2,437	2,416	2,397
3	0,950	3,708	3,587	3,490	3,411	3,344	3,287	3,239	3,197	3,160	3,127
3	0,975	4,826	4,630	4,474	4,347	4,242	4,153	4,077	4,011	3,954	3,903
3	0,990	6,552	6,217	5,953	5,739	5,564	5,417	5,292	5,185	5,092	5,010
4	0,900	2,605	2,536	2,480	2,434	2,395	2,361	2,333	2,308	2,286	2,266
4	0,950	3,478	3,357	3,259	3,179	3,112	3,056	3,007	2,965	2,928	2,895
4	0,975	4,468	4,275	4,121	3,996	3,892	3,804	3,729	3,665	3,608	3,559
4	0,990	5,994	5,668	5,412	5,205	5,035	4,893	4,773	4,669	4,579	4,500
5	0,900	2,522	2,451	2,394	2,347	2,307	2,273	2,244	2,218	2,196	2,176
5	0,950	3,326	3,204	3,106	3,025	2,958	2,901	2,852	2,810	2,773	2,740
5	0,975	4,236	4,044	3,891	3,767	3,663	3,576	3,502	3,438	3,382	3,333
5	0,990	5,636	5,316	5,064	4,862	4,695	4,556	4,437	4,336	4,248	4,171
6	0,900	2,461	2,389	2,331	2,283	2,243	2,208	2,178	2,152	2,130	2,109
6	0,950	3,217	3,095	2,996	2,915	2,848	2,790	2,741	2,699	2,661	2,628
6	0,975	4,072	3,881	3,728	3,604	3,501	3,415	3,341	3,277	3,221	3,172
6	0,990	5,386	5,069	4,821	4,620	4,456	4,318	4,202	4,101	4,015	3,939
7	0,900	2,414	2,342	2,283	2,234	2,193	2,158	2,128	2,102	2,079	2,058
7	0,950	3,135	3,012	2,913	2,832	2,764	2,707	2,657	2,614	2,577	2,544
7	0,975	3,950	3,759	3,607	3,483	3,380	3,293	3,219	3,156	3,100	3,051
7	0,990	5,200	4,886	4,640	4,441	4,278	4,142	4,026	3,927	3,841	3,765
8	0,900	2,377	2,304	2,245	2,195	2,154	2,119	2,088	2,061	2,038	2,017
8	0,950	3,072	2,948	2,849	2,767	2,699	2,641	2,591	2,548	2,510	2,477
8	0,975	3,855	3,664	3,512	3,388	3,285	3,199	3,125	3,061	3,005	2,956
8	0,990	5,057	4,744	4,499	4,302	4,140	4,004	3,890	3,791	3,705	3,631
9	0,900	2,347	2,274	2,214	2,164	2,122	2,086	2,055	2,028	2,005	1,984
9	0,950	3,020	2,896	2,796	2,714	2,646	2,588	2,538	2,494	2,456	2,423
9	0,975	3,779	3,588	3,436	3,312	3,209	3,123	3,049	2,985	2,929	2,880
9	0,990	4,942	4,632	4,388	4,191	4,030	3,895	3,780	3,682	3,597	3,523

v_1	$1 - \alpha$	v_2									
		20	21	22	23	24	25	26	27	28	29
1	0,900	2,975	2,961	2,949	2,937	2,927	2,918	2,909	2,901	2,894	2,887
1	0,950	4,351	4,325	4,301	4,279	4,260	4,242	4,225	4,210	4,196	4,183
1	0,975	5,871	5,827	5,786	5,750	5,717	5,686	5,659	5,633	5,610	5,588
1	0,990	8,096	8,017	7,945	7,881	7,823	7,770	7,721	7,677	7,636	7,598
2	0,900	2,589	2,575	2,561	2,549	2,538	2,528	2,519	2,511	2,503	2,495
2	0,950	3,493	3,467	3,443	3,422	3,403	3,385	3,369	3,354	3,340	3,328
2	0,975	4,461	4,420	4,383	4,349	4,319	4,291	4,265	4,242	4,221	4,201
