Sozialwissenschaftliche Datenanalyse

Andreas Engel
Michael Möhring
Klaus G. Troitzsch

25. Oktober 2001

UNIVERSITÄT KOBLENZ–LANDAU · ABT. KOBLENZ
INSTITUT FÜR SOZIALWISSENSCHAFTLICHE INFORMATIK

Vorwort

Der Einsatz von Computern hat in den vergangenen Jahrzehnten die Arbeitsweise von Sozialwissenschaftlern nachhaltig verändert. Es ist wohl nicht übertrieben zu behaupten, daß die Verbreitung und zunehmende Bedienungsfreundlichkeit von Rechenanlagen und statistischen Auswertungsprogrammen der empirischen Sozialforschung als vorherrschender Forschungsstrategie erst zum Durchbruch verholfen haben.

Mit der steigenden Bedeutung von Computern im sozialwissenschaftlichen Forschungsprozeß ist auch eine neue Wissenschaftsdisziplin entstanden, die sich schwerpunktmäßig mit der Adaption und Anwendung von Werkzeugen und Methoden der Informatik in der sozialwissenschaftlichen Forschung beschäftigt: die Sozialwissenschaftliche Informatik. Das vorliegende Skript ist aus einer Reihe von Vorlesungen hervorgegangen, die wir für Studierende dieses Anwendungsfaches im Studiengang Diplom–Informatik an der Universität Koblenz–Landau gehalten haben.

Im Gegensatz zu vielen anderen Einführungen in die uni- und multivariate Datenanalyse richtet sich die vorliegende Darstellung der statistischen Verfahren in erster Linie an fortgeschrittene Anwender der “klassischen” statistischen Methoden, die sich einen Überblick über die mathematischen Grundlagen der angewandten Methoden verschaffen und bei der Interpretation ihrer Analyseergebnisse die Fallstricke rezeptbuchartiger Einführungen vermeiden möchten.

Kenntnisse der Handhabung eines statistischen Auswertungsprogramms setzen wir dabei voraus; soweit wir SPSS-Beispiele bringen, dienen sie nur dazu, die Korrespondenz zwischen SPSS-Prozeduren und mathematisch–statistischen Verfahren herzustellen.

An dieser Stelle herzlich bedanken möchten wir uns bei Raphael Ostermann für die kritische Durchsicht, die zahlreichen Verbesserungsvorschläge und die Unterstützung bei der mühsamen textlichen und graphischen Aufbereitung des Manuskriptes mit dem Textformatierungsprogramm LaTeX. Nicht unerwähnt bleiben soll, daß er für die Generierung zahlreicher Graphiken und die vollständige Aufbereitung des Indexes eigens die Programme TeXKurve und MakeTIE ([LA-]TeX-Index-Environment) entwickelt und zur Anwendungsreife gebracht hat.

Ebenso geht ein Dank an Gabriele Cremer und Christa Paul für die TeXnische Umsetzung etlicher Schaubilder und einer früheren Version dieses Textes.

Koblenz, im Oktober 1994

Andreas Engel
Michael Möhring
Klaus G. Troitzsch

Inhaltsverzeichnis

Kapitel 1

Die Datenanalyse im empirischen Forschungsprozeß

1.1 Übersicht

Die Anwendung statistischer Methoden auf sozialwissenschaftliche Daten muß im Zusammenhang des gesamten sozialwissenschaftlichen Forschungsprozesses gesehen werden, in den diese Methoden eingebettet sind und durch den die der Analyse zugrunde liegenden Daten erst produziert werden. Abbildung 1.1 zeigt eine stark vereinfachte — für diese Übersicht jedoch ausreichende — Beschreibung des Verlaufs eines Forschungsprozesses unter besonderer Berücksichtigung der Rolle der sozialwissenschaftlichen Datenanalyse.[1]

Grundsätzlich kann die statistische Datenanalyse im sozialwissenschaftlichen Forschungsprozeß unter zwei Zielrichtungen eingesetzt werden: entweder um sozialwissenschaftliche Theorien empirisch zu überprüfen (konfirmatorische Datenanalyse) oder um auf der Grundlage gegebener Datenbestände empirisch begründete Konzepte, Hypothesen oder Theorieansätze zu finden (exploratorische Datenanalyse) ([Schnell/Hill/Esser 1988, S. 109]).

Die konfirmatorische Analysestrategie setzt Theorie- und Modellbildung bezüglich des zu untersuchenden Forschungsproblems voraus. Von besonderer Bedeutung sind hier die Definition und Operationalisierung der in der zu überprüfenden Theorie enthaltenen Begriffe sowie die Formulierung von Hypothesen über zu beobachtende Zusammenhänge zwischen den durch die theoretischen Begriffe beschriebenen Sachverhalte. Beides ist Voraussetzung einer systematischen

[1] Ausführlichere Beschreibungen dazu finden sich z.B. bei [Friedrichs 1973, S. 51] und [Schnell/Hill/Esser 1988, S. 110].

Datenerhebungs- und Datenauswertungsstrategie. Nach der Erhebungs- und Erfassungsphase erfolgt dann die statistische Analyse der gewonnenen Datensätze, deren Ergebnisse — im Sinne einer Theorieüberprüfung — interpretiert werden.

In der Forschungspraxis haben für diese Auswertungsstrategie vor allem statistische bi- und multivariate Kausalmodelle eine große Bedeutung gewonnen. Als Analysemethoden für diese Modelle werden vor allem Regressions- und Varianzanalysen eingesetzt, mit denen lineare Abhängigkeiten zwischen metrischen und zum Teil auch qualitativen Daten geschätzt bzw. berechnet werden.

Da die Ergebnisse der Datenanalyse schon allein aus Gründen unvermeidbarer Meß- und Schätzfehler niemals zu einer vollständigen Bestätigung theoretischer Annahmen führen, wird der in Abbildung 1.1 dargestellte Forschungszyklus in der Praxis meist mehrfach — mit verschiedenen Varianten der Anwendung statistischer Analyseverfahren — durchlaufen, was in der Regel auch zur Modifizierung der Ausgangstheorie führt.

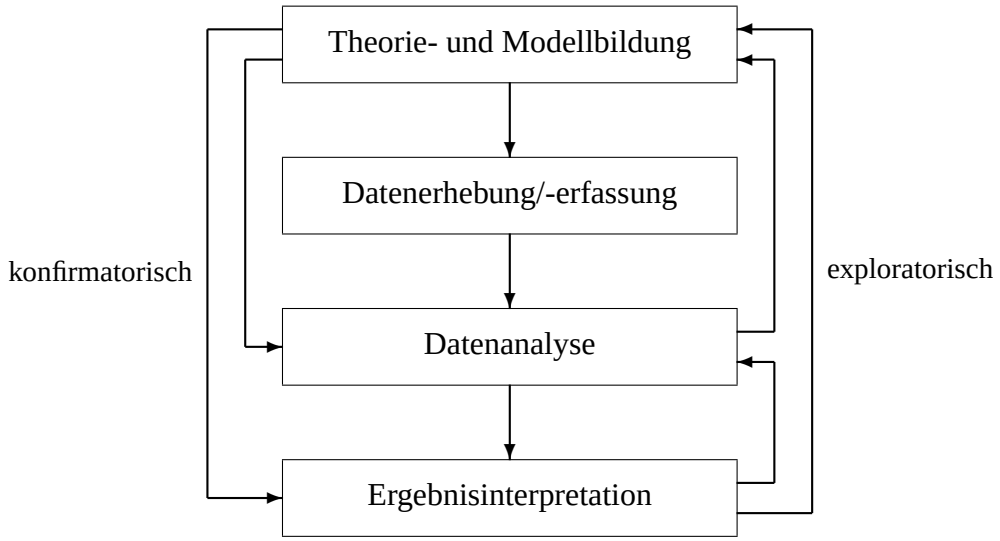


Abb. 1.1: Empirische Forschung als iterativer Prozeß

Die exploratorische Datenanalyse geht nicht von einer ausformulierten Theorie bzw. einem genau spezifizierten Analysemodell aus, sondern von einer vorhandenen Datenbasis, die einen hinreichend engen inhaltlichen Bezug zur Fragestellung der Untersuchung aufweist. Ziel der exploratorischen Anwendung statistischer Analysemethoden ist das Entdecken empirisch begründeter Datenstrukturen — erkennbar an signifikanten oder “aus Erfahrung guten” statistischen Kennwerten, die auf neue Konzepte oder bisher nicht vermutete Hypothesen hindeuten.

Zur großen Verbreitung der exploratorischen Auswertungsstrategie hat die Verfügbarkeit statistischer Analyseprogramme vor allem auch auf Personalcomputern erheblich beigetragen. Die Erweiterung der Statistikpakete um Komponenten zum Datenmanagement, d.h. zur Variablendefinition, zur Gewichtung und Indexbildung, aber auch zur graphischen Aufbereitung der Primärdaten oder zur Visualisierung von Analyseergebnissen, regen geradezu zum Experimentieren mit statistischen Methoden an. Darüber hinaus fördern die leichte Änderbarkeit der Struktur statistischer Modelle und die Auswertung auch großer Datensätze ohne nennenswerte Zeitverzögerung das Testen von Auswertungsvarianten.

Als statistische Methoden für die exploratorische Analyse werden sehr oft Faktor-, Diskriminanz- und Clusteranalyse eingesetzt, die zusammen mit der für die konfirmatorische Analyse typischen Regressionsanalyse in der vorliegenden Einführung in die Methoden der statistischen Datenanalyse behandelt werden.

1.2 Theorie- und Modellbildung

Theorie- und Modellbildung kann sowohl Ausgangspunkt als auch Ziel der Anwendung statistischer Methoden im sozialwissenschaftlichen Forschungsprozeß sein. Im ersten Fall steht die Übersetzung einer Theorie in ein statistisches Analysemodell im Vordergrund, im zweiten Fall die Interpretation statistischer Auswertungsergebnisse im Lichte zu überprüfender bzw. zu entdeckender theoretischer Annahmen. Beide möglichen Übergänge zwischen Theorie- und Modellbildung und der statistischen Analyse sollen hier am Beispiel einer rudimentären Theorie des Wählerverhaltens beschrieben werden, die auch den Interpretationsrahmen für die Berechnungsbeispiele in diesem Band darstellt.

Nach einem weit verbreiteten Erklärungsansatz wird die Entscheidung eines Wählers für eine Partei als rationale Wahl interpretiert, in der — sehr verkürzt — derjenigen Partei die Stimme gegeben wird, die in bezug auf zentrale politische Streitfragen bzw. Wertvorstellungen Positionen vertritt, die denen des Wählers am nächsten liegen. Die Entwicklung politisch relevanter Wertvorstellungen kann darüber hinaus mit der sozialen Gruppenzugehörigkeit der Wähler erklärt werden. Demnach werden im Prozeß der individuellen politischen Sozialisation bestimmte kollektive Erfahrungen in bezug auf politische Auseinandersetzungen als individuelle politische Einstellungen verankert.

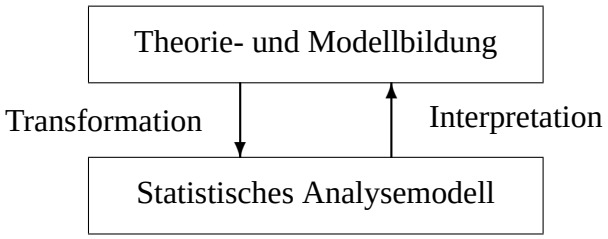


Abb. 1.2: Wechselwirkung zwischen Theorie- und Modellbildung und statistischer Datenanalyse

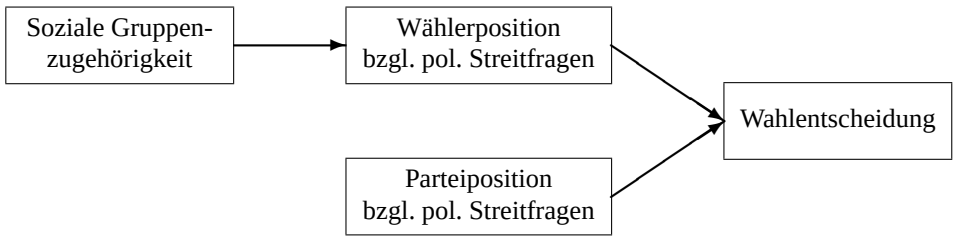


Abb. 1.3: Erklärungsskizze zum Wählerverhalten

Identifikation zentraler theoretischer Konzepte

Der erste Schritt des Übersetzens einer Theorie in ein statistisches Analysemodell besteht in der Identifikation der zentralen theoretischen Konzepte, mit denen Aussagen über die zu untersuchenden Realitätsausschnitte formuliert werden. Im Beispiel der dargestellten Theorie sind dies etwa die Begriffe “individuelle Position bezüglich politischer Streitfragen”, “Wahrnehmung der Position von Parteien bezüglich politischer Streitfragen”, “soziale Gruppenzugehörigkeit” und “Wahlentscheidung”. Abb. 1.3 stellt die Beziehungen zwischen diesen zentralen Begriffen im Rahmen der Erklärung des Wählerverhaltens graphisch dar.

Definition der Intension und Extension theoretischer Konzepte

Im allgemeinen sind die zentralen theoretischen Konstrukte zur Erklärung sozialwissenschaftlicher Phänomene nicht unmittelbar meßbar. Es handelt sich also um *latente*, nicht unmittelbar beobachtbare Variablen. Die durch sie beschriebenen Eigenschaften werden also erst anhand einer Mehrzahl empirischer Beobachtungen (Indikatoren) faßbar, die zur Unterscheidung von latenten Variablen als direkt meßbare, *manifeste* Variablen bezeichnet werden.

Um die für ein theoretisches Konzept relevanten Indikatoren zu bestimmen, besteht daher der zweite Schritt zu einem statistischen Analysemodell in der Definition der Intension und Extension der zentralen Erklärungskonzepte. In der intensionalen Definition werden die Beschreibungsdimensionen festgelegt, auf denen ein Erklärungsbegriff gemessen werden kann. Die Extension eines Begriffs beinhaltet die Klasse der Sachverhalte, auf die dieser Begriff angewandt werden kann. Im Fall der erläuterten Wählertheorie kann z.B. die Intension des Begriffs “Soziale Gruppenzugehörigkeit” definiert werden als Zugehörigkeit zu verschiedenen Alters-, Geschlechts-, Konfessionsgruppen oder sozialen Schichten. Die Wähler bzw. alle wahlberechtigten Bürger einer bestimmten Population bilden die Begriffsextension.

Definition von Operationalisierungsregeln

Im dritten Schritt werden die in ihrer Extension und Intension definierten theoretischen Begriffe operationalisiert, d.h. es wird angegeben, durch welche Meßvorschriften die durch einen theoretischen Begriff bezeichneten Sachverhalte konkret erfaßt werden können. Die in den nachfolgenden Tabellen aufgestellten Operationalisierungsregeln legen fest, welche Daten (manifeste Variablen, Indikatoren) den jeweiligen theoretischen Begriff messen.

Soziale Gruppenzugehörigkeit

- Definition:
  Einteilung einer Gesellschaft in Gruppen, deren Mitglieder aufgrund ähnlicher Lebensbedingungen und -erfahrungen untereinander häufigere Kontakte haben als zu Mitgliedern anderer sozialer Gruppen.

- Operationalisierung:
  Relevante Meßdimensionen der latenten Variablen $\Longleftrightarrow$ Indikatoren[2]

Variable	Indikatoren
Altersgruppenzugehörigkeit	V335
Geschlechtsgruppen	V334, V319
Kirchganghäufigkeit	V347
Schichtzugehörigkeit	V338–V345, V237, V240
Einbindung ins Wohnumfeld	V350, V352

Wertvorstellungen/Positionen von Wählern

- Definition:
  Unter einer Wertvorstellung wird die Neigung verstanden, das eigene und fremde Handeln anhand gewisser Normen, Standards oder Auswahlkriterien zu beurteilen.

- Operationalisierung:
  Relevante Meßdimensionen der latenten Variablen $\Longleftrightarrow$ Indikatoren

Variable	Indikatoren (=Einstellungen)
Konserv.-christl. Wertvorstellung	Einfluß der Kirchen auf die Politik: V315 Paragraph 218: V387
Soziale Wertvorstellung	Arbeitslosigkeit bekämpfen: V245 Einfluß der Gewerkschaften auf die Politik: V316
Ökonomische Wertvorstellung	Wirtschaft ankurbeln: V244
Nicht-materielle Wertvorstellung	Den Bürgern mehr Einfluß: V253 Ruhe und Ordnung als Staatsaufgabe: V246
Ökologische Wertvorstellung	Umweltschutz als Staatsaufgabe: V249 Kernenergieausbau: V303

(Wahrgenommene) Wertvorstellungen/Positionen von Parteien

- Definition:
  Wertvorstellungen von Parteien lassen sich analog zu denen der Wähler einteilen, wobei hier die Einschränkung wesentlich ist, daß nur die vom Wähler wahrgenommenen Wertvorstellungen auch relevant sind.

[2]Die in den Tabellen genannten Indikatoren (V...) beziehen sich auf die Bundestagswahlstudie von 1987, die vom Zentralarchiv für empirische Sozialforschung, Studien–Nr. 1537, zur Verfügung gestellt wurde [Wahlstudie 1987].

- Operationalisierung:
  Relevante Meßdimensionen der latenten Variablen $\Longleftrightarrow$ Indikatoren

Variable	Indikatoren (=Einstellungen)
Konserv.-christl. Wertvorstellung	Vermutete Einstellung der Parteien zur Abtreibung: V388–V391
Ökologische Wertvorstellung	Vermutete Einstellung der Parteien zur Kernenergie: V304–V308

Wahl des statistischen Analyseansatzes

Der vierte und letzte Schritt in der Herleitung besteht dann in der Auswahl eines dem Erklärungsansatz adäquaten statistischen Analysemodells. Die Entscheidung für ein statistisches Analysemodell ist einerseits abhängig vom Skalenniveau der Indikatoren bzw. der latenten Variablen und zum anderen von der Art der Beziehungen, die zwischen den Erklärungsvariablen unter theoretischen Gesichtspunkten gefordert wird.

Die hier beschriebenen statistischen Analyseansätze gehen im Grundsatz von linearen Zusammenhängen zwischen abhängigen und unabhängigen Erklärungsvariablen aus. Insofern ist die Gültigkeit der hier erläuterten statistischen Analysemethoden in bezug auf die Überprüfung sozialwissenschaftlicher Theorien auf lineare Kausalmodelle beschränkt. Im Rahmen dieser Modellklasse beinhaltet jedoch die vorgestellte Auswahl statistischer Analyseansätze sowohl Modelle mit metrischen und nicht–metrischen manifesten Variablen (bi- und multivariate Regressionsanalyse, Diskriminanzanalyse) als auch solche mit quantitativen und qualitativen latenten Variablen (Faktoren- und Clusteranalyse).

Entdecken der Dimensionen eines Erklärungskonzepts

Steht nicht die Theorieüberprüfung sondern die Entdeckung theoretischer Konzepte im Vordergrund, beginnt die statistische Analyse mit einer Menge manifester Variablen, von denen angenommen wird, daß sie ein theoretisches Konzept beschreiben. Ziel der Anwendung statistischer Methoden ist es dann, die Dimensionalität eines Erklärungskonzepts wie zum Beispiel die Dimensionalität des Raums politischer Streitfragen zu beschreiben. Entsprechend dem Skalenniveau der zu rekonstruierenden Erklärungskonzepte bieten Faktoren- und Clusteranalysen adäquate statistische Analysemodelle.

1.3 Datenerhebung

Eine wesentliche Voraussetzung für die Durchführung einer Datenerhebung ist die Festlegung der *Grundgesamtheit*, d.h.

> "die Gesamtheit der Einheiten (Personen, Institutionen, Gegenstände), die in einer statistischen Untersuchung auf ihre besonderen Eigenheiten hin zu beschreiben sind." [Leiner 1989, S. 1]

Beispiel: "Alle Personen mit deutscher Staatsbürgerschaft, die in der Bundesrepublik und West-Berlin in Privathaushalten leben und spätestens am 1. Januar 1962 geboren wurden (wobei als Privathaushalt jede Gemeinschaft von Personen gilt, die zusammen wohnen und gemeinsam wirtschaften, ohne notwendigerweise auch miteinander verwandt sein zu müssen" [Porst 1985, S. 89]

Die Grundgesamtheit bezeichnet also den Personenkreis, über den mit Hilfe der empirischen Untersuchung Aussagen gemacht werden sollen. Allgemein können bei der Datenerhebung Untersuchungsobjekte wie folgt unterschieden werden:

1. *Aussageeinheiten:*
   Objekte, über die geforscht werden soll, d.h. über die theoretische Schlußfolgerungen gemacht werden
   (z.B. die Wählerschaft der Bundesrepulik Deutschland zur Bundestagswahl 1987)

2. *Erhebungseinheiten:*
   Objekte, an denen Messungen (Beobachtungen) vorgenommen werden
   (z.B. eine repräsentativ ausgewählte Stichprobe von wahlberechtigten Bürgern der Bundesrepublik Deutschland für eine Vorwahlbefragung in 1987)

3. *Analyseeinheiten/Untersuchungseinheiten:*
   Objekte, die den statistischen Berechnungen zugrunde liegen
   (z.B. die Teilnehmer der Vorwahlbefragung zur Bundestagswahl 1987, die in bezug auf die analysierten Daten gültige Werte besitzen)

Die Aufgabe der Datenerhebung besteht nun darin, die Analyseeinheiten für die Durchführung einer statistischen Datenanalyse zu erzeugen. Dazu läßt sich der Datenerhebungsprozeß insgesamt in folgende Phasen gliedern:

1. Festlegung der Aussageeinheiten
2. Bestimmung der Erhebungseinheiten, an denen die für die Aussageeinheiten relevanten Daten erhoben werden können
   (Stichprobenziehung oder Wahl der Grundgesamtheit als Erhebungseinheiten)
3. Festlegung der Untersuchungsform
   (z.B. Befragung, Beobachtung, Inhaltsanalyse)
4. Entwicklung von Datenerhebungsplänen und -instrumenten
   (z.B. Fragebogen, Interviewleitfaden)
5. Durchführung der Datenerhebung
   (z.B. postalische Befragung, Interviews)
6. Erfassung und Korrektur der Daten
   (Übertragung der Ergebnisse aus der Datenerhebung (z.B. ausgefüllte Fragebögen) in eine für die Datenanalyse geeignete Form (= Datenkodierung))

Ergebnis der Erfassungsphase ist die *Datenmatrix* und das *Codebuch*, die beide Voraussetzung für eine statistische Datenanalyse sind.

Die *Datenmatrix* faßt in tabellarischer Form die Rohwerte der empirisch gewonnenen Daten so zusammen, daß allen Untersuchungseinheiten die erhobenen Variablenwerte nach einem einheitlichen Muster zugeordnet werden. In ihr sind somit die Analyseeinheiten zusammengefaßt.

		Variable, Merkmale, Stimuli (z.B. Interviewfragen)					
		S_1	S_2	…	S_j	…	S_m
Erhebungs-,	O_1	R_{11}	R_{12}	…	R_{1j}	…	R_{1m}
Untersuchungs-	O_2	R_{21}	R_{22}	…	R_{2j}	…	R_{2m}
einheiten,	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Merkmalsträger,	O_i	R_{i1}	R_{i2}	…	R_{ij}	…	R_{im}
Objekte	⋮	⋮	⋮	⋮	⋮	⋮	⋮
(z.B. Befragte)	O_n	R_{n1}	R_{n2}	…	R_{nj}	…	R_{nm}

Das *Codebuch* enthält für eine sozialwissenschaftliche Untersuchung die Variablenübersicht, die Zuordnung von Variablen und Variablennamen und für jede Variable die Zuordnung von Merkmalsausprägungen zu numerischen oder alphabetischen Symbolen.

Traditionell konzentriert sich die softwaretechnische Unterstützung des empirischen Forschungsprozesses auf die eigentliche Datenanalyse mittels statistischer Auswertungssysteme (vgl. Kapitel 1.4). Die Entwicklung von Softwarewerkzeugen für die Daten-Erhebungs- und Erfassungsphase mit dem Ziel, die Zeitspanne von der Theorie- und Modellbildung bis zum Vorliegen analysefähiger Daten zu verkürzen, hat erst in den 80er Jahren verstärkt eingesetzt. Neben dem Kostenaspekt spielt dabei insbesondere der immer stärker steigende Bedarf an *kurzfristig* verfügbaren Analyseergebnissen (z.B. Konsumentenforschung, Wahlforschung) eine wichtige Rolle.

Die im Rahmen der Datenerhebung hierfür eingesetzten Softwaresysteme lassen sich wie folgt charakterisieren:

- Unterstützung der Datenerfassung im engeren Sinne.
  Hierunter sind Programme (z.B. DATA ENTRY II für SPSS) zu verstehen, mit deren Hilfe — im Gegensatz zu der "klassischen" Datenerfassung mittels Texteditor — speziell auf eine Erhebung abgestimmte, maskengesteuerte Datenerfassungsprogramme erstellt werden. Dies umfaßt auch die Definition von Regeln für eine antwortgesteuerte Fragebogenabarbeitung (skip and fill rule) oder die Angabe von Gültigkeitsbereichen für Antworten (range), deren Verletzung eine Fehlermeldung erzeugt und direkt die Möglichkeit zur Korrektur vorsieht.

- Unterstützung des gesamten Datenerhebungsprozesses.
  Eine weitere Beschleunigung der Datenerhebungsphase läßt sich durch den Einsatz von Hard- und Softwaresystemen erzielen, die insbesondere Datenerhebung und Datenerfassung direkt miteinander verbinden, d.h. ohne Datenerhebungsinstrumente "aus Papier" auskommen. Der Schwerpunkt der Entwicklung liegt dabei in der Unterstützung von Telephoninterviews ([Frey/Kunz/Lüschen 1990], [Schneid 1991]). Allgemein lassen sich folgende Anwendungsbereiche unterscheiden [Joop/De Bier/De Leeuw 1990]:

  CAPI (Computer-Assisted Personal Interviewing):
  Die Antworten werden während des Interviews direkt durch den Interviewer in den Rechner eingegeben.

  CASAQ (Computer-Assisted Self-Administered Questionnaire):
  Der Befragte führt die Befragung selbständig im Dialog mit dem Rechner durch.

  CATI (Computer-Assisted Telephone Interviewing):
  Softwaretechnische Unterstützung für alle Aufgaben zur

Durchführung von Telefoninterviews, wie zum Beispiel automatische Stichprobenziehungen, die Durchführung und Verwaltung von Telefonkontakten (Telefonscheduling), der Fragebogenentwurf, die Durchführung von Interviews, Komponenten zur Interview(er)verwaltung und -kontrolle, etc.

1.4 Datenanalyse

1.4.1 Statistische Auswertungssysteme

Die Durchführung empirischer Studien hat mit der Entwicklung von statistischen Auswertungssystemen seit den 60er Jahren einen großen Aufschwung genommen und dem empirischen Erklärungsansatz als vorherrschendem Forschungsparadigma in den Sozialwissenschaften zum entscheidenden Durchbruch verholfen ([Engel/Möhring 1994]). Insbesondere die Anwendung rechenintensiver Auswertungsmethoden, das Auswerten großer Datenbestände sowie die mehrfache Auswertung eines Datenbestandes (vgl. [Haag/Haux/Kieser 1992, S. 1]) unter veränderten Fragestellungen ist ohne statistische Auswertungssysteme praktisch nicht durchführbar.

In einer ersten Phase der Entwicklung stand dabei ausschließlich die Anwendung statistischer Methoden im Vordergrund, sowohl in Form individuell erstellter Programme als auch durch die Verfügbarkeit erster Programmbibliotheken. Die Entwicklung stapelorientierter Softwaresysteme (z.B. P-STAT, SPSS, SAS, BMDP) mit einer eigenen Kommandosprache ab Mitte der 60er Jahre, die außer statistischen Prozeduren bereits erste Hilfsfunktionen zur Datendefinition und zur Datenaufbereitung (z.B. Variablendefinition, Rekodierung) enthielten, führte zu einer ersten Standardisierung und durch die vereinfachte Handhabung auch zu einer starken Verbreitung. Neben der Weiterentwicklung des statistischen Methodenspektrums stand in den folgenden Jahren gerade der Ausbau der Datenverwaltungskomponente im Vordergrund, insbesondere unter dem Aspekt einer Annäherung an Datenbanktechnologien[3]. Der nächste große Entwicklungsschub erfolgte Anfang der 80er Jahre durch die Portierung der existierenden bzw. die Entwicklung neuer Softwaresysteme (z.B. S-PLUS, GAUSS, SYSTAT, NSDstat+) auf Arbeitsplatzrechner. Die mit dieser Rechnergeneration verbundenen technischen Möglichkeiten (z.B. hochauflösende Farbgraphik, Menü-/ Fenstertechnik) unterstützten nicht nur die Entwicklung neuer Analyse- und Präsentationstechniken (z.B. innerhalb der exploratorischen Datenanalyse) oder auch stärker be-

[3] Entwicklung statistischer Auswertungssysteme auf Datenbankbasis (z.B. SIR/DBMS).

nutzerorientierter Sprachkonzepte, sondern sie veränderten auch den Vorgang der Datenanalyse selbst hin zu einem interaktiven Prozeß mit kurzen Antwortzeiten, direkter Ergebnispräsentation und der Möglichkeit der Weiterverarbeitung von Analyseergebnissen in nachfolgenden Bearbeitungsschritten. Die Vielfalt und Komplexität des von statistischen Analysesystemen mittlerweile angebotenen Methodenspektrums bringt zusätzliche Auswahlprobleme für konkrete Fragestellungen, insbesondere unter dem Aspekt einer steigenden Benutzerzahl ohne mathematisch-statistischem Hintergrundwissen. Neueste Entwicklungen gehen daher in die Richtung wissensbasierter Systeme (z.B. CADEMO, GLIMPSE, REX), um Anwender bei einer für ihren Problembereich angemessenen Methodenauswahl zu unterstützen.

Nachfolgende Auflistung enthält das von der Mehrzahl statistischer Analysesysteme[4] bereitgestellte Methodenspektrum. Die dabei jeweils in Klammern genannten Bezeichner stammen aus der Kommandosprache von SPSS [Uehlinger/Hermann/Huebner/Benke 1992], auf deren Grundlage auch stärker anwenderorientierte Eingabeformen realisiert sind (vgl. SPSS für Windows [SPSS 1993]).

1. *Datendefinition/-erfassung:*
   In diese Gruppe gehören alle Möglichkeiten des Anwenders zur Beschreibung von Variablen und Variablenstrukturen (`DATA LIST`, `VARIABLE LABELS`, `VALUE LABELS`, `MISSING VALUES`) sowie zum Einlesen (`BEGIN DATA ... END`, `GET FILE`) und zum Abspeichern (`SAVE`) von Daten.

2. *Datenmanagement/-modifikation:*
   Für die Durchführung von statistischen Datenanalysen ist es häufig notwendig, Variablen der Rohdatenmatrix unter Berücksichtigung der zugrundeliegenden Fragestellungen und zu überprüfender Hypothesen aufzubereiten:

   - Bildung neuer, nicht direkt gemessener Variablen durch Kombination von Merkmalen der Rohdaten (`COMPUTE`)
     (vgl. Anhang A.1)

     Beispiel ([Hoffmeyer-Zlotnik 1993, S. 135]): Sozialökonomischer Status (SES)

[4] Haag, Haux und Kieser sprechen von insgesamt mehr als 200 kommerziell verfügbaren Systemen ([Haag/Haux/Kieser 1992, S.41]).

SES = $f($ Bildung ($\rightarrow$ Schulabschluß) ,
Beruf ($\rightarrow$ berufliche Stellung) ,
Einkommen ,
ethnische Zugehörigkeit
)

- Modifikation existierender Variablen für Auswertungszwecke (`RECODE`)

  Beispiel:

  Generationen = Bildung von Altersklassen aus dem Alter der Befragten (z.B. 18–25, 26–40, 41–60, ≥ 61)

Ebenfalls in diese Kategorie gehören allgemeine Datenmanagementfunktionen zur Vorbereitung von Datenanalysen. Beispiele hierfür sind

- das Sortieren von Beobachtungen (`SORT`),
- das Zusammenfügen mehrerer Eingabedateien (`JOIN`),
- die bedingte Auswahl (`SELECT IF`) oder
- das Gewichten von Beobachtungen (`WEIGHT`).

3. *Statistische Datenanalyse:*
   Die statistische Datenanalyse umfaßt die Anwendung statistischer Methoden auf die Daten, d.h. die Datenanalyse im eigentlichen Sinne. Im Rahmen einer ersten, *deskriptiven Datenanalyse* werden dabei Zusammenhänge ausschließlich zwischen den erfaßten Analyseeinheiten untersucht. Für eine bessere Überschaubarkeit der Vielzahl von angebotenen Methoden existieren in der Literatur eine ganz Reihe von Unterscheidungskriterien, von denen die wichtigsten nachfolgend aufgeführt sind:

Kriterium	Methode
Anzahl der Variablen	univariat, bivariat, multivariat
Meßniveau der Variablen	nominal, ordinal, intervall, ratio
Beziehung zwischen den Variablen	symmetrisch (interdependent), asymmetrisch (kausal)
Funktionaler Zusammenhang zwischen den Variablen	linear, nichtlinear
Abstraktionsebene der Variablen	empirisch (manifest), theoretisch (latent)
Zeitstruktur der Datensätze	Querschnitt, Längsschnitt

In der Regel stammen die in der Datenmatrix enthaltenen Daten aus einer (repräsentativen) Stichprobe der Grundgesamtheit, da eine Vollerhebung aus Zeit- und Kostengründen meist viel zu aufwendig ist (z.B. Wahlforschung)[5]. Von Interesse sind aber natürlich die "wahren" Zusammenhänge auf der Ebene der Grundgesamtheit und weniger die Analyseergebnisse in der Stichprobe. So hat beispielsweise das Ergebnis der Sonntagsfrage ("Wenn nächsten Sonntag Bundestagswahl wäre, welche Partei würden Sie wählen?") nur bezogen auf alle Wahlberechtigten der Bundesrepublik Deutschland einen Aussagewert. Die Methoden der *schließenden* oder auch *induktiven Statistik* erlauben sowohl das *Schätzen* unbekannter Parameter (z.B. Mittelwert) als auch die *Prüfung von statistischen Hypothesen* über Sachverhalte in der Grundgesamtheit, falls die hierfür verwendete, einzige Stichprobe nach festgelegten, wissenschaftlich begründeten Regeln gezogen worden ist.

In den folgenden Kapiteln werden die wichtigsten Methoden deskriptiver Datenanalyse (univariat, bivariat, multivariat) vorgestellt. Aufbauend auf eine kurze Einführung in die Grundlagen der Inferenzstatistik (vgl. Kapitel 4) folgt dann die Durchführung inferenzstatistischer Verfahren jeweils methodenspezifisch (z.B. die Durchführung eines Hypothesentest für den χ^2-Wert).

4. *Ergebnisaufbereitung/-präsentation:*
   Eine vor allem im Hinblick auf die Vermittlung der Ergebnisse empirischer Forschung wichtige Rolle spielt die Aufbereitung und Präsentation von Analyseergebnissen. Im Mittelpunkt stehen dabei graphische Darstellungsmöglichkeiten (z.B. Histogramme, Scatterplots, Boxplots), die insbesondere durch die technische Weiterentwicklung sehr stark vorangetrieben wurden (z.B. Darstellung und Bearbeitung von dreidimensionalen Punktwolken aus verschiedenen Perspektiven). Ein weiterer Punkt ist die Unterstützung bei der Gestaltung von Tabellen (TABLES).

1.4.2 Kurzcharakteristik deskriptiver Analyseverfahren

Für eine erste Einordnung der in den folgenden Kapiteln ausführlich beschriebenen Datenanalyseverfahren werden diese zuvor anhand der vorgestellten Klassifikationskriterien kurz charakterisiert:

[5] Weiterhin gibt es — zumindest außerhalb der Sozialwissenschaften — auch Erhebungsmethoden, bei denen die Untersuchungsobjekte zerstört werden (z.B. chemische Analysen bei der Qualitätsüberprüfung von Produkten). Eine Vollerhebung ist hier also grundsätzlich ausgeschlossen.

- Univariate Verfahren

  Gegeben: eine manifeste, nichtmetrische oder metrische Variable

  Frage: Wie verteilen sich die Werte dieser Variablen innerhalb der Stichprobe bzw. der Gesamtpopulation?

  Beispiel: Wahlentscheidung (Sonntagsfrage)

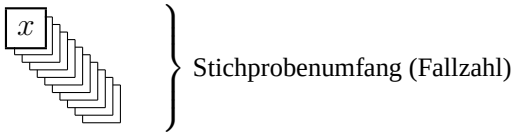


Abb. 1.4: Univariate Modelle

- Bivariate Verfahren

  – Zusammenhangsmaße

  Gegeben: zwei Variablen

  Frage: Wie stark ist der Zusammenhang zwischen diesen Variablen (symmetrisch)?
  Wie verbessert die Kenntnis einer (als unabhängig definierten Variablen) die Vorhersage der anderen (abhängigen) Variablen (asymmetrisch)?

  Beispiel: Konfession $\longleftrightarrow$ Wahlentscheidung
  Konfession $\longrightarrow$ Parteisympathie

nominal/nominal	:	χ^2, λ
ordinal/ordinal	:	τ_B, τ_C, γ
metrisch/metrisch	:	r^2
nominal/metrisch	:	η^2

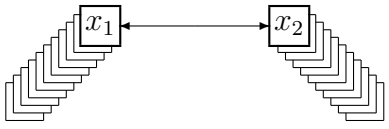


Abb. 1.5: Bivariate Assoziationsmaße

  – Bivariate Regression

Gegeben: zwei metrische Variablen:
eine unabhängige Variable (Regressor),
eine abhängige Variable (Regressand)

Frage: Wie stark steht der Regressand in einem kausalen Abhängigkeitsverhältnis zu dem Regressor?

Beispiel: Kanzlerkandidatensympathie $\longrightarrow$ Parteisympathie

Abb. 1.6: Bivariate Regression

- Multivariate Verfahren

  – Multiple Regression

  Gegeben: mehrere metrische, unabhängige Variablen,
  eine metrische, abhängige Variable

  Frage: Wie groß ist der Einfluß der einzelnen unabhängigen Variablen (im Vergleich zueinander) auf die abhängige Variable?
  Wie gut erklärt das Modell den Zusammenhang zwischen unabhängigen und abhängiger Variablen?

  Beispiel: Sympathie zu einzelnen Spitzenpolitikern $\longrightarrow$ Parteisympathie

  – Faktorenanalyse

  Gegeben: mehrere metrische Variablen

  Frage: Welche Variablen messen wie gut eine bzw. mehrere latente Variablen (Faktoren) ?

  Beispiel: Sympathie zu einzelnen Spitzenpolitikern $\longleftarrow$ Parteienspektrum (Links-Rechts-Spektrum)

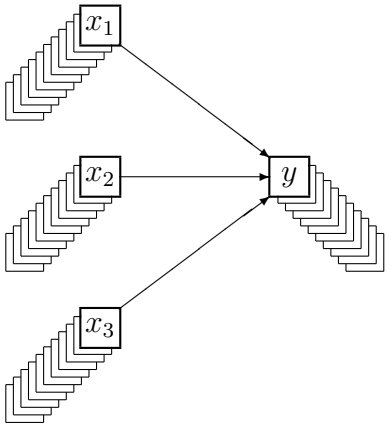


Abb. 1.7: Multiple Regression mit einem Regressanden

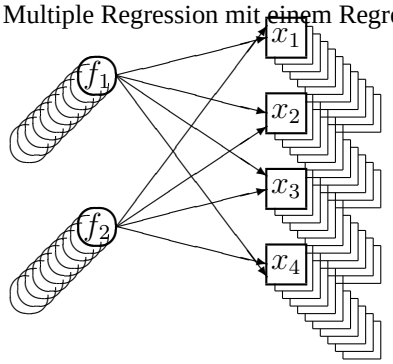


Abb. 1.8: Faktorenanalyse

– Diskriminanzanalyse

  Gegeben: mehrere metrische, unabhängige Variablen,
  eine nominal skalierte, abhängige Variable

  Frage: Wie groß ist der Einfluß der einzelnen unabhängigen Variablen (im Vergleich zueinander) auf die abhängige Variable?
  Wie gut erklärt das Modell den Zusammenhang zwischen unabhängigen und abhängiger Variablen?
  Wie gut lassen sich die durch die (nominal skalierte) abhängige Variable gebildeten Gruppen trennen (rekonstruieren)?

  Beispiel: Sympathie zu Spitzenpolitikern und Parteien $\longrightarrow$ Wahlverhalten

– Clusteranalyse

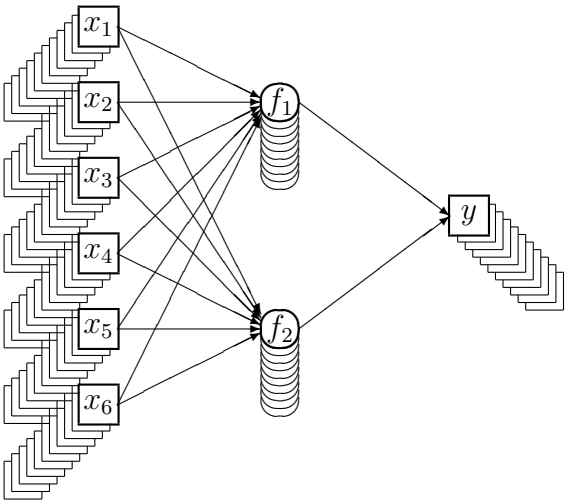


Abb. 1.9: Diskriminanzanalyse ($f_{1/2}$ kennzeichnen die Diskriminanzfunktionen)

Gegeben: mehrere metrische oder anders skalierte Variablen, die es erlauben anzugeben, wie ähnlich die Fälle untereinander sind

Frage: Lassen sich die Fälle auf Grund der gemessenen Variablen zu Gruppen zusammenfassen?

Beispiel: Sympathie zu Spitzenpolitikern und Parteien ⟵ Parteianhängerschaften

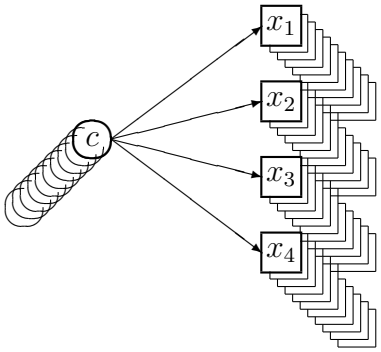


Abb. 1.10: Clusteranalyse (c enthält die Zugehörigkeit zu den neugebildeten Gruppen)

Kapitel 2

Meßniveaus und Skalentypen

Jede Modellierung — und also auch die Messung — repräsentiert Arten von anscheinend ähnlichen Dingen der Wirklichkeit durch Klassen von gleichartigen Objekten und die Eigenschaften von Dingen durch Attribute von Objekten. Messen ist dann die Operation, vermittels derer — nach Beobachtung einzelner Dinge der Wirklichkeit — einzelnen Objektinstanzen, die diese Dinge repräsentieren, Attributwerte zugeordnet werden.

Messen: Zuordnung von Symbolen (z.B. Zahlenwerten) zu den Ausprägungen von Eigenschaften (Merkmalen) bestimmter Merkmalsträger auf einer bestimmten Attributdimension (Variable) nach festen Regeln, so daß gleichen Merkmalsausprägungen gleiche Symbole (als Attributwerte) zugeordnet werden.

Die *Arten des Messens* unterscheiden sich in ihren Zuordnungsregeln (vgl. [Besozzi/Zehnpfennig 1976, S. 10ff]):

- *Fundamentales Messen*:
  Ableitung der Zuordnungsregeln durch kontrollierte Beobachtungen und Experimente aus Naturgesetzen — etwa bei der Messung von *Kräften* unter Zugrundelegung des Hebelgesetzes; freilich wird auch hier (wie beim "theoriegeleiteten Messen") vorausgesetzt, daß die Messung der Länge des Hebelarms bereits definiert ist.

- *Theoriegeleitetes Messen* (abgeleitetes Messen):
  Anwendung einer theoretisch definierten Zuordnungsregel — etwa bei der Messung der *Masse* unter Zugrundelegung des Newtonschen Gesetzes $F = ma$; hier wird vorausgesetzt, daß Kräfte und Beschleunigungen bereits ohne das Konzept der Masse meßbar sind.

- *Messen durch Abzählen* (Messen 'by counting'):
  Häufigkeiten des Auftretens gleicher Objekte/Objekteigenschaften in einer bestimmten Zeit und einem bestimmten Raum — auch hier wird eine Theorie vorausgesetzt, "die der Bildung homogener Klassen von Ereignissen implizit ist" [Besozzi/Zehnpfennig 1976, S. 11].

- *Messen durch zweckmäßiges Definieren und Vereinbaren einer Zuordnungsregel* (Messen 'by fiat'):
  Zuordnung eines Zahlenwertes zu einer bestimmten definierten Beobachtung — die Definition der Zuordnungsregel ist hier zunächst arbiträr; erst die Aufdeckung stabiler Beziehungen — wenn es solche denn gibt — zwischen 'by fiat' gemessenen Variablen erlaubt später, die Messungen theoretisch zu untermauern.

Eng verbunden mit dem Begriff der Messung ist der Begriff der Skala:

Skala: Homomorphe Abbildung von Merkmalsausprägungen auf eine Menge von Symbolen.

Skalen werden nach den auf sie anwendbaren Operationen klassifiziert, diese Klassifikationen werden *Skalenniveaus* oder auch *Meßniveaus* genannt. Skalentypen werden nach der *Art der Abbildung* von Merkmalsausprägungen auf eine Menge von Symbolen unterschieden:

- *Nichtmetrische (qualitative) Skalen*
  - Nominalskala

    Beschreibung: Abbildung von Untersuchungseinheiten nach der *Äquivalenz* von Merkmalsausprägungen in ungeordnete, sich gegenseitig ausschließende und den Wertebereich des Merkmals vollständig abdeckende Kategorien. Gibt es nur zwei solcher Kategorien, wird auch von einer *dichotomen* Skala gesprochen.