2	0,990	5,849	5,780	5,719	5,664	5,614	5,568	5,526	5,488	5,453	5,420
3	0,900	2,380	2,365	2,351	2,339	2,327	2,317	2,307	2,299	2,291	2,283
3	0,950	3,098	3,072	3,049	3,028	3,009	2,991	2,975	2,960	2,947	2,934
3	0,975	3,859	3,819	3,783	3,750	3,721	3,694	3,670	3,647	3,626	3,607
3	0,990	4,938	4,874	4,817	4,765	4,718	4,675	4,637	4,601	4,568	4,538
4	0,900	2,249	2,233	2,219	2,207	2,195	2,184	2,174	2,165	2,157	2,149
4	0,950	2,866	2,840	2,817	2,796	2,776	2,759	2,743	2,728	2,714	2,701
4	0,975	3,515	3,475	3,440	3,408	3,379	3,353	3,329	3,307	3,286	3,267
4	0,990	4,431	4,369	4,313	4,264	4,218	4,177	4,140	4,106	4,074	4,045
5	0,900	2,158	2,142	2,128	2,115	2,103	2,092	2,082	2,073	2,064	2,057
5	0,950	2,711	2,685	2,661	2,640	2,621	2,603	2,587	2,572	2,558	2,545
5	0,975	3,289	3,250	3,215	3,183	3,155	3,129	3,105	3,083	3,063	3,044
5	0,990	4,103	4,042	3,988	3,939	3,895	3,855	3,818	3,785	3,754	3,725
6	0,900	2,091	2,075	2,060	2,047	2,035	2,024	2,014	2,005	1,996	1,988
6	0,950	2,599	2,573	2,549	2,528	2,508	2,490	2,474	2,459	2,445	2,432
6	0,975	3,128	3,090	3,055	3,023	2,995	2,969	2,945	2,923	2,903	2,884
6	0,990	3,871	3,812	3,758	3,710	3,667	3,627	3,591	3,558	3,528	3,499
7	0,900	2,040	2,023	2,008	1,995	1,983	1,971	1,961	1,952	1,943	1,935
7	0,950	2,514	2,488	2,464	2,442	2,423	2,405	2,388	2,373	2,359	2,346
7	0,975	3,007	2,969	2,934	2,902	2,874	2,848	2,824	2,802	2,782	2,763
7	0,990	3,699	3,640	3,587	3,539	3,496	3,457	3,421	3,388	3,358	3,330
8	0,900	1,999	1,982	1,967	1,953	1,941	1,929	1,919	1,909	1,900	1,892
8	0,950	2,447	2,420	2,397	2,375	2,355	2,337	2,321	2,305	2,291	2,278
8	0,975	2,913	2,874	2,839	2,808	2,779	2,753	2,729	2,707	2,687	2,669
8	0,990	3,564	3,506	3,453	3,406	3,363	3,324	3,288	3,256	3,226	3,198
9	0,900	1,965	1,948	1,933	1,919	1,906	1,895	1,884	1,874	1,865	1,857
9	0,950	2,393	2,366	2,342	2,320	2,300	2,282	2,265	2,250	2,236	2,223
9	0,975	2,837	2,798	2,763	2,731	2,703	2,677	2,653	2,631	2,611	2,592
9	0,990	3,457	3,398	3,346	3,299	3,256	3,217	3,182	3,149	3,120	3,092

v_1	$1 - \alpha$	v_2									
		30	31	32	33	34	35	36	37	38	39
1	0,900	2,881	2,875	2,869	2,864	2,859	2,855	2,850	2,846	2,842	2,839
1	0,950	4,171	4,160	4,149	4,139	4,130	4,121	4,113	4,105	4,098	4,091
1	0,975	5,568	5,549	5,531	5,515	5,499	5,485	5,471	5,458	5,446	5,435
1	0,990	7,562	7,530	7,499	7,471	7,444	7,419	7,396	7,373	7,353	7,333
2	0,900	2,489	2,482	2,477	2,471	2,466	2,461	2,456	2,452	2,448	2,444
2	0,950	3,316	3,305	3,295	3,285	3,276	3,267	3,259	3,252	3,245	3,238
2	0,975	4,182	4,165	4,149	4,134	4,120	4,106	4,094	4,082	4,071	4,061
2	0,990	5,390	5,362	5,336	5,312	5,289	5,268	5,248	5,229	5,211	5,194
3	0,900	2,276	2,270	2,263	2,258	2,252	2,247	2,243	2,238	2,234	2,230
3	0,950	2,922	2,911	2,901	2,892	2,883	2,874	2,866	2,859	2,852	2,845
3	0,975	3,589	3,573	3,557	3,543	3,529	3,517	3,505	3,493	3,483	3,473
3	0,990	4,510	4,484	4,459	4,437	4,416	4,396	4,377	4,360	4,343	4,327
4	0,900	2,142	2,136	2,129	2,123	2,118	2,113	2,108	2,103	2,099	2,095
4	0,950	2,690	2,679	2,668	2,659	2,650	2,641	2,634	2,626	2,619	2,612
4	0,975	3,250	3,234	3,218	3,204	3,191	3,179	3,167	3,156	3,145	3,135
4	0,990	4,018	3,993	3,969	3,948	3,927	3,908	3,890	3,873	3,858	3,843
5	0,900	2,049	2,042	2,036	2,030	2,024	2,019	2,014	2,009	2,005	2,001
5	0,950	2,534	2,523	2,512	2,503	2,494	2,485	2,477	2,470	2,463	2,456
5	0,975	3,026	3,010	2,995	2,981	2,968	2,956	2,944	2,933	2,923	2,913
5	0,990	3,699	3,675	3,652	3,630	3,611	3,592	3,574	3,558	3,542	3,528
6	0,900	1,980	1,973	1,967	1,961	1,955	1,950	1,945	1,940	1,935	1,931
6	0,950	2,421	2,409	2,399	2,389	2,380	2,372	2,364	2,356	2,349	2,342
6	0,975	2,867	2,851	2,836	2,822	2,808	2,796	2,785	2,774	2,763	2,754
6	0,990	3,473	3,449	3,427	3,406	3,386	3,368	3,351	3,334	3,319	3,305
7	0,900	1,927	1,920	1,913	1,907	1,901	1,896	1,891	1,886	1,881	1,877
7	0,950	2,334	2,323	2,313	2,303	2,294	2,285	2,277	2,270	2,262	2,255
7	0,975	2,746	2,730	2,715	2,701	2,688	2,676	2,664	2,653	2,643	2,633
7	0,990	3,304	3,281	3,258	3,238	3,218	3,200	3,183	3,167	3,152	3,137
8	0,900	1,884	1,877	1,870	1,864	1,858	1,852	1,847	1,842	1,838	1,833
8	0,950	2,266	2,255	2,244	2,235	2,225	2,217	2,209	2,201	2,194	2,187
8	0,975	2,651	2,635	2,620	2,606	2,593	2,581	2,569	2,558	2,548	2,538
8	0,990	3,173	3,149	3,127	3,106	3,087	3,069	3,052	3,036	3,021	3,006
9	0,900	1,849	1,842	1,835	1,828	1,822	1,817	1,811	1,806	1,802	1,797
9	0,950	2,211	2,199	2,189	2,179	2,170	2,161	2,153	2,145	2,138	2,131
9	0,975	2,575	2,558	2,543	2,529	2,516	2,504	2,492	2,481	2,471	2,461
9	0,990	3,067	3,043	3,021	3,000	2,981	2,963	2,946	2,930	2,915	2,901

v_1	$1 - \alpha$	v_2										
		40	50	60	70	80	90	100	120	150	200	∞
1	0,900	2,835	2,809	2,791	2,779	2,769	2,762	2,756	2,748	2,739	2,731	2,706
1	0,950	4,085	4,034	4,001	3,978	3,960	3,947	3,936	3,920	3,904	3,888	3,841
1	0,975	5,424	5,340	5,286	5,247	5,218	5,196	5,179	5,152	5,126	5,100	5,024
1	0,990	7,314	7,171	7,077	7,011	6,963	6,925	6,895	6,851	6,807	6,763	6,635