    Operationen: Gleichheit, Ungleichheit

Beispiele: - Geschlecht (männlich, weiblich)
- Konfession (katholisch, protestantisch, ...)

– Ordinalskala

Beschreibung: Abbildung von Untersuchungseinheiten auf eine Skala, die eine Rangordnung zwischen den Untersuchungseinheiten darstellt (*Ordnungsrelation*)

Operationen: alle Vergleichsoperationen

Beispiele: - Schulabschluß (o. Abschluß, Hauptschule, ...)
- Schichtzugehörigkeit (Unter-, Mittelschicht,...)

- *Metrische (quantitative) Skalen*

– Intervallskala

Beschreibung: Abbildung von Untersuchungseinheiten auf eine Skala gleicher Distanzen zwischen den einzelnen Merkmalsausprägungen

Operationen: alle Vergleichsoperationen und alle linearen Transformationen

Beispiele: - Kalender
- Temperatur
- Sympathieskalometer ($+5, \ldots, -5$)

– Ratioskala

Beschreibung: Abbildung von Untersuchungseinheiten auf eine Skala, bei der das Verhältnis der Maßeinheiten gleich ist, d.h. die Ratioskala besitzt einen absoluten Nullpunkt

Operationen: alle arithmetischen Operationen

Beispiele: - Gewicht
- Alter

Die Skalenniveaus haben aufgrund der unterschiedlich mächtigen Operationen eine bestimmte Rangordnung:

$$\begin{array}{c} \text{Skalenniveau} \\ \text{qualitativ} \quad < \quad \text{metrisch} \\ \text{nominal} \; < \; \text{ordinal} \; < \; \text{intervall} \; < \; \text{ratio} \end{array}$$

Kapitel 3

Univariate Datenanalyse

3.1 Die Beschreibung von Häufigkeitsverteilungen

Grundlage einer statistischen Analyse ist die Verteilung der Untersuchungseinheiten auf die einzelnen Antwortkategorien der Variablen. Mit Hilfe geeigneter Darstellungsformen erlauben Häufigkeitsverteilungen einen ersten Einblick in die Struktur der erhobenen Daten.

Zur Charakterisierung der Häufigkeiten werden dabei für jeden Skalenpunkt folgende Kennzahlen verwendet:

- *Absolute Häufigkeiten* der Merkmalsausprägungen (`Frequency`):

  $$f_k, \qquad (k = 1, \ldots, l)$$

  Anzahl der Untersuchungseinheiten in einer Merkmalsklasse k
  ($l =$ Anzahl der Klassen bzw. der unterschiedenen Meßwerte)

- *Relative Häufigkeiten* der Merkmalsausprägungen (`Percent`, `Valid Percent`):

  $$h_k = \frac{f_k}{n}$$

  Verhältnis der Anzahl f_k der Merkmalsträger in einer Merkmalsklasse k zur Gesamtheit n aller Analyseeinheiten

- *Kumulierte Häufigkeiten* der Merkmalsausprägungen (`Cum Percent`):

$$c_t = \frac{1}{n}\sum_{k=1}^{t} f_k = \sum_{k=1}^{t} h_k, \qquad (1 \le t \le l)$$

  Summation aller relativen Häufigkeiten h_k bis zu einer festgelegten Merkmalsklasse t

Als Beispiel seien in tabellarischer Form die Häufigkeitsverteilung und die wesentlichen statistischen Kennzahlen für die Frage nach der Sympathie für die CDU aufgeführt:

> "Und was halten Sie — so ganz allgemein — von der CDU?
> Sagen Sie es bitte anhand einer Skala von $+5$ bis -5:
> $+5$ bedeutet, daß Sie sehr viel von der CDU halten.
> -5 bedeutet, daß Sie überhaupt nichts von ihr halten."

V213 SKALOMETER: CDU

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
	-5	27	5.4	5.4	5.4
	-4	20	4.0	4.0	9.4
	-3	35	7.0	7.0	16.5
	-2	32	6.4	6.4	22.9
	-1	30	6.0	6.0	28.9
	0	51	10.2	10.2	39.2
	1	51	10.2	10.2	49.4
	2	69	13.8	13.9	63.3
	3	69	13.8	13.9	77.1
	4	70	14.0	14.1	91.2
	5	44	8.8	8.8	100.0
	.	2	.4	Missing	
	Total	500	100.0	100.0	

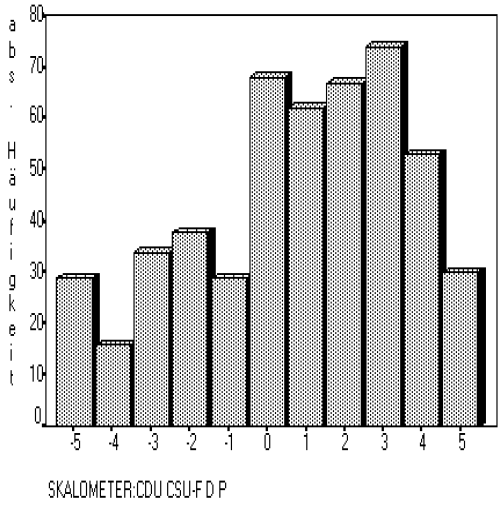
Mean	.968	Std err	.130	Median	2.000
Mode	4.000	Std dev	2.896	Variance	8.389
Kurtosis	-.793	S E Kurt	.218	Skewness	-.499
S E Skew	.109	Range	10.000	Minimum	-5.000
Maximum	5.000	Sum	482.000		

Diese Form der Darstellung verliert ihre Übersichtlichkeit, wenn die Anzahl der Antwortkategorien sehr hoch ist (z.B. Alter der Befragten) bzw. eine echt kontinuierliche Antwortskala wie bei physikalischen Meßgrößen vorliegt (z.B. Tem-

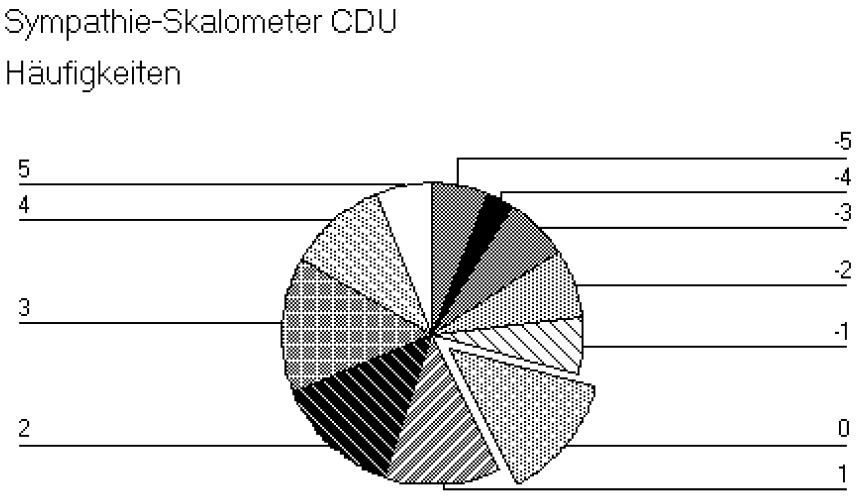
peratur). In diesem Fall empfiehlt sich eine vorherige Zusammenfassung der Variablenausprägungen in Gruppen (z.B. Altersklassen), die ihrerseits dann die Antwortkategorien für eine neu zu bildende Variable sind. Hierbei ist zu berücksichtigen, daß eine derartige Gruppenbildung immer einen von der gewählten Klassengröße abhängigen Informationsverlust darstellt, da mit steigender Klassengröße, d.h. einer Verringerung der Klassenzahl, auch die Unbestimmtheit des ursprünglichen, individuellen Variablenwertes zunimmt (vgl. [Böker 1993, S. 15ff]). Zudem eröffnet die Wahl der Klassengrenzen Möglichkeiten zur Manipulation von Häufigkeitsverteilungen.

Eine weitere Art, “das Wesentliche” einer Häufigkeitsverteilung schnell aufzunehmen bzw. zu vermitteln, sind graphische Darstellungen in verschiedenen Formen, von denen nachfolgend einige beispielhaft aufgeführt sind. Hierbei ist zu berücksichtigen, daß sich durch die Hinzunahme von Graphiken die Gefahr von verzerrten, d.h. bewußt oder unbewußt manipulierten Darstellungen von Häufigkeitsverteilungen vergrößert (vgl. [Krämer 1992, S. 29ff]).

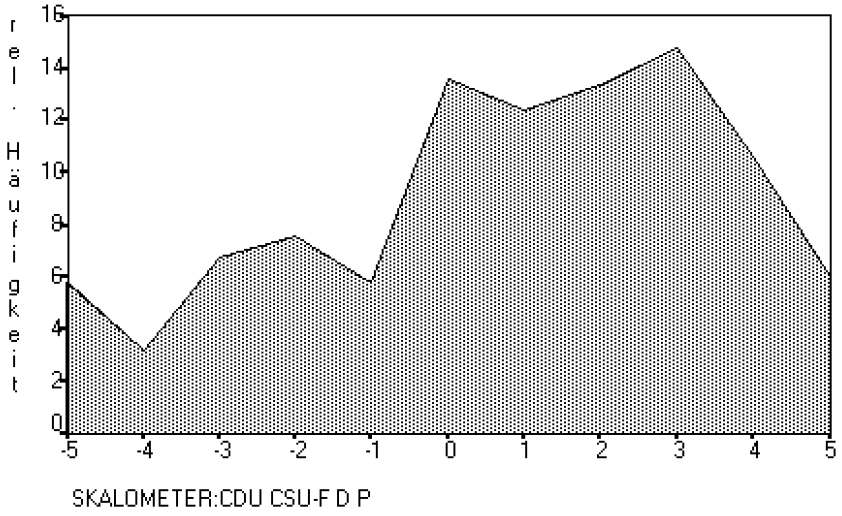
1. Balkendiagramm/Histogramm

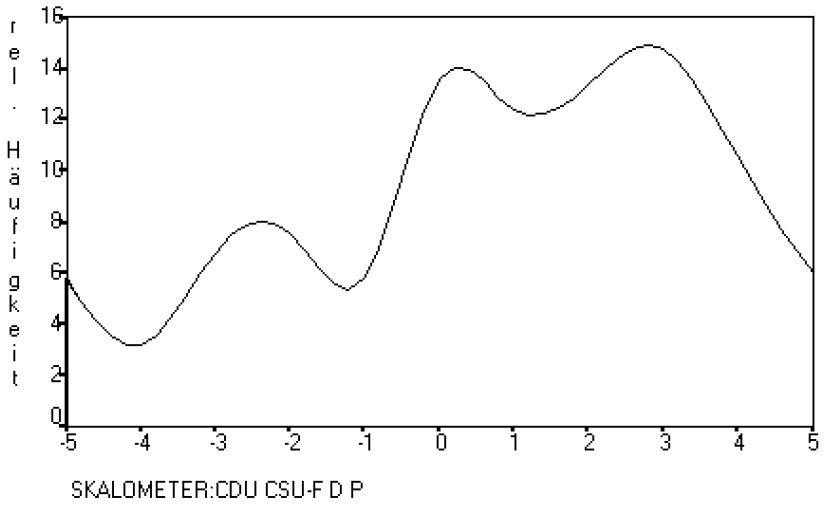


2. Tortendiagramm



3. Abschnitts-/Kurvendiagramm





Neben den eher übersichtsartigen, tabellarisch bzw. graphisch beschriebenen Häufigkeitsverteilungen ermöglicht die deskriptive Statistik auch spezifischere Charakterisierungen von Variablen. Hierzu existieren eine Reihe von Maßzahlen, die sich zum einen auf die zentrale Tendenz, den Schwerpunkt oder auch die Konzentration von Variablenwerten und zum anderen auf die Art ihrer Verteilung, ihrer Streuung bzw. ihrer Homogenität über alle Merkmalskategorien hinweg beziehen.

In den folgenden Abschnitten wird auf diese Maßzahlen eingegangen; sie sind nach dem Skalenniveau klassifiziert, das die zugrundeliegende Variable für eine Anwendung wenigstens aufweisen muß. Die Anwendung auf Variablen mit höherem Skalenniveau ist dabei immer möglich.

3.2 Kennzahlen für die "zentrale Tendenz" einer Variablen

Modus (häufigster/dichtester Wert, Modalwert, mode)

Skalenniveau: nominal

Definition: Der Modus h einer Verteilung bezeichnet den Variablenwert, der die höchste Fallzahl hat. Ist das Skalenniveau metrisch, so ergibt sich der Modus bei mehreren, nebeneinander liegenden Werten mit gleicher, höchster Fallzahl aus deren arithmetischem Mittel. Bei gruppierten Merkmalen wird der mittlere Variablenwert aus der Klasse mit der höchsten Fallzahl genommen.

Bemerkung: Der Vorteil des Modus liegt in seiner einfachen Definition und in seiner Verwendungsmöglichkeit auch für nominalskalierte Variablen. Weiterhin erlaubt er im Gegensatz zu den anderen Kennzahlen die Charakterisierung mehrgipfliger Verteilungen (z.B. bimodal), da mehrere nichtbenachbarte, hohe Häufigkeitszahlen aufweisende Merkmalsausprägungen als selbständige Modalwerte aufgefaßt werden, d.h. einer Verteilung können durchaus mehrere Modalwerte zugewiesen werden.

Beispiele: Der Modus der (intervallskalierten) Sympathieskalometerfrage für die CDU ist +4 ($\hat{=}$ 14.1%). Für den Fall, daß der Wert für die Merkmalsklasse +3 ebenfalls 14.1% wäre, ergäbe sich ein Modus von +3.5.
Sei in einer gruppierten Altersvariablen die Gruppe der 30- bis 39-jährigen am stärksten besetzt, so wäre der Modus dieser Variablen 34.5.

Median (Zentralwert, median)

Skalenniveau: ordinal

Definition: Der Median $\tilde{x}$ einer Verteilung ist der Variablenwert, der in der Mitte der nach der Größe sortierten Variablenwerte steht. Dabei muß zwischen einer geraden und einer ungeraden Anzahl n von Untersuchungseinheiten unterschieden werden.

Berechnung:

$$\tilde{x} = \begin{cases} x_i,\ i = \dfrac{n+1}{2} & \textit{falls n ungerade} \\ \dfrac{x_{n/2} + x_{(n/2)+1}}{2} & \textit{falls n gerade} \end{cases}$$

Bemerkung: Bei gruppierten Variablenwerten wird ein *fiktiver Medianwert* berechnet, basierend auf der Annahme, daß die Variablenwerte innerhalb der Klassenintervalle gleichverteilt sind. Sowohl obige Berechnung als auch die Interpolation bei gruppierten Variablenwerten (siehe Beispiel) lassen sich strenggenommen nicht auf ordinalen Daten durchführen. Daher wird bei einer geraden Anzahl von Untersuchungseinheiten häufig auch anstatt eines fiktiven mittleren Wertes einer der beiden mittleren (z.B. $\tilde{x} = x_{n/2}$) bzw. bei gruppierten Variablenwerten der Wert dieser Merkmalsklasse als Median angenommen.

Beispiel: Der Median der Sympathieskalometerfrage für die CDU ergibt sich zu

$$\tilde{x} = \frac{x_{250} + x_{251}}{2} = \frac{2+2}{2} = 2.$$

Eine Medianberechnung für gruppierte Variablenwerte läßt sich wie folgt durchführen ([Benninghaus 1985, S. 41]):

Klassenintervall	Häufigkeit	kumulierte Häufigkeit
6 – 8	5	5
9 – 11	10	15
12 – 14	14	29
15 – 17	13	42
18 – 20	11	53
21 – 23	16	69
24 – 26	19	88
27 – 29	12	100

Der Median liegt im Klassenintervall 18 – 20 (= Klasse m) mit den exakten Grenzen $x_u = 17.5$ und $x_o = 20.5$. Die Berechnung des Medians ergibt sich durch Interpolation:

$$\begin{aligned}
\tilde{x} &= x_u + \frac{n/2 - c_{m-1}}{f_m} * (x_o - x_u) \\
&= 17.5 + \frac{50 - 42}{11} * (20.5 - 17.5) \\
&= 17.5 + 0.73 * 3 \\
&= 19.7
\end{aligned}$$

Arithmetisches Mittel (Mittelwert, Durchschnittswert, arithmetic average, arithmetic mean)

Skalenniveau: metrisch

Definition: Das arithmetische Mittel $\bar{x}$ einer Verteilung ergibt sich aus der Summe der Variablenwerte, dividiert durch ihre Anzahl.

Berechnung:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{k=1}^{\ell} f_i x_i$$

Bemerkung: Das arithmetische Mittel als bekannteste Kennzahl für die "zentrale Tendenz" einer Variablen eignet sich insbesondere für die Beschreibung symmetrischer, unimodaler Verteilungen und zeichnet sich durch folgende Eigenschaften aus:

1. Die Addition (Subtraktion) einer Zahl zu allen Variablenwerten vergrößert (verkleinert) $\bar{x}$ um diese Zahl:

$$x_i \pm c,\ (i = 1, \ldots, n) \Longrightarrow \bar{x}' = \bar{x} \pm c$$

2. Die Summe der Abweichungen aller Variablenwerte von $\bar{x}$ ist gleich Null:

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

3. Die Summe der quadrierten Abweichungen aller Variablenwerte von $\bar{x}$ (=Variation) ist ein Minimum, d.h. sie ist kleiner als die Summe der quadrierten Abweichungen von einem beliebigen anderen Wert:

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \textit{Minimum}$$

Beispiel: Das arithmetische Mittel der Sympathieskalometerfrage für die CDU ergibt sich zu

$$\bar{x} \ = \ 0.968$$

3.3 Kennzahlen für die Streuung einer Variablen

Entropie

Skalenniveau: nominal

Definition: Bei dem aus der Thermodynamik und der Informationstheorie bekannten Entropie-Maß H wird die Streuung eines Merkmals als Konzentration der Untersuchungseinheiten auf bestimmte Merkmalsklassen bzw. als Abweichung vom Modell der Gleichverteilung verstanden, bei der jede Merkmalsklasse gleich stark besetzt ist.

Berechnung:[1]

$$H = \sum_{i=1, h_i \neq 0}^{k} h_i \text{ lb } \frac{1}{h_i}$$

Bemerkung: Die Entropie hat folgende Eigenschaften:

1. Falls *alle* Untersuchungseinheiten sich auf *eine* Merkmalsklasse konzentrieren ($\exists j,\ 1 \leq j \leq k:\ h_j = 1, h_i = 0$ für $i \neq j$), ist die Entropie minimal, d.h. 0:

$$H = h_j \text{ lb } \frac{1}{h_j} = 1 \text{ lb } 1 = 0$$

2. Falls die Untersuchungseinheiten sich gleichmäßig auf alle Merkmalsklassen verteilen ($\forall i,\ 1 \leq i \leq k:\ h_i = \frac{1}{k}$), ist die Entropie maximal, d.h. lb k:

$$H = \sum_{k=1}^{\ell} \frac{1}{k} \text{ lb } k = k \, \frac{1}{k} \text{ lb } k = \text{lb } k$$

Die Entropie einer Variablen liegt somit immer zwischen 0 und lb k.

[1] lb bezeichnet den Zweierlogarithmus: lb $= \log_2$

1. $H = 0$ 2. $H = \text{lb}\, k$

Abb. 3.1: Extremsituationen für die Entropie H

Quartilabstand (interquartile range)

Skalenniveau: ordinal

Definition: Grundlage für die Berechnung des Quartilabstandes ist der Begriff der Quantile. Quantile teilen eine nach ihrer Größe geordnete Menge von Untersuchungseinheiten in q Klassen mit einer jeweils *gleichen Anzahl* von Meßwerten ein. Je nach Anzahl der Quantile wird von Tertilen (q = 3), Quartilen (q = 4), . . . , Oktilen (q = 8), etc. gesprochen. Der Quartilabstand Q berechnet sich aus der Differenz zwischen dem Index des kleinsten Meßwerts des vierten Quartils (oder des größten des dritten Quartils) und dem Index des größten Meßwertes des ersten Quartils. Eine Erweiterung stellt der *mittlere Quartilabstand* (quartile deviation) Q_m dar, bei dem der Quartilabstand zusätzlich noch durch 2 dividiert wird.

Berechnung:

$$Q = Q_3 - Q_1 \qquad \text{bzw.} \qquad Q_m = \frac{Q_3 - Q_1}{2}$$

mit:

Q_1 : Schnittpunkt zwischen den ersten 25 Prozent der Verteilung und dem Rest

Q_3 : Schnittpunkt zwischen den ersten 75 Prozent der Verteilung und dem Rest

Bemerkung: Der Quartilabstand dient zur Beschreibung der Streubreite eines Merkmals, ohne daß auf die Extremwerte bezug genommen wird. Der Wert für Q_2, d.h. der Schnittpunkt zwischen den beiden Hälften einer Verteilung, entspricht dabei dem Median $\tilde{x}$ dieser Verteilung.

Beispiel: Bei der Skalometerfrage zur CDU-Sympathie fällt das erste Quartil Q_1 auf den Skalometerwert -1, das dritte (Q_3) auf den Skalometerwert 3. Somit beträgt der Quartilabstand:

$$Q = Q_3 - Q_1 = 3 - (-1) = 4$$

Die Berechnung des Quartilabstands für gruppierte Variablenwerte läßt sich wie folgt durchführen
([Benninghaus 1985, S. 52–53]):

Klassenintervall	Häufigkeit	kumulierte Häufigkeit	
6 – 8	5	5	
9 – 11	10	15	
12 – 14	14	29	$\Rightarrow Q_1$
15 – 17	13	42	
18 – 20	11	53	
21 – 23	16	69	
24 – 26	19	88	$\Rightarrow Q_3$
27 – 29	12	100	

Nach Festlegung der Merkmalsklassen, in die Q_1 und Q_3 fallen, erfolgt ihre konkrete Berechnung durch Interpolation entsprechend der Medianberechnung:

$$\begin{aligned} Q_1 &= 11.5 + \frac{1/4 * 100 - 15}{14} * 3 = 11.5 + 2.1 = 13.6 \\ Q_3 &= 23.5 + \frac{3/4 * 100 - 69}{19} * 3 = 23.5 + 0.9 = 24.4 \\ Q &= Q_3 - Q_1 = 24.4 - 13.6 = 10.8 \\ Q_m &= \frac{Q_3 - Q_1}{2} = \frac{10.8}{2} = 5.4 \end{aligned}$$

Variationsbreite/Spannweite (range)

Skalenniveau: metrisch

Definition: Die Variationsbreite R beschreibt die Differenz zwischen der größten (x_{max}) und der kleinsten (x_{min}) Ausprägung eines Merkmals.

Berechnung:

$$R = x_{max} - x_{min}$$

Bemerkung: Die Variationsbreite ist ein sehr einfaches Streuungsmaß, das die Verteilung eines Merkmals (z.B. im Gegensatz zum Quartilabstand) lediglich durch die zwei extremsten Merkmalsausprägungen beschreibt. Dies ist um so ungenauer, je untypischer diese Extremwerte für die Merkmalsverteilung sind. Die Variationsbreite wird deswegen weniger zur Beschreibung von Verteilungseigenschaften eingesetzt als vielmehr zur Kontrolle, ob die Variablenwerte in einem erwarteten Bereich liegen (z.B. Kontrolle auf Eingabe- bzw. Kodierungsfehler).

Beispiel: Die Variationsbreite für die Sympathieskalometerfrage bezüglich der CDU ergibt sich zu

$$R = +5 - (-5) = 10,$$

da alle zur Auswahl stehenden Skalenpunkte von den Befragten auch angekreuzt wurden.

Durchschnittliche Abweichung (average deviation)

Skalenniveau: metrisch

Definition: Die durchschnittliche Abweichung AD wird gemessen als die Summe der Absolutbeträge aller Abweichungen der Meßwerte vom arithmetischen Mittel, dividiert durch die Gesamtzahl aller Untersuchungseinheiten.

Berechnung:

$$AD = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}| = \frac{1}{n}\sum_{k=1}^{\ell} f_i|x_i - \bar{x}|$$

Bemerkung: Bei der Berechnung der durchschnittlichen Abweichung geht jeder Variablenwert mit dem gleichen Gewicht ein. Die Maßeinheit der durchschnittlichen Abweichung ist dabei mit derjenigen der Variablenwerte identisch. AD wird zugunsten der Standardabweichung s (siehe Folgeseite) in der Praxis kaum noch verwendet.

Beispiel: Die durchschnittliche Abweichung beträgt für die CDU-Sympathie (mit dem Mittelwert $\bar{x} = 0.968$):

$$AD = 2.39$$

Varianz (variance)/ Standardabweichung (standard deviation)

Skalenniveau: metrisch

Definition: Die Varianz s^2 einer Variablen ist definiert als die Summe der quadrierten Abweichungen der Variablenwerte von ihrem Mittelwert, dividiert durch die Gesamtanzahl der Variablenwerte minus 1. Die Standardabweichung s, auch mit SD bezeichnet, ergibt sich als Wurzel aus der Varianz.

Berechnung:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{1}{n-1}\sum_{k=1}^{\ell} f_i(x_i-\bar{x})^2$$

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2} = \sqrt{\frac{1}{n-1}\sum_{k=1}^{\ell} f_i(x_i-\bar{x})^2}$$

Bemerkung: Im Gegensatz zur durchschnittlichen Abweichung werden zur Berechnung der Varianz/Standardabweichung die Abweichungen der Variablenwerte vom Mittelwert mit ihrem Quadrat gewichtet. Dies führt dazu, daß zur Beschreibung der Streuung einer Variablen dem Betrag nach größere Abweichungen stärker berücksichtigt werden als kleinere. Bei gleicher durchschnittlicher Abweichung variiert demnach das Ergebnis für die Varianz je nach dem Anteil von Variablenwerten mit großen Abweichungen vom Mittelwert. Varianz und Standardabweichung sind grundsätzlich als gleichwertige Streuungsmaße anzusehen, wobei die Standardabweichung für die Beschreibung insofern vorzuziehen ist, da sie die gleiche Maßeinheit besitzt wie die zugrundeliegenden Variablenwerte.

Beispiel: Die Varianz bzw. Standardabweichung für die CDU-Sympathie ergibt sich bei einem Mittelwert von $\bar{x} = 0.968$ wie folgt (Zahlen sind gerundet):

$$s^2 = 8.389$$

$$s = 2.896$$

Bei der Berechnungsformel für s^2 finden im Nenner sowohl $n-1$ als auch — wie vielleicht erwartet — n Verwendung, wobei der Unterschied im Ergebnis lediglich

bei kleinen Stichproben ins Gewicht fällt (sei $\hat{\sigma}^2$ erwartungstreuer (unverzerrter) Schätzer für die Varianz; vgl. S. 63):

$$\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 \qquad\qquad |\cdot\frac{n}{n}\\
&= \frac{n}{n-1}\cdot\underbrace{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}_{=s^2 \text{ mit } n \text{ statt } n-1}\\
&= \frac{n}{n-1}s^2
\end{aligned}$$

Die Gründe für die unterschiedlichen Berechnungsformeln liegen in inferenzstatistischen Betrachtungen, d.h. der Übertragung von Ergebnissen aus einer Stichprobe auf die Grundgesamtheit, und hängen mit dem Begriff der *Erwartungstreue* zusammen. Wird inferenzstatistisch nicht von einer einzigen sondern von einer Vielzahl von Zufallsstichproben ausgegangen, die (hypothetisch) aus einer Grundgesamtheit gezogen werden, so läßt sich der Stichprobenmittelwert als eine Zufallsvariable $\bar{X}$ auffassen, deren Erwartungswert $E(\bar{X}) = \mu$ ist (vgl. Kapitel 4), d.h. die Mittelwerte der Stichproben verteilen sich "gleichmäßig" um den wahren Mittelwert μ der Grundgesamtheit, $\bar{X}$ ist erwartungstreu. Diese Eigenschaft ist auch für den Erwartungswert der Zufallsvariablen für die Stichprobenvarianz s^2 wünschenswert. Die Herleitung des Erwartungswertes führt aber zu dem Ergebnis

$$E(s^2) = \frac{n-1}{n}\sigma^2$$

d.h. der Erwartungswert ist in Abhängigkeit von der Stichprobengröße n verzerrt. Erst die Verwendung von $n-1$ anstatt n in der Berechnungsformel führt zu der Erwartungstreue für die Varianz:

$$\begin{aligned}
E(\hat{\sigma}^2) &= E(\frac{n}{n-1}s^2)\\
&= \frac{n}{n-1}E(s^2)\\
&= \frac{n}{n-1}\cdot\frac{n-1}{n}\sigma^2\\
&= \sigma^2
\end{aligned}$$

3.4 Kennzahlen für den Test einer Häufigkeitsverteilung auf Normalverteilung

Für die Charakterisierung einer Variablen ist ebenfalls die Frage von Bedeutung, inwieweit sich die empirische Häufigkeitsverteilung der Normalverteilung annähert.

Definition: (nach [Kreyszig 1979, S. 126ff])
Eine stetige Verteilung mit der Wahrscheinlichkeitsdichte

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

heißt *Gauß-Verteilung* oder *Normalverteilung* und wird graphisch durch die sogenannte *Glockenkurve* beschrieben.

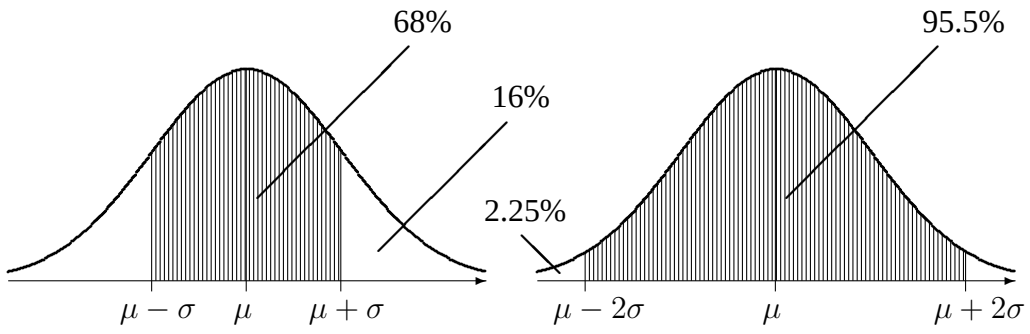


Abb. 3.2: Normalverteilung

Interpretation: Die Fläche unter der Kurve gibt die Wahrscheinlichkeit an, mit der sich ein Wert der Zufallsvariablen in einem gegebenem Intervall befindet. Die Gesamtfläche ist gleich 1.

Die herausragende Bedeutung der Normalverteilung und den aus ihr abgeleiteten Wahrscheinlichkeitsverteilungen (z.B. χ^2-, F-, t-Verteilung) ergibt sich aus der Rolle, die sie bei der Durchführung inferenzstatistischer Schlüsse, d.h. der Übertragung von Ergebnissen aus der Stichprobe auf die Grundgesamtheit, spielen. Grundlage hierfür ist der *zentrale Grenzwertsatz*, dessen wesentliche Aussagen wie folgt zusammengefaßt werden können (nach [Patzelt 1985, S. 196–197])[2]:

[2] Für die genaue Definition siehe auch [Kreyszig 1979, S. 201].

1. Werden m Zufallsstichproben der Größe n aus einer Grundgesamtheit gezogen, deren interessierendes Merkmal metrisch skaliert und durch einen Mittelwert μ und eine Varianz σ^2 beschreibbar ist, und werden die arithmetischen Mittel $\bar{x}_i$ ($i = 1, \ldots, m$) der m Zufallsstichproben berechnet, so nähert sich die *Häufigkeitsverteilung der Mittelwerte* $\bar{x}_i$ einer *Normalverteilung* an, und zwar völlig unabhängig davon, welche Verteilung das Merkmal selbst hat.

2. Die Verteilung der Stichprobenmittelwerte $\bar{x}_i$ nähert sich um so stärker einer Normalverteilung an, je größer der Stichprobenumfang n ist.

3. Der Mittelwert dieser angenäherten Normalverteilung der Stichprobenmittelwerte $\bar{x}_i$ ist genau der Mittelwert μ des in der Grundgesamtheit interessierenden Merkmals.

4. Die Varianz der angenäherten Normalverteilung der Stichprobenmittelwerte $\bar{x}_i$ ist $\frac{\sigma^2}{n}$. Ihre Wurzel $\frac{\sigma}{\sqrt{n}}$ wird *Standardfehler* (standard error, SE) genannt, da sie wie eine Standardabweichung die Streuung der Stichprobenmittelwerte $\bar{x}_i$ um den Grundgesamtheitsmittelwert μ ausdrückt.

Dies hat für die Inferenzstatistik folgende Konsequenzen:

1. Das Ausmaß der Annäherung der Stichprobenmittelwerte hängt ausschließlich vom Stichprobenumfang n ab, wobei der Unterschied ab $n = 30$ vernachlässigbar ist.

2. Mit wachsendem n wird die Streuung der Mittelwerte um μ immer geringer ($s_{\bar{x}_i} = \frac{\sigma}{\sqrt{n}}$).

3. Da die $\bar{x}_i$ praktisch normalverteilt sind, kann angegeben werden, mit welcher Wahrscheinlichkeit μ in welchem Intervall um *einen* Stichprobenmittelwert $\bar{x}_i$ liegt, d.h. es genügt ein einziger Mittelwert aus einer Stichprobe mit $n \geq 30$, um eine solche Schätzung vorzunehmen:

$$P(\bar{x} - k\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + k\frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

mit:

k : Intervallgrenzen
α : Irrtumswahrscheinlichkeit

Analoges gilt für die Schätzung von σ^2 aus einer einzigen Stichprobenvarianz s^2.

Die konkrete Durchführung eines inferenzstatistischen Schlusses auf der Basis einer einzigen Stichprobe ist somit auch davon abhängig, ob die Ausgangsvariable(n) einer Normalverteilung oder einer ihr verwandten Wahrscheinlichkeitsverteilung genügt. Nach dem Gauß'schen Fehlerverteilungstest lassen sich folgende Bedingungen für die Annahme einer Normalverteilung als Modell einer Häufigkeitsverteilung annehmen:

- Viele zufällige Faktoren wirken unabhängig voneinander und additiv auf das fragliche Merkmal.
- Der Einfluß jedes einzelnen Faktors ist dabei relativ gering.
- Verstärkende und abschwächende Effekte sind in etwa gleich wahrscheinlich.

Während für eine Reihe von Variablen, wie zum Beispiel die Körpergröße oder das Gewicht, bekannt ist, daß sie normalverteilt bzw. annähernd normalverteilt sind, gibt es für die Überprüfung einer Variablen auf Normalverteilung weitere Kennzahlen, die nachfolgend behandelt werden.

Schiefe (Skewness)

Skalenniveau: metrisch

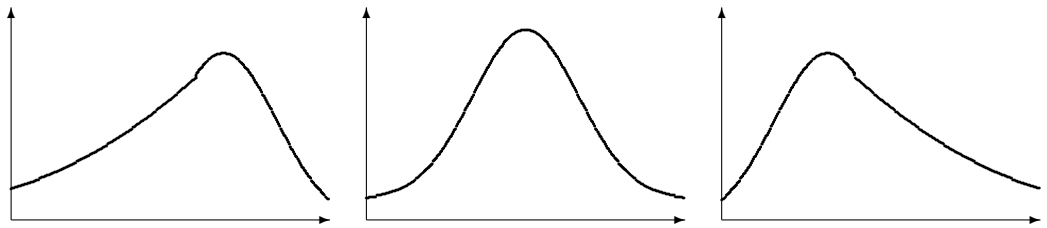
Definition: Bei der Schiefe γ_1 (= 3. Moment einer Verteilung) wird überprüft, ob eine eingipflige Verteilung gemessen an der Normalverteilungskurve nach rechts oder nach links verzogen ist.

Berechnung:

$$\gamma_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{s^3}$$

Beispiele: Abbildung 3.3 zeigt die unterschiedlichen Verteilungstypen, die sich aus den Werten für γ_1 erkennen lassen.

Die Schiefe für die Verteilung der CDU-Sympathie beträgt $\gamma_1 = -0.449$, sie ist linksschief und rechtssteil (vgl. den Graphen in der Zusammenfassung auf Seite 46).



linksschiefe, rechtssteile Verteilung $\gamma_1 < 0$ | symmetrische Verteilung $\gamma_1 = 0$ | rechtsschiefe, linkssteile Verteilung $\gamma_1 > 0$

Abb. 3.3: Graphische Bedeutung der Schiefe

Exzeß, Wölbung (Kurtosis)

Skalenniveau: metrisch

Definition: Bei der Wölbung γ_2 (= 4. Moment einer Verteilung) wird überprüft, ob eine eingipflige Verteilung gemessen an der Normalverteilungskurve flacher oder steiler ist.

Berechnung:

$$\gamma_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{s^4}$$

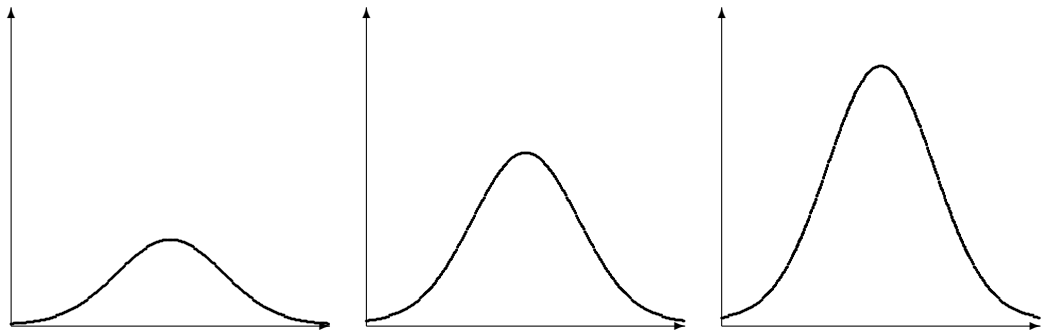
Bemerkung: Beim Vorliegen einer Normalverteilung hat γ_2 den Wert 3. Deshalb wird meist

$$\gamma_2' = \gamma_2 - 3$$

verwendet.

Beispiele: Die Verteilungstypen, die sich aus den Werten für γ_2 erkennen lassen, sind in Abbildung 3.4 dargestellt.

Der Graph zur CDU-Sympathieverteilung (s. S. 46) ist mit einem Wert von $\gamma_2 = -0.793$ flach gewölbt gegenüber der Normalverteilung.



flache Wölbung $\gamma_2' < 0$ | normale Wölbung $\gamma_2' = 0$ | steile Wölbung $\gamma_2' > 0$

Abb. 3.4: Graphische Bedeutung der Wölbung

3.5 Zusammenfassung

Maßzahlen zur Beschreibung von Häufigkeitsverteilungen	
absolute Häufigkeit	$f_k, \quad (k = 1, \ldots, l)$
relative Häufigkeit	$h_k = \dfrac{f_k}{n}$
kumulierte Häufigkeit	$c_t = \dfrac{1}{n}\sum_{k=1}^{t} f_k = \sum_{k=1}^{t} h_k, \quad 1 \le t \le l$

Kennzahl	Skalenniveau[3]				Berechnung
	qual.		metr.		
	n	o	i	r	
Kennzahlen für die "zentrale Tendenz" einer Variablen					
Modus	+	+	+	+	
Median	-	+	+	+	$\tilde{x} = \begin{cases} x_i,\ i = \dfrac{n+1}{2} & \textit{falls } n \textit{ ungerade} \\ \dfrac{x_{n/2} + x_{(n/2)+1}}{2} & \textit{falls } n \textit{ gerade} \end{cases}$
Mittelwert	-	-	+	+	$\bar{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i = \dfrac{1}{n}\sum_{k=1}^{\ell} f_i x_i$
Kennzahlen für die Streuung einer Variablen					
Entropie	+	+	+	+	$H = \sum_{i=1, h_i \neq 0}^{k} h_i \text{ lb } \dfrac{1}{h_i}$
Quartilabstand	-	+	+	+	$Q = Q_3 - Q_1 \ \text{ bzw. } \ Q_m = \dfrac{Q_3 - Q_1}{2}$
Variationsbreite	-	-	+	+	$R = x_{max} - x_{min}$
Durchschnittliche Abweichung	-	-	+	+	$AD = \dfrac{1}{n}\sum_{i=1}^{n} \lvert x_i - \bar{x} \rvert$
Varianz	-	-	+	+	$s^2 = \dfrac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2$
Standardabweichung	-	-	+	+	$s = \sqrt{\dfrac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2}$
Kennzahlen für den Test einer Häufigkeitsverteilung					
Schiefe	-	-	+	+	$\gamma_1 = \dfrac{1}{s^3} \cdot \dfrac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^3$
Kurtosis	-	-	+	+	$\gamma_2 = \dfrac{1}{s^4} \cdot \dfrac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^4$

[3] qual. = qualitativ, metr. = metrisch; n = nominal, o = ordinal, i = intervall, r = ratio

Anhand der Häufigkeitsverteilungen zweier Variablen sollen die spezifischen Eigenschaften der beschriebenen Kennzahlen vergleichend diskutiert werden.

Dazu werden zusätzlich zu der bereits auf Seite 24 vorgestellten Frage nach der Sympathie für die CDU die Häufigkeiten der Sympathieskalometerfrage für Strauß betrachtet.

V229 SKALOMETER: F J STRAUSS

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
	-5	63	12.6	12.7	12.7
	-4	25	5.0	5.0	17.7
	-3	38	7.6	7.6	25.3
	-2	20	4.0	4.0	29.3
	-1	25	5.0	5.0	34.3
	0	46	9.2	9.2	43.6
	1	67	13.4	13.5	57.0
	2	54	10.8	10.8	67.9
	3	57	11.4	11.4	79.3
	4	56	11.2	11.2	90.6
	5	47	9.4	9.4	100.0
	.	2	.4	Missing	
		500	100.0	100.0	
	Total				

Mean	.424	Std err	.145	Median	1.000
Mode	1.000	Std dev	3.234	Variance	10.462
Kurtosis	-1.107	S E Kurt	.218	Skewness	-.348
S E Skew	.109	Range	10.000	Minimum	-5.000
Maximum	5.000	Sum	211.000		

Die graphische Darstellung (in Abbildung 3.5) macht die unterschiedliche Verteilung der Merkmalsträger auf die einzelnen Skalenpunkte deutlich: Die stark polarisierte Meinung über Strauß in der Öffentlichkeit führt auch zu einer zweigipfligen Verteilung, während die Sympathie für die CDU eine “normale” Häufigkeitsverteilung mit einem Maximum zeigt.

Inwiefern spiegeln sich die Differenzen zwischen den empirischen Verteilungen in den zur Verfügung stehenden univariaten Kennzahlen wider? Beschreiben diese das Befragungsergebnis in adäquater Art und Weise, oder können sie durch Besonderheiten des Kurvenverlaufs der Stichprobenwerte verzerrt bzw. verfälscht werden? Die nachfolgende Tabelle listet für beide Variablen die wichtigsten Kennwerte auf.

Abb. 3.5: Gegenüberstellung: Sympathie CDU – Sympathie Strauß

Kennzahl		CDU	Strauß
Modus	h	4	1
Median	$\tilde{x}$	2	1
Mittelwert	$\bar{x}$	0.968	0.424
Quartilabstand	Q	4	6
Varianz	s^2	8.389	10.462
Standardabweichung	s	2.896	3.234
Schiefe	γ_1	-0.499	-0.348
Kurtosis	γ_2	-0.793	-1.107

Die Maßzahlen zur Beschreibung der zentralen Tendenz sind vor allem bei mehrgipfligen Verteilungen wenig aussagekräftig. Ebenso verleitet die *alleinige* Betrachtung von Schiefe und Kurtosis bei multiplen Gipfeln zu einer falschen Einschätzung (z.B. ähnelt die Strauß-Variable in bezug auf die Symmetrie eher einer Normalverteilung als die CDU-Variable), da der Vergleichsmaßstab eine eingipflige Verteilung ist.

Somit bleibt festzuhalten, daß die Reduzierung der Verteilungsinformationen auf einen einzigen Kennwert, insbesondere bei “nicht normalen” Verteilungen, zu falschen Schlüssen führen kann. Die Berechnung und Verwendung solcher Kennzahlen entbindet den Anwender daher nicht von der Notwendigkeit, sich grundsätzlich ein umfassendes Bild von den zugrundeliegenden empirischen Verteilungen zu machen.

Kapitel 4

Grundlagen der Inferenzstatistik

4.1 Zufallsvariable, Erwartungswert, Wahrscheinlichkeitsverteilung

Empirische Untersuchungen erfolgen in der Regel auf der Basis einer Stichprobe, da eine vollständige Einbeziehung der Grundgesamtheit meist nicht möglich ist (z.B. Kosten, Zerstörung von Untersuchungsobjekten). Da andererseits die Zusammenhänge in der Grundgesamtheit von eigentlichem Interesse sind,[1] ergibt sich die Notwendigkeit eines *Inferenzschlusses* von den Ergebnissen in der Stichprobe auf die in der Grundgesamtheit, d.h. es ist zu überprüfen, inwieweit die berechneten statistischen Kennzahlen aus der Stichprobe (Mittelwert, Varianz, ...) auch für die Grundgesamtheit gültig (*signifikant*) sind. Das Instrument für diese Überprüfung ist der sogenannte Signifikanz- oder auch *Hypothesentest*. Daraus abgeleitet lassen sich die unbekannten Parameter der Grundgesamtheit (Mittelwert (μ), Varianz (σ^2), ...) mit Hilfe der Stichprobenparameter (Mittelwert ($\bar{x}$), Varianz (s^2), ...) abschätzen, d.h. für ihre Werte können — bei Angabe einer Irrtumswahrscheinlichkeit — Vertrauensintervalle (*Konfidenzintervalle*) angegeben werden. Die wahrscheinlichkeitstheoretischen Grundlagen für die Durchführung von Hypothesentests und für die Berechnung von Vertrauensintervallen sollen in diesem Kapitel erläutert werden.

Grundsätzlich besteht das Problem, Aussagen über eine Menge von Untersuchungseinheiten treffen zu müssen (Grundgesamtheit), von der Informationen le-

[1] Bei einer repräsentativen Bevölkerungsumfrage mit einer Stichprobe von ca. 2000 Untersuchungseinheiten (z.B. [Wahlstudie 1987]) interessieren letztendlich Aussagen über die Gesamtbevölkerung und nicht über die befragten Individuen.

diglich über eine Teilmenge (Stichprobe) vorliegen. Der Lösungsweg hierfür läßt sich zusammenfassend wie folgt beschreiben:

1. Unter der Voraussetzung, daß es sich bei der Stichprobe um eine echte Zufallsauswahl handelt, läßt sich jeder aus der Stichprobe berechnete statistische Wert (z.B. $\bar{x}, s^2$) als die *Realisierung einer Zufallsvariablen* auffassen.

2. Mit Hilfe dieser Zufallsvariablen wird eine neue Zufallsvariable konstruiert (vgl. Anhang A.3), die einer bekannten, in Tabellenform vorliegenden *Wahrscheinlichkeitsverteilung* (Prüfverteilung) entspricht (z.B. Normalverteilung, χ^2-Verteilung).

3. Mit Hilfe der Stichprobe läßt sich für diese konstruierte Zufallsvariable eine einzelne Realisierung berechnen. Die Position dieses "empirischen" Wertes innerhalb der bekannten Wahrscheinlichkeitsverteilung der Zufallsvariablen erlaubt dann Aussagen darüber, ob der berechnete Wert auch für die Grundgesamtheit signifikant ist oder ob das Ergebnis im Bereich von Zufallsschwankungen liegt bzw. in welchem Intervall um den zu testenden Stichprobenparameter der wahre Wert der Grundgesamtheit liegt.

	Stichprobenwerte			
1. Stichprobe	$x_{11}, x_{12}, \ldots, x_{1n}$	$\Longrightarrow$	$\bar{x}_1 = \frac{1}{n}\sum_{i=1}^{n} x_i$	1. Schätzwert für μ
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
m. Stichprobe	$x_{m1}, x_{m2}, \ldots, x_{mn}$	$\Longrightarrow$	$\bar{x}_m = \frac{1}{n}\sum_{i=1}^{n} x_i$	m. Schätzwert für μ
Zufallsvariablen	$X_1, X_2, \ldots, X_n$	$\Longrightarrow$	$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$	

Die Tabelle erläutert die Betrachtung von Stichprobenwerten als Realisierung von Zufallsvariablen. Dabei wird davon ausgegangen, daß aus der Grundgesamtheit nicht nur eine sondern (hypothetisch) m Stichproben der Größe n gezogen werden. Der zu testende Stichprobenparameter sei dabei der Stichprobenmittelwert.

In reinen Zufallsstichproben aus einer Grundgesamtheit gilt für die einer Realisierung zugeordneten Zufallsvariablen:

$$E(X_i) = \mu$$

Für den Erwartungswert der Stichprobenmittelwerte gilt dann:

$$E(\bar{X}) = \mu$$

Damit ist der Stichprobenmittelwert in reinen Zufallsstichproben eine *erwartungstreue* (d.h. unverzerrte) Schätzung des Mittelwerts der Grundgesamtheit.

Herleitung:

$$\begin{aligned} E(\bar{X}) &= E(\frac{1}{n}\sum_{i=1}^{n} X_i) \\ &= \frac{1}{n}E(\sum_{i=1}^{n} X_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} E(X_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} \mu \\ &= \frac{1}{n}n\mu \\ &= \mu \end{aligned}$$

Für die Beurteilung der Güte einer Schätzfunktion für den "wahren" Parameter in einer Grundgesamtheit (hier: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$) reicht die Erwartungstreue allein nicht aus. Vielmehr ist weiterhin gefordert, daß die Schätzwerte nahe bei dem wahren Wert liegen, d.h. daß auch die Varianz der Stichprobenmittelwerte klein ist. Für identisch und unabhängig verteilte Zufallsvariablen X_i mit dem Erwartungswert μ und der Varianz σ^2 gilt für die Varianz des Stichprobenmittelwertes:

$$E([\bar{X} - \mu]^2) = V(\bar{x}) = \frac{\sigma^2}{n}$$

Herleitung:

$$\begin{aligned} V(\bar{x}) &= V(\frac{1}{n}\sum_{i=1}^{n} x_i) \\ &= \frac{1}{n^2}V(\sum_{i=1}^{n} x_i) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} V(x_i) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 \\ &= \frac{1}{n^2}n\sigma^2 \\ &= \frac{\sigma^2}{n} \end{aligned}$$

Daraus folgt für den Standardfehler, d.h. für die durchschnittliche Abweichung eines Stichprobenmittelwertes vom wahren Mittelwert:

$$SE(\bar{x}) = \frac{\sigma}{\sqrt{n}}$$

Weiterhin gilt nach dem zentralen Grenzwertsatz, daß die Stichprobenmittelwerte bei unabhängigen Stichproben gemäß der Normalverteilung streuen.

4.2 Die Durchführung von Hypothesentests

Die Überprüfung auf Signifikanz eines Stichprobenparameters für die Grundgesamtheit erfolgt durch die Anwendung eines Hypothesentests. Am Beispiel der Überprüfung des Stichprobenmittelwertes wird im folgenden das Prinzip eines Hypothesentests erläutert.

Frage: Ist der empirische Stichprobenmittelwert $\bar{x}$ auch für die Grundgesamtheit signifikant, oder liegt das Ergebnis im Bereich der Zufallsschwankungen?

1. Schritt: Hypothesenformulierung

$H_0 : \mu = 0$ (Nullhypothese)
d.h. der Stichprobenmittelwert $\bar{x}$ liegt im Bereich der Zufallsschwankungen (keine Signifikanz).

$H_A : \mu \neq 0$ (Alternativhypothese)
d.h. der Stichprobenmittelwert ist signifikant für die Grundgesamtheit.

Ziel des Hypothesentests ist es, die Nullhypothese zu widerlegen.

2. Schritt: Bestimmung der Prüfverteilung

Nach dem zentralen Grenzwertsatz verteilen sich die Stichprobenmittelwerte bei unabhängigen Stichproben gemäß der Normalverteilung. Standardisierung der Zufallsvariablen $\bar{X}$ mit dem wahren Mittelwert μ und dem wahren Standardfehler $SE(\bar{X})$ führt zu der standardnormalverteilten Zufallsvariable

$$\tilde{X} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

mit dem Mittelwert 0 und der Varianz 1. Die Existenz einer bekannten Verteilung für eine Zufallsvariable (hier: Stichprobenmittelwerte)

erlaubt den Hypothesentest auf der Basis lediglich einer Realisierung der Zufallsvariablen. Für die Prüfung anderer Stichprobenparameter lassen sich Zufallsvariablen konstruieren, die anderen charakteristischen Verteilungen entsprechen (z.B. χ^2-Verteilung, t-Verteilung, F-Verteilung).

3. Schritt: Bestimmung des Freiheitsgrads df

Im allgemeinen hängt der Kurvenverlauf einer Prüfverteilung von dem sogenannten Freiheitsgrad (df=degrees of freedom) ab. In diesem Beispiel gilt $df = n$, da bei der für die Mittelwertbildung notwendigen Summation der n Stichprobenwerte kein x_i festgelegt ist. Im Gegensatz dazu verringert sich beispielsweise der Freiheitsgrad bei der Variationsberechnung $\sum_{i=1}^{n}(x_i - \bar{x})^2$ um 1, da das bereits bekannte, sich aus allen x_i ($i = 1, \ldots, n$) ergebende $\bar{x}$ mitverwendet wird und somit mindestens *ein* x_i festgelegt ist. Der Freiheitsgrad bei der Varianzberechnung beträgt daher nur noch $n-1$ (vgl. Kapitel 3.3).

4. Schritt: Bestimmung des Signifikanzniveaus α

Die Existenz einer charakteristischen Verteilung für eine Zufallsvariable erlaubt es, innerhalb einer gewissen Fehlergrenze Schlüsse über die Grundgesamtheit auf der Basis lediglich einer Stichprobe zu ziehen. Die Festlegung dieser Fehlergrenze — oder auch Signifikanzniveau — α muß von dem Anwender vorgenommen werden und bezeichnet die Wahrscheinlichkeit dafür, daß H_0 fälschlicherweise verworfen wird (Fehler 1.Art). Mit welcher Fehlerwahrscheinlichkeit ist ein Anwender also bereit zu akzeptieren, daß er sich bei der Annahme der Signifikanz des Stichprobenmittelwertes für die Grundgesamtheit irrt (z.B. $\alpha = 0.05\ (5\%)$). Um Testverfälschungen zu vermeiden, sollte der Schritt 4 bereits vor Kenntnis der Daten durchgeführt werden.

5. Schritt: Ermittlung des theoretischen Verteilungswertes k und Berechnung des empirischen Verteilungswertes $\tilde{X}_{emp}$

Die Normalverteilung liegt tabelliert vor, so daß in Abhängigkeit vom Freiheitsgrad und dem gewählten Signifikanzniveau der “kritische” Wert k ermittelt werden kann:

$1-\alpha$	k
0.6827	1
0.95	1.96
0.9545	2
0.99	2.58
0.9973	3
0.999	3.3

Für die Berechnung des empirischen Verteilungswertes $\tilde{X}_{emp}$ aus $\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$ wird der Standardfehler $\frac{s}{\sqrt{n}}$ mit dem Schätzwert s für die Standardabweichung σ in der Grundgesamtheit verwendet.[2]

[2]Damit entspricht $\tilde{X}$ strenggenommen nicht mehr einer Normalverteilung sondern einer t-Verteilung mit $(n-1)$ Freiheitsgraden, d.h. es wird ein Hypothesentest für $\mu = 0$ auf der Basis einer t-Verteilung durchgeführt (vgl. t-Test in Abschnitt 5.4.5). Diese Unterscheidung ist aber lediglich für kleine Stichproben von Bedeutung, da die t-Verteilung sich für große n der Normalverteilung annähert.

6. Schritt: Entscheidung über die Ablehnung von H_0

Aufgrund der bekannten Eigenschaften der Standardnormalverteilung gilt folgende Beziehung für den Stichprobenmittelwert:

$$P(-k \leq \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq k) = P(-k \leq \tilde{X}_{emp} \leq k) = 1 - \alpha$$

$-k \leq \tilde{X}_{emp} \leq k$: Der Stichprobenmittelwert $\tilde{X}_{emp}$ befindet sich in dem gewählten Vertrauensbereich $(1 - \alpha)$, d.h. die Informationen reichen nicht aus, um die Nullhypothese zu verwerfen.

$|\tilde{X}_{emp}| > k$: Der Stichprobenmittelwert $\tilde{X}_{emp}$ befindet sich im Bereich der gewählten Irrtumswahrscheinlichkeit, d.h. es besteht nur noch eine Wahrscheinlichkeit von α, daß der $\tilde{X}_{emp}$-Wert größer als der Prüfwert ist und somit H_0 fälschlicherweise verworfen wird. (*Nicht:* H_A wird angenommen.)

Beispiel: Die beiden Sympathieskalometervariablen für die CDU bzw. Strauß aus dem vorhergehenden Kapitel ergeben in der Berechnung des empirischen Stichprobenmittelwertes $\tilde{X}_{emp}$:

$$\begin{aligned}
\tilde{X}_{emp} &= \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\bar{X} - 0}{\frac{s}{\sqrt{n}}} \\
\tilde{X}_{emp}^{CDU} &= \frac{0.968}{\frac{2.896}{\sqrt{500}}} = 7.474 \\
\tilde{X}_{emp}^{Strauss} &= \frac{0.424}{\frac{3.234}{\sqrt{500}}} = 2.932
\end{aligned}$$

Wird ein Signifikanzniveau von $\alpha = 5\%$ angenommen, so liegt der kritische Tabellenwert k bei 1.96 $(1 - \alpha = 0.95)$, d.h. die berechneten Werte befinden sich beide außerhalb des Vertrauensbereichs von H_0 und sind somit auf diesem Niveau signifikant. Ein Signifikanzniveau von $\alpha = 0.1\%$ $(1 - \alpha = 0.999)$ dagegen würde zu einem kritischen Wert k von 3.3 führen mit der Konsequenz, daß $\tilde{X}_{emp}^{Strauss}$ dann nicht mehr signifikant wäre. Die aus der unterschiedlichen Homogenität der beiden Variablen resultierenden Unterschiede für die Berechnung von $\tilde{X}_{emp}$ wirken sich somit auch auf die Signifikanz aus.

Aus der Berechnungsformel für $\tilde{X}_{emp}$ ergibt sich ebenso eine Abhängigkeit der Signifikanz von der Stichprobengröße n. So führt eine Berechnung auf der Basis von $n = 1534$ zu $\tilde{X}_{emp}^{CDU} = 16.29$ bzw. $\tilde{X}_{emp}^{Strauss} = 5.95$ und damit zu einer Signifikanz auf höherem Niveau.

4.3 Die Berechnung von Konfidenzintervallen

Für den Stichprobenmittelwert als standardnormalverteilte Zufallsvariable gilt:

$$P(-k \leq \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq k) = 1 - \alpha$$

Erläuterung: Die Wahrscheinlichkeit, daß der Wert der Zufallsvariablen sich zwischen den durch k und $-k$ definierten Intervallgrenzen befindet, beträgt $1 - \alpha$.

Die Intervallgröße und ihre jeweils zugehörigen Vertrauens- (Konfidenz-) Wahrscheinlichkeiten liegen dabei tabelliert vor (z.B. in [Kreyszig 1979, S. 423–424]).

Bemerkung: Während $1 - \alpha$ die Vertrauenswahrscheinlichkeit festlegt, bezeichnet α die Irrtumswahrscheinlichkeit (oder auch Signifikanzniveau), d.h. den Fehler, den ein Anwender bei gewählten Intervallgrenzen bereit ist zu akzeptieren (z.B. $\alpha = 0.05$ (5%)).

Mit Hilfe einiger Umformungen läßt sich das Konfidenzintervall für den Mittelwert der Grundgesamtheit μ bestimmen:

$$\begin{aligned}
P(-k \leq \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq k) &= 1 - \alpha \\
P(-k\frac{\sigma}{\sqrt{n}} \leq \bar{x} - \mu \leq k\frac{\sigma}{\sqrt{n}}) &= 1 - \alpha \\
P(-\bar{x} - k\frac{\sigma}{\sqrt{n}} \leq -\mu \leq -\bar{x} + k\frac{\sigma}{\sqrt{n}}) &= 1 - \alpha \\
P(\bar{x} + k\frac{\sigma}{\sqrt{n}} \geq \mu \geq \bar{x} - k\frac{\sigma}{\sqrt{n}}) &= 1 - \alpha \\
P(\bar{x} - k\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + k\frac{\sigma}{\sqrt{n}}) &= 1 - \alpha
\end{aligned}$$

Damit liegt μ im Intervall

$$[\bar{x} - d_{\bar{x}}, \bar{x} + d_{\bar{x}}], \quad \text{mit } d_{\bar{x}} = k\frac{\sigma}{\sqrt{n}} = k\ SE(\bar{x})$$

Das heißt, daß sich bei bekannter Varianz der Grundgesamtheit (bzw. bei Verwendung einer Stichprobenvarianz (z.B. aus einem Vortest) als Schätzwert), gewünschter Genauigkeit und akzeptierter Irrtumswahrscheinlichkeit ein benötigter Stichprobenumfang exakt vorausberechnen läßt.

Beispiel: Bei einer angenomenen Irrtumswahrscheinlichkeit von 5% ($\longrightarrow k = 1.96$) ergeben sich bezüglich der beiden Sympathieskalometervariablen folgende Werte für $d_{\bar{x}}$:

$$\begin{aligned} d_{\bar{x}}^{CDU} &= 1.96 * \frac{2.846}{\sqrt{500}} = 0.25 \\ d_{\bar{x}}^{Strauss} &= 1.96 * \frac{3.234}{\sqrt{500}} = 0.28 \end{aligned}$$

Wie erwartet, ergibt sich für die CDU-Sympathie aufgrund der größeren Homogenität ein geringerer Wert, d.h. das resultierende Vertrauensintervall um den wahren Mittelwert der Grundgesamtheit ist kleiner (0.5) als das für die Strauß-Sympathie (0.56).

Die — hier nachträglich durchgeführten — Berechnungen für die notwendigen Stichprobenumfänge ($k = 1.96$, $\sigma = 2.896/3.234$) ergeben bei einer höheren Genauigkeit (z.B. $d_{\bar{x}} = 0.2$):

$$\begin{aligned} n &= (k\frac{\sigma}{d_{\bar{x}}})^2 \\ n_{CDU} &= (1.96 * \frac{2.896}{0.2})^2 \approx 806 \\ n_{Strauss} &= (1.96 * \frac{3.234}{0.2})^2 \approx 1005 \end{aligned}$$

Die gewünschte Verkleinerung des Vertrauensintervalls hätte also eine erheblich Erhöhung des Stichprobenumfanges notwendig gemacht.

4.4 Parameterschätzungen aus geschichteten Stichproben

4.4.1 Schätzung des Mittelwerts

Der Standardfehler der Schätzung des Mittelwerts ist von der Varianz des Merkmals und vom Stichprobenumfang abhängig; daher wird bei Merkmalen großer Varianz eine große Stichprobe erforderlich, um den Mittelwert hinreichend genau zu schätzen. Wenn es gelingt, ein *Schichtungsmerkmal* zu finden, das die Grundgesamtheit in mehrere in sich bezüglich des Untersuchungsmerkmals ziemlich homogene, aber untereinander sehr verschiedene Schichten (*strata*) zerlegt, so genügt eine kleinere Stichprobe bereits den bestehenden Genauigkeitsanforderungen, und zwar um so eher, je höher das Schichtungsmerkmal mit dem Untersuchungsmerkmal korreliert ist, was gleichbedeutend damit ist, daß sich die Mittelwerte des Untersuchungsmerkmals in den einzelnen Schichten sehr stark voneinander unterscheiden. Hinzu kann kommen, daß in den einzelnen Schichten die Erhebungskosten je Untersuchungseinheit unterschiedlich sind. In diesem Fall wird versucht, den Standardfehler der Schätzung bei konstanten Kosten durch geeignete Wahl schichtspezifischer Auswahlsätze zu minimieren. Folgende Schichtungsverfahren sind zu unterscheiden:

- *Proportionale Schichtung:*
  Aus jeder Schicht wird ein fester Anteil $f_j = \frac{n_j}{N_j} = f$ in die Stichprobe aufgenommen.

- *Minimum-Varianz-Schichtung:*
  Aus jeder Schicht wird ein variabler Anteil $f_j = \frac{n_j}{N_j}$ in die Stichprobe aufgenommen; die f_j werden so gewählt, daß der Standardfehler der Schätzung minimal wird.

- *Optimale Schichtung:*
  Aus jeder Schicht wird ein variabler Anteil $f_j = \frac{n_j}{N_j}$ in die Stichprobe aufgenommen; die f_j werden so gewählt, daß der Standardfehler der Schätzung bei vorgegebenen Erhebungskosten minimal wird.

Der mittlere Fall ist ein Spezialfall des letzteren, dort sind die Erhebungskosten schichtunabhängig. Auch der erste Fall ist ein Spezialfall der beiden anderen: dort sind außerdem die Varianzen schichtunabhängig. Schließlich ist auch die einfache Zufallsauswahl ein "Spezialfall" aller komplizierteren Verfahren: dort sind auch noch die Schichtenmittelwerte gleich (Schichtungs- und Untersuchungsmerkmal sind nicht miteinander korreliert).

Für die weiteren Ableitungen in diesem Kapitel werden die nachstehenden Bezeichnungen verwendet:

N_j	:	Anzahl der Untersuchungseinheiten in der j-ten Schicht der Grundgesamtheit
N	:	Anzahl der Untersuchungseinheiten in der Grundgesamtheit
n_j	:	Anzahl der Untersuchungseinheiten in der j-ten Schicht der Stichprobe
n	:	Anzahl der Untersuchungseinheiten in der Stichprobe
c_j	:	Erhebungskosten je Untersuchungseinheit in der j-ten Schicht der Grundgesamtheit
x_{ji}	:	Merkmalswert der i-ten Untersuchungseinheit in der j-ten Schicht
μ	:	Mittelwert der Grundgesamtheit
μ_j	:	Mittelwert in der j-ten Schicht der Grundgesamtheit
$\hat{\mu}$, $\hat{\mu}_j$	:	entsprechende Schätzwerte aus der Stichprobe
σ	:	Standardabweichung der Grundgesamtheit
σ_j, σ_j^2	:	Standardabweichung bzw. Varianz in der j-ten Schicht der Grundgesamtheit
$E\langle\cdot\rangle$	:	Erwartungswert einer Stichprobenschätzung

Damit gilt

$$\begin{aligned}
\mu &= \frac{1}{N}\sum_{j=1}^{J} N_j\mu_j \\
\mu_j &= \frac{1}{N_j}\sum_{i=1}^{N_j} x_{ji} \\
\hat{\mu}_j &= \frac{1}{n_j}\sum_{i=1}^{n_j} x_{ji} \\
\hat{\mu} &= \frac{1}{N}\sum_{j=1}^{J}\sum_{i=1}^{n_j} \frac{N_j}{n_j} x_{ji}
\end{aligned}$$

Gesucht und zu minimieren ist

$$E\langle(\hat{\mu}-\mu)^2\rangle = E\langle[\frac{1}{N}\sum_{j=1}^{J} N_j(\frac{1}{n_j}\sum_{i=1}^{n_j} x_{ji}) - \frac{1}{N}\sum_{j=1}^{J} N_j\mu_j]^2\rangle$$

$$
\begin{aligned}
&= E\langle \frac{1}{N^2}[\sum_{j=1}^{J} N_j(\underbrace{\frac{1}{n_j}\sum_{i=1}^{n_j} x_{ji} - \mu_j)}_{=S_j}]^2\rangle \\
&= \frac{1}{N^2} E\langle [\sum_{j=1}^{J} N_j S_j]^2\rangle \\
&= \frac{1}{N^2} E\langle N_1^2 S_1^2 + 2N_1 N_2 S_1 S_2 + \ldots + N_J^2 S_J^2\rangle
\end{aligned}
$$

wobei

$$
\begin{aligned}
E\langle N_j^2 S_j^2\rangle &= N_j^2 E\langle(\hat{\mu}_j - \mu_j)^2\rangle = N_j^2 \frac{\sigma_j^2}{n_j} \\
E\langle N_j N_k S_j S_k\rangle &= N_j N_k E\langle(\hat{\mu}_j - \mu_j)(\hat{\mu}_k - \mu_k)\rangle = 0
\end{aligned}
$$

so daß

$$
\begin{aligned}
E\langle(\hat{\mu} - \mu)^2\rangle &= \frac{1}{N^2} E\langle \sum_{j=1}^{J} N_j^2 \frac{\sigma_j^2}{n_j}\rangle \\
&= \sum_{j=1}^{J} \frac{N_j^2 \sigma_j^2}{N^2 n_j}
\end{aligned}
$$

Für $J = 1$ ergibt sich speziell $SE[\hat{\mu}] = \frac{\sigma}{\sqrt{n}}$. Von den in dieser Gleichung vorkommenden Termen sind die n_j frei wählbar bei fixierten Kosten $c = \sum_{j=1}^{J} c_j n_j$. Es gilt also, $E\langle(\hat{\mu} - \mu)^2\rangle$ in Abhängigkeit von allen n_j unter Beachtung der Nebenbedingung $c = \sum_{j=1}^{J} c_j n_j = \text{const.}$ zu minimieren:

$$
H(n_1, \ldots, n_j, \ldots, n_J) = \sum_{j=1}^{J} \frac{N_j^2 \sigma_j^2}{N^2 n_j} - \lambda(c - \sum_{j=1}^{J} c_j n_j)
$$

Die zugehörigen Ableitungen nach n_j lauten:

$$
\frac{\partial H}{\partial n_j} = \frac{N_j^2 \sigma_j^2}{N^2} \cdot \frac{-1}{n_j^2} + \lambda c_j
$$

Diese sind 0 für

$$
\begin{aligned}
\lambda n_j^2 &= N_j^2 \frac{\sigma_j^2}{N^2 c_j} \\
n_j &= \frac{1}{N\sqrt{\lambda}} \cdot \frac{N_j \sigma_j}{\sqrt{c_j}}
\end{aligned}
$$

Nun gilt aber

$$\begin{aligned}\sum_{j=1}^{J} n_j &= \frac{1}{N\sqrt{\lambda}} \sum_{j=1}^{J} \frac{N_j \sigma_j}{\sqrt{c_j}} = n \\ \frac{1}{N\sqrt{\lambda}} &= \frac{n}{\sum_{j=1}^{J} \frac{N_j \sigma_j}{\sqrt{c_j}}}\end{aligned}$$

so daß sich schließlich ergibt:

$$\begin{aligned}n_j &= \frac{n}{\sum_{j=1}^{J} \frac{N_j \sigma_j}{\sqrt{c_j}}} \cdot \frac{N_j \sigma_j}{\sqrt{c_j}} \\ &= n \frac{N_j \sigma_j / \sqrt{c_j}}{\sum_{j=1}^{J} (N_j \sigma_j / \sqrt{c_j})}\end{aligned}$$

Dies ist die Formel für die Teilstichprobenumfänge bei der optimalen Schätzung. Bei schichtunabhängigen Kosten ($c_j = \text{const.}$) ergibt sich für die Minimum-Varianz-Schichtung

$$n_j = n \frac{N_j \sigma_j}{\sum_{j=1}^{J} N_j \sigma_j}$$

und bei schichtunabhängiger Varianz ($\sigma_j = \text{const.}$) für die proportionale Schichtung

$$n_j = n \frac{N_j}{\sum_{j=1}^{J} N_j}$$

4.4.2 Schätzung der Varianz

Die Varianz der Grundgesamtheit ergibt sich wie üblich aus der Summe der quadrierten Abweichungen der Merkmalswerte aller Untersuchungseinheiten vom gemeinsamen Mittelwert nach Division durch die Anzahl der Untersuchungseinheiten der Grundgesamtheit; hier wird diese Summe lediglich Schicht für Schicht gebildet:

$$\sigma^2 = \frac{1}{N}\sum_{j=1}^{J}\sum_{i=1}^{N_j}(x_{ji}-\mu)^2$$

Durch die folgenden Umformungen ergibt sich

$$\begin{aligned}
\sigma^2 &= \frac{1}{N}\sum_{j=1}^{J}[\sum_{i=1}^{N_j}(x_{ji}-\mu_j+\mu_j-\mu)^2] \\
&= \frac{1}{N}\sum_{j=1}^{J}[\sum_{i=1}^{N_j}[(x_{ji}-\mu_j)+(\mu_j-\mu)]^2] \\
&= \frac{1}{N}\sum_{j=1}^{J}\sum_{i=1}^{N_j}[(x_{ji}-\mu_j)^2+2(x_{ji}-\mu_j)(\mu_j-\mu)+(\mu_j-\mu)^2] \\
&= \frac{1}{N}\sum_{j=1}^{J}[\sum_{i=1}^{N_j}(x_{ji}-\mu_j)^2+2(\mu_j-\mu)\sum_{i=1}^{N_j}(x_{ji}-\mu_j)+N_j(\mu_j-\mu)^2] \\
&= \frac{1}{N}[\sum_{j=1}^{J}\sum_{i=1}^{N_j}(x_{ji}-\mu_j)^2+\sum_{j=1}^{J}N_j(\mu_j-\mu)^2]
\end{aligned}$$

Schließlich kann die Varianz der Grundgesamtheit dargestellt werden als Summe der mit den Schichtenumfängen gewichteten schichtinternen Varianzen und der ebenfalls mit den Schichtenumfängen gewichteten quadrierten Abweichungen der Schichtenmittelwerte vom Gesamtmittelwert. Die beiden Summanden werden auch als "Varianz innerhalb der Schichten" und "Varianz zwischen den Schichten" bezeichnet.

$$\sigma^2 = \frac{1}{N}\sum_{j=1}^{J}N_j\sigma_j^2+\frac{1}{N}\sum_{j=1}^{J}N_j(\mu_j-\mu)^2$$

Wie bei einfachen Zufallsstichproben ist auch bei geschichteten die Varianz der Stichprobe kein erwartungstreuer Schätzer für die Varianz der Grundgesamtheit. Zunächst wird die Varianz der Stichprobe als durch die Stichprobenlänge dividierte Summe der angemessen gewichteten quadrierten Abweichungen der Merkmalswerte der Untersuchungseinheiten vom Stichprobenmittelwert (letzterer ist

ein erwartungstreuer Schätzer) ermittelt:

$$\begin{aligned} s^2 &= \frac{1}{n}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\frac{N_j n}{N n_j}(x_{ji}-\hat{\mu})^2 \\ &= \frac{1}{N}\sum_{j=1}^{J}\frac{N_j}{n_j}\sum_{i=1}^{n_j}[(x_{ji}-\hat{\mu}_j)+(\hat{\mu}_j-\hat{\mu})]^2 \\ &= \frac{1}{N}\sum_{j=1}^{J}\frac{N_j}{n_j}[\sum_{i=1}^{n_j}(x_{ji}-\hat{\mu}_j)^2+2(\hat{\mu}_j-\hat{\mu})\sum_{i=1}^{n_j}(x_{ji}-\hat{\mu}_j)+n_j(\hat{\mu}_j-\hat{\mu})^2] \end{aligned}$$

Dies ist aufgrund der gleichen Überlegungen wie zuvor wieder die Summe aus "Varianz innerhalb der Schichten" und "Varianz zwischen den Schichten":

$$s^2 = \frac{1}{N}\sum_{j=1}^{J}N_j s_j^2 + \frac{1}{N}\sum_{j=1}^{J}N_j(\hat{\mu}_j-\hat{\mu})^2$$

Hiervon ist nun der Erwartungswert zu nehmen.

$$\begin{aligned} E\langle s^2\rangle &= \frac{1}{N}\sum_{j=1}^{J}N_j E\langle s_j^2\rangle + \frac{1}{N}\sum_{j=1}^{J}N_j E\langle(\hat{\mu}_j-\hat{\mu})^2\rangle \\ E\langle s^2\rangle &= \frac{1}{N}\sum_{j=1}^{J}\frac{N_j(n_j-1)}{n_j}\sigma_j^2 + \frac{1}{N}\sum_{j=1}^{J}N_j E\langle(\hat{\mu}_j-\hat{\mu})^2\rangle \end{aligned}$$

Der zweite Summand der letzten Gleichung wird umgeformt:

$$\frac{1}{N}\sum_{j=1}^{J}N_j E\langle(\hat{\mu}_j-\hat{\mu})^2\rangle = \frac{1}{N}\sum_{j=1}^{J}N_j E\langle[(\hat{\mu}_j-\mu_j)-(\hat{\mu}-\mu)+(\mu_j-\mu)]^2\rangle$$

Die Summe in eckigen Klammern ist zu quadrieren, ihr Erwartungswert ist zu bilden, und schließlich ist über j zu summieren. Beim Quadrieren der Summe in eckigen Klammern entstehen drei quadratische Terme und drei gemischte Terme, und zwar die letzteren wie folgt:

$$\begin{aligned} E\langle\frac{1}{N}\sum_{j=1}^{J}N_j(\hat{\mu}_j-\mu_j)(\hat{\mu}-\mu)\rangle &= E\langle(\hat{\mu}-\mu)\frac{1}{N}\sum_{j=1}^{J}N_j(\hat{\mu}_j-\mu_j)\rangle \\ &= E\langle(\hat{\mu}-\mu)^2\rangle \end{aligned}$$

denn

$$\frac{1}{N}\sum_{j=1}^{J}N_j(\hat{\mu}_j-\mu_j) = (\hat{\mu}-\mu)$$

Die beiden anderen gemischten Terme verschwinden bei Erwartungswert- und Summenbildung:

$$
\begin{aligned}
E\langle \frac{1}{N}\sum_{j=1}^{J} N_j(\mu_j-\mu)(\hat{\mu}-\mu)\rangle &= \overbrace{E\langle(\hat{\mu}-\mu)\rangle}^{=0}\overbrace{\frac{1}{N}\sum_{j=1}^{J} N_j(\mu_j-\mu)} = 0 \\
E\langle \frac{1}{N}\sum_{j=1}^{J} N_j(\hat{\mu}_j-\mu_j)(\mu_j-\mu)\rangle &= \frac{1}{N}\sum_{j=1}^{J} \underbrace{N_j(\mu_j-\mu)}\,\underbrace{E\langle(\hat{\mu}_j-\mu_j)\rangle}_{=0} = 0
\end{aligned}
$$

Zusammenfassend läßt sich der zweite Summand (s. Vorseite) nun folgendermaßen schreiben:

$$
\begin{aligned}
\frac{1}{N}\sum_{j=1}^{J} N_j E\langle(\hat{\mu}_j-\hat{\mu})^2\rangle &= \frac{1}{N}\sum_{j=1}^{J} N_j E\langle(\hat{\mu}_j-\mu_j)^2+(\mu_j-\mu)^2-(\hat{\mu}-\mu)^2\rangle \\
&= \frac{1}{N}\sum_{j=1}^{J} N_j E\langle(\hat{\mu}_j-\mu_j)^2\rangle + \frac{1}{N}\sum_{j=1}^{J} N_j(\mu_j-\mu)^2 \\
&\quad - E\langle(\hat{\mu}-\mu)^2\rangle\frac{1}{N}\sum_{j=1}^{J} N_j \\
&= \frac{1}{N}\sum_{j=1}^{J} N_j\frac{\sigma_j^2}{n_j} + \frac{1}{N}\sum_{j=1}^{J} N_j(\hat{\mu}-\mu)^2 - \sum_{j=1}^{J}\frac{N_j^2\sigma_j^2}{N^2 n_j}
\end{aligned}
$$

Damit kann der Erwartungswert der Stichprobenvarianz notiert werden:

$$
\begin{aligned}
E\langle s^2\rangle &= \frac{1}{N}[\sum_{j=1}^{J}\frac{N_j(n_j-1)}{n_j}\sigma_j^2 + \sum_{j=1}^{J} N_j\frac{\sigma_j^2}{n_j} + \sum_{j=1}^{J} N_j(\hat{\mu}-\mu)^2] - \sum_{j=1}^{J}\frac{N_j^2\sigma_j^2}{N^2 n_j} \\
&= \underbrace{\frac{1}{N}\sum_{j=1}^{J} N_j\sigma_j^2 + \frac{1}{N}\sum_{j=1}^{J} N_j(\hat{\mu}-\mu)^2}_{=\sigma^2} - \sum_{j=1}^{J}\frac{N_j^2\sigma_j^2}{N^2 n_j}
\end{aligned}
$$

Die beiden ersten Summanden der rechten Gleichungsseite ergeben zusammen σ^2 (siehe Seite 60); im dritten Summanden wird σ_j^2 durch seinen erwartungstreuen Schätzer ersetzt:

$$E\langle s^2\rangle = \sigma^2 - \frac{1}{N^2}\sum_{j=1}^{J}(\frac{N_j^2}{n_j}\cdot\frac{n_j}{n_j-1}E\langle s_j^2\rangle)$$

Hieraus folgt als erwartungstreue Schätzung für die Varianz in der Grundgesamtheit:

$$\hat{\sigma}^2 = s^2 + \frac{1}{N^2}\sum_{j=1}^{J}\frac{N_j^2}{n_j-1}s_j^2$$

Für $J = 1$ ergibt sich speziell:

$$\hat{\sigma}^2 = s^2(1+\frac{1}{n-1}) = s^2\frac{n}{n-1}$$

Kapitel 5

Bivariate Datenanalyse

5.1 Bivariate Häufigkeitsverteilungen

Im Rahmen der bivariaten Datenanalyse wird der Zusammenhang zwischen je zwei Variablen (= Spalten in der Datenmatrix) untersucht. Allgemein läßt sich der Begriff *Zusammenhang* hier im Sinne einer *Beeinflussung* der Merkmalswerte beider Variablen interpretieren, wobei es insbesondere um *Stärke* und *Richtung* dieser Beeinflussung geht. Analog zur univariaten Datenanalyse bildet die empirische Häufigkeitsverteilung von zwei Variablen die Grundlage für die Bestimmung von bivariaten Zusammenhangsmaßen, die sich nach unterschiedlichen Kriterien charakterisieren lassen:

1. *Skalenniveau*
   Entsprechend dem zugrunde liegenden Skalenniveau der beiden untersuchten Variablen wird (auch begrifflich) unterschieden zwischen:

   - *Kontingenzkoeffizienten (nominal)*
   - *Assoziationskoeffizienten (ordinal)*
   - *Korrelationskoeffizienten (metrisch)*

   Zusätzlich gibt es Zusammenhangsmaße auch für zwei Variablen mit unterschiedlichen Skalenniveaus (z.B. nominal/metrisch).

2. *Richtung*

   - *symmetrisch*
     Das Zusammenhangsmaß bezeichnet einen allgemeinen Zusammenhang zwischen beiden Variablen, ohne Bevorzugung einer bestimmten Richtung.
   - *asymmetrisch*
     Das Zusammenhangsmaß gilt nur in einer bestimmten Wirkungsrichtung, d.h. es gibt eine unabhängige und eine abhängige Variable.

3. *Wertebereich*

   0 (=kein Zusammenhang) $\rightarrow$ 1 (=perfekter Zusammenhang)
   -1 (=negativer Zusammenhang) $\rightarrow$ +1 (=positiver Zusammenhang)

Grundlage für die Berechnung von Zusammenhangsmaßen für nominal- bzw. ordinalskalierte Variablen ist die *Kreuztabelle* als gemeinsame Häufigkeitsverteilung zweier Variablen:

```
* SPSS-Aufruf: CROSSTABS /TABLES V334 BY V202.
V334 GESCHLECHT  by  V202 BTW:ZWEITSTIMME
                 V202
      Count     |CDU    SPD    FDP    Gruene NPD    Sonst.          Row
                |   1  |   2  |   3  |   4  |   5  |   6  |   9  | Total
V334 ----------+------+------+------+------+------+------+------+
          1     |  89  |  99  |  14  |  24  |   1  |   2  |   9  |   238
maennlich       |      |      |      |      |      |      |      |  48.5
                +------+------+------+------+------+------+------+
          2     |  96  | 116  |  11  |  18  |      |   2  |  10  |   253
weiblich        |      |      |      |      |      |      |      |  51.5
                +------+------+------+------+------+------+------+
      Column      185    215     25     42      1      4     19      491
       Total    37.7   43.8    5.1    8.6     .2     .8    3.9    100.0
```

Allgemein läßt sich eine Kreuztabelle wie folgt darstellen:

		Variable x					
		x_1	$\ldots$	x_j	$\ldots$	x_c	
	y_1	h_{11}	$\ldots$	h_{1j}	$\ldots$	h_{1c}	$h_{1\bullet}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Variable y	y_i	h_{i1}	$\ldots$	h_{ij}	$\ldots$	h_{ic}	$h_{i\bullet}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	y_r	h_{r1}	$\ldots$	h_{rj}	$\ldots$	h_{rc}	$h_{r\bullet}$
		$h_{\bullet 1}$	$\ldots$	$h_{\bullet j}$	$\ldots$	$h_{\bullet c}$	N

mit:

x_j, y_i	:	Merkmalsklassen der Variablen x und y
h_{ij}	:	Anzahl der Untersuchungseinheiten, die *gleichzeitig* in der Variable x den Merkmalswert x_j und in der Variable y den Merkmalswert y_i haben
$h_{i\bullet}, h_{\bullet j}$	:	Anzahl der Untersuchungseinheiten, die in der Variable y den Merkmalswert y_i bzw. in der Variable x den Merkmalswert x_j haben (*Randhäufigkeiten, Randauszählungen*)
N	:	Anzahl der Untersuchungseinheiten
c	:	Anzahl der Merkmalsklassen für die Variable x (c = *columns* = Anzahl der Spalten)
r	:	Anzahl der Merkmalsklassen für die Variable y (r = *rows* = Anzahl der Zeilen)

5.2 Zusammenhangsmaße zwischen zwei nominalskalierten Variablen

5.2.1 Maßzahlen auf der Basis von χ^2

Die in einer Kreuztabelle dargestellte empirische Verteilung der Untersuchungseinheiten bezüglich zweier Variablen enthält auch den zu ermittelnden Zusammenhang zwischen diesen beiden Variablen. Grundlage für die Berechnung des χ^2-Maßes ist dabei die Ermittlung einer weiteren Kreuztabelle unter der Annahme, daß es keinen Zusammenhang gibt (*Indifferenztabelle*). Eine solche Indifferenztabelle wird allgemein wie folgt berechnet:

Sei $P(x_j, y_i)$ die Wahrscheinlichkeit dafür, daß eine Untersuchungseinheit sich in der Kategorie x_j (der Variablen x) und gleichzeitig in der Kategorie y_i (der Variablen y) befindet. Unter der Annahme, daß zwischen beiden Ereignissen kein Zusammenhang besteht, d.h. daß beide Ereignisse unabhängig voneinander sind, gilt:

$$\begin{aligned} P(x_j, y_i) &= P(x_j)\ P(y_i) \\ &= \frac{\sum_{i=1}^{r} h_{ij}}{N} \cdot \frac{\sum_{j=1}^{c} h_{ij}}{N} \end{aligned}$$

$$= \frac{h_{\bullet j}}{N} \cdot \frac{h_{i\bullet}}{N}$$

$$= \frac{h_{i\bullet} h_{\bullet j}}{N^2}$$

Die *erwarteten* Besetzungen $\hat{h}_{ij}$ in der Indifferenztabelle ergeben sich dann aus:

$$\hat{h}_{ij} = P(x_j, y_i)N$$

$$= \frac{h_{i\bullet} h_{\bullet j}}{N^2} N$$

$$= \frac{h_{i\bullet} h_{\bullet j}}{N}$$

Somit sind die erwarteten Zellenbesetzungen in der Indifferenztabelle das Produkt der jeweiligen Randverteilungen, dividiert durch die Anzahl der Untersuchungseinheiten.

Aus obigem Beispiel resultiert danach folgende Indifferenztabelle, wobei jeweils auf ganze Zahlen auf- bzw. abgerundet wurde:

Geschlecht	Zweitstimme (BTW 1987)							
	CDU	SPD	FDP	Grüne	NPD	Sonst.		
männlich	90	104	12	20	0	2	9	237
weiblich	95	111	13	22	1	2	10	254
	185	215	25	42	1	4	19	491

mit:

$$\begin{aligned}
\hat{h}_{11} &= \frac{238 * 185}{491} = 90, & \hat{h}_{21} &= \frac{253 * 185}{491} = 95 \\
\hat{h}_{12} &= \frac{238 * 215}{491} = 104, & \hat{h}_{22} &= \frac{253 * 215}{491} = 111 \\
\hat{h}_{13} &= \frac{238 * 25}{491} = 12, & \hat{h}_{23} &= \frac{253 * 25}{491} = 13 \\
\hat{h}_{14} &= \frac{238 * 42}{491} = 20, & \hat{h}_{24} &= \frac{253 * 42}{491} = 22 \\
\hat{h}_{15} &= \frac{238 * 1}{491} = 0, & \hat{h}_{25} &= \frac{253 * 1}{491} = 1 \\
\hat{h}_{16} &= \frac{238 * 4}{491} = 2, & \hat{h}_{26} &= \frac{253 * 4}{491} = 2 \\
\hat{h}_{17} &= \frac{238 * 19}{491} = 9, & \hat{h}_{27} &= \frac{253 * 19}{491} = 10
\end{aligned}$$

Chi-Quadrat (χ^2)

Skalenniveau: nominal

Definition: Der Wert von χ^2 ergibt sich dadurch, daß die (quadrierten) Differenzen zwischen den erwarteten und den empirischen Häufigkeiten addiert werden, d.h. Zusammenhang wird hier definiert als der Unterschied zwischen der tatsächlich festgestellten Häufigkeitsverteilung und einer angenommenen unabhängigen Verteilung.

Berechnung:

$$\chi^2 = \sum_{j=1}^{c} \sum_{i=1}^{r} \frac{(h_{ij} - \hat{h}_{ij})^2}{\hat{h}_{ij}}$$

Bemerkung: Einige ungünstige Eigenschaften von χ^2 haben zur Entwicklung von Varianten geführt:

1. χ^2 ist proportional abhängig von der Anzahl der Untersuchungseinheiten, d.h. χ^2 wächst oder verringert sich mit der Größe der Stichprobe, auch wenn der Grad der Kontingenz gleich bleibt.
   Beispiel:

24	16	40
16	24	40
40	40	80

$\Longrightarrow \chi^2 = 3.2$

48	32	80
32	48	80
80	80	160

$\Longrightarrow \chi^2 = 6.4$

$\Longrightarrow$ *Phi-Koeffizient* Φ^2:

$$\Phi^2 = \frac{\chi^2}{N} \qquad \text{bzw.} \qquad \Phi = \sqrt{\frac{\chi^2}{N}}$$

2. Während das Minimum von χ^2 gleich 0 ist, liegt das Maximum bei N, in Abhängigkeit von der Anzahl der Zellen, d.h. von der Anzahl der Merkmalsausprägungen der beiden Variablen. Hierdurch wird ein Vergleich zwischen Kreuztabellen unterschiedlicher Größe erschwert. Diese Eigenschaft gilt ebenso für die Maßzahl Φ^2, die lediglich für 2×2-Tabellen eine Obergrenze von 1 erreichen kann:

   Beispiel:

20	0	20
0	20	20
20	20	40

bzw.

10	10	20
10	10	20
20	20	40

   Der Zusammenhang in obigem Beispiel ist maximal, da zwei Diagonalzellen unbesetzt sind.

   Zusammen mit der zugehörigen Indifferenztabelle ergibt sich für χ^2 und Φ^2:

$$\begin{aligned} \chi^2 &= \frac{(10)^2 + (10)^2 + (-10)^2 + (-10)^2}{10} = 40 \\ \Phi^2 &= \frac{40}{40} = 1 \end{aligned}$$

   $\Longrightarrow$ *Cramers* V^2:

$$V^2 = \frac{\chi^2}{N \cdot \min(r-1, c-1)}$$

   Gegenüber Φ^2 zeichnet sich V^2 dadurch aus, daß auch für beliebige Kreuztabellen das Maximum gleich 1 ist.