2	0,900	2,440	2,412	2,393	2,380	2,370	2,363	2,356	2,347	2,338	2,329	2,303
2	0,950	3,232	3,183	3,150	3,128	3,111	3,098	3,087	3,072	3,056	3,041	2,996
2	0,975	4,051	3,975	3,925	3,890	3,864	3,844	3,828	3,805	3,781	3,758	3,689
2	0,990	5,178	5,057	4,977	4,922	4,881	4,849	4,824	4,787	4,749	4,713	4,605
3	0,900	2,226	2,197	2,177	2,164	2,154	2,146	2,139	2,130	2,121	2,111	2,084
3	0,950	2,839	2,790	2,758	2,736	2,719	2,706	2,696	2,680	2,665	2,650	2,605
3	0,975	3,463	3,390	3,343	3,309	3,284	3,265	3,250	3,227	3,204	3,182	3,116
3	0,990	4,313	4,199	4,126	4,074	4,036	4,007	3,984	3,949	3,915	3,881	3,782
4	0,900	2,091	2,061	2,041	2,027	2,016	2,008	2,002	1,992	1,983	1,973	1,945
4	0,950	2,606	2,557	2,525	2,503	2,486	2,473	2,463	2,447	2,432	2,417	2,372
4	0,975	3,126	3,054	3,008	2,975	2,950	2,932	2,917	2,894	2,872	2,850	2,786
4	0,990	3,828	3,720	3,649	3,600	3,563	3,535	3,513	3,480	3,447	3,414	3,319
5	0,900	1,997	1,966	1,946	1,931	1,921	1,912	1,906	1,896	1,886	1,876	1,847
5	0,950	2,449	2,400	2,368	2,346	2,329	2,316	2,305	2,290	2,274	2,259	2,214
5	0,975	2,904	2,833	2,786	2,754	2,730	2,711	2,696	2,674	2,652	2,630	2,566
5	0,990	3,514	3,408	3,339	3,291	3,255	3,228	3,206	3,174	3,142	3,110	3,017
6	0,900	1,927	1,895	1,875	1,860	1,849	1,841	1,834	1,824	1,814	1,804	1,774
6	0,950	2,336	2,286	2,254	2,231	2,214	2,201	2,191	2,175	2,160	2,144	2,099
6	0,975	2,744	2,674	2,627	2,595	2,571	2,552	2,537	2,515	2,494	2,472	2,408
6	0,990	3,291	3,186	3,119	3,071	3,036	3,009	2,988	2,956	2,924	2,893	2,802
7	0,900	1,873	1,840	1,819	1,804	1,793	1,785	1,778	1,767	1,757	1,747	1,717
7	0,950	2,249	2,199	2,167	2,143	2,126	2,113	2,103	2,087	2,071	2,056	2,010
7	0,975	2,624	2,553	2,507	2,474	2,450	2,432	2,417	2,395	2,373	2,351	2,288
7	0,990	3,124	3,020	2,953	2,906	2,871	2,845	2,823	2,792	2,761	2,730	2,639
8	0,900	1,829	1,796	1,775	1,760	1,748	1,739	1,732	1,722	1,712	1,701	1,670
8	0,950	2,180	2,130	2,097	2,074	2,056	2,043	2,032	2,016	2,001	1,985	1,938
8	0,975	2,529	2,458	2,412	2,379	2,355	2,336	2,321	2,299	2,278	2,256	2,192
8	0,990	2,993	2,890	2,823	2,777	2,742	2,715	2,694	2,663	2,632	2,601	2,511
9	0,900	1,793	1,760	1,738	1,723	1,711	1,702	1,695	1,684	1,674	1,663	1,632
9	0,950	2,124	2,073	2,040	2,017	1,999	1,986	1,975	1,959	1,943	1,927	1,880
9	0,975	2,452	2,381	2,334	2,302	2,277	2,259	2,244	2,222	2,200	2,178	2,114
9	0,990	2,888	2,785	2,718	2,672	2,637	2,611	2,590	2,559	2,528	2,497	2,407