Eine weitere, ältere Maßzahl auf der Basis von χ^2 ist der von *Pearson* entwickelte Kontingenzkoeffizient C mit

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}}$$

Der Hauptnachteil dieses Koeffizienten liegt darin, daß er ein unter 1 liegendes, von der Zeilen- und Spaltenanzahl abhängiges Maximum hat, dessen konkreter Wert sich jeweils nur für quadratische Kreuztabellen ermitteln läßt (vgl. Anhang A.4):

$$\begin{aligned} & C_{max} = \sqrt{\frac{r-1}{r}} \\ \Longrightarrow \quad & 2 \times 2 : \quad C_{max} = 0.707 \\ & 3 \times 3 : \quad C_{max} = 0.816 \\ & 4 \times 4 : \quad C_{max} = 0.866 \\ & 5 \times 5 : \quad C_{max} = 0.894 \end{aligned}$$

Für den Vergleich auch unterschiedlich großer, quadratischer Kreuztabellen ist daher eine Korrektur von C notwendig:

$$C_{korr} = \frac{C}{C_{max}}$$

5.2.2 Der Signifikanztest von χ^2

Analog zu der in Kapitel 4.2 beschriebenen Vorgehensweise bei einem Hypothesentest soll die Signifikanz des χ^2-Wertes aus einer Stichprobe für die Grundgesamtheit überprüft werden.

Frage: Ist der durch den χ^2-Wert beschriebene Zusammenhang zwischen zwei Variablen auch für die Grundgesamtheit signifikant, oder liegt das Ergebnis im Bereich der Zufallsschwankungen?

1. Schritt: Hypothesenformulierung

$H_0 : \chi^2 = 0$

$H_A : \chi^2 \neq 0$

2. Schritt: Bestimmung der Prüfverteilung

Bei den Erläuterungen zur Normalverteilung und zum zentralen Grenzwertsatz (Kapitel 3.4) wurde festgestellt, daß die Mittelwerte von Stichproben aus einer Grundgesamtheit sich gemäß einer Standardnormalverteilung um den Mittelwert verteilen. Die Existenz einer bekannten Verteilung für eine Zufallsvariable erlaubt dann den Hypothesentest auf der Basis lediglich einer Realisierung der Zufallsvariablen. Analog dazu läßt sich χ^2 als eine Zufallsvariable auffassen, die ebenfalls einer charakteristischen, gleichnamigen Verteilung entspricht (vgl. Anhang A.3).

3. Schritt: Bestimmung des Freiheitsgrads df

Wie aus der χ^2-Verteilung ersichtlich, hängt der Kurvenverlauf jeweils von der Anzahl der Freiheitsgrade (df=degrees of freedom) ab. Grundsätzlich ermittelt sich die Anzahl der Freiheitsgrade für χ^2 durch die Zellenzahl in der Kreuztabelle ($r \times c$). Da für die Berechnung der Häufigkeiten in der Indifferenztabelle ($\hat{h}_{ij}$) alle Randhäufigkeiten mit eingehen, verringern sich die Freiheitsgrade entsprechend:

$$df = (c-1)\,(r-1)$$

4. Schritt: Bestimmung des Signifikanzniveaus α

5. Schritt: Ermittlung des theoretischen Verteilungswertes k und Berechnung des empirischen Verteilungswertes χ^2_{emp}

Die χ^2-Verteilung liegt tabelliert vor, so daß in Abhängigkeit vom Freiheitsgrad und dem gewählten Signifikanzniveau der "kritische" Wert k ermittelt werden kann. Der empirische Verteilungswert berechnet sich nach

$$\chi^2_{emp} = \sum_{j=1}^{c} \sum_{i=1}^{r} \frac{(h_{ij} - \hat{h}_{ij})^2}{\hat{h}_{ij}}$$

6. Schritt: Entscheidung über die Ablehnung von H_0

Im Gegensatz zum Mittelwerttest mit Hilfe der Normalverteilung werden Tests auf Signifikanz von χ^2 nur einseitig durchgeführt, da es keine negativen Werte für χ^2 gibt. Somit gilt folgende Beziehung:

$$P(\chi^2 \leq k) = 1 - \alpha$$

$\chi^2_{emp} \leq k$: Der empirische χ^2-Wert befindet sich in dem gewählten Vertrauensbereich $(1 - \alpha)$, d.h. die Informationen reichen nicht aus, um die Nullhypothese zu verwerfen.

$\chi^2_{emp} > k$: Der empirische χ^2-Wert befindet sich im Bereich der gewählten Irrtumswahrscheinlichkeit, d.h. es besteht nur noch eine Wahrscheinlichkeit von α, daß χ^2_{emp} größer als der Prüfwert ist und somit H_0 fälschlicherweise verworfen wird.

Bemerkungen:

- Als wesentliche Voraussetzung für die Anwendung des χ^2-Tests wird eine Häufigkeit von mindestens 5 Untersuchungseinheiten in den Zellen der Indifferenztabelle genannt. Der Grund hierfür liegt darin, daß aufgrund der Berechnungsformel für χ^2 bei geringen $\hat{h}_{ij}$ der jeweilige Summand groß und somit der Wert für χ^2 künstlich angehoben wird. Der Grenzwert 5 ist nicht theoretisch begründet und ebenso ein (umstrittener) Erfahrungswert wie die Maßgabe, nach der ein χ^2-Test bei mehr als 20% aller Zellen mit einer Häufigkeit kleiner als 5 nicht mehr verwendet werden sollte. Die Anzahl der Zellen, die diese Zahl unterschreiten, wird von SPSS gesondert ausgegeben (vgl. Kapitel 5.2.4). Als Maßnahmen für diesen Fall sind das (inhaltlich begründbare) Zusammenlegen von benachbarten Zellen (SPSS: `RECODE`) oder die Anwendung alternativer Tests (z.B. Fisher-Test) möglich.

- Bei der Verwendung statistischer Analysesysteme (z.B. SPSS/PC+) entfällt die Notwendigkeit, den Schwellenwert k aus Tabellen zu ermitteln; es wird eine Irrtumswahrscheinlichkeit berechnet, die direkt mit dem vom Benutzer festgelegten Wert α verglichen werden kann. So ergibt sich für die Kreuztabelle zwischen der Zweitstimme und dem Geschlecht bei einem χ^2-Wert von 3.42377 eine Signifikanz von 0.75408 (`SIGNIFICANCE` = 0.75408), d.h. die berechnete Irrtumswahrscheinlichkeit liegt bei über 75%.

5.2.3 Das Modell der proportionalen Fehlerreduktion (PRE)

Die Grundidee dieses Modells, für das es Zusammenhangsmaße auf allen Skalierungsebenen gibt, besteht darin, den Zusammenhang zwischen zwei Variablen durch das Ausmaß zu definieren, mit dem die Vorhersage einer Variablen y aufgrund der Kenntnis einer anderen Variablen x verbessert wird. Die berechnete Maßzahl drückt dann aus, um wieviel sich der Fehler bei der Vorhersage einer (abhängigen) Variablen y reduziert, wenn die Informationen über eine (unabhängige) Variable x berücksichtigt werden.

Die im folgenden am Beispiel des Zusammenhangsmaßes Lambda (λ) für zwei nominalskalierte Variablen demonstrierte, typische Vorgehensweise bei der Berechnung eines PRE(=proportional reduction of error)-Maßes ist ebenso auf Variablen höheren Skalenniveaus übertragbar.

Lambda (λ)

Skalenniveau: nominal

Berechnung: Die Berechnung von λ läßt sich in drei Schritte aufteilen:

1. Vorhersage von y *ohne* Berücksichtigung der Informationen über x und Ermittlung des Vorhersagefehlers E_1:

   Für eine möglichst gute Vorhersage, für welche Merkmalsklasse sich eine Untersuchungseinheit entschieden hat, ist der Modus h der beste Schätzer, da diese Klasse die größte Häufigkeit enthält. Der daraus resultierende Schätzfehler ergibt sich dann zu

   $$E_1 = N - \max(h_{i\bullet})$$

   mit:

N	:	Anzahl der Untersuchungseinheiten
$\max(h_{i\bullet})$	:	Anzahl der Untersuchungseinheiten in der modalen Kategorie der Zeilenvariablen y

   d.h. in $N - \max(h_{i\bullet})$ Fällen erweist sich die Vorhersage als Irrtum, wenn die Modalkategorie gewählt wurde. Bezogen auf eine konkrete Kreuztabelle ist $\max(h_{i\bullet})$ die größte Randhäufigkeit aller Merkmalsklassen von y.

2. Vorhersage von y *mit* Berücksichtigung der Informationen über x und Ermittlung des Vorhersagefehlers E_2:

Die zusätzlichen Informationen bestehen darin, daß neben der Häufigkeitsverteilung auf die Merkmalsklassen von y auch die Verteilung bezüglich x bekannt ist. Der beste Schätzer ist daher nicht der allgemeine Modus von y, sondern der Modus, der sich jeweils für y bezogen auf *jede* Merkmalsklasse von x festlegen läßt. Der daraus resultierende Schätzfehler ergibt sich dann zu

$$E_2 = \sum_{j=1}^{c} (h_{\bullet j} - \max(h_{ij}))$$

mit:

$h_{\bullet j}$: Anzahl der Untersuchungseinheiten der j-ten Spalte

$\max(h_{ij})$: maximale Häufigkeit in der j-ten Spalte

3. Berechnung einer Maßzahl λ_r, um wieviel bei Kenntnis von x die Vorhersage verbessert (d.h. der Vorhersagefehler reduziert) wird. Der Index r bedeutet dabei, daß in der Kreuztabelle die Zeilenvariable als abhängig und die Spaltenvariable als unabhängig betrachtet wird:

$$\begin{aligned}
\lambda_r &= \frac{E_1 - E_2}{E_1} \\
&= \frac{N - \max(h_{i\bullet}) - \sum_{j=1}^{c} (h_{\bullet j} - \max(h_{ij}))}{N - \max(h_{i\bullet})} \\
&= \frac{N - \max(h_{i\bullet}) - (N - \sum_{j=1}^{c} \max(h_{ij}))}{N - \max(h_{i\bullet})} \\
&= \frac{\sum_{j=1}^{c} \max(h_{ij}) - \max(h_{i\bullet})}{N - \max(h_{i\bullet})}
\end{aligned}$$

Analog läßt sich für den umgekehrten Fall, daß die Zeilenvariable unabhängig und die Spaltenvariable abhängig ist, λ_c berechnen:

$$\lambda_c = \frac{\sum_{i=1}^{r} \max(h_{ij}) - \max(h_{\bullet j})}{N - \max(h_{\bullet j})}$$

Für ein symmetrisches Zusammenhangsmaß λ_s existiert folgende kombinierte Berechnungsformel:

$$\begin{aligned}\lambda_s &= \lambda \\ &= \frac{\sum_{j=1}^{c}\max(h_{ij}) - \sum_{i=1}^{r}\max(h_{ij})}{2N - \max(h_{i\bullet}) - \max(h_{\bullet j})} \\ &\quad - \frac{\max(h_{i\bullet}) + \max(h_{\bullet j})}{2N - \max(h_{i\bullet}) - \max(h_{\bullet j})}\end{aligned}$$

Bemerkung: Eigenschaften von $\lambda, \lambda_r, \lambda_c$:

- λ ist ein symmetrisches Maß, während λ_r und λ_c asymmetrisch sind, d.h. abhängig von der Kausalrichtung.
- Der Wertebereich liegt zwischen 0 und 1 und läßt sich als *prozentuale Reduzierung* des Vorhersagefehlers von y durch die Kenntnis von x interpretieren (z.B. λ = 0.8 $\Longrightarrow$ die Fehlerreduktion beträgt 80%).

  $\lambda_{r,c}$ = 0 : Die unabhängige Variable leistet keinen Beitrag zur Vorhersage der abhängigen, d.h. es gibt keinen Zusammenhang.

  $\lambda_{r,c}$ = 1 : Die abhängige Variable kann zu 100% durch die unabhängige vorausgesagt werden, d.h. es besteht ein maximaler Zusammenhang.

Beispiel: Ausgehend von der Hypothese, daß das Geschlecht eines Befragten dessen Wahlabsicht für die Zweitstimme bei der Bundestagswahl beeinflußt, läßt sich mit Hilfe der Kreuztabelle auf Seite 66 ein Wert für λ berechnen, der eine Hypothesenüberprüfung aufgrund der vorliegenden Daten erlaubt. Dabei wird das Geschlecht als unabhängige und die Wahlabsicht als abhängige Variable be-

trachtet.

$$\begin{aligned} E_1 &= N - \max(h_{i\bullet}) \\ &= 491 - 215 \\ &= 276 \end{aligned}$$

$$\begin{aligned} E_2 &= \sum_{j=1}^{c}(h_{\bullet j} - \max(h_{ij})) \\ &= (238 - 99) + (253 - 116) \\ &= 276 \end{aligned}$$

$$\begin{aligned} \lambda &= \frac{E_1 - E_2}{E_1} \\ &= \frac{276 - 276}{276} \\ &= 0 \end{aligned}$$

Die Hinzunahme der Verteilung der Variable Geschlecht hat die Vorhersage der Wahlabsicht nicht verbessert, da die Berücksichtigung der geschlechtsspezifischen Häufigkeitsmaxima zum gleichen Vorhersagefehler führt wie die Verwendung des Maximums für die Wahlabsicht allein (vgl. auch Kapitel 5.2.4). Somit wird die Ausgangshypothese aufgrund der vorliegenden Daten nicht gestützt.

5.2.4 Zusammenfassung

Durch den SPSS-Befehl `CROSSTABS /TABLES V334 BY V202 /STATISTICS CHISQ PHI CC LAMBDA` wird neben der Kreuztabelle, die auf Seite 66 angegeben ist, eine Auswahl von statistischen Maßzahlen berechnet und ausgegeben:

```
      Chi-Square             Value           DF               Significance
--------------------     -----------        ----             ------------

Pearson                     3.42377           6                  .75408
Likelihood Ratio            3.81210           6                  .70209
Mantel-Haenszel test for     .30106           1                  .58322
      linear association

Minimum Expected Frequency -     .485
Cells with Expected Frequency < 5 -     4 OF    14 ( 28.6%)

                                                                 Approximate
     Statistic               Value        ASE1      T-value    Significance
--------------------       ---------    --------    -------    ------------

Phi                          .08350                               .75408 *1
Cramer's V                   .08350                               .75408 *1
Contingency Coefficient      .08322                               .75408 *1

Lambda :
   symmetric                 .01946       .01631     1.18018
   with V334     dependent   .04202       .03490     1.18018
   with V202     dependent   .00000       .00000
Goodman & Kruskal Tau :
   with V334     dependent   .00697       .00621                .75500 *2
   with V202     dependent   .00114       .00167                .76530 *2

*1 Pearson chi-square probability
*2 Based on chi-square approximation

Number of Missing Observations:  9
```

In dem Beispiel liegt für einen χ^2-Wert von 3.42377 die Signifikanz bei 75%, d.h. beim Verwerfen der Nullhypothese, daß es keinen Zusammenhang zwischen V202 und V334 gibt, akzeptiert der Anwender eine Irrtumswahrscheinlichkeit von 75%. Die auf χ^2-basierten Kennzahlen Phi, Cramer's V sowie der Kontingenzkoeffizient weisen für die Stichprobe alle einen sehr geringen Wert aus. Da auch λ als

PRE-Maß lediglich einen Wert von 0.01946 (=1.9%) berechnet, kann davon ausgegangen werden, daß die vorliegenden Daten keinen Zusammenhang zwischen der Zweitstimmenabgabe und dem Geschlecht belegen.

Die Angabe, daß 4 von 14 Zellen der Häufigkeitstabelle auf Seite 66 eine erwartetete Häufigkeit kleiner 5 haben, weist daraufhin, daß der Wert für χ^2 nicht korrekt berechnet wurde. Es würde sich daher in diesem Fall anbieten, die Kategorien 5 (NPD), 6 (Sonst.) und 7 (keine Angaben) bei der Zweitstimmenabgabe — ohne wesentlichen Informationsverlust — zu einer Kategorie zusammenzufassen und die auf dieser Basis berechneten Kennzahlen mit den vorliegenden zu vergleichen.

5.3 Zusammenhangsmaße zwischen zwei ordinalskalierten Variablen

5.3.1 Das Prinzip der Paarvergleiche

Im Unterschied zu nominalskalierten Variablen besteht zwischen den Kategorien einer ordinalskalierten Variablen eine Rangordnung. Es können nun Vergleiche darüber angestellt werden, wie sich jeweils zwei Untersuchungseinheiten bezüglich der Merkmalswerte zweier Variablen unterscheiden.

Beispiel:

Gegeben: Sei x die Meinung zu den Gewerkschaften allgemein (sehr schlecht, ..., sehr gut) und y der Schulabschluß des Befragten (Hauptschule ohne Lehre, ..., Hochschule mit Abschluß).

Frage: Wenn eine Untersuchungseinheit A einen höheren (niedrigeren) Schulabschluß hat als eine Untersuchungseinheit B, besitzt sie dann auch eine höhere (niedrigere) Meinung von den Gewerkschaften, d.h. gilt

$$(x_A > x_B \wedge y_A > y_B) \vee (x_A < x_B \wedge y_A < y_B)$$

Die Größe des Anteils von diesen (hier: *konkordanten*) Paaren von Untersuchungseinheiten an allen möglichen Paaren wird dabei für die Beschreibung eines Zusammenhangs zwischen zwei Variablen verwendet. Neben den obengenannten konkordanten Paaren gibt es eine Reihe weiterer Klassen von Paaren, die sich aus den Merkmalswerten zweier Variablen ermitteln lassen. Die bivariaten Zusammenhangsmaße für ordinalskalierte Variablen beruhen letztendlich darauf, die Häufigkeiten für die unterschiedlichen Paartypen untereinander in Beziehung zu setzen.

Entsprechend den möglichen Paarvergleichen sind insgesamt fünf Klassen von Paaren zu unterscheiden:

Seien $(x_i, y_i), (x_j, y_j)$ die Merkmalswerte zweier Untersuchungseinheiten i und j $(i \neq j)$ bezüglich der zwei Variablen x und y. Dann gilt:

$$C = \{(x_i, y_i), (x_j, y_j) \mid (x_i > x_j \wedge y_i > y_j) \vee (x_i < x_j \wedge y_i < y_j)\}$$

ist die Menge aller *konkordanten Paare*,

$$D = \{(x_i, y_i), (x_j, y_j) \mid (x_i > x_j \wedge y_i < y_j) \vee (x_i < x_j \wedge y_i > y_j)\}$$

ist die Menge aller *diskordanten Paare*,

$$T_x = \{(x_i, y_i), (x_j, y_j) \mid (x_i = x_j \wedge y_i < y_j) \vee (x_i = x_j \wedge y_i > y_j)\}$$

ist die Menge aller *mit gleichen x verbundenen ("tied") Paare*,

$$T_y = \{(x_i, y_i), (x_j, y_j) \mid (x_i > x_j \wedge y_i = y_j) \vee (x_i < x_j \wedge y_i = y_j)\}$$

ist die Menge aller *mit gleichen y verbundenen Paare* und

$$T_{xy} = \{(x_i, y_i), (x_j, y_j) \mid (x_i = x_j \wedge y_i = y_j)\}$$

ist die Menge aller *gleichen Paare*.

Insgesamt lassen sich aus N Untersuchungseinheiten $\begin{pmatrix} N \\ 2 \end{pmatrix}$ Paare bilden und es gilt:

$$\begin{pmatrix} N \\ 2 \end{pmatrix} = \frac{N(N-1)}{2} = N_C + N_D + N_x + N_y + N_{xy}$$

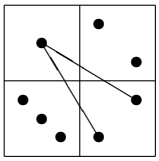
mit:

N_C	:	Anzahl der konkordanten Paare
N_D	:	Anzahl der diskordanten Paare
N_x	:	Anzahl aller mit gleichen x verbundenen Paare
N_y	:	Anzahl aller mit gleichen y verbundenen Paare
N_{xy}	:	Anzahl aller gleichen Paare

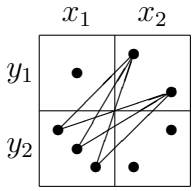
Voraussetzung für die Berechnung von Zusammenhangsmaßen ist die Ermittlung der Häufigkeiten für alle beschriebenen Paartypen. Wie Abbildung 5.1 (nach [Benninghaus 1985, S. 148]) zeigt, lassen sich diese Paarhäufigkeiten direkt aus einer gegebenen Kreuztabelle ermitteln, vorausgesetzt die Kategorien in der Kreuztabelle stimmen mit der ordinalen Reihenfolge und der Richtung für beide Variablen überein.

1	2	3
3	2	5
4	4	8

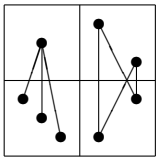
$$\begin{pmatrix} N \\ 2 \end{pmatrix} = \frac{N(N-1)}{2} = \frac{8(8-1)}{2}$$
$$= 2 + 6 + 7 + 8 + 5 = 28$$



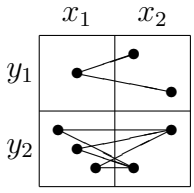
$N_C = (1)(2) = 2$



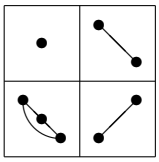
$N_D = (2)(3) = 6$



$N_x = (1)(3) + (2)(2) = 7$



$N_y = (1)(2) + (3)(2) = 8$



Die Anzahl der Paare in (x_1, y_2) ist $\frac{n(n-1)}{2} = \frac{3(3-1)}{2} = 3$, in (x_1, y_1) $\frac{1(1-1)}{2} = 0$ und in (x_2, y_2) und (x_2, y_1) je $\frac{2(2-1)}{2} = 1$. Folglich beträgt $N_{xy} = 0 + 1 + 3 + 1 = 5$.

Abb. 5.1: Ermittlung von Paarhäufigkeiten aus einer Kreuztabelle

5.3.2 Die Ermittlung von Zusammenhangsmaßen auf der Basis von Paarvergleichen

Die Ermittlung der Häufigkeiten für alle Klassen von Paaren (N_D, N_C, N_x, N_y, N_{xy}) auf der Basis einer gegebenen Kreuztabelle bildet die Grundlage der bivariaten Zusammenhangsmaße für ordinalskalierte Variablen. Dabei unterscheiden sich die einzelnen Kennzahlen dadurch, auf welche Art und Weise die Paarhäufigkeiten miteinander in Beziehung gesetzt sind.

Eine gemeinsame Eigenschaft der hier vorgestellten Zusammenhangsmaße ist ein Wertebereich zwischen -1 und $+1$, der sie insbesondere von den Zusammenhangsmaßen für nominalskalierte Variablen unterscheidet. Die "ordinalen" Zusammenhangsmaße sollten daher — wann immer möglich — den "nominalen" vorgezogen werden, da sie einen größeren Interpretationsspielraum besitzen.

Kendalls τ_a, τ_b, τ_c

Skalenniveau: ordinal

Definition: Diesen Zusammenhangsmaßen liegt die Differenz zwischen der Anzahl der konkordanten und der diskordanten Paare, jeweils bezogen auf eine unterschiedliche "Gesamtzahl" von Paaren, zugrunde.

Berechnung:

$$\tau_a = \frac{N_C - N_D}{N_C + N_D + N_x + N_y + N_{xy}}$$

$$\tau_b = \frac{N_C - N_D}{\sqrt{(N_C + N_D + N_x)(N_C + N_D + N_y)}}$$

$$\tau_c = \frac{N_C - N_D}{\frac{1}{2}N^2\frac{m-1}{m}} = \frac{2(N_C - N_D)}{N^2\frac{m-1}{m}}$$

mit: $m = \min(r, c)$

Bemerkung: Die Art des Zusammenhanges wird hier allein durch den Zähler festgelegt:

- $N_C > N_D$: Ein *positiver* Zusammenhang liegt vor, da die Anzahl der konkordanten Paare die der diskordanten Paare überwiegt.
  $\Longrightarrow \tau_{a,b,c} > 0$
- $N_C < N_D$: Es liegt ein *negativer* Zusammenhang vor, da die diskordanten Paare überwiegen.
  $\Longrightarrow \tau_{a,b,c} < 0$
- $N_C = N_D$: Es besteht *kein* Zusammenhang, da beide Paarhäufigkeiten gleich sind.
  $\Longrightarrow \tau_{a,b,c} = 0$

Die Grundlage für die Berechnung von τ_a ist die Annahme, daß gleiche Paare (T_x, T_y, T_{xy}) keinen Einfluß auf den Zusammenhang haben. Ein einfaches Beispiel zeigt aber, daß sich die Existenz von gleichen Paaren durchaus auf den Wert von τ_a auswirkt:

1			
	1		
		1	
			3

$$\Longrightarrow \tau_a = \frac{(1+1+1)-0}{3} = 1$$

10			
	10		
		10	
			30

$$\Longrightarrow \tau_a = \frac{(10*10+10*10+10*10)-0}{300+0+0+0+135} = \frac{300}{435} = 0.69$$

Offensichtlich ist der Zusammenhang in beiden Kreuztabellen maximal, er erreicht aber nur im ersten Fall, der keine gleichen Paare enthält, seinen Maximalwert 1. Da ein niedriger Wert von τ_a somit sowohl durch eine hohe Anzahl von verknüpften Paaren als auch durch eine niedrige Differenz zwischen konkordanten und diskordanten Paaren verursacht werden kann, wird τ_a nur selten verwendet [1].

Für das zweite Beispiel ergibt sich folgender Wert für τ_b:

$$\tau_b = \frac{300-0}{\sqrt{(300+0+0)(300+0+0)}} = \frac{300}{\sqrt{300*300}} = 1$$

τ_b kann also auch bei $N_{xy} \neq 0$ sein Maximum ereichen. Dies gilt aber nicht für beliebige, sondern lediglich für quadratische Kreuztabellen mit symmetrischen Randhäufigkeiten, wie folgendes Beispiel zeigt:

10				10
	10			10
		10	10	20
10	10	10	10	40

$$\Longrightarrow \tau_b = \frac{500-0}{\sqrt{(500+0+0)(500+0+100)}} = \frac{500}{\sqrt{500*600}}$$

[1] τ_a wird beispielsweise von SPSS/PC+ erst gar nicht angeboten.

$$= \quad 0.91$$

τ_c bezieht die minimale Spalten-/Zeilenzahl m in die Berechnungsformel mit ein und nimmt somit eine Korrektur von τ_b vor. Andererseits ergibt sich hierdurch wiederum eine Abhängigkeit von m, so daß die theoretischen Maxima ($+1$, -1) auch bei τ_c nicht immer erreicht werden können.

Von diesen drei *symmetrischen* Maßzahlen hat τ_b die breiteste Verwendung gefunden.

Gamma γ

Skalenniveau: ordinal

Definition: Der Assoziationskoeffizient γ ist definiert als das Verhältnis des Überschusses bzw. des Defizites konkordanter Paare im Vergleich zu diskordanten Paaren zur Gesamtheit aller konkordanten und diskordanten Paare, d.h. Paare mit gleichen x_i und/oder y_i werden als irrelevant angesehen.

Berechnung:

$$\gamma = \frac{N_C - N_D}{N_C + N_D}$$

Bemerkung: Ebenso wie bei den τ-Zusammenhangsmaßen wird die Richtung des Zusammenhanges hier durch den Zähler festgelegt, und es besteht die Ordnung

$$|\tau_a| \leq |\tau_b| \leq |\gamma|,$$

da für den Nenner der Berechnungsvorschrift von τ_b gilt:

$$\sqrt{(N_C + N_D + N_x)(N_C + N_D + N_y)} \begin{cases} = & N_C + N_D, \\ & \textit{wenn } N_x = 0 \\ & \textit{und } N_y = 0 \\ > & N_C + N_D, \\ & \textit{sonst} \end{cases}$$

γ liefert somit tendenziell immer höhere Werte als zum Beispiel τ_b.

γ ist aber auch abhängig von der Kategorienzahl, d.h. sein Wert kann durch das Zusammenlegen von Kategorien künstlich angehoben werden, da sich in diesem Fall die Zahl der unberücksichtigten gleichen Paare erhöht.

Beispiel ([Benninghaus 1985, S. 163–166]):
In einer Untersuchung an einer amerikanischen Hochschule wurde sowohl die Einstellung (14 Skalenitems) als auch das Verhalten (11 Skalenitems) gegenüber Farbigen erfragt. Bei der Berechnung von γ zwischen diesen beiden Variablen konnte der Wert von γ = 0.322 für die Originalvariablen bis auf γ = 0.929 für die dichotomisierten Variablen gesteigert werden.

Hieraus ergeben sich folgende Konsequenzen für die Auswahl eines geeigneten Assoziationskoeffizienten:

- Verwendung von γ bei der Berechnung auf der Basis der Originalvariablen.
- Alternative Verwendung von τ_b, falls Kategorien zusammengefaßt werden (z.B. aufgrund zu geringer Zellenhäufigkeiten).

Interpretation von γ als PRE-Maß:

1. Vorhersage für die jeweils erste Untersuchungseinheit eines jeden nicht verbundenen Paares: Die erste Untersuchungseinheit ist im Hinblick auf die abhängige Variable größer als die zweite:
$$E_1 = \frac{N_C + N_D}{2}$$
2. $N_C > N_D$:
Vorhersage: Die Untersuchungseinheiten haben bezüglich der abhängigen Variablen *dieselbe* Rangordnung wie bezüglich der unabhängigen.
$N_C < N_D$:
Vorhersage: Die Untersuchungseinheiten haben bezüglich der abhängigen Variablen die *umgekehrte* Rangordnung wie bezüglich der unabhängigen.
$$E_2 = \min(N_C, N_D)$$

3. Berechnung einer Maßzahl γ, um wieviel bei Kenntnis von x die Vorhersage verbessert, d.h. der Vorhersagefehler reduziert wird:

$$\begin{aligned}
\gamma &= \frac{E_1 - E_2}{E_1} \\
&= \frac{\frac{1}{2}(N_C + N_D) - \min(N_C, N_D)}{\frac{1}{2}(N_C + N_D)} \\
&= \frac{N_C + N_D - 2 \cdot \min(N_C, N_D)}{N_C + N_D} \\
N_C > N_D \Rightarrow \gamma &= \frac{N_C + N_D - 2N_D}{N_C + N_D} = \frac{N_C - N_D}{N_C + N_D} \\
N_C < N_D \Rightarrow \gamma &= \frac{N_C + N_D - 2N_C}{N_C + N_D} = \frac{N_D - N_C}{N_C + N_D}
\end{aligned}$$

4. Zusammengefaßt gilt: Bei Anwendung der Formel

$$\gamma = \frac{N_C - N_D}{N_C + N_D}$$

gibt das Vorzeichen von γ die Richtung der Assoziation an. $|\gamma|$ beschreibt die proportionale Fehlerreduktion bei der Vorhersage der Rangordnung der Paare.

Somers' d

Skalenniveau: ordinal

Definition: Bei Somers' d handelt es sich um eine asymmetrische Variante von γ, bei der für die jeweils abhängige Variable die einseitig verbundenen Paare mitberücksichtigt werden. Zusätzlich existiert eine symmetrische Version von Somers' d.

Berechnung:

$$\begin{aligned}
\textit{y abhängig} \quad &\Longrightarrow \quad d_{yx} \;=\; \frac{N_C - N_D}{N_C + N_D + N_y} \\
\textit{x abhängig} \quad &\Longrightarrow \quad d_{xy} \;=\; \frac{N_C - N_D}{N_C + N_D + N_x} \\
\textit{symmetrisch} \quad &\Longrightarrow \quad d_s \;=\; \frac{N_C - N_D}{N_C + N_D + \frac{1}{2}(N_x + N_y)}
\end{aligned}$$

Bemerkung: Im Vergleich mit γ gilt

$$|d_{yx}| \leq |\gamma|, \quad |d_{xy}| \leq |\gamma|, \quad |d_s| \leq |\gamma|,$$

da sich für den Nenner der Berechnungsvorschrift von Somers' d ableiten läßt (als Beispiel d_{yx}):

$$N_C + N_D + N_y \begin{cases} = N_C + N_D, & \textit{wenn } N_y = 0 \\ > N_C + N_D, & \textit{sonst} \end{cases}$$

5.3.3 Zusammenfassung

Mit dem Aufruf `CROSSTABS /TABLES V217 BY V225 /STATISTICS BTAU CTAU GAMMA D` wurde nachfolgende Kreuztabelle erzeugt und die in diesem Kapitel behandelten Kennwerte selektiert:

```
V217   SKALOMETER:CDU CSU-F D P  by  V225   SKALOMETER: H  KOHL
```

V217 Count \ V225	-5	-4	-3	-2	-1	0	1	2	3	4	5	Row Total
-5	24	2	1		2							29
												5.8
-4	4	7	2	1	2							16
												3.2
-3	11	6	6	7	1	2				1		34
												6.8
-2	7	5	7	4	9	2	2		2			38
												7.6
-1	2	2	6	3	6	6	3				1	29
												5.8
0	5	2	4	10	9	17	11	3	3	2	2	68
												13.6
1	2	1	7	6	7	7	21	8	2	1		62
												12.4
2	1	1	1	2	1	4	16	18	19	2	1	66
												13.2
3				1			4	16	28	18	7	74
												14.8
4							1	14	23	15		53
												10.6
5								1	4	25		30
												6.0
Column	56	26	34	34	37	38	57	46	69	51	51	499
Total	11.2	5.2	6.8	6.8	7.4	7.6	11.4	9.2	13.8	10.2	10.2	100.0

Statistic		Value	ASE1	T-value	Approximate Significance
Kendall's Tau-b		.73174	.01777	40.66409	
Kendall's Tau-c		.72278	.01777	40.66409	
Gamma		.79882	.01839	40.66409	
Somers' D :					
symmetric		.73173	.01777	40.66410	
with V217	dependent	.72802	.01811	40.66409	
with V225	dependent	.73549	.01753	40.66409	

Das Beispiel zeigt für alle berechneten Kennwerte (τ_a wird von SPSS nicht angeboten) einen annähernd gleich hohen Zusammenhang zwischen der Sympathie für Kohl und der Sympathie für die CDU/CSU/FDP-Regierung. Da das PRE-Maß γ auf der Basis der Originaldaten berechnet wird, eignet sich der Wert von 0.79882 (=80%) am besten zur Interpretation.

5.4 Zusammenhangsmaße zwischen zwei metrischen Variablen

5.4.1 Die Analyse bivariater Zusammenhänge mittels Streuungsdiagrammen

Während die Basis für die Berechnung von Zusammenhangsmaßen bisher

- die kategoriale Zugehörigkeit (nominal), bzw.
- die Rangordnung (ordinal)

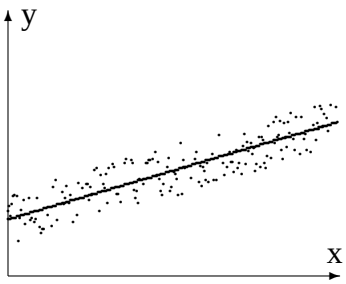
der jeweils untersuchten Variablen war, bildet bei metrischen Variablen die *Größe* der Merkmalswerte die Berechnungsgrundlage.

```
* SPSS-Aufruf: REGRESSION /VARIABLES V217 V225 /DEPENDENT V217
*                         /METHOD ENTER /SCATTERPLOT=(V217,V225).
Equation Number 1    Dependent Variable..   V217    SKALOMETER:CDU CSU-FDP
Variable(s) Entered on Step Number
   1..   V225        SKALOMETER: H  KOHL
Standardized Scatterplot
Across - V225       Down - V217
Out ++-----+-----+-----+-----+-----+-----++
  3 +                                     +   Symbols:
    |                                     |
    |                                     |      Max N
  2 +                                     +
    |                       . .*          |   .      7.0
    |                     . : **          |   :     14.0
  1 +                                     +   *     28.0
    |              .     . * * *.         |
    |        . . .. . . * * * ..         |
  0 +        . . .. . . * : . .          +
    |        . . .: : * : . . ..         |
    |        . . .. . . .      .         |
 -1 +        . . .. : . .   .            +
    |        : . .. . .       .          |
    |        . . .. .                    |
 -2 +        * . .  .                    +
    |                                     |
    |                                     |
 -3 +                                     +
Out ++-----+-----+-----+-----+-----+-----++
    -3    -2    -1     0     1     2     3 Out
```

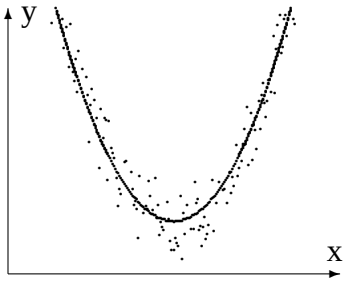
Einen ersten Überblick über die Verteilung der Merkmalswerte liefert zum einen wiederum die Kreuztabelle (vgl. Seite 66ff), wobei ratio-skalierte Variablen (z.B. Alter) sinnvollerweise zuvor gruppiert werden, zum anderen das Streuungsdiagramm, d.h. die Darstellung aller Merkmalswerte als Punkte in einem durch die Variablen aufgespannten Koordinatensystem.

Die Form der durch die Merkmalswerte dargestellten Punktwolke (vgl. Abbildung 5.2 nach [Patzelt 1985, S. 81]) liefert erste Anhaltspunkte über die Art des vermuteten Zusammenhanges:

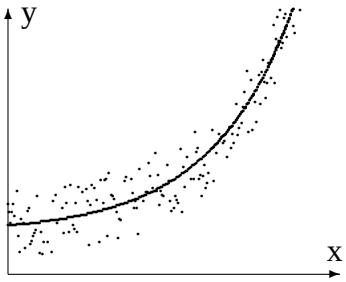
- linear
- kurvilinear: parabolisch, exponentiell, logarithmisch oder sinusförmig



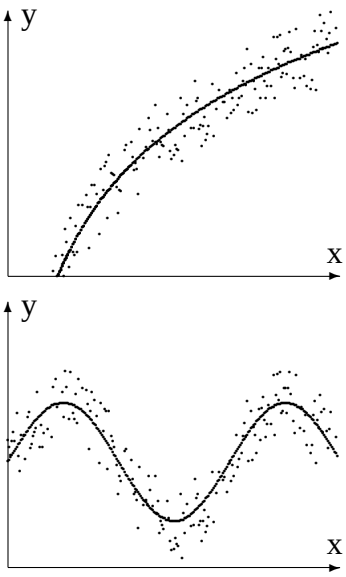
linear



kurvilinear: parabolisch



kurvilinear: exponentiell



kurvilinear: logarithmisch

kurvilinear: sinusförmig

Abb. 5.2: Graphischer Zusammenhang von Merkmalswerten

Im weiteren Verlauf erfolgt eine Beschränkung auf die Betrachtung linearer Zusammenhänge, deren Vorteil in der vergleichsweise einfachen Berechnung und Interpretation liegt. Für die Untersuchung nichtlinearer Zusammenhänge existieren ebenso entsprechende Verfahren bzw. es gibt in bestimmten Fällen die Möglichkeit einer Transformation auf den linearen Fall. Hierauf wird im Rahmen der multiplen Regression (Kapitel 6.2) näher eingegangen.

5.4.2 Das Modell der linearen, bivariaten Regression

Während bei der Betrachtung eines Streuungsdiagramms lediglich Vermutungen über die Art des Zusammenhangs zwischen einer abhängigen Variablen y und einer unabhängigen Variablen x angestellt werden konnten, erlaubt das lineare Regressionsmodell eine analytische Untersuchung.

Hierfür werden einige Bezeichner in ihrer Bedeutung festgelegt mit:

y	:	abhängige, endogene, zu erklärende Variable = Regressand y
x	:	unabhängige, exogene, erklärende Variable = Regressor x
y_i, x_i	:	i-ter Merkmalswert der abhängigen bzw. unabhängigen Variablen ($i = 1, \ldots, n$)
$\hat{y}_i$	:	Schätzwerte von y_i (y-Koordinatenwerte auf der festzulegenden Gerade)
e	:	Störvariable, Residuum
e_i	:	i-ter Störvariablenwert (= vertikaler Abstand zwischen dem empirischen Variablenwert y_i und dem durch die Regressionsgerade geschätzten Variablenwert $\hat{y}_i$)
b_0, b_1	:	Regressionskoeffizienten
$\bar{x}, \bar{y}$	:	Variablenmittelwerte

Zwei Aufgaben lassen sich damit für die Untersuchung formulieren:

1. Herleitung eines formalen Modells über den Zusammenhang zwischen x und y.
   Im bivariaten Fall und unter der Annahme eines linearen Zusammenhanges besteht das formale Modell aus der einfachen Geradengleichung:

$$y = b_0 + b_1 x$$

2. Optimale Anpassung dieses Modells an die empirischen Daten.
   Unter einer optimalen Anpassung wird hier das Ermitteln einer Geradengleichung verstanden, die die empirischen Daten möglichst gut beschreibt, d.h. in dem hier vorgestellten Verfahren werden letztlich die optimalen Geradenparameter b_0 und b_1 berechnet.

Die Anwendung obiger Geradengleichung ergibt für jeden Befragten ($i = 1, \ldots, n$):

$$y_i = b_0 + b_1 x_i$$

Dies bedeutet, daß der zwischen y und x angenommene lineare Zusammenhang für alle Untersuchungseinheiten gleichermaßen gilt wie die Regressionskoeffizienten b_0 und b_1.

Bemerkung: Die statistische Bestätigung eines linearen Zusammenhangs zwischen der unabhängigen und der abhängigen Variablen sagt noch nichts über die Gültigkeit einer kausalen Beziehung aus. Auch kausal unsinnige Variablenkombinationen können statistisch korrelieren (z.B. "Je mehr Störche in einem Gebiet nisten, umso höher ist die Zahl der Kindergeburten.").

Das mit der Geradengleichung formulierte deterministische Modell der bivariaten, linearen Regression ist in den Sozialwissenschaften in der Regel unangemessen.

Allgemein wird von der Existenz eines Fehlerterms e_i $(i = 1, \ldots, n)$ ausgegangen, in dem alle Schätzfehler der y_i-Werte aufgrund der x_i-Werte zusammengefaßt sind. Mögliche Schätzfehlerursachen sind:

- Einflüsse nicht berücksichtigter Erklärungsvariablen ($\Longrightarrow$ multiple Regression)
- Meßfehler für y_i
- fehlerhafte Spezifikation der funktionalen Beziehung zwischen y_i und x_i.

Insgesamt wird e als eine Zufallsvariable betrachtet, deren Variablenwerte e_i die y_i zusätzlich beeinflussen, deren Wirkung sich im Mittel aber aufhebt ($E(e) = 0$).

Die vollständige Regressionsgleichung lautet daher:

$$y_i = b_0 + b_1 x_i + e_i$$

In diesem Zusammenhang wird auch von einem *stochastischen Regressionsmodell* gesprochen.

Lediglich für die geschätzten Variablenwerte auf der Regressionsgerade gilt:

$$\hat{y}_i = b_0 + b_1 x_i$$

Dies bedeutet für den durch die Schätzung zu minimierenden Fehler innerhalb der Regressionsgleichung:

$$\begin{aligned} e_i &= y_i - \hat{y}_i \\ &= y_i - b_0 - b_1 x_i \qquad (i = 1, \ldots, n) \end{aligned}$$

Das Minimierungskriterium wird durch die *Summation aller quadrierten Fehlerwerte* e_i gewonnen ($\Longrightarrow$ Kleinstquadratschätzung (*O*rdinary *L*east *S*quare)):

$$\begin{aligned} S(e) &= \sum_{i=1}^{n} e_i^2 \\ &= \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 \stackrel{!}{=} \textit{Minimum} \end{aligned}$$

Die Bestimmungsgleichungen für die beiden Regressionskoeffizienten b_0 und b_1 ergeben sich durch Lösen der Extremwertaufgabe für $S(e)$:

- Partielle Ableitung von $S(e)$ nach b_0 und b_1
- Lösen der Normalengleichung

Partielle Ableitung nach b_0:

$$\frac{\delta S}{\delta b_0} = \sum_{i=1}^{n} 2(y_i - b_0 - b_1 x_i)(-1)$$

Normalengleichung für b_0:

$$\begin{aligned} 0 &= -\sum_{i=1}^{n} y_i + \sum_{i=1}^{n} b_0 + \sum_{i=1}^{n} b_1 x_i \\ 0 &= -\sum_{i=1}^{n} y_i + n b_0 + b_1 \sum_{i=1}^{n} x_i \\ 0 &= -\frac{1}{n}\sum_{i=1}^{n} y_i + b_0 + b_1 \frac{1}{n}\sum_{i=1}^{n} x_i \end{aligned}$$

Wegen

$$\frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y} \qquad \text{bzw.} \qquad \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$

folgt für b_0:

$$b_0 = \bar{y} - b_1 \bar{x}$$

Partielle Ableitung nach b_1:

$$\begin{aligned}\frac{\delta S}{\delta b_1} &= \sum_{i=1}^{n} 2(y_i - b_0 - b_1 x_i)(-x_i) \\ &= 2\sum_{i=1}^{n}(-y_i x_i + b_0 x_i + b_1 x_i^2)\end{aligned}$$

Normalengleichung für b_1:

$$0 = -\sum_{i=1}^{n} y_i x_i + b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2$$

Einsetzen von b_0 ergibt dann:

$$0 = -\sum_{i=1}^{n} y_i x_i - (\bar{y} - b_1 \bar{x}) \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2$$

Nach Division durch n und Ersetzen von $\frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$ gilt:

$$\begin{aligned}0 &= -\frac{1}{n}\sum_{i=1}^{n} y_i x_i + (\bar{y} - b_1 \bar{x})\bar{x} + b_1 \frac{1}{n}\sum_{i=1}^{n} x_i^2 \\ \frac{1}{n}\sum_{i=1}^{n} y_i x_i &= \bar{y}\,\bar{x} - b_1 \bar{x}^2 + b_1 \frac{1}{n}\sum_{i=1}^{n} x_i^2 \\ \frac{1}{n}\sum_{i=1}^{n} y_i x_i - \bar{y}\,\bar{x} &= b_1 \left(\frac{1}{n}\sum_{i=1}^{n}(x_i^2) - \bar{x}^2\right)\end{aligned}$$

Somit folgt für b_1:

$$b_1 = \frac{\frac{1}{n}\sum_{i=1}^{n} y_i x_i - \bar{y}\,\bar{x}}{\frac{1}{n}\sum_{i=1}^{n}(x_i^2) - \bar{x}^2}$$

Unter Berücksichtigung von[2]

$$\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{y}\,\bar{x} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \text{cov}(x, y)$$

[2]Mit cov(x, y) = *Kovarianz* : gemeinsame Varianz der Variablen x und y (vgl. Anhang A.5).

und

$$\frac{1}{n}\sum_{i=1}^{n}(x_i^2) - \bar{x}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \text{var}(x)$$

gilt für b_1:

$$b_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\text{cov}(x,y)}{\text{var}(x)}$$

Folgerung: Die Anwendung des Kleinstquadratschätzers setzt voraus, daß die Varianz der unabhängigen Variablen ungleich Null ist.

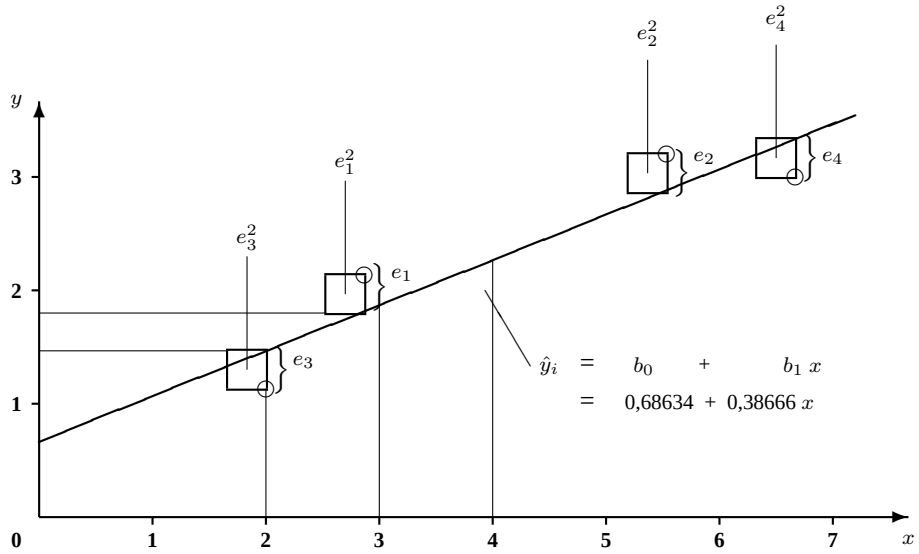


Abb. 5.3: Graphische Interpretation des bivariaten Regressionsmodells

Abbildung 5.3 (nach [Gruber 1981, S. 71]) veranschaulicht noch einmal das Prinzip der Kleinstquadratschätzung und die Interpretation der Koeffizienten, wobei:

b_0 = Schnittpunkt der Regressionsgeraden mit der y–Achse $\hat{=}$
Prognosewert für y_i, falls $x_i = 0$.

b_1 = Steigungswinkel der Regressionsgeraden $\hat{=}$
Grad des Zusammenhangs zwischen unabhängiger und abhängiger Variablen

Für standardisierte Variablen ergibt sich eine einfachere Berechnung für b_0 und b_1 mit:

$$x^* = \frac{x - \bar{x}}{s_x}, \qquad s_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Daraus folgt:

$$\bar{x}^* = 0 \qquad \text{und} \qquad \text{var}(x^*) = s_{x*}^2 = 1$$

Nach Standardisierung beider Variablen x ($\rightarrow x^*$) und y ($\rightarrow y^*$) gilt für b_0 und b_1:

$$\begin{aligned} b_0^* &= \bar{y}^* - b_1^{(*)}\bar{x}^* \\ &= 0 \end{aligned}$$

$$\begin{aligned} b_1^* &= \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i^* - \bar{x}^*)(y_i^* - \bar{y}^*)}{\frac{1}{n}\sum_{i=1}^{n}(x_i^* - \bar{x}^*)^2} \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i^* y_i^* \end{aligned}$$

Weiterhin gilt für b_1^*:

$$b_1^* = b_1 \frac{s_x}{s_y}$$

Herleitung:

$$
\begin{aligned}
b_1 &= \frac{\text{cov}(x,y)}{\text{var}(x)} \\
&= \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2} \\
b_1^* &= \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i^*-\bar{x}^*)(y_i^*-\bar{y}^*)}{\frac{1}{n}\sum_{i=1}^{n}(x_i^*-\bar{x}^*)^2} \\
&= \frac{\frac{1}{n}\sum_{i=1}^{n}\frac{(x_i-\bar{x})}{s_x}\cdot\frac{(y_i-\bar{y})}{s_y}}{\frac{1}{n}\sum\left(\frac{x_i-\bar{x}}{s_x}\right)^2} \\
&= \frac{\frac{1}{n}\cdot\frac{1}{s_x s_y}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\frac{1}{n}\cdot\frac{1}{s_x^2}\sum_{i=1}^{n}(x_i-\bar{x})^2} \\
&= \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}\cdot\frac{s_x}{s_y} \\
&= b_1\frac{s_x}{s_y}
\end{aligned}
$$

Regressionskoeffizient b_1

Skalenniveau: metrisch

Berechnung: x, y unstandardisiert:

$$b_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\text{cov}(x, y)}{\text{var}(x)}$$

x, y standardisiert:

$$b_1^* = \frac{1}{n}\sum_{i=1}^{n} x_i^* y_i^* = \text{cov}(x^*, y^*)$$

Bemerkung:

1. b_1 ist asymmetrisch, wobei y die abhängige und x die unabhängige Variable bezeichnet.
2. b_1^* liegt im Intervall $[-1, +1]$
3. Interpretation:

   "Nimmt unter sonst gleichen Bedingungen die Variable x um eine Einheit ab (zu), so nimmt die Variable y um b_1 Einheiten ab (zu)".

Beispiel: Nachfolgende SPSS-Ausgabe zeigt das Ergebnis einer bivariaten Regression, bei der die Sympathie für die CDU/CSU/FDP-Regierungskoalition in Abhängigkeit von der Sympathie für den Kanzler untersucht wurde. Für die Regressionskoeffizienten b_0 und b_1 ergeben sich dabei die Werte 2.076811 und 0.720843 (`B`). Im Fall standardisierter Variablen hat b_1 (`Beta`) den Wert 0.839476, d.h.: steigt die Sympathie für Kohl um eine Einheit, dann steigt die Zufriedenheit mit der Regierung um 0.839476.

```
* Ausschnitt aus: REGRESSION /VARIABLES V217 V225 /DEPENDENT V217
*                            /METHOD ENTER.

Equation Number 1    Dependent Variable..   V217    SKALOMETER:CDU CSU-FDP
Variable(s) Entered on Step Number
   1..    V225       SKALOMETER: H  KOHL

------------------ Variables in the Equation ------------------

Variable              B        SE B       Beta         T  Sig T

V225            .720843     .020930    .839476    34.441  .0000
(Constant)     2.076811     .150789               13.773  .0000
```

5.4.3 Das Bestimmtheitsmaß R^2 und der Korrelationskoeffizient r

Die Kleinstquadratschätzung (*OLS*) ist ein mathematisches Verfahren zur *optimalen* Anpassung eines Modells (im bivariaten Fall: einer Geraden) an empirische Daten.

Frage: *Wie gut* ist die Vorhersagekraft einer Regressionsfunktion bei gegebenen Werten von unabhängigen Variablen?

Wie in den vorhergegangenen Kapiteln gezeigt, eignen sich für die Beantwortung dieser Frage insbesondere Zusammenhangsmaße, die nach dem Prinzip der proportionalen Fehlerreduktion konzipiert sind (PRE). Für den Bereich zweier metrischer Variablen existiert hierzu das Bestimmtheitsmaß R^2.

Bestimmtheitsmaß R^2 (Coefficient of Determination, R Square)

Skalenniveau: metrisch

Berechnung: Das Bestimmtheitsmaß R^2 basiert ebenso wie λ, γ und η^2 auf dem Prinzip der proportionalen Fehlerreduktion (PRE):

1. Bester Vorhersagewert für y *ohne* Kenntnis von x ist der Gesamtmittelwert $\bar{y}$. Der daraus resultierende Schätzfehler ergibt sich dann zu:

$$E_1 = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

d.h. die Summe aller quadrierten Abweichungen der Merkmalswerte von y von dem Vorhersagewert $\bar{y}$ (Gesamtvariation).

2. Mit Kenntnis der unabhängigen Variablen x sind die besten Schätzer für y_i die geschätzten Werte auf der Regressionsgeraden $\hat{y}_i = b_0 + b_1 x_i$. Daraus ergibt sich der Vorhersagefehler
$$E_2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$
d.h. die Summe aller quadrierten Abweichungen der Merkmalswerte von y von dem jeweiligen Schätzwert.

3. Daraus ergibt sich für R^2, definiert als PRE-Maß:
$$\begin{aligned} R^2 &= \frac{E_1 - E_2}{E_1} \\ &= \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \\ &= 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \end{aligned}$$

Bemerkung:

1. R^2 ist ein symmetrisches, vorzeichenloses Maß im Intervall $[0, 1]$.

2. Mit Hilfe der Varianzzerlegung von y läßt sich R^2 interpretieren als Prozentsatz der Varianz von y, der durch die Regression, d.h. durch x erklärt wird (vgl. Anhang A.6).
Beispiel: $R^2 = 0.71$, d.h. 71% der Gesamtvarianz von y werden durch x erklärt.

3. R^2 berücksichtigt extreme Merkmalswerte stärker, d.h. es gibt hier eine Ungleichbehandlung von Merkmalswerten.

Für eine einfachere Berechnung des Bestimmtheitsmaßes ist folgende Umformung nützlich:

$$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i^2 - 2\hat{y}_i\bar{y} + \bar{y}^2)$$

$$
\begin{aligned}
&= \sum_{i=1}^{n}\hat{y}_i^2 - 2n\bar{y}\sum_{i=1}^{n}\hat{y}_i + n\bar{y}^2 \text{ wegen } \sum_{i=1}^{n}\hat{y}_i = n\bar{\hat{y}},\ \bar{\hat{y}} = \bar{y} \\
&= \sum_{i=1}^{n}\hat{y}_i^2 - 2n\bar{y}^2 + n\bar{y}^2 \\
&= \sum_{i=1}^{n}\hat{y}_i^2 - n\bar{y}^2
\end{aligned}
$$

Somit ergibt sich für R^2:

$$R^2 = \frac{\sum_{i=1}^{n}\hat{y}_i^2 - n\bar{y}^2}{\sum_{i=1}^{n}y_i^2 - n\bar{y}^2}$$

Eine weitere Berechnungsformel für R^2 ergibt sich aus der Varianzzerlegung von y in Anhang A.6.

Produkt-Moment-Korrelationskoeffizient r

Skalenniveau: metrisch

Definition: Der Produkt-Moment-Korrelationskoeffizient r wird durch die Kombination der beiden asymmetrischen Zusammenhangsmaße b_{1y} (y abhängig) bzw. b_{1x} (x abhängig) definiert.

Berechnung:

$$
\begin{aligned}
r &= \sqrt{b_{1y}b_{1x}} \\
&= \sqrt{\frac{\text{cov}(x,y)}{\text{var}(x)} \cdot \frac{\text{cov}(x,y)}{\text{var}(y)}} \\
&= \frac{\text{cov}(x,y)}{\sqrt{\text{var}(x)}\sqrt{\text{var}(y)}}
\end{aligned}
$$

Falls x, y standardisiert sind, gilt

$$r = \text{cov}(x^*, y^*)$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i^* y_i^* = b_{1y} = b_{1x}$$

d.h. der Wert von r ist identisch mit dem der Regressionskoeffizienten b_{1y} und b_{1x}.

Bemerkung:

1. r ist ein symmetrisches Maß im Intervall $[-1, +1]$.

2. Die Namensgebung für das Bestimmtheitsmaß (R^2) und den Korrelationskoeffizienten (r) ist nicht zufällig, sondern r läßt sich durch Wurzelziehen aus dem Bestimmtheitsmaß ermitteln. Dagegen ist eine inhaltliche Übertragung nur schwer möglich, da sich "das Quadrat der Steigung" der Regressionsgerade kaum als "Prozentsatz der Varianzerklärung" interpretieren läßt.

5.4.4 Ein weiteres Gütemaß zur Beurteilung von Regressionsschätzungen

Wie auch das Bestimmtheitsmaß R^2 gibt der *Standardfehler der Regression* σ_e, d.h. die Höhe der durchschnittlichen Abweichung der vorhergesagten von den tatsächlichen Werten der abhängigen Variablen, genauere Auskunft über die Vorhersagekraft einer Regressionsfunktion bei vorliegenden Werten unabhängiger Variablen.

Als Ausgangspunkt dient die Summe der Abweichungsquadrate der geschätzten Werte von den tatsächlichen Werten (= Minimierungskriterium $S(e)$)

$$S(e) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Standardfehler der Regression σ_e

Skalenniveau: metrisch

Berechnung: Analog zur allgemeinen Herleitung der Standardabweichung ergibt sich für den Standardfehler der Regression σ_e (korrigiert nach Anzahl der Freiheitsgrade):

$$\sigma_e = \sqrt{\frac{1}{n-k}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

(k = Anzahl der unabhängigen Variablen)

Bemerkung: In die Berechnung von σ_e gehen — im Gegensatz zu R^2 — alle Untersuchungseinheiten mit gleichem Gewicht ein. Daneben weist dieses Gütemaß aber einige Eigenschaften auf, die gegen eine Verwendung sprechen:

1. Der Wert ist nach oben unbegrenzt.
2. Der Wert ist abhängig von der Anzahl der Untersuchungseinheiten und der Anzahl der unabhängigen Variablen.
3. Der Betrag steht im umgekehrten Verhältnis zur inhaltlichen Aussage.

5.4.5 Inferenzstatistik in der bivariaten Regressionsanalyse

Inferenzstatistische Eigenschaften des OLS-Schätzers b

Die im Rahmen der bisherigen deskriptiven Regressionsanalyse ermittelten Werte (Regressionskoeffizienten (b), Bestimmtheitsmaß (R^2)) beschreiben lediglich die Zusammenhänge innerhalb der vorliegenden Stichprobe.

Frage: Wie weit und unter welchen Bedingungen lassen sich diese Zusammenhänge auch auf die Grundgesamtheit übertragen?

Da in der Regel immer nur eine Stichprobe vorliegt, sind die auf der Basis dieser Stichprobe geschätzten Werte “fehlerhaft”, d.h. sie weichen von dem “wahren” Wert der Grundgesamtheit ab. Die Verfahren der Inferenzstatistik erlauben es nun, unter Berücksichtigung bestimmter, vom Anwender auf Plausibilität zu

überprüfender Annahmen über die verschiedenen Modellparameter und die Stichprobe, die *Signifikanz* von Schätzwerten für die Grundgesamtheiten innerhalb bestimmter Fehlerwahrscheinlichkeiten zu bestimmen. Dazu wird analog zu Kapitel 4.2 ein Hypothesentest durchgeführt.

Der im Rahmen einer Regressionsanalyse ermittelte Regressionskoeffizient für standardisierte Ausgangsvariablen[3]

$$b = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$$

beruht auf einer einzigen Stichprobe aus der Grundgesamtheit.

Andere Betrachtungsweise für b:

> Der berechnete Wert des empirischen Regressionskoeffizienten ist *die Realisierung einer Zufallsvariablen*, d.h. jede Regressionsanalyse auf der Basis einer weiteren Stichprobe ergäbe eine weitere Realisierung.

Frage: Welche Eigenschaften (Erwartungswerte, Varianzen bzw. Kovarianzen) hat diese Zufallsvariable bezogen auf n (hypothetische) Stichproben?

1. b ist im klassischen linearen Modell erwartungstreu (bzw. unverzerrt):

$$E(b) = \beta$$

Dies bedeutet, daß sich die Schätzwerte aus den einzelnen Stichproben "gleichmäßig" um den wahren Wert β für die Grundgesamtheit verteilen, der OLS-Schätzer b im klassischen linearen Modell somit der varianzminimale Schätzer (*b*est *l*inear *u*nbiased *e*stimator = *blue*) ist.

Herleitung:

$$E(b) = E(\frac{1}{n}\sum_{i=1}^{n} x_i y_i)$$

[3] Im weiteren Verlauf werden standardisierte Variablen nicht mehr durch "Sternchen" gekennzeichnet (statt x_i^* also x_i) und b anstatt b_1 verwendet.

Durch Einsetzen der Regressionsgleichung $y_i = x_i\beta + e_i$ — mit dem "wahren" Regressionskoeffizienten β anstatt des Schätzwertes — ergibt sich:

$$\begin{aligned} E(b) &= E(\ \frac{1}{n}\sum_{i=1}^{n} x_i(x_i\beta + e_i)\ ) \\ &= E(\ \frac{1}{n}\sum_{i=1}^{n} x_i^2\beta + \frac{1}{n}\sum_{i=1}^{n} x_i e_i\ ) \\ &= \underbrace{\frac{1}{n}\sum_{i=1}^{n} x_i^2}_{=1} E(\beta) + \frac{1}{n}\sum_{i=1}^{n} x_i \underbrace{E(e_i)}_{=0} \\ &= \beta \end{aligned}$$

2. Für die Varianz der Zufallsvariablen b gilt:

$$\text{var}(b) = (\frac{1}{n}\sum_{i=1}^{n} x_i)^2 s_e^2$$

Herleitung:

$$\begin{aligned} \text{var}(b) &= E\left[(b - E(b))(b - E(b))\right] \\ &= E\left[(b - \beta)(b - \beta)\right] \\ &= E\left[(\underbrace{\frac{1}{n}\sum_{i=1}^{n} x_i^2}_{=1}\beta + \frac{1}{n}\sum_{i=1}^{n} x_i e_i - \beta)\ (\underbrace{\frac{1}{n}\sum_{i=1}^{n} x_i^2}_{=1}\beta + \frac{1}{n}\sum_{i=1}^{n} x_i e_i - \beta)\right] \\ &= E\left[(\frac{1}{n}\sum_{i=1}^{n} x_i e_i)(\frac{1}{n}\sum_{i=1}^{n} x_i e_i)\right] = (\frac{1}{n}\sum_{i=1}^{n} x_i)^2 E[e_i\ e_i] \\ &= (\frac{1}{n}\sum_{i=1}^{n} x_i)^2 E[(e_i - E(e_i))(e_i - E(e_i))] \\ &= (\frac{1}{n}\sum_{i=1}^{n} x_i)^2\ \text{var}(e_i) \end{aligned}$$

Da die e_i als Realisation eines Vektors von Zufallsvariablen ($\boldsymbol{e}$) aufgefaßt werden können, wird die Varianz dieses Vektors auch nicht durch einen Wert ausgedrückt, sondern durch die Kovarianzmatrix $\Sigma_{\boldsymbol{e}}$ aller Varianzen und Kovarianzen zwischen den einzelnen Variablen beschrieben, die durch das dyadische Produkt von $E(\boldsymbol{e}\ \boldsymbol{e}^T)$ entsteht.

Für die Varianz von $\boldsymbol{e}$ gilt dann:

$$\begin{aligned}
\text{var}(\boldsymbol{e}) &= E(\ (\boldsymbol{e} - E(\boldsymbol{e}))(\boldsymbol{e} - E(\boldsymbol{e}))^T) \\
&= E(\boldsymbol{e}\ \boldsymbol{e}^T) \\
&= \Sigma_{\boldsymbol{e}} \\
&= \begin{pmatrix} \text{var}(e_1) & \text{cov}(e_1, e_2) & \cdots & \text{cov}(e_1, e_n) \\ \text{cov}(e_2, e_1) & & \cdots & \text{cov}(e_2, e_n) \\ \vdots & \ddots & \vdots & \vdots \\ \text{cov}(e_n, e_1) & & \cdots & \text{var}(e_n) \end{pmatrix}
\end{aligned}$$

Die konkrete Berechnung dieser Kovarianzmatrix bereitet in der Praxis Schwierigkeiten, da lediglich *eine* Stichprobe zur Verfügung steht. Um trotzdem Berechnungen durchführen zu können, werden über die Fehlervariablen $\boldsymbol{e}$ zusätzliche Annahmen gemacht, durch die sich die Kovarianzmatrix erheblich vereinfacht:

1. Annahme: *Homoskedastizität*,
d.h. die Varianz der einzelnen Fehlervariablen ist für alle möglichen Werte konstant:

$$\text{var}(e_i) = \sigma_e^2, \quad (i = 1, \ldots, n)$$

Interpretation: Die Restschwankungen dürfen nicht davon abhängig sein, zu welchem Zeitpunkt die Erhebung erfolgte.

2. Annahme: *keine Autokorrelation*,
d.h. die Residuen sind untereinander nicht korreliert:

$$\text{cov}(e_i, e_j) = 0, \quad (i, j = 1, \ldots, n; i \neq j)$$

Interpretation: Der Wert einer Restschwankung darf nicht von den Restschwankungen vorhergehender Beobachtungswerte abhängig sein, wie es zum Beispiel häufig in Zeitreihen der Fall ist.

Unter Berücksichtigung dieser Annahmen ergibt sich für die Struktur der Kovarianzmatrix:[4]

$$\text{var}(\boldsymbol{e}) = \begin{pmatrix} \sigma_e^2 & 0 & \cdots & 0 \\ 0 & & \cdots & \\ \vdots & \vdots & \ddots & \vdots \\ 0 & & \cdots & \sigma_e^2 \end{pmatrix} = \sigma_e^2\ \boldsymbol{E_n}$$

[4]Hierbei bezeichnet $\boldsymbol{E_n}$ die Einheitsmatrix mit n Zeilen bzw. Spalten und der mit 1 besetzten Diagonalen (e_{ii}) für $i = 1, \ldots, n$.

Auch die gesamte Fehlervarianz von $\boldsymbol{e}$ über n Stichproben ist natürlich nicht bekannt. Bekannt ist aber die Varianz der Residuen aus der vorliegenden Stichprobe, korrigiert nach Freiheitsgraden:

$$s_e^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Diese wird hier als Schätzwert für die Gesamtvarianz σ_e^2 verwendet. Damit läßt sich die Herleitung der Varianz von b wie folgt vervollständigen:

$$\begin{aligned}
\text{var}(b) &= (\frac{1}{n}\sum_{i=1}^{n}x_i)^2 \ \text{var}(e_i) \\
&= (\frac{1}{n}\sum_{i=1}^{n}x_i)^2\sigma_e^2 \\
&= (\frac{1}{n}\sum_{i=1}^{n}x_i)^2 s_e^2
\end{aligned}$$

Daraus folgt für die Standardabweichung der Residuen (die häufig auch als "Standardfehler" bezeichnet wird):

$$\begin{aligned}
SE(b) &= \sqrt{s_e^2(\frac{1}{n}\sum_{i=1}^{n}x_i)^2} \\
&= s_e(\frac{1}{n}\sum_{i=1}^{n}x_i)
\end{aligned}$$

Signifikanztest für b

Entsprechend den grundlegenden Ausführungen zu inferenzstatistischen Schlußverfahren in Kapitel 4.2 ist es notwendig, für einen zu testenden Modellparameter b eine Zufallsvariable zu konstruieren, die einer bekannten Wahrscheinlichkeitsverteilung genügt.

Hierzu wird eine weitere Annahme über die Residuen benötigt:

3. Annahme: Die Residuen der Regressionsfunktion sind normalverteilt:

$$e_i \sim N(0, \sigma_e^2)$$

b und var(b) können als linear abhängige Funktionen der Zufallsvariablen y ebenfalls als Zufallsvariablen betrachtet werden, für die nach obiger Annahme und der Herleitung von var(b) gilt:

$$\begin{array}{lcll} b & & \sim & N(\beta, s_e^2(\frac{1}{n}\sum_{i=1}^{n} x_i)^2) \\ \text{var}(b) & = s_b^2 & \sim & \chi^2(n-1) \end{array}$$

d.h. b ist normalverteilt und var(b) χ^2-verteilt.

Damit läßt sich (vgl. Anhang A.3) eine t-verteilte Zufallsvariable

$$t = \frac{b - \beta}{s_b}$$

mit $(n-1)$ Freiheitsgraden konstruieren, mit deren Hilfe ein Hypothesentest (t-Test) möglich ist:

1. Schritt: Hypothesenformulierung

$H_0 : \beta = 0$ (Zweiseitiger Test)

$H_A : \beta \neq 0$

allgemein: $H_0 : \beta = \beta^*$ mit β^* als reeller Zahl
$H_A : \beta \neq \beta^*$

2. Schritt: Bestimmung der Prüfverteilung

Die mit Hilfe von b und s_b konstruierte Zufallsvariable t

$$t = \frac{b - \beta}{s_b}$$

ist t-verteilt.

3. Schritt: Bestimmung des Freiheitsgrads df

$df = n - 1$

4. Schritt: Bestimmung des Signifikanzniveaus α

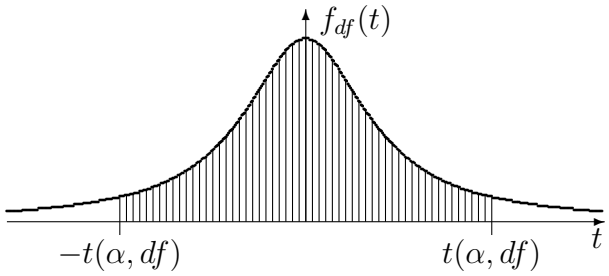
5. Schritt: Ermittlung des theoretischen Verteilungswertes t und Berechnung des empirischen Verteilungswertes t_{emp}

Während der theoretische Verteilungswert t in Abhängigkeit von α und df aus Tabellen ermittelt werden kann, läßt sich t_{emp} nach obiger Formel berechnen ($\beta = 0$).

6. Schritt: Entscheidung über die Ablehnung von H_0

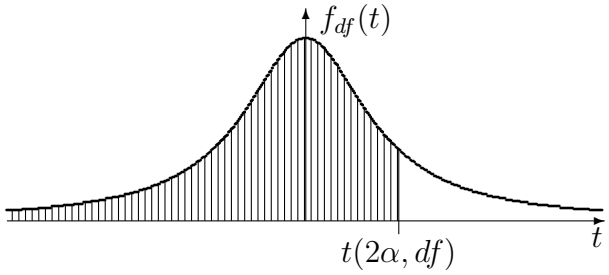
Allgemein sind auf der Basis des t-Tests folgende Testvarianten möglich:

Zweiseitiger Test: $H_0 : \beta = 0$



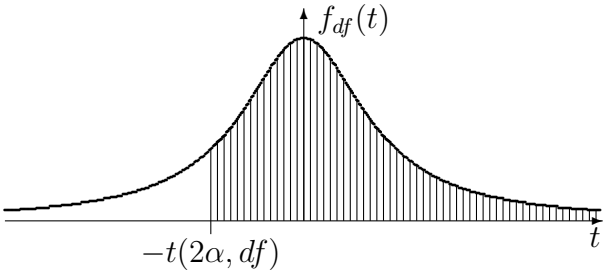
$$P(-t(\alpha, df) < t_{emp} < t(\alpha, df)) \; = \; 1 - \alpha$$

Rechts-einseitiger Test: $H_0 : \beta \leq 0$



$$P(t_{emp} \leq t(2\alpha, df)) \; = \; 1 - \alpha$$

Links-einseitiger Test: $H_0 : \beta \geq 0$



$$P(t_{emp} \geq t(2\alpha, df)) \; = \; 1 - \alpha$$

Abb. 5.4: Graphische Bedeutung des t-Tests

Bemerkung: In den Graphiken werden die Bereiche für Annahme/Ablehnung der Nullhypothese repräsentiert durch:

$\boxed{\alpha}$ Ablehnungsbereich:
Mit einer Wahrscheinlichkeit von α % (einseitig) bzw. $2\frac{\alpha}{2}$ % (zweiseitig) wird H_0 irrtümlich verworfen.

Annahmebereich:
Mit $(1-\alpha)$ % Wahrscheinlichkeit kann H_0 vertraut werden.

Nullhypothese	Test	Ablehnungs-bereich von H_0
$H_0 : \beta \; = \; 0$	zweiseitig	$\lvert t_{emp} \rvert \geq t(\alpha, df)$
$H_0 : \beta \leq 0$	rechts-einseitig	$t_{emp} > t(2\alpha, df)$
$H_0 : \beta \geq 0$	links-einseitig	$t_{emp} < t(2\alpha, df)$

Interpretation: Bei Ablehnung der Nullhypothese ist b mit einer Irrtumswahrscheinlichkeit α statistisch signifikant von Null verschieden.

Beispiel: Im Rahmen einer Regressionsanalyse mit Hilfe statistischer Analysesysteme wird der t-Wert zusammen mit den daraus resultierenden Signifikanzniveaus (`Sig T`) standardmäßig ausgegeben. So bedeutet ein Signifikanzniveau `Sig T` = 0.023, daß die gesamte Regressionsschätzung mit einer Fehlerwahrscheinlichkeit von 2.3% signifikant ist. Ein vorab festgelegtes Niveau von 5% würde danach zur Ablehnung der Nullhypothese führen. Die Zusammenfassung in Kapitel 5.4.6 diskutiert ein konkretes Beispiel.

Berechnung des Konfidenzintervalls für b

Die Eigenschaften von Prüfverteilungen (hier t-Verteilung) ermöglichen ebenso, ein Intervall um den zu testenden Modellparameter (hier b) zu berechnen, innerhalb dessen sich der "wahre" Wert (hier β) befindet:

$$
\begin{array}{lrcccl}
P(& -t(\alpha,df) & \leq & t_{emp} & \leq & t(\alpha,df) \quad) = 1-\alpha \\
P(& -t(\alpha,df) & \leq & \frac{b-\beta}{s_b} & \leq & t(\alpha,df) \quad) = 1-\alpha \\
P(& -t(\alpha,df)s_b & \leq & b-\beta & \leq & t(\alpha,df)s_b \quad) = 1-\alpha \\
P(& -b-t(\alpha,df)s_b & \leq & -\beta & \leq & -b+t(\alpha,df)s_b \quad) = 1-\alpha \\
P(& b+t(\alpha,df)s_b & \geq & \beta & \geq & b-t(\alpha,df)s_b \quad) = 1-\alpha \\
P(& b-t(\alpha,df)s_b & \leq & \beta & \leq & b+t(\alpha,df)s_b \quad) = 1-\alpha
\end{array}
$$

Damit liegt β im Intervall

$$[b-t(\alpha,df)s_b, b+t(\alpha,df)s_b]$$

d.h. bei gegebenen Freiheitsgraden und einem angenommenen Signifikanzniveau läßt der Fehler sich exakt berechnen, der bei dem Schluß von der Stichprobe auf die Grundgesamtheit entsteht.

5.4.6 Zusammenfassung

Anhand der Untersuchung des Zusammenhangs zwischen der Zufriedenheit mit der Regierung und der Sympathie zu Kohl wurden unter anderem die folgenden Werte ermittelt:

```
* Aus: REGRESSION /VARIABLES V217 V225 /DEPENDENT V217
*                 /METHOD ENTER.

Equation Number 1    Dependent Variable..   V217    SKALOMETER:CDU CSU-FDP
Variable(s) Entered on Step Number
   1..    V225       SKALOMETER: H  KOHL

Multiple R           .83948
R Square             .70472
Adjusted R Square    .70413
Standard Error      1.50164

Analysis of Variance
                    DF      Sum of Squares      Mean Square
Regression           1          2674.68574       2674.68574
Residual           497          1120.70103          2.25493

F =    1186.14936       Signif F =  .0000

------------------ Variables in the Equation ------------------

Variable              B        SE B       Beta         T  Sig T

V225            .720843     .020930    .839476    34.441  .0000
(Constant)     2.076811     .150789               13.773  .0000
```

Die Korrelation kann dabei ebenfalls über die Ermittlung der Kreuztabelle berechnet werden:

```
* Ausschnitt aus: CROSSTABS /TABLES V217 BY V225 /STATISTICS CORR.

                                                          Approximate
     Statistic              Value        ASE1      T-value   Significance
--------------------       ---------    --------    -------   ------------

Pearson's R                 .83948       .01606    34.44052     .00000
Spearman Correlation        .85194       .01632    36.27099     .00000
```

Der t-Test für den Regressionskoeffizienten b_1 ergibt bei einem t-Wert von 34.441 eine Signifikanz von 0.0% für die Nullhypothese, d.h. b_1 ist bei einer angenommenen Irrtumswahrscheinlichkeit von 5% auch für die Grundgesamtheit signifikant. Die Güte der Gesamtschätzung R^2 liegt bei 0.70472, d.h. über 70% der Varianz der Regierungszufriedenheit wird bereits durch die Sympathie für den Kanzler erklärt. Weiterhin werden die bei standardisierten Variablen existierenden Beziehungen zwischen den Kennwerten R^2 (`R Square`), r (`Pearson's R`) und b_1 (`Beta`) bestätigt:

$$r = b_1 = 0.83948 = \sqrt{0.70472} = \sqrt{R^2}$$

5.5 Zusammenhangsmaße für unterschiedliche Skalenniveaus

Die bisher erläuterten Zusammenhangsmaße erlauben keine Untersuchung des Zusammenhanges zwischen zwei Variablen, die ein unterschiedliches Skalenniveau besitzen.

Fragestellungen:

Sympathieskalometer (metrisch)	$\longleftrightarrow$	Geschlecht (nominal)
Häufigkeit religiöser Aktivitäten (metrisch)	$\longleftrightarrow$	Konfession (nominal)
Wöchentliche Arbeitszeit (metrisch)	$\longleftrightarrow$	Berufsgruppenzugehörigkeit (nominal)

Grundsätzlich lassen sich bei der Untersuchung derartiger Zusammenhänge die Zusammenhangsmaße für die jeweils niedrigere Variablenskalierung verwenden. Dies beinhaltet aber einen Informationsverlust, da die aus der höheren Skalierung resultierenden, zusätzlichen Informationen (Rangfolge der Merkmalsklassen, Größe der Merkmalswerte) nicht einfließen. Das Zusammenhangsmaß η^2 beschreibt den Zusammenhang zwischen einer metrischen und einer beliebig skalierten Variablen, wobei hier natürlich besonders nominale und ordinale Variablen von Interesse sind, da bei einer ebenfalls metrisch skalierten Variablen die Zusammenhangsmaße für zwei metrische Variablen besser geeignet sind (vgl. Kapitel 5.4).

Eta η^2

Skalenniveau: metrisch, beliebig

Berechnung: Das Zusammenhangsmaß η^2 setzt eine metrisch skalierte, abhängige Variable y und eine nominal oder ordinal skalierte, unabhängige Variable x voraus. Für die weiteren Ausführungen werden folgende Festlegungen getroffen:

- y_{ij} : Wert der Variablen y, gemessen an der j-ten Untersuchungseinheit in der Merkmalsklasse i der unabhängigen (nominalen, ordinalen) Variablen x
- $\bar{y}_i$: Mittelwert von y über alle Untersuchungseinheiten in der i-ten Merkmalsklasse von x
- $\bar{y}$: Gesamtmittelwert über alle Merkmalswerte von y
- n_i : Anzahl der Merkmalswerte von y in der i-ten Merkmalsklasse von x

η^2 basiert ebenso wie λ und γ auf dem Prinzip der proportionalen Fehlerreduktion (PRE):

1. Bester Vorhersagewert für y *ohne* Kenntnis der Verteilung von y auf die Merkmalsklassen von x ist der Gesamtmittelwert $\bar{y}$. Der daraus resultierende Schätzfehler ergibt sich dann zu:
$$E_1 = \sum_{k=1}^{\ell} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2$$
d.h. die Summe aller quadrierten Abweichungen der Merkmalswerte von y von dem Vorhersagewert $\bar{y}$ (Gesamtvariation).
2. Mit Kenntnis der unabhängigen Variablen x sind die besten Schätzer für y die Klassenmittelwerte $\bar{y}_i$. Daraus ergibt sich der Vorhersagefehler
$$E_2 = \sum_{k=1}^{\ell} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$
d.h. die Summe aller quadrierten Abweichungen der Merkmalswerte von y von den jeweiligen Klassenmittelwerten.
3. Somit folgt für η^2, definiert als PRE-Maß:
$$\begin{aligned} \eta^2 &= \frac{E_1 - E_2}{E_1} \\ &= \frac{\sum_{k=1}^{\ell} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 - \sum_{k=1}^{\ell} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2}{\sum_{k=1}^{\ell} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2} \\ &= 1 - \frac{\sum_{k=1}^{\ell} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2}{\sum_{k=1}^{\ell} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2} \end{aligned}$$

Bemerkung: η^2 ist ein asymmetrisches, vorzeichenloses Maß im Intervall [0,1].

Die Festlegung, daß die unabhängige Variable metrisch skaliert sein muß, hat ihre Bedeutung darin, daß bei der Berechnung von η^2 die metrischen Maßzahlen $\bar{y}$ und $\bar{y}_i$ benötigt werden, während von x lediglich die Häufigkeit in den einzelnen Merkmalsklassen verwendet wird.

Interpretation:

$\eta^2 = 0 \implies$ die Variation in den einzelnen Kategorien ist identisch mit der Gesamtvariation
$\implies$ x trägt nichts zur Erklärung von y bei.

$\eta^2 = 1 \implies$ die Variation in den einzelnen Kategorien ist 0, d.h. es gibt dort insgesamt keine Streuung der Variablenwerte
$\implies$ x erklärt y vollständig.

Die Höhe von η^2 ist abhängig von der Anzahl der Kategorien von x:

1. Jede Kategorie enthält nur einen Wert ($k = n$), d.h. die Variation in den einzelnen Kategorien ist gleich 0.
   $\implies \eta^2 = 1$

2. Es gibt nur eine Kategorie, d.h. die Variation in dieser Kategorie ist gleich der Gesamtvariation.
   $\implies \eta^2 = 0$

Beispiel: Anhand der Skalometervariablen zur Sympathie von Kohl (metrisch skaliert) wird der Zusammenhang mit dem Geschlecht (nominal skaliert) überprüft (Aufruf: `CROSSTABS /TABLES V334 BY V225 /STATISTICS ETA.`):

```
V334 GESCHLECHT  by  V225 SKALOMETER: H  KOHL

              V225                                                Page 1 of 2
      Count   |
              |                                                        Row
              |    -5|    -4|    -3|    -2|    -1|     0|     1| Total
V334 ---------+------+------+------+------+------+------+------+
           1  |    29|    15|    18|    16|    14|    10|    24|   242
maennlich     |      |      |      |      |      |      |      |  48.5
              +------+------+------+------+------+------+------+
           2  |    27|    11|    16|    18|    23|    28|    33|   257
weiblich      |      |      |      |      |      |      |      |  51.5
              +------+------+------+------+------+------+------+
      Column       56     26     34     34     37     38     57    499
(Cont.)Total     11.2    5.2    6.8    6.8    7.4    7.6   11.4  100.0

                    V225                    Page 2 of 2
            Count   |
                    |                              Row
                    |     2|     3|     4|     5| Total
V334        --------+------+------+------+------+
                 1  |    20|    51|    24|    21|   242
  maennlich         |      |      |      |      |  48.5
                    +------+------+------+------+
                 2  |    26|    18|    27|    30|   257
  weiblich          |      |      |      |      |  51.5
                    +------+------+------+------+
            Column       46     69     51     51    499
             Total      9.2   13.8   10.2   10.2  100.0
                                                          Approximate
     Statistic                Value         ASE1       T-value   Significance
--------------------        ---------     --------     -------   ------------

Eta :
   with V334     dependent   .24911
   with V225     dependent   .00545

Number of Missing Observations:  1
```

Da SPSS nicht das Skalenniveau einer Variablen überprüft, berechnet es den Wert für η^2 sowohl für V334 als auch V225 als abhängige Variable. Die Herleitung dieses Kennwertes hat aber deutlich gemacht, daß für die abhängige Variable die metrischen Eigenschaften unbedingte Voraussetzung sind, d.h. in diesem Beispiel ist ausschließlich der Wert mit der Sympathie für Kohl (V225) als abhängige Variable sinnvoll interpretierbar, der allerdings lediglich eine Abhängigkeit von 0.00545 (=0.545%) ausweist.

5.6 Auswahlkriterien für bivariate Zusammenhangsmaße

Erfahrungsgemäß liegen die größten Probleme bei der Durchführung statistischer Analysen mit Hilfe von Datenanalysesystemen nicht in der Anwendung von Methoden auf Variablen, d.h. in dem Aufruf einer konkreten Statistikprozedur, sondern eher in der "Umsetzung" von Hypothesen, Fragestellungen etc. aus der zugrunde gelegten Theorie in geeignete, zu berechnende statistische Kennzahlen und in einer korrekten und angemessenen Interpretation der berechneten Kennwerte.

Beispiel: "Das Ausmaß der Zufriedenheit mit der Regierung insgesamt wird in hohem Maße durch die Meinung zur Person des Kanzlers bestimmt. Andere Variablen, wie zum Beispiel die Meinung zu anderen Regierungspolitikern, spielen hierbei lediglich eine untergeordnete Rolle."

Unabhängig von der Tatsache, daß jede der beschriebenen statistischen Kennzahlen charakteristische Merkmale mit daraus resultierenden Konsequenzen für ihre Anwendung aufweist, lassen sich für die Vorgehensweise bei der Auswahl einige Faustregeln aufstellen, die etwas Ordnung in die Vielzahl der angebotenen Methoden bringen:[5]

1. Festlegung der an der Analyse beteiligten Variablen, Identifikation der Skalenniveaus

   Beispiel: Sympathieskalometer von Regierungspolitikern (metrisch):
   z.B. Genscher (V224), Kohl (V225), Stoltenberg (V228)
   Zufriedenheit mit der Regierung (metrisch): V217

2. Art der Beziehung zwischen den Variablen (symmetrisch/asymmetrisch)

   Beispiel: V224/V225/V228/... $\longrightarrow$ V217

3. Bevorzugte Verwendung von PRE-Maßen, da diese eine direkte Interpretation ermöglichen (Prozentsatz der Varianzerklärung) und die Vergleichbarkeit untereinander erlauben:

[5] Die in den vorangegangenen Kapiteln vorgestellte Auswahl konzentriert sich dabei nur auf die wichtigsten der von Datenanalysesystemen angebotenen bivariaten Zusammenhangsmaße.

Nominales Skalenniveau	:	λ
Ordinales Skalenniveau	:	γ
Metrisches/beliebiges Skalenniveau	:	η^2
Metrisches Skalenniveau	:	R^2

Beispiel: Da es sich generell um metrische Variablen handelt, empfiehlt sich hier die Auswahl des Bestimmtheitsmaßes R^2, das mit V217 als abhängiger und jeweils einem Sympathieskalometer als unabhängiger Variablen berechnet werden kann. Ein Vergleich der ermittelten Werte erlaubt dann Aussagen im Sinne der Ausgangshypothese. Unter Beachtung der Richtung der Asymmetrie lassen sich auch entsprechende η^2-Werte berechnen und zu Vergleichszwecken heranziehen.

Nachfolgende Tabelle enthält für alle unabhängigen Variablen die jeweiligen Bestimmtheitsmaße R^2 bei der Durchführung einer bivariaten Regression. Hieraus ergibt sich die eindeutige Unterstützung der Ausgangshypothese.

	R^2
Kohl (V225)	0.70460
Stoltenberg (V228)	0.44321
Genscher (V224)	0.19592

Kapitel 6

Multivariate Datenanalyse

6.1 Die Komplexität des sozialwissenschaftlichen Analysegegenstandes

Die bisherigen Betrachtungen beschränkten sich im wesentlichen auf zwei Bereiche:

- Verteilung von empirischen Daten: Mittelwert, Standardabweichung, Varianz, Schiefe, Exzeß etc.
  $\Longrightarrow$ *univariate* Datenanalyse
- Zusammenhangsmaße für zwei Variablen mit gleichen oder unterschiedlichen Skalenniveaus
  $\Longrightarrow$ *bivariate* Datenanalyse

Die Problematik dieser einfachen statistischen Analysemodelle liegt darin, daß sie in der Regel zu stark von der Komplexität des sozialwissenschaftlichen Gegenstandsbereiches abstrahieren. Dies bezieht sich sowohl auf

1. die Komplexität der Ursache–Wirkungsverhältnisse in sozialwissenschaftlichen Erklärungsmodellen (z.B. Erklärungsmodelle der Wahlentscheidung) als auch auf
2. die Komplexität der Indikatorenstruktur in sozialwissenschaftlichen Meßmodellen, d.h. sozialwissenschaftliche Erklärungsbegriffe lassen sich meist nicht nur durch einen Indikator messen (z.B. Parteisympathie $\longleftarrow$ Sonntagsfrage).

Die Behandlung von Problemen dieser Art erfolgt durch Methoden der multivariaten Datenanalyse.

Definition: *Multivariate Datenanalyse:*
"die gemeinsame, gleichzeitige Analyse mehrerer Merkmale bzw. deren Ausprägungen" [Hartung/Elpelt 1984, S. 2].

Als Beispiele hierzu werden im weiteren Verlauf einige zentrale Verfahren der multivariaten Datenanalyse vorgestellt und die Anwendungsbedingungen für ihren Einsatz aufgezeigt. Im Bereich komplexer Erklärungsmodelle für kausale Zusammenhänge (Erklärungsanalyse) zählen hierzu

- die *multiple Regression*, mit einer metrischen, abhängigen Variablen und mehreren metrischen, unabhängigen Variablen sowie
- die *Diskriminanzanalyse*, bei der die abhängige Variable nominal skaliert ist.

Für die Entdeckung sozialwissenschaftlicher Begriffe auf der Basis von metrischen, empirischen Variablen (Strukturanalyse) dienen

- die *Faktorenanalyse*, bei der die resultierenden, theoretischen Variablen (Faktoren) ebenfalls metrisch skaliert sind, und
- die *Clusteranalyse*, die als Ergebnis einen nominal skalierten Faktor erzeugt.

6.2 Multiple Regressionsanalyse

6.2.1 Das allgemeine Modell der multiplen Regression

In Erweiterung des bivariaten Regressionsansatzes (vgl. Kapitel 5.4.2) erlaubt die multiple Regressionsanalyse die Untersuchung der Abhängigkeit einer Variablen von mehreren anderen Variablen, d.h. Ausgangspunkt ist folgende Variablenkonfiguration:

- *eine metrisch skalierte, abhängige, endogene, zu erklärende Variable* y (Regressand y)
- *mehrere metrisch skalierte, unabhängige, exogene, erklärende Variablen* x_j $(j = 1, \ldots, m)$ (Regressoren)

Ebenso wie im bivariaten Fall wird aufgrund theoretischer bzw. meßtheoretischer Annahmen von der Existenz eines *linearen* Zusammenhangs zwischen y und den x_j ausgegangen:

$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_m x_m$$

Für den einzelnen Befragten $(i = 1, \ldots, n)$ ergibt sich daraus:

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_m x_{im}$$

Durch die Einführung einer Hilfsvariablen x_0, deren Variablenwerte alle den Wert 1 erhalten ($x_{i0} = 1, i = 1, \ldots, n$), kann der lineare Zusammenhang zwischen y und den x_j auf der Ebene aller empirischen Werte in Matrizenschreibweise formuliert werden:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b}$$

wobei

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

den n–dimensionalen Spaltenvektor der Messungen von y,

$$\boldsymbol{b} = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_m \end{pmatrix}$$

den $(m+1)$–dimensionalen Spaltenvektor der Regressionskoeffizienten b_j und

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1j} & \cdots & x_{1m} \\ 1 & x_{21} & \cdots & x_{2j} & \cdots & x_{2m} \\ \vdots & \vdots & & \vdots & & \vdots \\ 1 & x_{i1} & \cdots & x_{ij} & \cdots & x_{im} \\ \vdots & \vdots & & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{nj} & \cdots & x_{nm} \end{pmatrix}$$

die $(n \times m+1)$–dimensionale Matrix der Werte der unabhängigen Variablen x_j bezeichnet.

Im Fall der bivariaten Regression hat die Matrix der unabhängigen Variablen $\boldsymbol{X}$ genau zwei Spalten ($m = 0, 1$) und der Vektor der Regressionskoeffizienten zwei Werte (b_0, b_1).

Beispiel: Überprüfung einer (Meß-)Theorie der Parteisympathie

Hypothese: Parteisympathie ist eine lineare Projektion der persönlichen Sympathien und Antipathien gegenüber einzelnen Parteipolitikern auf eine Partei.

Ausgangspunkt für die Überprüfung der Hypothese mit Hilfe der multiplen Regressionsanalyse ist die oben eingeführte Matrizengleichung, wobei

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

den Vektor der Antworten auf die Sympathieskalometerfrage zu einer Partei (z.B. CDU) bezeichnet, während

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1j} & \cdots & x_{1m} \\ 1 & x_{21} & \cdots & x_{2j} & \cdots & x_{2m} \\ \vdots & \vdots & & \vdots & & \vdots \\ 1 & x_{i1} & \cdots & x_{ij} & \cdots & x_{im} \\ \vdots & \vdots & & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{nj} & \cdots & x_{nm} \end{pmatrix}$$

die $(n \times m + 1)$–Matrix der Antworten auf m Sympathieskalometerfragen bezüglich Spitzenpolitikern aus dieser Partei ist.

Die formulierte Grundgleichung der multiplen Regression beschreibt einen deterministischen Zusammenhang zwischen abhängiger und unabhängigen Variablen, der in den Sozialwissenschaften in der Regel unangemessen ist.

Analog zum bivariaten Ansatz wird die Existenz eines Fehlerterms e_i $(i = 1, \ldots, n)$ angenommen, in dem alle Schätzfehler der y_i-Werte aufgrund der x_{ij}-Werte zusammengefaßt sind (z.B. Einflüsse nicht berücksichtigter Erklärungsvariablen, Meßfehler für y_i, fehlerhafte Spezifikation der funktionalen Beziehung zwischen y_i und x_{ij}).

Da e als Zufallsvariable aufgefaßt werden kann, deren Werte e_i die y_i zusätzlich beeinflussen — wobei sich deren Wirkung im Mittel aufhebt ($E(e) = 0$) —, wird bei Einbeziehung der e_i in die Regressionsgleichung auch von einem *stochastischen Regressionsmodell* gesprochen:

$$y_i = b_0 x_{i0} + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_m x_{im} + e_i, \qquad (i = 1, \ldots, n)$$

bzw.

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e}$$

Lediglich die geschätzten Variablenwerte $\hat{y}_i$ erfüllen die ursprüngliche Regressionsgleichung mit

$$\hat{y}_i = b_0 x_{i0} + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_m x_{im}, \qquad (i = 1, \ldots, n)$$

bzw.

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{b}$$

Somit läßt sich das stochastische Regressionsmodell auch wie folgt modellieren:

$$y_i = \hat{y}_i + e_i, \qquad (i = 1, \ldots, n)$$

bzw.

$$\boldsymbol{y} = \hat{\boldsymbol{y}} + \boldsymbol{e}$$

Hieraus folgt für den Fehlerterm:

$$e_i = y_i - \hat{y}_i, \qquad (i = 1, \ldots, n)$$

bzw.

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}}$$

6.2.2 Die Verallgemeinerung der Kleinstquadratschätzung für m Regressoren ($m > 1$)

Ebenso wie im bivariaten Fall besteht die Aufgabe bei der multiplen Regression darin, den Regressionskoeffizienten $\boldsymbol{b}$ durch Minimierung des Fehlerterms $\boldsymbol{e}$ mit Hilfe der Kleinstquadratmethode zu schätzen.

Die Regressionsgleichung, die Prognosegleichung und der Fehlerterm lauten für den allgemeinen Fall:

$$\begin{aligned} y_i &= b_0 x_{i0} + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_m x_{i(m)} + e_i \\ \hat{y}_i &= b_0 x_{i0} + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_m x_{im} \\ e_i &= y_i - \hat{y}_i \end{aligned}$$

bzw. in Matrizenschreibweise:

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e} \\ \hat{\boldsymbol{y}} &= \boldsymbol{X}\boldsymbol{b} \\ \boldsymbol{e} &= \boldsymbol{y} - \boldsymbol{X}\boldsymbol{b} \end{aligned}$$

Die Matrix $\boldsymbol{X}$ und der Vektor der abhängigen Variablen ($\boldsymbol{y}$) sind gegeben. Für die Schätzung der Regressionskoeffizienten wird von der Summe der Fehlerquadrate der e_i $(i = 1, \ldots, n)$ ausgegangen:

$$\begin{aligned} S(e) &= \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} (y_i - (b_0 x_0 + b_1 x_{i1} + \ldots + b_m x_{im}))^2 \quad \stackrel{!}{=} \quad \textit{Minimum} \end{aligned}$$

Die Durchführung der Kleinstquadratschätzung erfolgt in der Matrizenschreibweise. Der Übergang von der Summen- zur Matrizendarstellung für das Minimierungskriterium $S(e)$ läßt sich dabei an einem einfachen Beispiel erläutern ($n = 2$):

$$\sum_{i=1}^{2} e_i^2 = e_1^2 + e_2^2 = (e_1 \; e_2) \begin{pmatrix} e_1 \\ e_2 \end{pmatrix} = \boldsymbol{e}^T \boldsymbol{e}$$

Die zu minimierende Funktion lautet also in Matrizenschreibweise:

$$S(e) = \boldsymbol{e}^T \boldsymbol{e}$$

$$
\begin{aligned}
&= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}) \\
&= \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}\boldsymbol{b} - \boldsymbol{b}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{b}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{b} \\
&= \boldsymbol{y}^T\boldsymbol{y} - (\boldsymbol{X}\boldsymbol{b})^T\boldsymbol{y} - \boldsymbol{b}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{b}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{b} \\
&= \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{b}^T\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{b}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{b}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{b} \\
&= \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{b}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{b}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{b}
\end{aligned}
$$

Partielle Ableitungen nach $\boldsymbol{b}$*:*

$$
\begin{aligned}
\frac{\delta S(e)}{\delta \boldsymbol{b}} &= \frac{\delta(\boldsymbol{y}^T\boldsymbol{y})}{\delta \boldsymbol{b}} - 2\frac{\delta(\boldsymbol{b}^T\boldsymbol{X}^T\boldsymbol{y})}{\delta \boldsymbol{b}} + \frac{\delta(\boldsymbol{b}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{b})}{\delta \boldsymbol{b}} \\
&= -2\boldsymbol{X}^T\boldsymbol{y} + 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{b}
\end{aligned}
$$

Normalengleichung für $\boldsymbol{b}$*:*

$$
\begin{aligned}
0 &= -2\boldsymbol{X}^T\boldsymbol{y} + 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{b} \\
\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{b} &= \boldsymbol{X}^T\boldsymbol{y}
\end{aligned}
$$

Unter der Voraussetzung, daß $(\boldsymbol{X}^T\boldsymbol{X})$ invertierbar ist, kann die Gleichung von links mit $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ multipliziert werden:

$$
(\boldsymbol{X}^T\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{X})\boldsymbol{b} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}
$$

Da $(\boldsymbol{X}^T\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{X}) = \boldsymbol{E_n}$ und $\boldsymbol{E_n}\boldsymbol{b} = \boldsymbol{b}$ lautet die Lösung für die Schätzung des Regressionskoeffizientenvektors $\boldsymbol{b}$:

$$
\boldsymbol{b} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}
$$

Eigenschaften von b_j:

- Steigt (fällt) unter sonst gleichen Bedingungen der Wert des j-ten Regressionskoeffizienten um eine Einheit, dann steigt (fällt) der Wert der abhängigen Variablen um so viele Einheiten, wie b_j groß ist.

- Das Vorzeichen von b_j gibt die Richtung der Beziehung zwischen x_j und y an.

- b_j ist der relative Erklärungsbeitrag der Variablen x_j für y, wenn alle übrigen x_k $(k \neq j)$ konstant und die Maßeinheiten der betrachteten Variablen gleich sind. Dies wird durch die Standardisierung der b_j erreicht:

  Berechnung von standardisierten Regressionskoeffizienten b_j^* aus b_j (s.S. 99):

  $$b_j^* = b_j \frac{s_{x_j}}{s_y}, \qquad (j = 0, \ldots, m)$$

- Der standardisierte Regressionskoeffizient ist dimensionslos und auf das Intervall von $[-1, +1]$ beschränkt.

Die Schätzfunktion $S(e)$ ist — als lineares Gleichungssystem — nur dann eindeutig lösbar, wenn die Zahl der Untersuchungseinheiten größer oder gleich der Anzahl der unabhängigen Variablen ist ($n \geq m$), d.h. wenn es mindestens so viele Bestimmungsgleichungen wie unbekannte b_j gibt.

Bemerkung: Die Forderung nach Invertierbarkeit von $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ ist die Verallgemeinerung der Forderung im bivariaten Fall, daß die Varianz der erklärenden Variablen $x_{(j)}$ ungleich Null ist.

Voraussetzung für die Invertierbarkeit von $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ ist dabei die Annahme, daß *alle Variablen* x_j voneinander *linear unabhängig* sind:

x_j $(j = 0, \ldots, m)$ linear unabhängig
$\Longrightarrow Rang(\boldsymbol{X})$ ist maximal $(= n)$
$\Longrightarrow \boldsymbol{X}$ ist regulär
$\Longrightarrow \boldsymbol{X}^T\boldsymbol{X}$ ist regulär und invertierbar

6.2.3 Ein Gütemaß zur Beurteilung von Regressionsschätzungen

Ebenso wie im bivariaten Fall läßt sich auch hier das Bestimmtheitsmaß R^2 als Gütekriterium verallgemeinern (vgl. Kapitel 5.4.3). Danach gilt für R^2 :

$$R^2 = \frac{\sum_{i=1}^{n} \hat{y}_i^2 - n\bar{y}^2}{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2}$$

bzw. in Matrizenschreibweise:

$$R^2 = \frac{\hat{\boldsymbol{y}}^T\hat{\boldsymbol{y}} - n\bar{y}^2}{\boldsymbol{y}^T\boldsymbol{y} - n\bar{y}^2}$$

Weiterhin läßt sich umformen:

$$\begin{aligned} \hat{\boldsymbol{y}}^T\hat{\boldsymbol{y}} &= (\boldsymbol{X}\boldsymbol{b})^T\boldsymbol{X}\boldsymbol{b} \\ &= \boldsymbol{b}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{b} \\ &= \boldsymbol{b}^T\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \\ &= \boldsymbol{b}^T\boldsymbol{X}^T\boldsymbol{y} \end{aligned}$$

Somit läßt sich R^2 aus $\boldsymbol{y}$, $\boldsymbol{X}$ und $\boldsymbol{b}$ berechnen:

$$R^2 = \frac{\boldsymbol{b}^T\boldsymbol{X}^T\boldsymbol{y} - n\bar{y}^2}{\boldsymbol{y}^T\boldsymbol{y} - n\bar{y}^2}$$

Eigenschaften von R^2:

1. R^2 gibt den Anteil der durch alle unabhängigen Variablen x_j erklärten Varianz von y an (PRE-Maß: $0 \leq R^2 \leq 1$)

2. Analog zum bivariaten Fall ist R^2 abhängig von extremen Fällen, d.h. es werden nicht alle Variablenwerte gleich behandelt.

3. Weiterhin ist R^2 abhängig von der Zahl der unabhängigen Variablen, *nicht* aber von der Anzahl der Fälle — wenn die $(k+1)$-te Variable nicht linear von den k bisher in die Regression aufgenommenen Variablen abhängig ist, steigt R^2 durch die Hinzunahme der $(k+1)$-ten Variable an; bei n Fällen ist die n-te Variable auf jeden Fall von allen anderen Variablen abhängig, daher ist R^2 bei n Fällen und $n-1$ Variablen auf jeden Fall gleich 1.

4. Das korrigierte R^2: $\hat{R}^2 = \left(R^2 - \frac{k}{n-1}\right)\left(\frac{n-1}{n-k-1}\right)$ trägt diesem Effekt Rechnung; hier kann allerdings die Aufnahme einer zusätzlichen Variablen zu einer Verminderung des korrigierten R^2 führen.

Gründe für ein niedriges R^2:

- Spezifikationsfehler (z.B. fehlende Erklärungsvariablen)
- Meßfehler in den Indikatoren (wird etwa in dem folgenden Beispiel zu jedem Wert der abhängigen Variablen V217 eine normalverteilte Zufallsvariable mit Mittelwert 0 und Standardabweichung 0.2 addiert, so sinkt R^2 geringfügig ab; bei einer Standardabweichung von 1 sinkt R^2 von 0.70144 auf 0.61422; dabei wurde nur V225 als unabhängige Variable benutzt)

Das nachfolgende Beispiel zeigt die Ergebnisse einer multiplen Regression mit der Zufriedenheit mit der Regierung (V217) als abhängiger Variable und den Sympathieskalometern von führenden Politikern der Regierungsparteien CDU, CSU und FDP (Bangemann (V222), Genscher (V224), Kohl (V225), Stoltenberg (V228), Strauß (V229)) als unabhängige Variablen. Die Auswahl der Regressionsmethode `FORWARD` bedeutet, daß nacheinander immer die Variable in die Regressionsgleichung einbezogen wird, die jeweils den größten Anteil an der Varianz der abhängigen Variablen erklärt *und* die gleichzeitig auf dem 5%-Niveau (Default-Einstellung) signifikant ist (siehe `Sig T`-Werte).

Die Analyse zeigt, daß erwartungsgemäß Kohl mit einem Erklärungsanteil von über 70% (`R Square` = 0.70144) als erste Variable in die Gleichung einbezogen wird. Im zweiten Schritt erfolgt die Hinzunahme von Stoltenberg, so daß für beide Variablen zusammen ein Erklärungsanteil von 72.3% (`R Square` = 0.72328) ereicht wird. Ein dritter Regressionsschritt schließlich bringt eine Erweiterung durch Strauß (als Nichtregierungsmitglied!) und damit einen Gesamterklärungsanteil von 73.1% (`R Square` = 0.73075). Wie aus den `Sig T`-Werten für die Sympathieskalometer für Genscher und Bangemann ersichtlich (0.1495 bzw. 0.4621), sind diese beiden Variablen auf dem 5%-Niveau nicht signifikant und werden daher nicht in die Regressionsgleichung aufgenommen.

Insgesamt zeigt sich, daß der multiple Regressionsansatz die Hypothese von der herausragenden Bedeutung der Person des Kanzlers für die Zufriedenheit mit der Regierung noch stärker unterstützt, als dies bei der bivariaten Regression der Fall war.

```
* SPSS-Aufruf: REGRESSION /VARIABLES V217 V222 V224 V225 V228 V229
```

```
*                        /DEPENDENT V217 /METHOD FORWARD.

         * * * *   M U L T I P L E   R E G R E S S I O N   * * * *

Equation Number 1    Dependent Variable..   V217    SKALOMETER:CDU CSU-FDP
Variable(s) Entered on Step Number
   1..    V225       SKALOMETER: H  KOHL

Multiple R           .83752
R Square             .70144
Adjusted R Square    .70083
Standard Error      1.50646

Analysis of Variance
                    DF      Sum of Squares      Mean Square
Regression           1          2575.29474       2575.29474
Residual           483          1096.12794          2.26942

F =    1134.78300       Signif F =  .0000

------------------ Variables in the Equation ------------------

Variable              B        SE B       Beta         T  Sig T

V225            .720092     .021376    .837522    33.687  .0000
(Constant)     2.095243     .153430               13.656  .0000

------------- Variables not in the Equation -------------

Variable     Beta In  Partial  Min Toler         T  Sig T

V222         .005460  .008330    .694818      .183  .8550
V224         .072714  .118912    .798453     2.629  .0088
V228         .196943  .270467    .563085     6.168  .0000
V229         .181397  .229683    .478658     5.181  .0000
```

```
* * * *   M U L T I P L E   R E G R E S S I O N   * * * *

Equation Number 1    Dependent Variable..   V217    SKALOMETER:CDU CSU-FDP
Variable(s) Entered on Step Number
   2..    V228       SKALOMETER:G STOLTENBERG

Multiple R           .85046
R Square             .72328
Adjusted R Square    .72214
Standard Error      1.45182

Analysis of Variance
                    DF      Sum of Squares      Mean Square
Regression           2          2655.47899       1327.73949
Residual           482          1015.94370          2.10777

F =     629.92707       Signif F =  .0000

------------------ Variables in the Equation ------------------

Variable              B        SE B       Beta          T  Sig T

V225            .608166     .027454    .707344    22.153  .0000
V228            .242621     .039336    .196943     6.168  .0000
(Constant)     1.021944     .228353                4.475  .0000

------------- Variables not in the Equation -------------

Variable     Beta In  Partial  Min Toler         T  Sig T

V222        -.035880 -.055423    .498834    -1.217  .2241
V224         .018568  .029657    .497841      .651  .5155
V229         .130512  .164307    .407351     3.653  .0003
```

```
* * * *   M U L T I P L E   R E G R E S S I O N   * * * *

Equation Number 1    Dependent Variable..   V217    SKALOMETER:CDU CSU-FDP
Variable(s) Entered on Step Number
   3..    V229       SKALOMETER: F J  STRAUSS

Multiple R           .85484
R Square             .73075
Adjusted R Square    .72907
Standard Error      1.43357

Analysis of Variance
                    DF      Sum of Squares      Mean Square
Regression           3          2682.90607        894.30202
Residual           481           988.51661          2.05513

F =     435.15634       Signif F =  .0000

------------------ Variables in the Equation ------------------

Variable              B        SE B       Beta          T  Sig T

V225            .546933     .031872    .636125    17.160  .0000
V228            .199723     .040578    .162122     4.922  .0000
V229            .110945     .030369    .130512     3.653  .0003
(Constant)     1.024042     .225484                4.542  .0000

------------- Variables not in the Equation -------------

Variable     Beta In  Partial  Min Toler         T  Sig T

V222        -.042054 -.065748    .378548    -1.444  .1495
V224         .020739  .033574    .398552      .736  .4621

Equation Number 1    Dependent Variable..   V217    SKALOMETER:CDU CSU-FDP
End Block Number   1   PIN =     .050 Limits reached.
```

6.2.4 Signifikanztests in der multiplen Regressionsanalyse

In der multiplen Regressionsanalyse sind sowohl die berechneten Regressionskoeffizienten $\boldsymbol{b}$ als auch die Güte der Gesamtschätzung R^2 auf ihre Signifikanz bezogen auf die Grundgesamtheit zu testen.[1]

Mit Hilfe inferenzstatistischer Schlüsse lassen sich im Rahmen der multiplen Regression zum Beispiel folgende Fragen beantworten:

- Hat eine Gruppe von m Regressoren oder haben einzelne Regressoren einen signifikanten Einfluß auf y?
- Ist ein geschätztes R^2 signifikant? Insbesondere: Wird bei Hinzunahme einer weiteren unabhängigen Variablen (x_{m+1}) R^2 signifikant?
- In der Stichprobe wurden zwei Regressionskoeffizienten b_i und b_j geschätzt. Wie groß ist die Wahrscheinlichkeit, daß in der Grundgesamtheit b_i von b_j verschieden ist?

Dabei wird der Test von $\boldsymbol{b}$ analog zum bivariaten Fall für jeden einzelnen Regressionskoeffizienten b_j $(j = 1, \ldots, m)$ mittels t-Test (vgl. Kapitel 5.4.5) vorgenommen.

Der Signifikanztest für die Gesamtgüte der Schätzung erfolgt mit Hilfe des F-Tests. Als Vorbereitung hierzu ist es notwendig, für den zu testenden Parameter R^2 eine Zufallsvariable zu konstruieren, die einer bekannten Wahrscheinlichkeitsverteilung — hier der F-Verteilung — genügt (vgl. Anhang A.3).

Da sich R^2 aus der Summe der quadrierten Zufallsvariablen y_i bildet (vgl. $\boldsymbol{y}^T\boldsymbol{y}$ in der Berechnungsformel für R^2), handelt es sich bei diesem Anteil der erklärten Varianz von y ebenso um eine χ^2-verteilte Zufallsvariable wie bei dem Anteil der unerklärten Varianz $1 - R^2$. Somit läßt sich eine F-verteilte Zufallsvariable mit $df_1 = k - 1$ und $df_2 = n - k$ Freiheitsgraden konstruieren:

$$
\begin{aligned}
F &= \frac{\textit{Anteil erklärter Varianz von y}}{\textit{Anteil nicht erklärter Varianz von y}} \\
&= \frac{R^2/(k-1)}{(1-R^2)/(n-k)} \\
&= \frac{R^2(n-k)}{(1-R^2)(k-1)}
\end{aligned}
$$

[1] Der Test von R^2 erübrigt sich im bivariaten Fall, da für standardisierte Variablen $b_1 = r = \sqrt{R^2}$ gilt.

Bei dem hier durchgeführten Test handelt es sich um einen einseitigen Test, da — wie aus der graphischen Darstellung der F-Verteilung ersichtlich — F für alle Werte der Zufallsvariablen größer ist als 0.

Durchführung des Hypothesentests

1. Schritt: Hypothesenformulierung

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_m = 0$$
$$H_A : \beta_1 \neq 0 \vee \beta_2 \neq 0 \vee \ldots \vee \beta_m \neq 0$$

2. Schritt: Bestimmung einer Prüfverteilung

$$F = \frac{R^2(n-k)}{(1-R^2)(k-1)}$$

3. Schritt: Bestimmung der Freiheitsgrade

$df_1 = k - 1$
$df_2 = n - k$

4. Schritt: Bestimmung des Signifikanzniveaus α

5. Schritt: Ermittlung des theoretischen Verteilungswertes F und Berechnung des empirischen Verteilungswertes F_{emp}

Während der theoretische Verteilungswert F in Abhängigkeit von α, df_1 und df_2 aus Tabellen ermittelt werden kann, läßt sich F_{emp} nach obiger Formel berechnen.

6. Schritt: Entscheidung über die Ablehnung von H_0

Es gilt: $P(F_{emp} < F) = 1 - \alpha$

$F_{emp} < F$: Der empirische F-Wert befindet sich im gewählten Vertrauensbereich $(1 - \alpha)$, d.h. die Informationen reichen *nicht* aus, um H_0 zu verwerfen.

$F_{emp} \geq F$: Der empirische F-Wert befindet sich nicht im Bereich der gewählten Irrtumswahrscheinlichkeit, d.h. es besteht nur noch eine Wahrscheinlichkeit von α, daß der F-Wert größer als der Prüfwert ist und somit H_0 fälschlicherweise verworfen wird.

Interpretation: Bei Ablehnung der Nullhypothese hat das geschätzte Regressionsmodell einen von Null verschiedenen signifikanten Varianzerklärungsanteil.

Bemerkung: Im Rahmen einer multiplen Regressionsanalyse mit Hilfe von SPSS werden die t-Werte und der F-Wert zusammen mit den daraus resultierenden Signifikanzniveaus (`Sig T` bzw. `Signif F`) standardmäßig ausgegeben. So bedeutet ein Signifikanzniveau `Signif F` = 0.023, daß die gesamte Regressionsschätzung mit einer Fehlerwahrscheinlichkeit von 2.3% signifikant ist. Ein vorab festgelegtes Niveau von 5% würde danach zur Ablehnung der Nullhypothese führen.

In dem ausführlichen Beispiel auf Seite 132 wird ein Signifikanzniveau von `Signif F` = 0.0000 ermittelt, d.h. die für die Regressionsschätzung geltende Fehlerwahrscheinlichkeit liegt sehr nahe bei Null.

6.2.5 Annahmenüberprüfung im klassischen, linearen Regressionsmodell

Wie in den vorangegangenen Kapiteln gezeigt, ist die Anwendbarkeit von Verfahren der Regressionsanalyse an bestimmte Annahmen geknüpft (z.B. Homoskedastizität). Das Nichterfülltsein dieser Voraussetzungen verhindert nicht zwangsläufig die Anwendung dieser Verfahren, es kann aber dazu führen, daß

- ein geeigneteres Schätzverfahren anzuwenden ist (z.B. 2KQ-Schätzung, Aitken-Schätzung) bzw.
- Interpretationen nur in eingeschränktem Maße zulässig sind.

Die Kenntnis dieser Annahmen *und* ihre Überprüfung sind daher notwendiger Bestandteil einer konkreten Regressionsanalyse.

Ausgehend von der allgemeinen Regressionsgleichung

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e}$$

werden daher im folgenden

- die wesentlichen Annahmen des klassischen, linearen Regressionsmodells noch einmal zusammengestellt und

- Möglichkeiten zu ihrer Überprüfung und geeignete Abhilfmaßnahmen bei Nichterfülltsein aufgezeigt.

Dabei ist zu berücksichtigen, daß die geschilderten Überprüfungsmöglichkeiten einen erheblichen Interpretationsspielraum für den Anwender beinhalten, da es sich überwiegend um die visuelle Beurteilung von Graphiken (z.B. Verteilung der Residuen) handelt.

In der folgenden Tabelle sind die zu überprüfenden Annahmen nach den einzelnen Elementen der Regressionsgleichung aufgeschlüsselt:

allgemein	$\boldsymbol{y}$	$\boldsymbol{X}$	$\boldsymbol{b}$	$\boldsymbol{e}$
linearer Zusammenhang zwischen y und $x_j, (j = 1, \ldots, m)$	Normalverteilung	lineare Unabhängigkeit der x_j	konstante Regressionskoeffizienten b_j, $(j = 1, \ldots, m)$, für alle Untersuchungseinheiten	Normalverteilung; $E(\boldsymbol{e}) = 0$; konstante Varianz (Homoskedastizität); keine Autokorrelation

Die Überprüfung, ob die Regressionskoeffizienten $\boldsymbol{b}$ für alle Beobachtungen konstant sind, unterbleibt, da diese Annahme für Querschnittsanalysen, d.h. für Datenerhebungen zu einem Zeitpunkt, im allgemeinen plausibel ist. Für Längsschnittuntersuchungen (z.B. Zeitreihen) dagegen kommt diesem Aspekt Bedeutung zu, da die Gewichtung einer unabhängigen Variablen, ausgedrückt durch den jeweiligen Regressionskoeffizienten, sich durch unerwartete Ereignisse im Zeitverlauf verändern kann (z.B. die Sympathie für einen Minister aufgrund eines von ihm durchgesetzten Gesetzesvorhabens).

Die in den folgenden Kapiteln dargestellten Ergebnisse eines SPSS-Laufes beziehen sich alle auf das Regressionsbeispiel auf Seite 132ff, bei dem der Einfluß der Sympathie zu einzelnen, führenden Politikern der Regierungsparteien CDU, CSU und FDP (Bangemann, Genscher, Lambsdorff, Kohl, Blüm, Stoltenberg, Strauß) auf die Sympathie zur Regierung insgesamt untersucht wurde.

1. Linearität

Der lineare Zusammenhang zwischen y und den x_j wird anhand der Form der in *Streuungsdiagrammen* (SPSS: `REGRESSION .../SCATTERPLOT(...)`)

dargestellten Punktwolke überprüft. Dies ist im bivariaten Fall auf direktem Weg möglich:

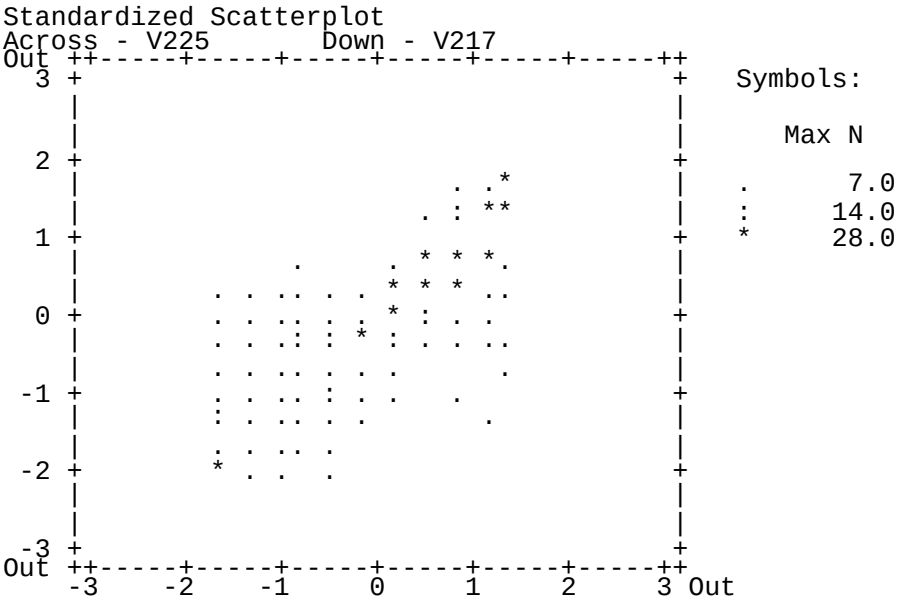


Abb. 6.1: Sympathie Regierung (V217) ↔ Sympathie Kohl (V225)

Im multiplen Fall erlaubt die Untersuchung des Streuungsdiagramms zwischen $\boldsymbol{y}$ und den geschätzten Werten auf der "Regressionsgerade" $\hat{\boldsymbol{y}}$ eine Aussage über die Linearität. Dem liegt folgende Überlegung zugrunde:

$$\begin{aligned} \boldsymbol{e} &= \boldsymbol{y} - \hat{\boldsymbol{y}} \\ &= \boldsymbol{y} - \boldsymbol{X}\boldsymbol{b} \end{aligned}$$

Im Fall $e = 0$, d.h. wenn die Anpassung der empirischen Werte von y an die Werte auf der "Regressionsgeraden" $\hat{y}$ optimal ist, gilt:

$$y = \hat{y}$$

Für das Streuungsdiagramm mit diesen beiden Variablen bedeutet dies, daß um so eher die Linearitätsannahme gerechtfertigt ist, je näher die Punkte an der Winkelhalbierenden ($y = \hat{y}$) liegen:

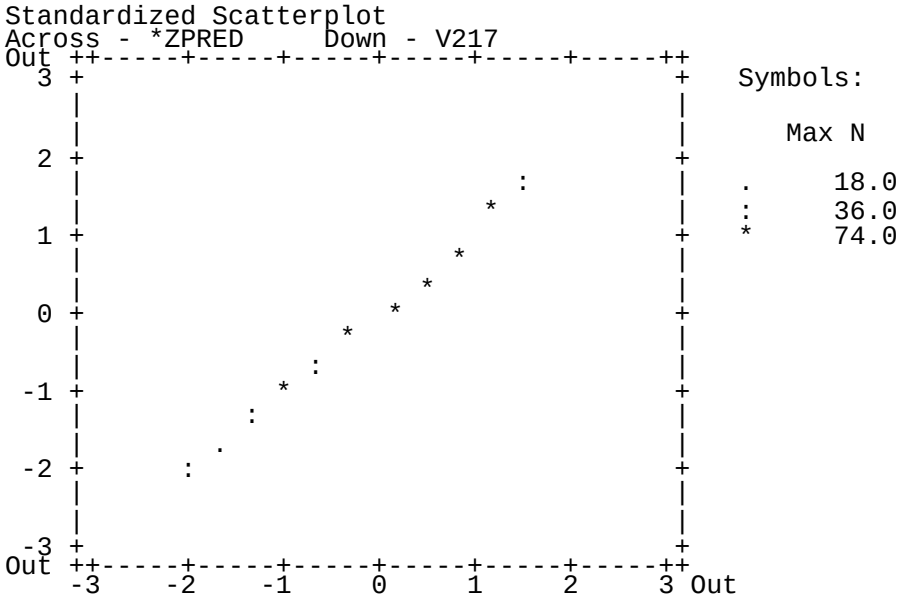
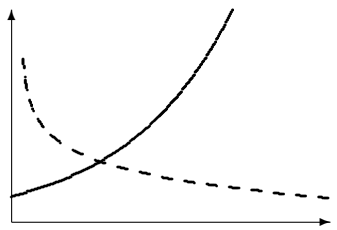


Abb. 6.2: Sympathie Regierung (V217) ↔ geschätzte Sympathie Regierung (*ZPRED)

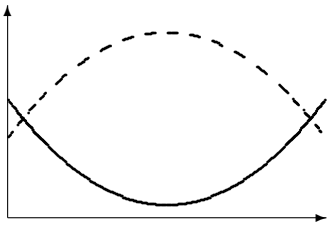
Wurde zuvor ein nichtlinearer, mathematischer Zusammenhang eindeutig identifiziert (z.B. Exponentialfunktion), besteht in bestimmten Fällen die Möglichkeit einer *Transformation* auf einen linearen Ansatz. Dabei ist zu berücksichtigen, daß durch eine Transformation Eigenschaften verlorengehen bzw. neu entstehen können (z.B. $x_{ij} > 0$ bei einer logarithmischen Transformation), die zusätzlich überprüft werden müssen.

Beispiele:

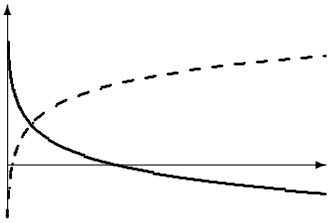
1. $y = ae^{bx}e'$
   $\Longrightarrow$
   $y' = a' + bx + e''$
   mit $y' = \ln y$
   $a' = \ln a$
   $e'' = \ln e'$



2. $y = a + b_1 x_1 + b_2 x_1^2 + e'$
   $\Longrightarrow$
   $y' = a + b_1 x_1 + b_2 x_2 + e'$
   mit $x_2 = x_1^2$



3. $y = a + b \log x + e'$
   $\Longrightarrow$
   $y' = a + b\, x' + e'$
   mit $x' = \log x$



4. $y = a + b\, \frac{1}{x} + e'$
   $\Longrightarrow$
   $y' = a + b\, x' + e'$
   mit $x' = \frac{1}{x}$

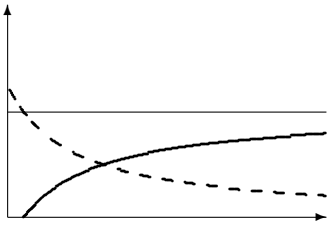


Abb. 6.3: Graphische Bedeutung verschiedener Transformationen

2. Annahmen bezüglich der Residuen

Die meisten Annahmen werden bezüglich der Residuen e getroffen. Insofern ist die Untersuchung der Residuen (= Residuenanalyse) ein Schwerpunkt bei der Annahmenüberprüfung (SPSS: `REGRESSION .../RESIDUALS`).

Der Mittelwert der Residuen ist dabei über die Residuenstatistik ermittelbar:

```
Equation Number 1    Dependent Variable..   V225    SKALOMETER: H  KOHL
Residuals Statistics:
                   Min           Max     Mean  Std Dev    N

*PRED            .8515       10.6278   6.4489   2.6989  499
*RESID         -6.6949        7.1932    .0000   1.7470  499
*ZPRED         -2.0739        1.5484    .0000   1.0000  499
*ZRESID        -3.8284        4.1133    .0000    .9990  499

Total Cases =      500
```

Die Überprüfung auf Normalverteilung der Residuen ist auf zwei verschiedene Arten möglich:

1. Ausgabe eines Residuenhistogramms (mit eingeblendeter Normalverteilung)

   (SPSS: `REGRESSION .../RESIDUALS HISTOGRAM(*ZRESID)`)

```
Histogram - Standardized Residual

 N Exp N      (* = 2 Cases,    . : = Normal Curve)
 2   .38   Out *
 4   .76  3.00 **
 0  1.95  2.67 .
 4  4.45  2.33 *:
 7  9.10  2.00 ****.
 4 16.68  1.67 **     .
22 27.38  1.33 **********  .
 0 40.25  1.00                    .
97 53.00   .67 *************************:***********************
94 62.51   .33 ******************************:****************
85 66.05   .00 *********************************:**********
86 62.51  -.33 *********************************:***********
 0 53.00  -.67                          .
49 40.25 -1.00 ******************:*****
 0 27.38 -1.33              .
20 16.68 -1.67 *******:**
13  9.10 -2.00 ****:**
 0  4.45 -2.33  .
 8  1.95 -2.67 :***
 0   .76 -3.00
 4   .38   Out **
```

2. Darstellung der Wahrscheinlichkeitsverteilung für die Residuen (in der SPSS-Grafik: `Observed`) zusammen mit der für die Normalverteilung (graphisch: `Expected`).

   (SPSS: `REGRESSION .../RESIDUALS NORMPROB(*ZRESID)`)

   Je näher dabei die Punkte an der Winkelhalbierenden liegen, um so besser ist die Annäherung der Residuen an die Normalverteilung. Aufgrund des funktionalen Zusammenhangs zwischen e und y in der Regressionsgleichung folgt aus der Normalverteilung von e auch die Normalverteilung von y.

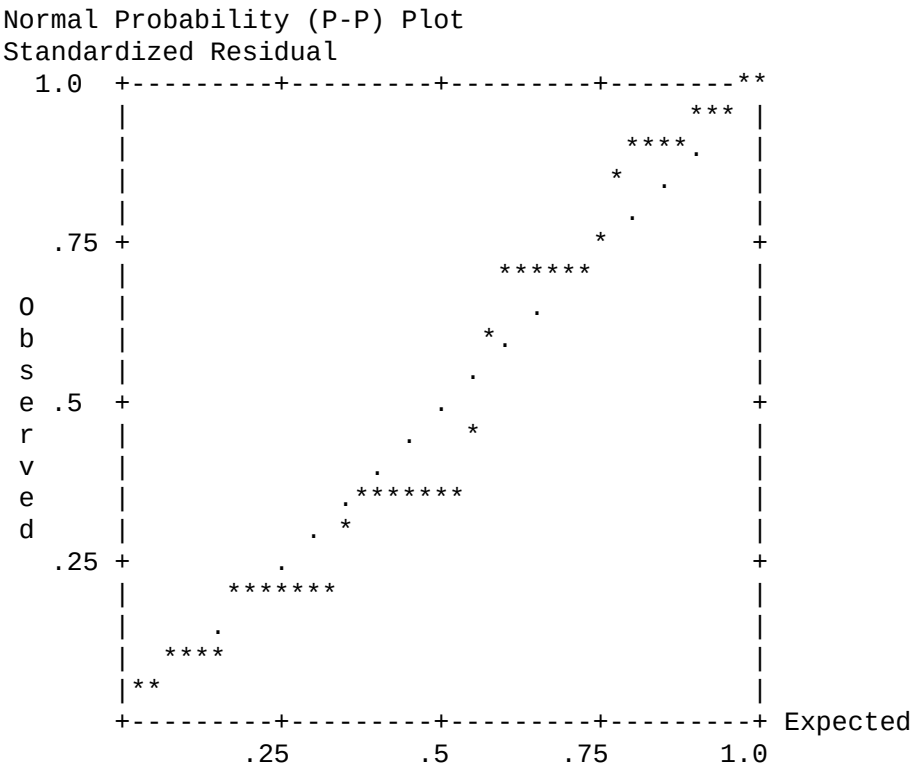


Abb. 6.4: Streudiagramm von beobachteter und erwarteter Wahrscheinlichkeitsverteilung

Bei der Überprüfung, ob die Varianz der Residuen konstant ist (Homoskedastizität), ergibt sich das Problem, daß lediglich eine Stichprobe vorliegt. Hinweise darauf, daß die Varianz bei mehreren Stichproben konstant bleiben könnte, ergeben sich durch die Untersuchung der Streuungsdiagramme zwischen dem Residuum und der unabhängigen Variablen oder dem Residuum und der geschätzten abhängigen Variablen (Abbildung 6.5)

(SPSS: `REGRESSION .../RESIDUALS /SCATTERPLOT(*RESID,V225)` bzw. `/RESIDUALS /SCATTERPLOT(*RESID,*PRED)`).

Falls die Spannweite der Residuen mit ansteigenden (fallenden) Werten der jeweils anderen Variablen ebenfalls ansteigt (fällt), weist dies darauf hin, daß die Variabilität bei kleinen (großen) Werten der abhängigen Variablen kleiner (größer) ist als bei großen (kleinen).

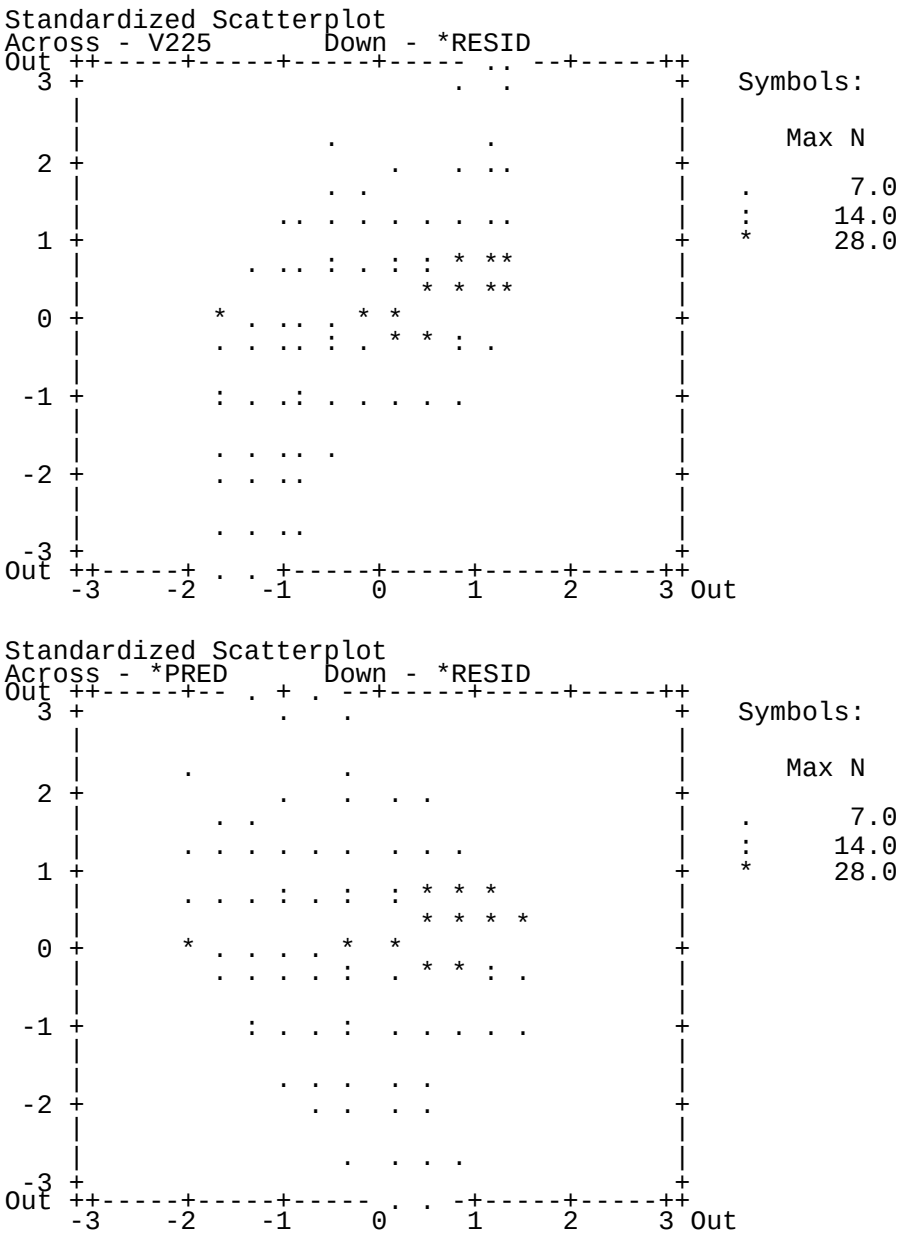


Abb. 6.5: Streudiagramm von Residuum und unabhängiger/geschätzter Variablen

Im obigen Beispiel ist dies nicht erkennbar, sondern die Punktwolke entspricht eher einem gleichmäßigen horizontalen Band. Ein typisches Punktwolkenmuster, das auf mögliche Homoskedastizität hinweist, zeigt das konstruierte Beispiel in Abbildung 6.6, in dem die Residuen mit steigender unabhängiger Variable einen Trichter bilden.

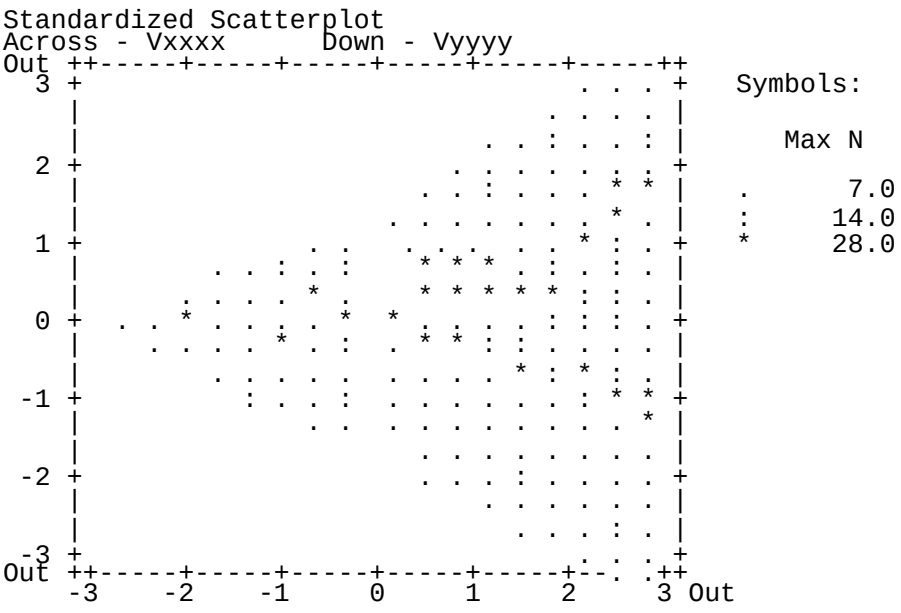


Abb. 6.6: Streudiagramm möglicher Homoskedastizität

Bei der Untersuchung der Residuen auf Autokorrelation besteht die Möglichkeit einer fallweisen Ausgabe der Residuen.

(SPSS: `REGRESSION .../CASEWISE OUTLIERS(3) PLOT(*ZRESID)`)

```
Casewise Plot of Standardized Residual

Outliers = 3.    *: Selected   M: Missing

          -6.     -3.  3.     6.
  Case #   O:.......:  :.......:O   V225       *PRED      *RESID
       5   .         *..          .      1      6.7173     -5.7173
      98   .       *  ..          .      1      7.6949     -6.6949
     194   .         *..          .      2      7.6949     -5.6949
     287   .          ..*         .     11      5.7397      5.2603
     323   .          .. *        .     11      4.7620      6.2380
     336   .         *..          .      1      6.7173     -5.7173
     376   .          ..   *      .     10      2.8068      7.1932
     418   .          ..*         .     11      5.7397      5.2603

       8 Outliers found.
```

```
Outliers - Standardized Residual

  Case #       *ZRESID

     376       4.11329
      98      -3.82835
     323       3.56704
     336      -3.26932
       5      -3.26932
     194      -3.25653
     418       3.00800
     287       3.00800
     190       2.98242
     150       2.98242
```

Ein erkennbares Muster bei der Ausgabe weist dabei auf eine gegenseitige Abhängigkeit von Residuenwerten hin. Die Möglichkeit der fallweisen Ausgabe von Residuen eignet sich sinnvollerweise nur für relativ kleine Stichproben. Ebenso ist darauf hinzuweisen, daß bei Querschnittsbefragungen die Erfassungsreihenfolge keine Rolle spielt und somit das Nichterkennen eines Musters Autokorrelation nicht automatisch ausschließt[2].

Im Zusammenhang mit der Residuenanalyse sind einige generelle Bemerkungen zur Behandlung von "Ausreißern" angebracht:

Als Ausreißer werden Fälle bezeichnet, die (relativ) weit von der Regressionsgeraden entfernt liegen (z.B. mehr als zwei Standardabweichungen), also durch diese schlecht geschätzt werden und die Güte der Schätzung insgesamt herabsetzen. Mit Hilfe einer Residuenanalyse (vgl. obiges Beispiel) lassen sich Ausreißer identifizieren, um beispielsweise auf der Ebene des einzelnen Datensatzes nach Ursachen zu suchen (z.B. Kodierfehler, generelles extremes Antwortverhalten) oder auch den relativen Einfluß dieser Fälle abschätzen zu können (z.B. Regressionsschätzung unter Ausschluß der Ausreißer).

3. Kollinearität zwischen Regressoren

Die Forderung nach linearer Unabhängigkeit zwischen den einzelnen erklärenden Variablen x_j läßt sich direkt aus dem linearen Regressionsmodell ableiten. Sei dazu beispielsweise

$$y = b_0 + b_1 x_1 + b_2 x_2 + e, \qquad \text{mit} \qquad x_2 = a x_1,$$

[2]Für die Untersuchung der Autokorrelation von Zeitreihen ist zusätzlich der Durbin-Watson-Test anwendbar (SPSS: `REGRESSION .../RESIDUALS DURBIN`).

d.h. x_2 ist linear von x_1 abhängig. Dann gilt:

$$\begin{aligned} y &= b_0 + b_1 x_1 + b_2(a x_1) + e \\ &= b_0 + (b_1 + b_2 a) x_1 + e \end{aligned}$$

Im Rahmen der Regressionsanalyse werden die Koeffizienten b_1 und b_2 geschätzt. Die Ermittlung des "wahren" Regressionskoeffizienten $(b_1 + b_2 a)$ ist nicht möglich.

Für das Vorliegen von Kollinearitäten zwischen einzelnen Variablen gibt es eine Reihe von Hinweisen:

- Hohe Korrelationen zwischen einzelnen x_j in der Korrelationsmatrix.
- Hoher Standardfehler für b_j (für nicht signifikante Variablen, bei gleichzeitiger signifikanter Gesamtschätzung), da dieser Wert direkt von der Kovarianz und damit auch von der Korrelation von x_j zu anderen Regressoren abhängt.
- Große Schwankungen der Werte von b_j, wenn Variablen aus dem Modell entfernt werden.

Als mögliche Abhilfemaßnahmen bei vorliegender Kollinearität sind zu nennen:

- Entfernung linear abhängiger Variablen aus der Regressionsgleichung. Je nach Ausmaß der Kollinearität kann es hierbei allerdings zu einem Informationsverlust kommen.
- Abschätzen der Stärke der Kollinearität mittels Durchrechnen verschiedener Variablenkombinationen.
- Vor der Durchführung einer multiplen Regressionsanalyse Anwendung von Verfahren, die linear voneinander abhängige Variablen "zusammenfassen" (vgl. dazu Kapitel 6.3).

6.3 Faktorenanalyse

6.3.1 Problemstellung

Im Rahmen der multiplen Regressionsanalyse führt die Kollinearität von Regressoren, d.h. lineare Abhängigkeiten zwischen einzelnen unabhängigen Variablen, zu Verzerrungen bei der Schätzung von Regressionskoeffizienten. Große lineare Abhängigkeiten zwischen Variablen lassen sich auch so interpretieren, daß die entsprechenden Variablen letztendlich *denselben* Sachverhalt messen. Hieraus ergeben sich folgende Ausgangsfragen:

- Lassen sich derartig "zusammengehörige" Variablen auf geeignete Art und Weise zu einem neuen "Konstrukt" zusammenfassen?
- Was messen diese Variablen dann schließlich?

Beispiel: Sympathieskalometer für Parteien

```
* SPSS-Aufruf: FACTOR      /VARIABLES V212 TO V216
*                          /PRINT UNIVARIATE CORRELATION EXTRACTION.

                Mean     Std Dev   Label

V212         7.42222     2.55892    SKALOMETER: SPD
V213          .97172     2.88708    SKALOMETER: CDU
V214         6.17778     3.13087    SKALOMETER: CSU
V215         6.11313     2.35411    SKALOMETER: FDP
V216         4.84848     2.94746    SKALOMETER: DIE GRUENEN

Number of Cases  =     495

Correlation Matrix:

             V212     V213     V214     V215     V216

V212      1.00000
V213      -.43295  1.00000
V214      -.44802   .83365  1.00000
V215      -.12926   .49995   .46390  1.00000
V216       .42477  -.45986  -.44019  -.17286  1.00000
```

Als ein Maß für die Beurteilung von linearer Abhängigkeit zwischen zwei Variablen kann der Produkt-Moment-Korrelationskoeffizient r angesehen werden.

Die Betrachtung einer Korrelationsmatrix gibt daher erste Anhaltspunkte für eine mögliche Zusammenfassung von Variablen.

So lassen sich anhand der Korrelationskoeffizienten Gruppen von Parteien zusammenfassen ((CDU, CSU, FDP) bzw. (SPD, Die Grünen)), die innerhalb einer Gruppe deutlich positiv, zu Mitgliedern der anderen Gruppe aber negativ korreliert sind. Dies läßt vermuten, daß die Wähler das Parteiensystem weitgehend *eindimensional* mit den beiden Polen *Regierung* und *Opposition* sehen. Eine ähnliche Konstellation ergibt sich auf der Ebene der Sympathieskalometer von Politikern, bei denen Regierungspolitiker bzw. Oppositionspolitiker untereinander eine hohe positive Korrelation aufweisen, während sie zu Politikern der jeweils anderen Gruppe negativ korreliert sind.

```
* SPSS-Aufruf: FACTOR      /VARIABLES V222 TO V231
*                          /PRINT UNIVARIATE CORRELATION EXTRACTION.
                 Mean     Std Dev   Label

V222          5.63333     2.40019    SKALOMETER: M  BANGEMANN
V223          6.41042     2.74457    SKALOMETER: W  BRANDT
V224          6.88333     2.34742    SKALOMETER:H D  GENSCHER
V225          6.43542     3.20620    SKALOMETER: H  KOHL
V226          4.90625     2.95364    SKALOMETER: O  SCHILY
V227          7.26875     2.75375    SKALOMETER: J  RAU
V228          7.38750     2.24302    SKALOMETER:G STOLTENBERG
V229           .39792     3.23956    SKALOMETER: F J  STRAUSS
V230          6.40417     2.48233    SKALOMETER: H J  VOGEL
V231          5.37708     2.53030    SKALOMETER: O LAMBSDORFF

Number of Cases  =     480
Correlation Matrix:

             V222     V223     V224     V225     V226     V227     V228

V222       1.00000
V223       -.15363  1.00000
V224        .58340  -.19767  1.00000
V225        .55007  -.50267   .45086  1.00000
V226       -.24693   .46085  -.19429  -.42843  1.00000
V227       -.15973   .64832  -.15500  -.44103   .32677  1.00000
V228        .50342  -.37925   .52405   .66042  -.33231  -.30723  1.00000
V229        .46396  -.41968   .37289   .72054  -.34628  -.36889   .62661
V230       -.08720   .58141  -.07322  -.26663   .31982   .70087  -.24753
V231        .60960  -.26734   .53675   .59347  -.23019  -.26805   .51566

             V229     V230     V231

V229       1.00000
```

```
V230      -.23526  1.00000
V231       .53560  -.12702  1.00000
```

Unabhängig davon, daß die Betrachtung von Korrelationskoeffizienten nur einen ersten Anhaltspunkt für die mögliche Strukturierung der vorliegenden Variablen geben kann, zeigt dieses zweite Beispiel auch, daß die Korrelationsmatrix für eine größere Anzahl von Variablen — und gerade dann machen Strukturierungen Sinn — schnell unübersichtlich wird.

Mit der Faktorenanalyse steht ein multivariates Verfahren zur Verfügung, das eine systematische Entdeckung von nicht direkt meßbaren Variablen aus einer Menge von gegebenen Variablen — auf der Basis einer Korrelationsmatrix — durchführt. Die Aufgabenstellung der Faktorenanalyse läßt sich wie folgt präzisieren:

Gegeben: mehrere *normalverteilte, metrisch skalierte, untereinander unkorrelierte Variablen* x_i (Observablen, Indikatoren)

Gesucht: eine *geringere* Anzahl *normalverteilter, metrisch skalierter, nicht unmittelbar beobachtbarer und in der Regel unkorrelierter Variablen* f_k (Faktoren), mit deren Hilfe sich die Indikatoren einfacher beschreiben lassen.

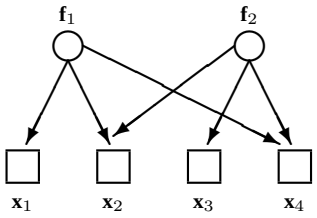


Abb. 6.7: Zuordnung von Faktoren zu Variablen

Ein weiteres Ziel der Faktorenanalyse ist daher neben einer Variablenstrukturierung die *Datenreduktion*.

Grundsätzlich lassen sich bei der Durchführung von Faktorenanalysen zwei Vorgehensweisen unterscheiden:

- *Konfirmatorische Faktorenanalyse:*
  Überprüfung *theoretisch begründeter* Annahmen, daß bestimmte Variablen hoch mit bestimmten Faktoren korreliert sind.

- *Exploratorische Faktorenanalyse:*
  Untersuchung und Interpretation der sich aus den Variablen ergebenden Faktorenmuster.

In der Praxis wird es dabei eher zu Mischformen kommen, indem auf der Grundlage von theoretischen Vorüberlegungen mit den Ausgangsvariablen experimentiert wird.

Die Faktorenanalyse bezeichnet kein einzelnes, vollständig festgelegtes Berechnungsverfahren, sondern sie umfaßt innerhalb einzelner Berechnungsstufen jeweils eine Vielzahl von Varianten, deren konkrete Auswahl in der Veranwortung eines Anwenders liegt.[3] Die nachfolgenden Kapitel spiegeln in ihrer Abfolge den grundsätzlichen Ablauf einer Faktorenanalyse unter Einbeziehung der notwendigen Anwenderentscheidungen wider.

6.3.2 Das Fundamentaltheorem der Faktorenanalyse

Die Faktorenanalyse geht von der Grundannahme aus, daß zwischen Indikatorvariablen und den nichtmeßbaren Faktoren ein *linearer Zusammenhang* besteht, d.h.

$$x_{ij} = f_{i1}a_{1j} + \ldots + f_{ik}a_{kj} + \ldots + f_{ir}a_{rj}$$

mit:

x_{ij} : Variablenwert des Indikators x_j $(j = 1, \ldots, m)$ in der i-ten Beobachtung $(i = 1, \ldots, n)$

f_{ik} : Wert des Faktors f_k $(k = 1, \ldots, r)$ in der i-ten Beobachtung

a_{kj} : Wert des Gewichtungskoeffizienten zwischen dem Indikator x_j und dem Faktor f_k

Für die weiteren Berechnungsschritte wird von standardisierten Indikatorvariablen z_{ij} ausgegangen, d.h. $\bar{z}_j = 0$ und $s^2_{z_j} = 1$. Für die Ausgangsgleichung der Faktorenanalyse gilt somit:

$$z_{ij} = f_{i1}a_{1j} + \ldots + f_{ik}a_{kj} + \ldots + f_{ir}a_{rj}$$

bzw. für alle Beobachtungen in Matrizenschreibweise

$$\boldsymbol{Z} = \boldsymbol{F}\boldsymbol{A}$$

[3] Die Tatsache, daß statistische Analysesysteme für die Durchführung einer Faktorenanalyse lediglich die Angabe der zu verwendenden Variablen benötigen, bedeutet nur, daß ohne Eingreifen des Anwenders eine Standardauswahl für die Analyse zugrunde gelegt wird.

mit:

$\boldsymbol{Z}$: $(n \times m)$–Matrix der standardisierten Indikatorenwerte
$\boldsymbol{F}$: $(n \times r)$–Matrix der Faktorwerte
$\boldsymbol{A}$: $(r \times m)$–Matrix der Gewichtungskoeffizienten (Faktorladungsmatrix) (in statistischen Analysesystemen wie SPSS wird aus technischen Gründen die $(m \times r)$–Matrix $\boldsymbol{A}^T$ ausgedruckt)

Weiterhin gilt:[4]

$$r_{x_j,x_{j^*}} = \frac{\text{cov}(x_j, x_{j^*})}{\sqrt{s^2_{x_j}}\sqrt{s^2_{x_{j^*}}}}$$

bzw.

$$\begin{aligned} r_{z_j,z_{j^*}} &= \text{cov}(z_j, z_{j^*}) \\ &= \frac{1}{n}\sum_{i=1}^{n}(z_{ij} - \bar{z}_j)(z_{ij^*} - \bar{z}_{j^*}) \\ &= \frac{1}{n}\sum_{i=1}^{n} z_{ij} z_{ij^*} \end{aligned}$$

Daraus ergibt sich für die Korrelationsmatrix aller standardisierten Indikatorvariablen z_j $(j = 1, \ldots, m)$:

$$\boldsymbol{R} = \frac{1}{n}\boldsymbol{Z}^T\boldsymbol{Z}$$

Zusätzlich soll gelten:

1. alle Faktoren f_k sind standardisiert, d.h.

$$\begin{aligned} \bar{f}_k &= 0 \qquad \text{und} \\ s^2_{f_k} &= \frac{1}{n}\sum_{i=1}^{n}(f_{ik} - \bar{f}_k)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} f^2_{ik} \\ &= 1 \qquad (k = 1, \ldots, r) \end{aligned}$$

2. alle Faktoren f_k sind untereinander unkorreliert (orthogonal), d.h.

$$\frac{1}{n}\sum_{i=1}^{n} f_{ik} f_{ik^*} = 0 \qquad (k = 1, \ldots, r)$$

[4]Der Index "j^*" kennzeichnet nur eine andere Indexausprägung als das "ungesternte" j, $(j, j^* = 1, \ldots, m)$.

Daraus folgt für die Korrelationsmatrix der standardisierten Faktoren f_k ($k = 1, \ldots, r$):

$$\frac{1}{n}\boldsymbol{F}^T\boldsymbol{F} = \boldsymbol{E_n}$$

Unter Verwendung der Orthogonalität führt die Ausgangsgleichung der Faktorenanalyse zu dem *Fundamentaltheorem der Faktorenanalyse* :

$$\begin{aligned}
\boldsymbol{R} &= \frac{1}{n}\boldsymbol{Z}^T\boldsymbol{Z} \\
&= \frac{1}{n}(\boldsymbol{F}\boldsymbol{A})^T(\boldsymbol{F}\boldsymbol{A}) \\
&= \frac{1}{n}\boldsymbol{A}^T\boldsymbol{F}^T\boldsymbol{F}\boldsymbol{A} \\
&= \boldsymbol{A}^T(\frac{1}{n}\boldsymbol{F}^T\boldsymbol{F})\boldsymbol{A} \\
&= \boldsymbol{A}^T\boldsymbol{E_n}\boldsymbol{A} \\
&= \boldsymbol{A}^T\boldsymbol{A}
\end{aligned}$$

Dies bedeutet, daß sich im Falle unkorrelierter Faktoren die Korrelationsmatrix der Indikatoren aus der Faktorladungsmatrix ($\boldsymbol{A}$) reproduzieren läßt.

Bemerkungen:

- Die Matrix der Faktorladungen (*factor loadings*) $\boldsymbol{A}$ wird auch als *Faktormuster* (*factor pattern*) bezeichnet.
- Die Reproduzierbarkeit der Korrelationsmatrix aus der Faktorladungsmatrix stützt den Untersuchungsansatz, mit Hilfe der Korrelationskoeffizienten etwas über die Existenz von nicht direkt meßbaren, theoretischen Variablen auszusagen.

Bezogen auf die hergeleiteten Gleichungen und ausgehend von der Grundgleichung $\boldsymbol{Z} = \boldsymbol{F}\boldsymbol{A}$ stellt sich für die Durchführung einer Faktorenanalyse folgende allgemeine Aufgabe:

> Bestimmung der Abhängigkeit der Indikatorvariablen z_j von den vermuteten theoretischen Variablen f_k durch Berechnung der Faktorladungsmatrix $\boldsymbol{A}$ unter Zuhilfenahme des Fundamentaltheorems $\boldsymbol{R} = \boldsymbol{A}^T\boldsymbol{A}$.

Nachfolgende SPSS-Ausgaben zeigen die Faktorladungsmatrizen für die beiden bereits vorgestellten Beispiele. Während im ersten Fall lediglich ein Faktor ermittelt wird, extrahiert das in diesem Fall angewandte Verfahren der Hauptkomponentenanalyse (vgl. Kapitel 6.3.4) im zweiten Beispiel die Ladungskoeffizienten für zwei Faktoren.

1. Sympathieskalometer für Parteien

```
* SPSS-Aufruf: FACTOR      /VARIABLES V212 TO V216
*                          /PRINT UNIVARIATE CORRELATION EXTRACTION.

              - - - -   F A C T O R   A N A L Y S I S   - - - -

Analysis Number  1  Listwise deletion of cases with missing values

                Mean     Std Dev   Label

V212         7.42222     2.55892    SKALOMETER: SPD
V213          .97172     2.88708    SKALOMETER: CDU
V214         6.17778     3.13087    SKALOMETER: CSU
V215         6.11313     2.35411    SKALOMETER: FDP
V216         4.84848     2.94746    SKALOMETER: DIE GRUENEN

Number of Cases  =     495

Extraction  1  for Analysis  1, Principal-Components Analysis (PC)
   PC Extracted   1 factors.

Factor Matrix:

              FACTOR  1

V212           -.63801
V213            .90098
V214            .89162
V215            .59067
V216           -.65739
```

2. Sympathieskalometer für Politiker

```
* SPSS-Aufruf: FACTOR      /VARIABLES V222 TO V231
*                          /PRINT UNIVARIATE CORRELATION EXTRACTION.

              - - - -   F A C T O R   A N A L Y S I S   - - - -
```

```
                 Mean     Std Dev   Label

V222         5.63333     2.40019    SKALOMETER: M  BANGEMANN
V223         6.41042     2.74457    SKALOMETER: W  BRANDT
V224         6.88333     2.34742    SKALOMETER:H D  GENSCHER
V225         6.43542     3.20620    SKALOMETER: H  KOHL
V226         4.90625     2.95364    SKALOMETER: O  SCHILY
V227         7.26875     2.75375    SKALOMETER: J  RAU
V228         7.38750     2.24302    SKALOMETER:G STOLTENBERG
V229          .39792     3.23956    SKALOMETER: F J  STRAUSS
V230         6.40417     2.48233    SKALOMETER: H J  VOGEL
V231         5.37708     2.53030    SKALOMETER: O LAMBSDORFF

Number of Cases  =     480

Extraction  1  for Analysis  1, Principal-Components Analysis (PC)
   PC Extracted   2 factors.

Factor Matrix:

              FACTOR  1     FACTOR  2

V222            .65929        .48454
V223           -.66284        .54161
V224            .61148        .45474
V225            .85626        .10544
V226           -.55227        .25383
V227           -.61964        .60310
V228            .77662        .20798
V229            .77988        .12126
V230           -.49355        .68987
V231            .71278        .37806
```

6.3.3 Das Problem der Kommunalitäten

Faktoren lassen sich in unterschiedliche Typen klassifizieren:

- *Allgemeiner Faktor (general factor):*
  Die Faktorladungen sind für *alle* Observablen hoch.
- *Gemeinsamer Faktor (common factor):*
  Die Faktorladungen sind für *mindestens zwei* Observablen hoch.
- *Einzelrestfaktor (unique factor):*
  Die Faktorladung für lediglich *eine* Observable ist hoch.

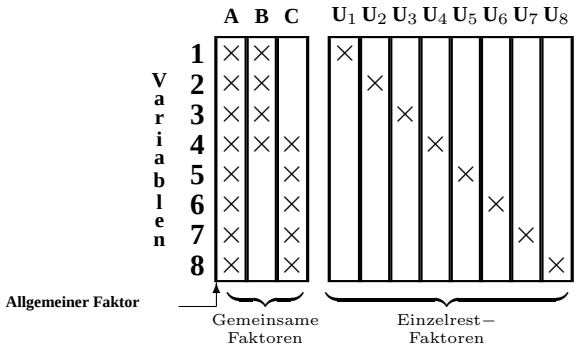


Abb. 6.8: Faktortypen (nach [Überla 1977, S. 55])

Im Rahmen der Faktorenanalyse sind insbesondere die gemeinsamen Faktoren — inklusive des Spezialfalls "allgemeiner Faktor"— von Interesse, da nur hierbei eine echte Datenreduktion erzielt werden kann. Bei der im nachfolgenden Beispiel aus den Sympathieskalometervariablen zu Politikern erzeugten Faktorladungsmatrix zeigt sich, daß beide extrahierte Faktoren gemeinsame Faktoren sind.

```
Factor Matrix:

               FACTOR  1      FACTOR  2

V222            .65929         .48454
V223           -.66284         .54161
V224            .61148         .45474
V225            .85626         .10544
V226           -.55227         .25383
V227           -.61964         .60310
V228            .77662         .20798
V229            .77988         .12126
V230           -.49355         .68987
V231            .71278         .37806
```

Die in der Ausgangsgleichung (vgl. S. 152) formulierte Beziehung zwischen Observablen und Faktoren erinnert an den Modellansatz der multiplen Regressionsanalyse, mit dem Unterschied, daß neben den Faktorladungen auch die Faktorwerte an dieser Stelle noch nicht bekannt sind. Ebenso wie in der multiplen Regression interessiert hier der Anteil der Varianz der Observablen, der durch die Faktoren erklärt wird.

Allgemein gilt für die Varianz der Variablen z_j:

$$
\begin{aligned}
s_{z_j}^2 &= \frac{1}{n}\sum_{i=1}^{n} z_{ij}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(\sum_{k=1}^{r}(f_{ik}a_{kj})^2) \\
&= \frac{1}{n}\sum_{i=1}^{n}(f_{i1}^2 a_{1j}^2 + \ldots + f_{ir}^2 a_{rj}^2 \\
&\qquad +2(f_{i1}a_{1j}f_{i2}a_{2j} + \ldots + f_{ir-1}a_{r-1j}f_{ir}a_{rj})) \\
&= a_{1j}^2 \underbrace{\frac{1}{n}\sum_{i=1}^{n} f_{i1}^2}_{=1} + \ldots + a_{rj}^2 \underbrace{\frac{1}{n}\sum_{i=1}^{n} f_{ir}^2}_{=1} \\
&\qquad +2(a_{1j}a_{2j}\underbrace{\frac{1}{n}\sum_{i=1}^{n} f_{i1}f_{i2}}_{=0} + \ldots + a_{r-1j}a_{rj}\underbrace{\frac{1}{n}\sum_{i=1}^{n} f_{ir-1}f_{ir}}_{=0}) \\
&= a_{1j}^2 + \ldots + a_{rj}^2 \\
&= 1
\end{aligned}
$$

Die Gesamtvarianz einer Observablen z_j ergibt sich somit aus der Summe der Quadrate *aller* Faktorladungen. Die durch die Faktoren maximal zu erklärende Varianz aller Observablen beträgt damit:

$$s^2 = \sum_{j=1}^{m} s_{z_j}^2 = \sum_{j=1}^{m} 1 = m$$

Wird davon ausgegangen, daß sich unter den Faktoren — realistischerweise — auch Einzelrestfaktoren befinden, die zur Erklärung der (gemeinsamen Varianz aller) Observablen nichts beitragen, so lassen sich die bisherigen Gleichungen für z_{ij} und s_{z_j} wie folgt umformen:

$$z_{ij} = \underbrace{f_{i1}a_{1j} + \ldots + f_{ir}a_{rj}}_{\text{gemeinsame Faktoren}} + \underbrace{f_{i(r+1)}a_{(r+1)j} + \ldots + f_{iq}a_{qj}}_{\text{Einzelrestfaktoren}}$$

und

$$s_{z_j}^2 = (a_{1j}^2 + \ldots + a_{rj}^2) + \underbrace{b_j^2 + e_j^2}_{u_j^2} = 1$$

mit:

b_j^2 : Varianz der Einzelrestfaktoren, die nichts mit den Observablen zu tun hat (*Spezifität*)

e_j^2 : Anteil der Varianz einer Observablen, der auf Meßfehler in der Observablen zurückgeht (*Restvarianz*)

u_j^2 : *Fehlervarianz*

Die Unterscheidung zwischen Spezifität und Restvarianz wird im folgenden vernachlässigt und allgemein von Fehlervarianz gesprochen.

Definition: Die *Kommunalität* einer Observablen z_j bezeichnet den Varianzanteil von z_j, der durch *gemeinsame* Faktoren erklärt wird, d.h. es gilt

$$\begin{aligned} h_j^2 &= a_{1j}^2 + \ldots + a_{rj}^2 \\ &= 1 - u_j^2 \end{aligned}$$

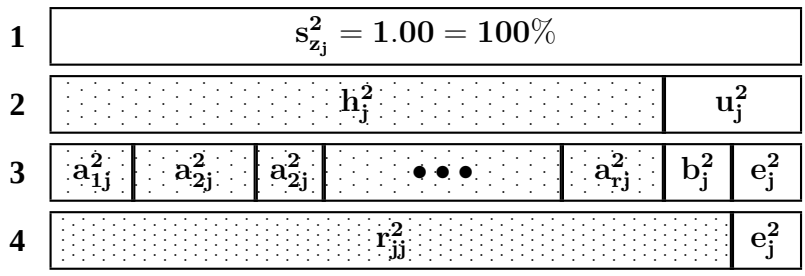


Abb. 6.9: Beziehungen zwischen Observablen, Kommunalität und Faktorladungen (nach [Überla 1977, S. 57])

Die Beziehungen zwischen Observablen, Kommunalität und Faktorladungen sind in Abbildung 6.9 noch einmal graphisch dargestellt. Dabei gilt zusätzlich:

r_{jj}^2 : Teil der Einheitsvarianz, der nicht auf Irrtum beruht, d.h der gemessen werden kann (*Reliabilität*).

Bemerkungen:

- Die Kommunalität h_j^2 ist — ähnlich wie das Bestimmtheitsmaß r^2 in der multiplen Regressionsanalyse — ein Maß für die Anpassungsgüte der Faktoren an die Observablen. So bedeutet zum Beispiel eine Kommunalität von 0.8, daß 80% der Gesamtvarianz einer Observablen z_j durch gemeinsame Faktoren erklärbar sind.

- $h_j^2 = 1$:
  Die gesamte Varianz einer Observablen wird von gemeinsamen Faktoren erklärt, d.h. es gibt keine Fehlervarianzen.

- $h_j^2 < 1$:
  Die Varianz einer Observablen wird nicht vollständig durch gemeinsame Faktoren erklärt, d.h. es gibt Fehlervarianzen (z.B. in Form von Einzelrestfaktoren)

Wie wirkt sich nun die ausschließliche Betrachtung von *gemeinsamen Faktoren* auf das Fundamentaltheorem aus?

Für jeden Variablenwert z_{ij} gilt:

$$z_{ij} = f_{i1}a_{1j} + \ldots + f_{ir}a_{rj} + f_{i(r+1)}a_{(r+1)j} + \ldots + f_{iq}a_{qj}$$

Hieraus folgt für die gesamte Faktorladungsmatrix:

$$\boldsymbol{A}_{ges} = \begin{pmatrix} a_{11} & \ldots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{r1} & \ldots & a_{rm} \\ a_{(r+1)1} & \ldots & a_{(r+1)m} \\ \vdots & \ddots & \vdots \\ a_{q1} & \ldots & a_{qm} \end{pmatrix}$$

Wird angenommen, daß es sich bei den $f_{r+1}, \ldots, f_q$ um Einzelrestfaktoren handelt, so sind deren Faktorladungen bis auf einen Wert $u_{r+1}, \ldots, u_q$ alle gleich 0. Bei entsprechender Sortierung ergibt sich dann:

$$
\begin{aligned}
\boldsymbol{A}_{ges} &= \begin{pmatrix} a_{11} & \dots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{r1} & \dots & a_{rm} \\ u_{r+1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & u_q \end{pmatrix} \\
&= \begin{pmatrix} a_{11} & \dots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{r1} & \dots & a_{rm} \\ 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{pmatrix} + \begin{pmatrix} 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \\ u_{r+1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & u_q \end{pmatrix} \\
&= \qquad \boldsymbol{A} \qquad + \qquad \boldsymbol{U}
\end{aligned}
$$

Für das Fundamentaltheorem folgt daraus:

$$
\begin{aligned}
\boldsymbol{R} &= \boldsymbol{A}_{ges}^T \boldsymbol{A}_{ges} \\
&= (\boldsymbol{A} + \boldsymbol{U})^T (\boldsymbol{A} + \boldsymbol{U}) \\
&= (\boldsymbol{A}^T + \boldsymbol{U}^T)(\boldsymbol{A} + \boldsymbol{U}) \\
&= \boldsymbol{A}^T \boldsymbol{A} + \underbrace{\boldsymbol{A}^T \boldsymbol{U}}_{=0} + \underbrace{\boldsymbol{U}^T \boldsymbol{A}}_{=0} + \boldsymbol{U}^T \boldsymbol{U} \\
&= \boldsymbol{A}^T \boldsymbol{A} + \boldsymbol{U}^T \boldsymbol{U}
\end{aligned}
$$

Wenn also $\boldsymbol{A}$ die Faktorladungsmatrix der gemeinsamen Faktoren bezeichnet, dann reproduziert sie nicht die ursprüngliche Korrelationsmatrix, sondern lediglich eine — um die Fehlervarianzen — reduzierte:

$$
\begin{aligned}
\boldsymbol{R}_h &= \boldsymbol{A}^T \boldsymbol{A} \\
&= \boldsymbol{R} - \boldsymbol{U}^T \boldsymbol{U}
\end{aligned}
$$

Ausgehend von obiger Gleichung läßt sich $\boldsymbol{R}_h$ weiter umformen:

$$
\begin{aligned}
\boldsymbol{R}_h &= \boldsymbol{R} - \boldsymbol{U}^T\boldsymbol{U} \\
&= \begin{pmatrix} 1 & \dots & r_{1m} \\ \vdots & \ddots & \vdots \\ r_{m1} & \dots & 1 \end{pmatrix} - \begin{pmatrix} u_1^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & u_m^2 \end{pmatrix} \\
&= \begin{pmatrix} 1-u_1^2 & \dots & r_{1m} \\ \vdots & \ddots & \vdots \\ r_{m1} & \dots & 1-u_m^2 \end{pmatrix} \\
&= \begin{pmatrix} h_1^2 & \dots & r_{1m} \\ \vdots & \ddots & \vdots \\ r_{m1} & \dots & h_m^2 \end{pmatrix}
\end{aligned}
$$

Allgemein läßt sich also $\boldsymbol{R}_h$ aus der ursprünglichen Korrelationsmatrix $\boldsymbol{R}$ konstruieren, indem anstatt der 1 die entsprechenden Kommunalitäten in die Hauptdiagonale eingesetzt werden. Wird davon ausgegangen, daß die Fehlervarianz real eher ungleich Null ist, so stellt die reduzierte Korrelationsmatrix den besseren Ausgangspunkt für eine Analyse dar. Hieraus ergibt sich für die konkrete Durchführung folgendes Problem:

- Zur Bestimmung von $\boldsymbol{R}_h$ werden die Kommunalitäten h_j^2 und damit die Faktorladungen a_{kj} ($j = 1, \dots, m; k = 1, \dots, r$) benötigt.
- Gleichzeitig können die Faktorladungen a_{kj} nach dem Fundamentaltheorem erst mit Hilfe von $\boldsymbol{R}_h$ berechnet werden.

Die Lösung dieses Problems liegt in einer expliziten Schätzung der Kommunalitäten am Beginn des Verfahrens. Hierzu bieten sich folgende Möglichkeiten:

- Willkürliche Auswahl einer Zahl zwischen 0 und 1.
- Der jeweils höchste Korrelationskoeffizient zwischen z_j und allen anderen Observablen:

Beispiel:

Correlation Matrix:

	V212	V213	V214	V215	V216
V212	*.46209*				
V213	⇑ **-.46486**	*.83226*			
V214	**-.47908**	⇑ **.83226**	⇒ *.83226*		
V215	**-.19249**	**.52722**	**.46150**	⇒ *.52722*	
V216	**.46209**	**-.46273**	**-.44435**	**-.20346**	⇒ *.46209*

- Das Quadrat des multiplen Korrelationskoeffizienten (gemeinsame Varianz der Variablen z_j mit allen übrigen Observablen; im SPSS-Analyseergebnis: `MULTIPLE R`).

In statistischen Analysesystemen werden die entweder vom Anwender explizit geschätzten oder vom System automatisch gewählten Kommunalitätenschätzungen als Startwerte für ein iteratives Verfahren verwendet:

1. Schätzung der Kommunalitäten (s.o.)
2. Faktorextraktion mit Hilfe des Fundamentaltheorems und der geschätzten Kommunalitäten, d.h. Berechnung der Faktorladungsmatrix $\boldsymbol{A}$ (vgl. Kapitel 6.3.4).
3. Berechnung der Kommunalitäten anhand der Faktorladungen
4. Vergleich von geschätzten und berechneten Kommunalitäten:
   (a) Falls beide annähernd gleich sind, stellt $\boldsymbol{A}$ die endgültige Faktorladungsmatrix dar ($\Rightarrow$ Ende des Verfahrens).
   (b) Ansonsten werden die berechneten Kommunalitäten als neue Kommunalitätenschätzung verwendet ($\Rightarrow$ Schritt 2).

Der Vorteil dieser iterativen Festlegung von Kommunalitäten und — in Abhängigkeit davon — Faktorladungen liegt darin, daß sich auch eine ungünstige Anfangsschätzung im Verlauf der Iteration weitgehend ausgleicht[5].

[5] Die Wahl einer willkürlichen Zahl zwischen 0 und 1 als Kommunalitätenschätzung hat daher auch eher Auswirkungen auf die Anzahl der benötigten Iterationsschritte.

6.3.4 Das Problem der Faktorextraktion

1. Die Bestimmung der Faktorladungsmatrix $\boldsymbol{A}$

Für die Bestimmung der Faktorladungsmatrix $\boldsymbol{A}$ existieren eine Vielzahl von Verfahren und Verfahrensvarianten, von denen hier nur die bekanntesten genannt werden[6]. Die rechentechnisch einfacher durchzuführende Zentroidmethode wird an dieser Stelle lediglich aus historischen Gründen aufgeführt, da sie im Zeitalter computerunterstützter Datenanalyse mittlerweile selten Verwendung findet. Im Gegensatz zu den anderen Verfahren läßt sie aber noch eine Berechnung "per Hand" zu.

- Hauptkomponenten-/Hauptachsenmethode (`PC`/`PAF`)
- ALPHA–Methode (`ALPHA`)
- IMAGE–Methode (`IMAGE`)
- Maximum-Likelihood–Methode (`ML`)
- Zentroidmethode

Den Schwerpunkt in diesem Kapitel bildet die Darstellung der Hauptkomponenten-/Hauptachsenmethode als das in der Praxis am häufigsten verwendete Verfahren. Der Doppelname kommt dadurch zustande, daß — ausgehend von unterschiedlichen Annahmen über die zugrunde gelegten Kommunalitäten — dasselbe *mathematische* Verfahren zur Faktorextraktion durchgeführt wird:

- *Hauptkomponentenmethode:*
  Die gemeinsamen Faktoren erklären *die gesamte Varianz* der Observablen z_j
  $\Longrightarrow h_j^2 = 1$,
  d.h. Ausgangspunkt der Berechnungen ist $\boldsymbol{R} = \boldsymbol{A}^T\boldsymbol{A}$
- *Hauptachsenmethode:*
  Die gemeinsamen Faktoren erklären lediglich *einen Teil der Varianz* der Observablen z_j
  $\Longrightarrow h_j^2 < 1$,
  d.h. Ausgangspunkt der Berechnungen ist $\boldsymbol{R}_h = \boldsymbol{A}^T\boldsymbol{A} - \boldsymbol{U}^T\boldsymbol{U}$

[6] In Klammern sind die hinter dem SPSS-Befehl `FACTOR .../EXTRACTION` zu kodierenden Optionen zur Methodenwahl angegeben.

Technisch besteht die Unterscheidung zwischen Hauptkomponenten- und Hauptachsenmethode also darin, ob in der Hauptdiagonalen der Korrelationsmatrix Werte gleich oder kleiner 1 verwendet werden.

Hauptkomponenten-/Hauptachsenmethode

Definition: Ermittlung von untereinander unkorrelierten Faktoren f_k $(k = 1, \ldots, r)$, deren Varianz *nacheinander* jeweils maximal ist. Die Varianz der Faktoren $s^2_{f_k}$ ergibt sich dabei analog zu der Observablenvarianz aus der Summe der quadrierten Faktorladungen zwischen den Faktoren f_k und den Observablen z_j:

$$s^2_{f_k} = a^2_{k1} + \ldots + a^2_{km} = \sum_{j=1}^{m} a^2_{kj}$$

Geometrisch bedeutet diese Form der Faktorermittlung das Auffinden eines neuen Koordinatensystems mit senkrecht aufeinander stehenden Achsen f_k $(k = 1, \ldots, r)$, die so gelegt werden, daß die Summe der quadrierten Achsenabschnittsquadrate (= Faktorenvarianz $s^2_{f_k}$) für alle Achsen nacheinander maximal werden.

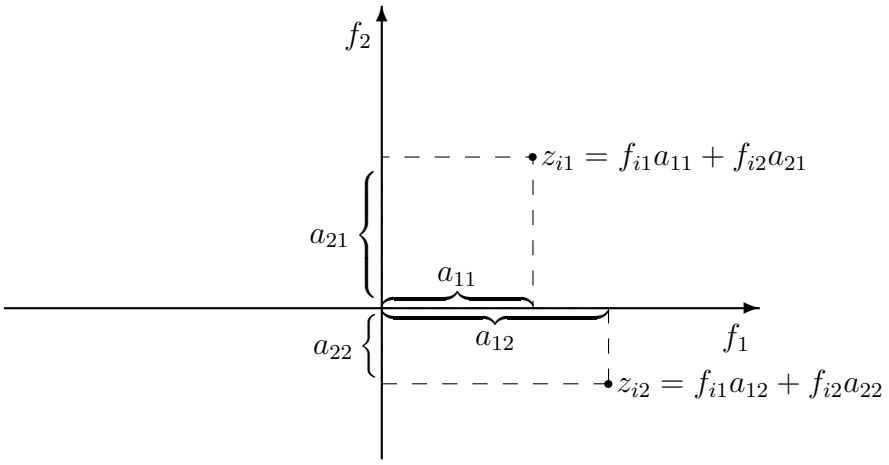


Abb. 6.10: Faktorermittlung (bezogen auf zwei Faktoren)

Die erste Hauptachse f_1 wird dabei so gezogen (vgl. Abbildung 6.10), daß

$$s^2_{f_1} = a^2_{11} + a^2_{12}$$

maximal wird.

Für eine *mathematische* Ermittlung des ersten Faktors sei f_1 als Linearkombination der Observablenwerte $\boldsymbol{Z}$ mit einem unbekannten, auf 1 normierten Koeffizientenvektor $\boldsymbol{b}_1$ zu betrachten:

$$\boldsymbol{f}_1 = \boldsymbol{Z}\boldsymbol{b}_1 \qquad \text{mit} \qquad \sqrt{\boldsymbol{b}_1^T\boldsymbol{b}_1} = 1$$

Die Forderung nach Maximierung der Faktorvarianz läßt sich dann wie folgt umsetzen:

$$\begin{aligned} s_{f_1}^2 &= \frac{1}{n}\boldsymbol{f}_1^T\boldsymbol{f}_1 \\ &= \frac{1}{n}(\boldsymbol{Z}\boldsymbol{b}_1)^T(\boldsymbol{Z}\boldsymbol{b}_1) \\ &= \frac{1}{n}\boldsymbol{b}_1^T\boldsymbol{Z}^T\boldsymbol{Z}\boldsymbol{b}_1 \\ &= \boldsymbol{b}_1^T(\frac{1}{n}\boldsymbol{Z}^T\boldsymbol{Z})\boldsymbol{b}_1 \\ &= \boldsymbol{b}_1^T\boldsymbol{R}\boldsymbol{b}_1 \\ &\stackrel{!}{=} \textit{Maximum} \end{aligned}$$

Somit bedeutet die Berechnung eines Faktors im Rahmen einer Hauptkomponenten-/Hauptachsenmethode letztendlich die Lösung einer Extremwertaufgabe ($\boldsymbol{b}_1^T\boldsymbol{R}\boldsymbol{b}_1$) mit einer Nebenbedingung ($\boldsymbol{b}_1^T\boldsymbol{b}_1 = 1$):

> Falls x_0 ein Extremwert einer Funktion $f(x)$ für die Werte von x ist, für die eine weitere Funktion $g(x) = 0$ gilt, so existiert eine Zahl l (=*Lagrange-Multiplikator*) mit
>
> $$\frac{\delta f}{\delta x_0} + l\frac{\delta g}{\delta x_0} = 0$$

Bezogen auf die gestellte Extremwertaufgabe gilt:

$$f(\boldsymbol{b}_1) = \boldsymbol{b}_1^T\boldsymbol{R}\boldsymbol{b}_1 \qquad \Longrightarrow \quad \frac{\delta f}{\delta \boldsymbol{b}_1} = 2\boldsymbol{R}\boldsymbol{b}_1$$

$$g(\boldsymbol{b}_1) = 1 - \boldsymbol{b}_1^T\boldsymbol{b}_1 = 0 \quad \Longrightarrow \quad \frac{\delta g}{\delta \boldsymbol{b}_1} = -2\boldsymbol{b}_1$$

Eingesetzt in die Ausgangsgleichung ergibt dies:

$$\begin{aligned} 2\boldsymbol{R}\boldsymbol{b}_1 - l_1 2\boldsymbol{b}_1 &= 0 \\ \boldsymbol{R}\boldsymbol{b}_1 &= l_1\boldsymbol{b}_1 \end{aligned}$$

Somit ist der gesuchte Koeffizientenvektor $\boldsymbol{b}_1$ ein *Eigenvektor* der Korrelationsmatrix $\boldsymbol{R}$ mit dem *Eigenwert* l_1.

Für die Ermittlung des zweiten Faktors wird unter Verwendung der weiteren Nebenbedingung

$$\boldsymbol{b}_2^T \boldsymbol{b}_1 = 0,$$

analog vorgegangen. Dies bedeutet das Lösen einer Extremwertaufgabe mit zwei Nebenbedingungen.

Durch die Reduzierung der Lösung der Extremwertaufgabe auf das allgemeine — und in der Mathematik gelöste — Eigenvektor-/Eigenwertproblem von Matrizen wird die Bestimmung der gesamten Faktorladungsmatrix $\boldsymbol{A}$ vereinfacht. Dazu wird die sogenannte *Spektralzerlegung* einer Matrix verwendet, d.h. ihre Darstellung als Summe aller *dyadischen Produkte* der Eigenvektoren, multipliziert jeweils mit den zugehörigen Eigenwerten:

$$\boldsymbol{R} = l_1 \boldsymbol{b}_1 \boldsymbol{b}_1^T + \ldots + l_r \boldsymbol{b}_r \boldsymbol{b}_r^T$$

bzw. in Matrizenschreibweise

$$\boldsymbol{R} = \boldsymbol{B}\boldsymbol{L}\boldsymbol{B}^T$$

mit:

$$\boldsymbol{B} = \begin{pmatrix} b_{11} & \ldots & b_{1r} \\ \vdots & \ddots & \vdots \\ b_{r1} & \ldots & b_{rr} \end{pmatrix}, \quad \boldsymbol{B}^T = \begin{pmatrix} b_{11} & \ldots & b_{r1} \\ \vdots & \ddots & \vdots \\ b_{1r} & \ldots & b_{rr} \end{pmatrix}, \quad \boldsymbol{L} = \begin{pmatrix} l_1 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & l_r \end{pmatrix}$$

Dann gilt:

$$\begin{aligned} \boldsymbol{R} &= \boldsymbol{B}\boldsymbol{L}\boldsymbol{B}^T \\ &= \boldsymbol{B}\boldsymbol{L}^{1/2}\boldsymbol{L}^{1/2}\boldsymbol{B}^T \\ &= (\boldsymbol{B}\boldsymbol{L}^{1/2})(\boldsymbol{L}^{1/2}\boldsymbol{B}^T) \end{aligned}$$

Gleichzeitig gilt mit dem Fundamentaltheorem:

$$\boldsymbol{R} = \boldsymbol{A}^T\boldsymbol{A}$$

Für die Faktorladungsmatrix $(j = 1, \ldots, m; k = 1, \ldots, r)$ ergibt sich

$$\boldsymbol{A} = \boldsymbol{L}^{1/2}\boldsymbol{B}^T$$

mit:

$$a_{kj} = \sqrt{l_k}\, b_{kj}$$

Die Faktorladungen lassen sich also direkt aus den Eigenvektoren und Eigenwerten der Korrelationsmatrix $\boldsymbol{R}$ berechnen.

Als Beispiel für die sukzessive Bestimmung der Eigenvektoren/-werte einer Matrix und der Faktorladungsmatrix $\boldsymbol{A}$ wird im folgenden das Verfahren nach *von Mises* (vgl. [Holm 1982]) erläutert:

Es gilt: $\boldsymbol{R}\boldsymbol{b}_k = l_k \boldsymbol{b}_k, \;\; (k = 1, \ldots, r)$

1. Schritt: Wahl eines beliebigen Ausgangsvektors $\boldsymbol{b}_0$ (z.B. konstant 1) und eines beliebigen Ausgangswertes l_0 (z.B. absolut größter Wert von $\boldsymbol{y}_0$)

2. Schritt: Berechnung von $\boldsymbol{y}_0 = \boldsymbol{R}\boldsymbol{b}_0 \; (= l_0 \boldsymbol{b}_0)$.

3. Schritt: Berechnung von $\boldsymbol{b}_1 = \dfrac{\boldsymbol{y}_0}{l_0}$.

4. Schritt:

- Falls: $\boldsymbol{b}_1 \approx \boldsymbol{b}_0 \implies \boldsymbol{b}_0 = \dfrac{\boldsymbol{y}_0}{l_0}$
  d.h. $\boldsymbol{b}_0$ ist Eigenvektor von $\boldsymbol{R}$ mit Eigenwert l_0.
- Sonst: $\boldsymbol{b}_0 = \boldsymbol{b}_1$ ($\Longrightarrow$ 2. Schritt)

5. Schritt: Berechnung der Faktorladungen $\boldsymbol{a}$ für einen Faktor (aus $\boldsymbol{b}_0$, l_0)

6. Schritt: Berechnung von $\boldsymbol{R}_{res} = \boldsymbol{R} - \boldsymbol{a}\boldsymbol{a}^T$, wobei $\boldsymbol{a}\boldsymbol{a}^T$ die reproduzierte Korrelationsmatrix für die Korrelationsanteile bezeichnet, die durch den Faktor bestimmt sind.

- Falls: $\boldsymbol{R}_{res} \approx 0$ ($\Longrightarrow$ Stop des Verfahrens)
- Sonst: Bestimmung des nächsten Eigenvektors/-wertes mit $\boldsymbol{R}_{res} = \boldsymbol{R}$ ($\Longrightarrow$ 2. Schritt)

Nachfolgendes Beispiel zeigt die Durchführung einer Hauptkomponentenanalyse (PC) auf den Sympathieskalometern für Politiker. Entsprechend den beschriebenen Eigenschaften dieses Verfahrens wird für alle Variablen eine Kommunalität von 1 angenommen, und es werden genausoviele Faktoren extrahiert wie Ausgangsvariablen vorhanden sind, d.h. bei Anwendung dieses Verfahrens findet erst einmal *keine Datenreduktion* statt.

```
* Auszug aus: FACTOR      /VARIABLES V222 TO V231 /CRITERIA FACTORS(10)
*                         /ROTATION NOROTATE /PRINT EXTRACTION INITIAL
*                         /PLOT EIGEN ROTATION(1,2).

            - - - -   F A C T O R   A N A L Y S I S   - - - -

Extraction  1  for Analysis  1, Principal-Components Analysis (PC)
Initial Statistics:
Variable     Communality  *  Factor   Eigenvalue   Pct of Var   Cum Pct
V222             1.00000  *      1       4.63308       46.3         46.3
V223             1.00000  *      2       1.85099       18.5         64.8
V224             1.00000  *      3        .76638        7.7         72.5
V225             1.00000  *      4        .66265        6.6         79.1
V226             1.00000  *      5        .50409        5.0         84.2
V227             1.00000  *      6        .42231        4.2         88.4
V228             1.00000  *      7        .34722        3.5         91.9
V229             1.00000  *      8        .31217        3.1         95.0
V230             1.00000  *      9        .27810        2.8         97.8
V231             1.00000  *     10        .22300        2.2        100.0
   PC Extracted  10 factors.

Factor Matrix:

         FACTOR  1     FACTOR  2     FACTOR  3     FACTOR  4     FACTOR  5

V222       .65929        .48454        .12463       -.23909       -.24198
V223      -.66284        .54161        .03727       -.00483       -.08788
V224       .61148        .45474        .24766       -.38362        .33908
V225       .85626        .10544       -.14418        .22716       -.05584
V226      -.55227        .25383        .65785        .41264        .07733
V227      -.61964        .60310       -.28940       -.02543        .10342
V228       .77662        .20798       -.05569        .16497        .41053
V229       .77988        .12126       -.16767        .43727       -.02133
V230      -.49355        .68987       -.29747        .13175       -.01442
V231       .71278        .37806        .17640        .00354       -.36584

         FACTOR  6     FACTOR  7     FACTOR  8     FACTOR  9     FACTOR 10

V222       .28709       -.28352       -.15789        .04416       -.09804
V223       .35066        .18313        .26236       -.17090        .06017
V224      -.19553       -.05330        .22751        .02473        .04872
V225      -.04010       -.15104       -.02331       -.21942        .32724
V226      -.04014       -.10549       -.08219        .03176        .03251
V227       .01336        .03918       -.15795        .30215        .19817
V228       .19753        .21551       -.21848       -.09911       -.10832
V229       .06234       -.06419        .28586        .24018       -.09949
V230      -.28624       -.13104       -.02072       -.19118       -.19583
V231      -.22464        .35672       -.06720        .04270        .00047
```

2. Die Festlegung der Faktorenzahl

Zwischen den Elementen der Faktorladungsmatrix a_{kj} und den Eigenwerten l_k läßt sich ein weiterer Zusammenhang ableiten:

$$\begin{aligned}
& a_{kj} && = && \sqrt{l_k}\, b_{kj} && \qquad (k = 1, \ldots, r; j = 1, \ldots, m) \\
\iff\ & a_{kj}^2 && = && l_k b_{kj}^2 \\
\iff\ & \sum_{j=1}^{m} a_{kj}^2 && = && \sum_{j=1}^{m} l_k b_{kj}^2 \\
\iff\ & \sum_{j=1}^{m} a_{kj}^2 && = && l_k \underbrace{\sum_{j=1}^{m} b_{kj}^2}_{=1 \text{ (s.S. 166)}}
\end{aligned}$$

Somit ist die Summe der quadrierten Ladungen eines Faktors gleich dem Eigenwert zu diesem Faktor, d.h. der Eigenwert l_k bezeichnet den durch *einen* Faktor erklärten Teil der Gesamtvarianz aller Observablen:

$$l_k = \sum_{j=1}^{m} a_{kj}^2$$

Zwischen Kommunalität und Eigenwert wird damit folgende Analogie deutlich:

Kommunalität: $h_j^2 = \sum_{k=1}^{r} a_{kj}^2$

Teil der Einheitsvarianz der Observablen z_j, der durch alle gemeinsamen Faktoren erklärt wird.

Eigenwert: $l_k = \sum_{j=1}^{m} a_{kj}^2$

Teil der Einheitsvarianz des Faktors f_k, der durch alle Observablen erklärt wird.

Sowohl Kommunalitäten als auch Eigenwerte lassen sich direkt aus der Faktorla-

dungsmatrix ermitteln:

$$\begin{array}{c|ccc|l} & f_1 & \dots & f_r & \\ \hline z_1 & a_{11} & \dots & a_{r1} & \\ \vdots & \vdots & \ddots & \vdots & \rightarrow h_j^2 = \sum_{k=1}^{r} a_{kj}^2 \\ z_m & a_{1m} & \dots & a_{rm} & \\ & & \downarrow & & \\ & & l_k = \sum_{j=1}^{m} a_{kj}^2 & & \end{array}$$

Beispiel:

$$\begin{aligned} l_1 &= \sum_{j=1}^{10} a_{1j}^2 \\ &= (.65929)^2 + (-.66284)^2 + (.61148)^2 + (.85626)^2 + (-.55227)^2 + \\ &\quad (-.61964)^2 + (.77662)^2 + (.77988)^2 + (-.49355)^2 + (.71278)^2 \\ &= 4.6481 \\ h_1^2 &= \sum_{k=1}^{10} a_{k1}^2 \\ &= (.65929)^2 + (.48454)^2 + (.12463)^2 + (-.23909)^2 + (-.24198)^2 + \\ &\quad (.28709)^2 + (-.28352)^2 + (-.15789)^2 + (.04416)^2 + (-.09804)^2 \\ &= 10 \end{aligned}$$

Da als Ergebnis der Hauptkomponenten- bzw. Hauptachsenmethode immer so viele Eigenvektoren (= Faktoren) wie Observablen erzeugt werden, stellt sich dem Anwender das Problem, die Anzahl der für seine Fragestellung geeigneten Faktoren zu bestimmen. Hierfür gibt es kein allgemein anerkanntes bzw. mathematisch beweisbares Verfahren. Im folgenden werden einige übliche Verfahren zur Festlegung der Faktorenanzahl aufgeführt.

1. Theoretische Vorannahmen über die Anzahl der Faktoren.

2. *Kaiser-Kriterium:*
   Auswahl aller Faktoren mit einem Eigenwert von mindestens 1.

Idee: Da jede Observable bereits einen Varianzanteil von 1 erklärt[7], sollte jeder Faktor mindestens ebensoviel Varianz (= Eigenwert) binden.

SPSS: `FACTOR .../CRITERIA MINEIGEN (1.0).` (Default-Einstellung)

3. Vorgabe eines *prozentualen Varianzanteils* der Observablen, der durch die Faktoren erklärt wird:

   (a) Anteil an der Gesamtvarianz oder

   (b) Anteil an der Kommunalität

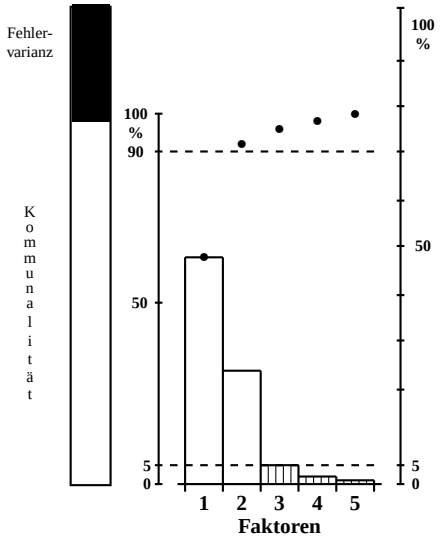


Abb. 6.11: Durch prozentualen Varianzanteil ermittelte Faktorenanzahl (nach [Überla 1977, S. 126])

4. *Scree-Test:*
   Darstellung der Eigenwerte in abfallender Reihenfolge in einem Diagramm. An die niedrigsten Eigenwerte wird eine Gerade gelegt. Es werden die Faktoren ausgewählt, deren Eigenwerte oberhalb der Geraden liegen.

   Idee: Eigenwerte von Korrelationsmatrizen normalverteilter, unabhängiger Zufallszahlen zeigen typischerweise einen flach und gleichmäßig abfallenden Verlauf. Die Eigenwerte auf der Gerade gehören demnach zu zufälligen Faktoren.

   SPSS: `FACTOR .../PLOT EIGEN.`

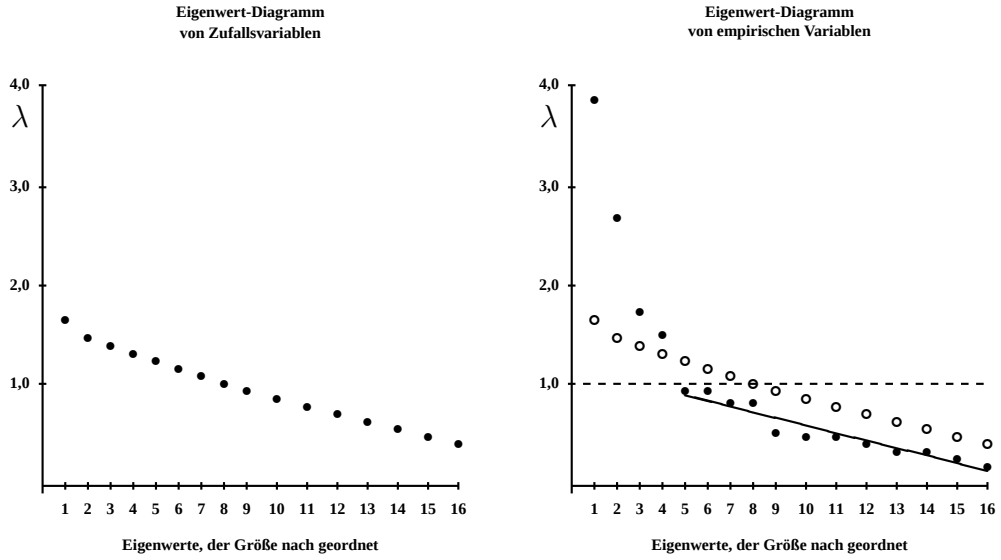


Abb. 6.12: Durch Scree-Test ermittelte Faktoren(-anzahl) (nach [Überla 1977, S. 127f])

5. *Residualmatrixverfahren:*
   Es werden solange Faktoren erzeugt, bis die Differenzen zwischen allen Elementen der Korrelationsmatrix $\boldsymbol{R}$ und der reproduzierten Korrelationsmatrix $\boldsymbol{A}^T\boldsymbol{A}$ nicht mehr signifikant von Null verschieden sind.

6. Ein Faktor wird dann extrahiert, wenn eine bestimmte Anzahl von Variablen hoch auf einen Faktor laden.

   Beispiel: Mindestens 3 Variablen müssen jeweils eine Ladung von mehr als 0.7 haben.
   $\Longrightarrow l = (0.7)^2 + (0.7)^2 + (0.7)^2 = 1.47$
   Ein Faktor wird dann extrahiert, wenn sein Eigenwert mindestens 1.47 beträgt.

Um bei der ermittelten Faktorenanzahl einigermaßen sichergehen zu können, ist die Anwendung verschiedener Verfahren und Vergleich der erhaltenen Werte zu empfehlen.

Nachfolgende Graphik zeigt die zehn Eigenwerte für die aus den 10 Beispielvariablen extrahierten Faktoren. Der Eigenwerteverlauf legt es nahe, lediglich die beiden ersten Faktoren als relevant auszuwählen. Gleichzeitig erfüllt diese Auswahl auch das Kaiserkriterium.

[7] Für standardisierte Variablen hat die Varianz den Wert 1.

```
* Auszug aus: FACTOR      /VARIABLES V222 TO V231 /CRITERIA FACTORS(10)
*                         /ROTATION NOROTATE /PRINT EXTRACTION INITIAL
*                         /PLOT EIGEN ROTATION(1,2).
      4.633 +   *
            |
            |
            |
 E          |
 I          |
 G          |
 E          |
 N          |
 V          |
 A          |
 L          |
 U    1.851 +       *
 E          |
 S          |
            |
       .766 +           *
       .504 +               *   *
       .278 +                       *   *   *   *
       .000 +---+---+---+---+---+---+---+---+---+---*
                1   2   3   4   5   6   7   8   9  10

   PC Extracted  10 factors.
```

6.3.5 Das Problem der Faktorrotation

Die Bestimmung der Faktorladungsmatrix nach der Hauptkomponenten-/ Hauptachsenanalyse erfolgt dadurch, daß die Varianz der einzelnen Faktoren über alle Observablen nacheinander maximiert wird, d.h. die Faktoren korrelieren jeweils mit *allen* Variablen *möglichst hoch*.

Wie Abbildung 6.13 zeigt, lassen sich die Observablen als Punkte in einem von den Faktoren gebildeten Koordinatensystem darstellen. Die Faktorladungen a_{kj} sind dabei die jeweiligen Koordinatenwerte, und es gilt:

$$h_j^2 = a_{1j}^2 + a_{2j}^2 = \sum_{k=1}^{r} a_{kj}^2$$

$$\cos\varphi = \frac{a_{1j}}{h_j} = r_{z_j f}$$

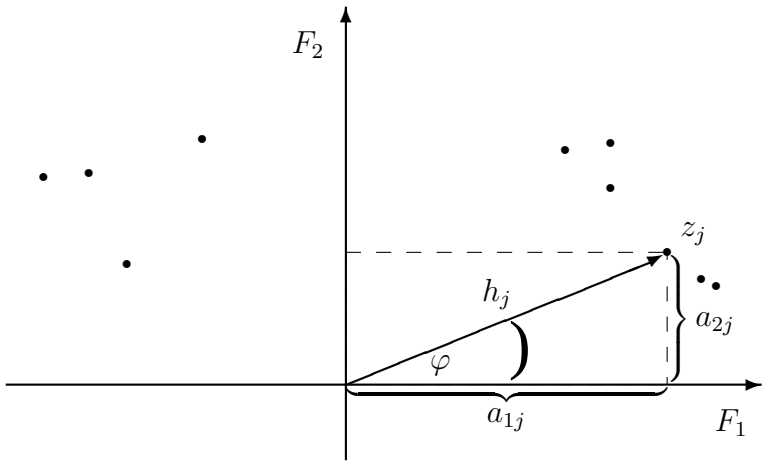


Abb. 6.13: Nicht optimale Faktorextraktion

Die Darstellung der Observablen z_j in dem durch die Faktoren f_r gebildeten Faktorraum zeigt auch, daß die nach einem mathematischen Kriterium durchgeführte Faktorextraktion — insbesondere für die Zwecke der Faktorinterpretation (vgl. Kapitel 6.3.6) — nicht automatisch zu optimalen Faktoren führt. Dies ist zum Beispiel dann der Fall, wenn jeder Faktor mit allen Observablen mittelmäßig korreliert ist (d.h. die Observablen liegen alle in einem mittleren Abstand von den Koordinatenachsen).

Wünschenswert dagegen im Sinne einer möglichst einfachen inhaltlichen Interpretation ist die sogenannte *Einfachstruktur* (nach [Thurstone 1947]) mit folgenden Eigenschaften:

- Einzelne Observable korrelieren möglichst nur mit einem Faktor hoch, mit allen anderen Faktoren nicht oder nur schwach (= lediglich *eine* hohe Ladung in jeder Zeile von $\boldsymbol{A}$).

- Einzelne Faktoren korrelieren möglichst entweder sehr hoch oder sehr niedrig mit den Observablen (= keine "mittelmäßigen" Ladungen in den Spalten von $\boldsymbol{A}$).

In der Realität sind diese Forderungen selten erfüllt. Mit Hilfe der *Faktorrotation* ist es aber möglich, bessere Annäherungen zu erreichen.

Definition: *Faktorrotation:*
Drehung der Koordinatenachsen des durch die Faktoren aufgespannten Faktorraumes um einen bestimmten Winkel φ

Die Rotation von Faktoren entspricht technisch einer Neuberechnung der Faktorladungsmatrix $\boldsymbol{A}$. Diese läßt sich mit Hilfe einiger trigonometrischer Umformungen *graphisch* erläutern:

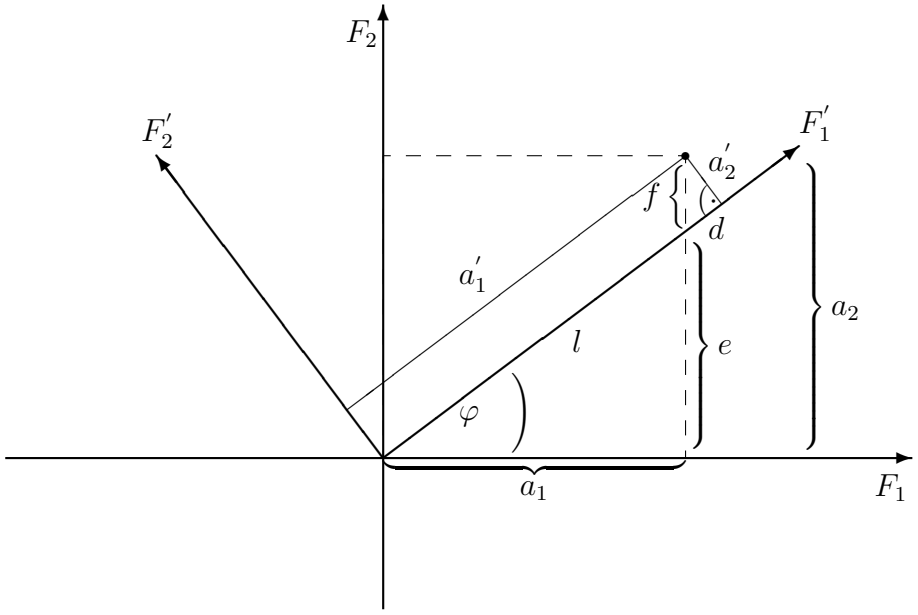


Abb. 6.14: Graphische Faktorrotation

Es gilt:

$$\begin{aligned} l &= \frac{a_1}{\cos\varphi} \\ d &= f\sin\varphi \\ &= (a_2 - a_1\tan\varphi)\sin\varphi \end{aligned}$$

Hieraus ergibt sich für $a_1^{'}$:

$$\begin{aligned}
a_1^{'} &= l + d \\
&= \frac{a_1}{\cos\varphi} + (a_2 - a_1 \tan\varphi)\sin\varphi \\
&= \frac{a_1}{\cos\varphi} + a_2 \sin\varphi - a_1 \tan\varphi \sin\varphi \\
&= a_2 \sin\varphi + \frac{a_1}{\cos\varphi} - a_1 \tan\varphi \sin\varphi \\
&= a_2 \sin\varphi + a_1(\frac{1}{\cos\varphi} - \tan\varphi \sin\varphi) \quad | \quad \tan\varphi = \frac{\sin\varphi}{\cos\varphi} \\
&= a_2 \sin\varphi + a_1(\frac{1}{\cos\varphi} - \frac{\sin^2\varphi}{\cos\varphi}) \\
&= a_2 \sin\varphi + a_1(\frac{1-\sin^2\varphi}{\cos\varphi}) \quad | \quad 1 - \sin^2\varphi = \cos^2\varphi \\
&= a_2 \sin\varphi + a_1(\frac{\cos^2\varphi}{\cos\varphi}) \\
&= a_1 \cos\varphi + a_2 \sin\varphi
\end{aligned}$$

Ebenso gilt:

$$\begin{aligned}
e &= a_1 \tan\varphi \\
f &= a_2 - e
\end{aligned}$$

Und damit für $a_2^{'}$:

$$\begin{aligned}
a_2^{'} &= f \cos\varphi \\
&= (a_2 - a_1 \tan\varphi)\cos\varphi \\
&= a_2 \cos\varphi - a_1 \tan\varphi \cos\varphi \quad | \quad \tan\varphi = \frac{\sin\varphi}{\cos\varphi} \\
&= a_2 \cos\varphi - a_1 \frac{\sin\varphi}{\cos\varphi} \cos\varphi \\
&= a_2 \cos\varphi - a_1 \sin\varphi \\
&= -a_1 \sin\varphi + a_2 \cos\varphi
\end{aligned}$$

Allgemein folgt daraus für die rotierte Faktorladungsmatrix $\boldsymbol{A}_{rot}$:

$$\begin{aligned}
\begin{pmatrix} a_1^{'} \\ a_2^{'} \end{pmatrix} &= \begin{pmatrix} \cos\varphi\, a_1 & + & \sin\varphi\, a_2 \\ -\sin\varphi\, a_1 & + & \cos\varphi\, a_2 \end{pmatrix} \\
\begin{pmatrix} a_1^{'} \\ a_2^{'} \end{pmatrix} &= \begin{pmatrix} \cos\varphi & \sin\varphi \\ -\sin\varphi & \cos\varphi \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \\
\boldsymbol{A}_{rot} &= \boldsymbol{T}\boldsymbol{A}
\end{aligned}$$

Die Rotation um einen Winkel φ entgegen dem Uhrzeigersinn bedeutet somit die Multiplikation der ursprünglichen Faktorladungsmatrix $\boldsymbol{A}$ mit einer charakteristischen Transformationsmatrix $\boldsymbol{T}$:

$$\boldsymbol{T} = \begin{pmatrix} \cos\varphi & \sin\varphi \\ -\sin\varphi & \cos\varphi \end{pmatrix}$$

Für Tranformationsmatrizen dieser Art gilt: $\boldsymbol{T}^T\boldsymbol{T} = \boldsymbol{E_n}$

Eigenschaften der rotierten Faktorladungsmatrix $\boldsymbol{A}_{rot}$:

1. Die Rotation beeinflußt nicht die Korrelationsmatrix, d.h. die Beziehungen der Observablen untereinander, sondern nur ihre Beziehungen zu den Faktoren.
$$\boldsymbol{A}_{rot}^T\boldsymbol{A}_{rot} = (\boldsymbol{T}\boldsymbol{A})^T(\boldsymbol{T}\boldsymbol{A}) = \boldsymbol{A}^T\underbrace{(\boldsymbol{T}^T\boldsymbol{T})}_{\boldsymbol{E_n}}\boldsymbol{A} = \boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{R}$$
2. Die Rotation ändert die Faktorladungen, jedoch nicht den jeweils durch die gemeinsamen Faktoren erklärten Varianzanteil der Observablen (= Kommunalität). Es ändert sich lediglich dessen Verteilung auf die Faktoren.
3. Mit der Rotation ändert sich die Varianz der einzelnen Faktoren (= Eigenwert).

Bei mehr als zwei Faktoren kann die Transformationsmatrix aus dem Produkt der nacheinander aus allen Faktorenpaaren berechneten Transformationsmatrizen bestimmt werden.

Beispielsweise gilt für drei Faktoren:

$$\boldsymbol{T} = \boldsymbol{T}_{12}\boldsymbol{T}_{23}\boldsymbol{T}_{13}$$

mit:

$$\boldsymbol{T}_{12} = \begin{pmatrix} \cos\varphi & \sin\varphi & 0 \\ -\sin\varphi & \cos\varphi & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$\boldsymbol{T}_{13} = \begin{pmatrix} \cos\varphi & 0 & \sin\varphi \\ 0 & 1 & 0 \\ -\sin\varphi & 0 & \cos\varphi \end{pmatrix}$$

$$\boldsymbol{T}_{23} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\varphi & \sin\varphi \\ 0 & -\sin\varphi & \cos\varphi \end{pmatrix}$$

Aufgabe *mathematischer Verfahren* zur Rotation in die Einfachstruktur ist daher die Bestimmung eines Rotationswinkels φ, so daß die Faktoren in den Schwerpunkt der Variablenpunktwolke gedreht werden.

Grundsätzlich lassen sich dabei folgende Rotationsverfahren unterscheiden:

- *Rechtwinklige (orthogonale) Rotation:*
  Alle in einem rechtwinkligen Koordinatensystem befindlichen Achsen werden um den jeweils gleichen Winkel φ in die gleiche Richtung gedreht.
  (im SPSS-Aufruf: `FACTOR .../ROTATION VARIMAX` oder `QUARTIMAX, EQUAMAX`)

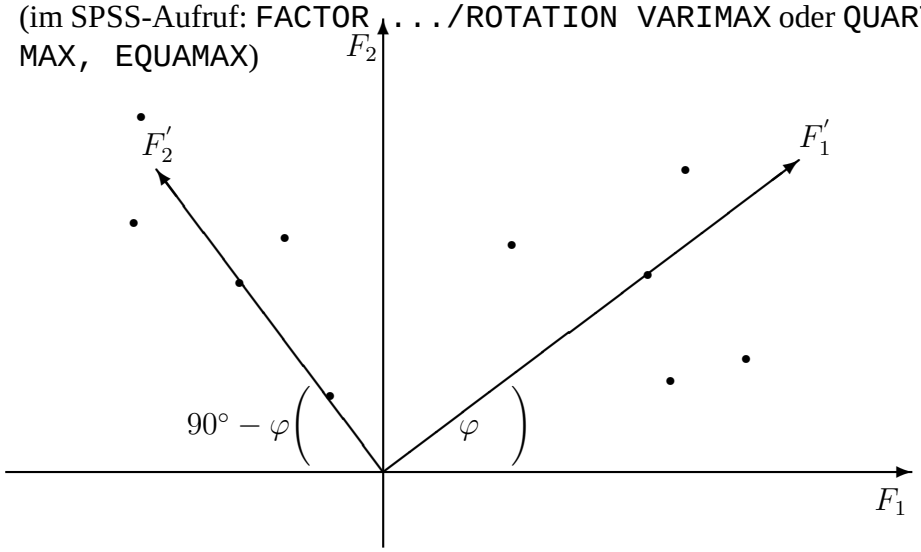


Abb. 6.15: Orthogonale Transformation

- *Schiefwinklige (oblique) Rotation:*
  Jede Koordinatenachse wird jeweils um einen bestimmten Winkel gedreht (s. Abb. 6.16).
  (im SPSS-Aufruf: `FACTOR .../ROTATION OBLIMIN.`)

Gemeinsames Kennzeichen aller dieser Verfahren ist jeweils die Annahme eines Maximierungs- bzw. Minimierungskriteriums, mit dem iterativ die Einfachstruktur angenähert wird.

Beispiel: `VARIMAX`

Kriterium: Maximierung der Varianz der quadrierten Faktorladungen,
d.h. es wird nach der Faktorladungsmatrix gesucht, bei der die Faktorladungen eines Faktors maximal differieren.

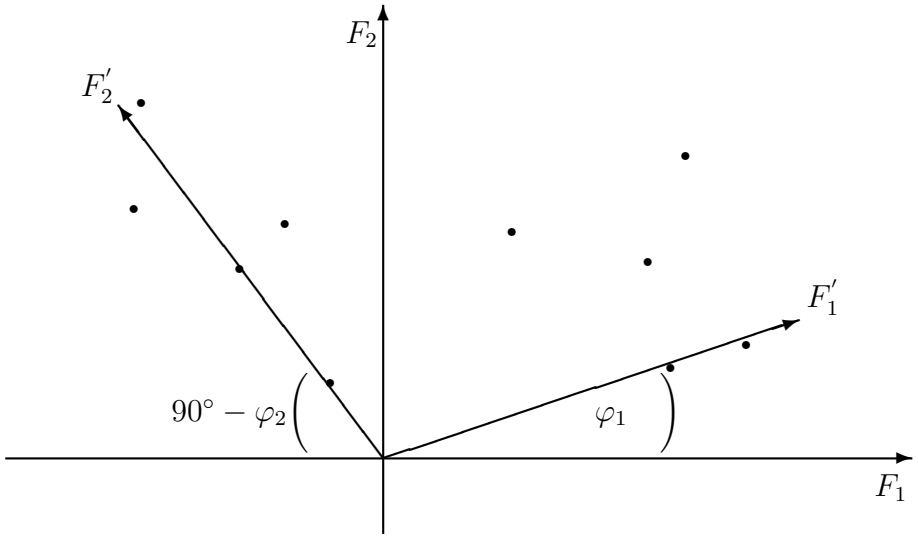


Abb. 6.16: Oblique Transformation

Die Varianz der quadrierten Faktorladungen errechnet sich aus:

$$
\begin{aligned}
s_k^2 &= \frac{1}{m}\sum_{j=1}^{m}(a_{kj}^2 - \bar{a}_k)^2, \qquad (k = 1, \ldots, r) \\
&= \frac{1}{m}\sum_{j=1}^{m}(a_{kj}^4 - 2a_{kj}^2\bar{a}_k + \bar{a}_k^2) \\
&= \frac{1}{m}\sum_{j=1}^{m} a_{kj}^4 - 2\underbrace{\frac{1}{m}\sum_{j=1}^{m} a_{kj}^2}_{\bar{a}_k}\bar{a}_k + \frac{1}{m}\sum_{j=1}^{m}\bar{a}_k^2 \\
&= \frac{1}{m}\sum_{j=1}^{m}(a_{kj}^4) - 2\bar{a}_k^2 + \bar{a}_k^2 \\
&= \frac{1}{m}\sum_{j=1}^{m} a_{kj}^4 - \bar{a}_k^2 \\
&= \frac{1}{m}\sum_{j=1}^{m} a_{kj}^4 - \left(\frac{1}{m}\sum_{j=1}^{m} a_{kj}^2\right)^2
\end{aligned}
$$

Somit ergibt sich:

$$
\sum_{k=1}^{r} s_k^2 = \frac{1}{m}\sum_{k=1}^{r}\sum_{k=1}^{m} a_{kj}^4 - \frac{1}{m^2}\sum_{k=1}^{r}\left(\sum_{j=1}^{m} a_{kj}^2\right)^2 \stackrel{!}{=} \textit{Maximum}
$$

Diese Maximierungsbedingung hat den Nachteil, daß Variablen mit hoher Kommunalität das Ergebnis sehr viel stärker beeinflussen als Variablen mit niedriger Kommunalität. Um diese Eigenschaft zu vermeiden, werden die Faktorladungen durch die Wurzel aus der Kommunalität dividiert. So erhält jede Variable für die Bestimmung des Rotationswinkels das gleiche Gewicht. Nach der Rotation wird diese Normierung durch Multiplikation mit h_j wieder rückgängig gemacht.

Damit gilt:

$$\sum_{k=1}^{r} s_k^2 = \frac{1}{m} \sum_{k=1}^{r} \sum_{k=1}^{m} z_{kj}^4 - \frac{1}{m^2} \sum_{k=1}^{r} \left(\sum_{j=1}^{m} z_{kj}^2 \right)^2 \stackrel{!}{=} \textit{Maximum}$$

mit

$$z_{kj} = \frac{a_{kj}}{h_j}$$

In dieses Maximierungskriterium werden nun die durch Multiplikation mit der Transformationsmatrix ermittelten "neuen" Achsenabschnitte (= Faktorladungen) $a_{kj}^{'}$ (bzw. $z_{kj}^{'}$) eingesetzt, d.h. im Falle von zwei Faktoren:

$$\begin{aligned} z_{1j}^{'} &= z_{1j} \cos\varphi + z_{2j} \sin\varphi \\ z_{2j}^{'} &= -z_{1j} \sin\varphi + z_{2j} \cos\varphi \end{aligned}$$

Zur Berechnung von $\boldsymbol{A}_{rot}$ werden anschließend folgende Schritte durchgeführt:

1. Berechnung der partiellen Ableitung von φ ($\delta(\sum_{k=1}^{r} s_k^2)/\delta\varphi \stackrel{!}{=} 0$)

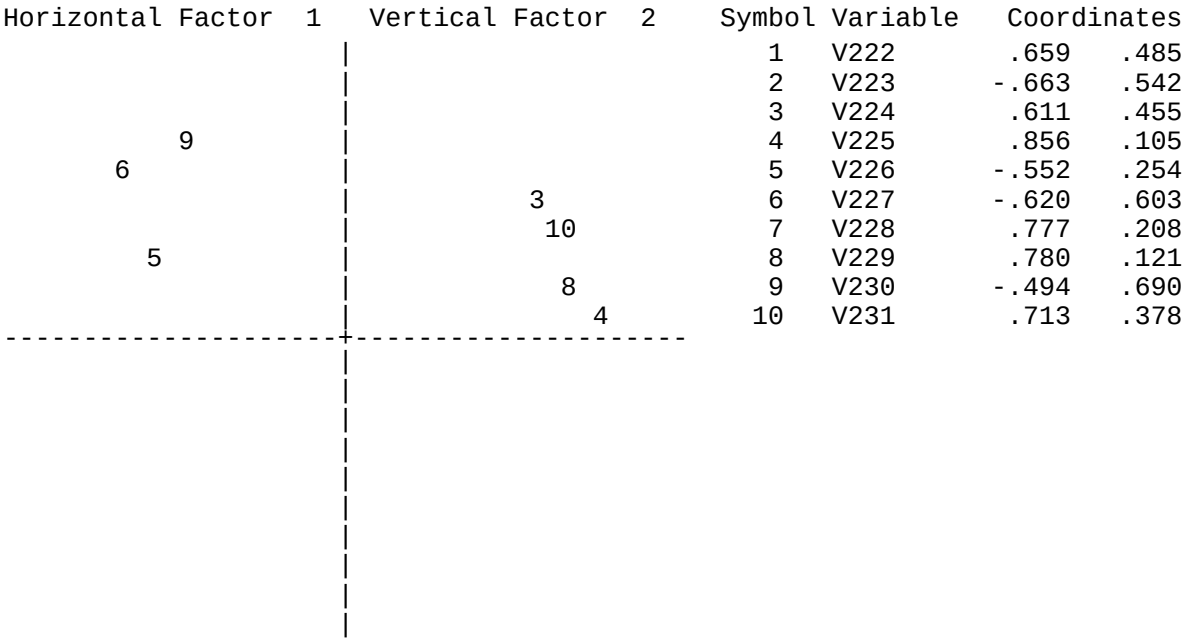
2. Berechnung von φ

3. Bestimmung der endgültigen Transformationsmatrix durch Einsetzen von φ

4. Berechnung von $\boldsymbol{A}_{rot} = \boldsymbol{T}\boldsymbol{A}$
5. "Entnormalisierung" der Faktorladungen von $\boldsymbol{A}_{rot}$ durch Multiplikation mit h_k

Das nachfolgende Beispiel zeigt die Ergebnisse einer Hauptkomponentenanalyse der Sympathieskalometer für Politiker mit anschließender VARIMAX-Rotation. Die graphische Darstellung der Observablen im Faktorraum vor der Rotation zeigt, daß offensichtlich keine Einfachstruktur vorliegt, da zum Beispiel die Variable 3 (V224) sowohl bezüglich des ersten als auch des zweiten Faktors eine hohe positive Faktorladung besitzt[8]. Nach Durchführung der VARIMAX-Rotation gegen den Uhrzeigersinn mit Hilfe der angegebenen Transformationsmatrix $\boldsymbol{T}$ zeigt sich hier eine deutliche Verbesserung, indem jede Observablengruppe stärker an eine Koordinatenachse herangerückt ist und sich jeweils von der anderen weiter entfernt hat. Die bessere Trennung läßt sich auch direkt an den Veränderungen bei den Faktorladungen ablesen (z.B Variable 1 (V222): $(0.65929, 0.48454) \longrightarrow (0.81818, -0.00636)$).

```
* Auszug aus: FACTOR      /VARIABLES V222 TO V231 /CRITERIA FACTORS(10)
*                         /ROTATION NOROTATE /PRINT EXTRACTION INITIAL
*                         /PLOT EIGEN ROTATION(1,2).
```



Symbol	Variable	Coordinates	
1	V222	.659	.485
2	V223	-.663	.542
3	V224	.611	.455
4	V225	.856	.105
5	V226	-.552	.254
6	V227	-.620	.603
7	V228	.777	.208
8	V229	.780	.121
9	V230	-.494	.690
10	V231	.713	.378

[8] Einige Observablen tauchen in der Blockgraphik nicht auf, z.B. 1 und 2 in der ersten Darstellung, die von 3 bzw. 5 überlagert werden. Dies liegt an der geringen Auflösung dieser Darstellungsform.

```
* Auszug aus: FACTOR      /VARIABLES V222 TO V231
*                         /PRINT EXTRACTION INITIAL ROTATION FSCORE
*                         /PLOT ROTATION(1,2).

Factor Matrix:

              FACTOR  1     FACTOR  2

V222            .65929        .48454
V223           -.66284        .54161
V224            .61148        .45474
V225            .85626        .10544
V226           -.55227        .25383
V227           -.61964        .60310
V228            .77662        .20798
V229            .77988        .12126
V230           -.49355        .68987
V231            .71278        .37806

Final Statistics:
Variable     Communality  *  Factor   Eigenvalue   Pct of Var   Cum Pct
V222             .66945  *     1       4.63308       46.3         46.3
V223             .73270  *     2       1.85099       18.5         64.8
V224             .58070  *
V225             .74430  *
V226             .36943  *
V227             .74768  *
V228             .64640  *
V229             .62292  *
V230             .71951  *
V231             .65098  *
Varimax  Rotation  1,  Extraction  1,  Analysis  1 - Kaiser Normalization.
 Varimax converged in    3 iterations.

Rotated Factor Matrix:
              FACTOR  1     FACTOR  2
V222            .81818       -.00636
V223           -.20691        .83059
V224            .76204       -.00163
V225            .74910       -.42796
V226           -.29055        .53387
V227           -.13550        .85400
V228            .74666       -.29815
V229            .69738       -.36958
V230            .01745        .84806
V231            .79730       -.12368
```

```
Factor Transformation Matrix:
                FACTOR  1     FACTOR  2
FACTOR  1         .80116        -.59845
FACTOR  2         .59845         .80116
```

```
Horizontal Factor  1   Vertical Factor  2
                       |
                  6   9|
                       |
                       |
                5      |
                       |
                       |
                       |
                       |
                       |
                       |
----------------------+----------------------
                       |               3 1
                       |                10
                       |
                       |               8
                       |               4
                       |
                       |
                       |
                       |
                       |
```

Symbol	Variable	Coordinates	
1	V222	.818	-.006
2	V223	-.207	.831
3	V224	.762	-.002
4	V225	.749	-.428
5	V226	-.291	.534
6	V227	-.136	.854
7	V228	.747	-.298
8	V229	.697	-.370
9	V230	.017	.848
10	V231	.797	-.124

6.3.6 Die Interpretation von Faktoren

Das Ergebnis der Faktorextraktion — und eventuell anschließender Rotation — ist eine Beschreibung der Beziehungsstruktur zwischen standardisierten Observablen z_j und theoretischen Faktoren f_r durch die Faktorladungsmatrix $\boldsymbol{A}$. Als wesentliche Aufgabe bleibt die Interpretation dieser funktionalen Abhängigkeiten, d.h. die Festlegung der inhaltlichen Bedeutung der Faktoren unter Berücksichtigung der ermittelten Faktorladungen, und die Überprüfung dieser Faktoren auf Übereinstimmung mit den theoretischen Grundannahmen.

Bei der im allgemeinen schwierigen Aufgabe der Interpretation von Faktoren hat sich das Prinzip der *Leitvariablen* durchgesetzt, d.h. die besondere Berücksichtigung von Observablen mit hohen Faktorladungen bei der Namensgebung von Faktoren. Im Sympathie-Beispiel sind aufgrund der klaren Trennung von Variablengruppen und des Vorwissens aus der Analyse von Parteisympathieskalometern die Interpretationsprobleme nicht groß. Es wird deutlich, daß der auf Parteiebene identifizierte bipolare Faktor sich auf der Politikerebene in zwei Faktoren aufspaltet:

f_1 : Sympathie Regierung
f_2 : Sympathie Opposition

```
* Auszug aus: FACTOR      /VARIABLES V222 TO V231
*                         /PRINT EXTRACTION INITIAL ROTATION FSCORE
*                         /PLOT ROTATION(1,2).
Horizontal Factor  1   Vertical Factor  2    Symbol Variable   Coordinates
                     |                          1   V222       .818  -.006
                  6  9                          2   V223      -.207   .831
                     |                          3   V224       .762  -.002
                     |                          4   V225       .749  -.428
                5    |                          5   V226      -.291   .534
                     |                          6   V227      -.136   .854
                     |                          7   V228       .747  -.298
                     |                          8   V229       .697  -.370
                     |                          9   V230       .017   .848
                     |                         10   V231       .797  -.124
---------------------+---------------------
                     |              3 1
                     |               10
                     |
                     |              8
                     |              4
                     |
                     |
                     |
                     |
                     |
```

6.3.7 Die Berechnung der Faktorwerte

Ausgehend von der Gleichung $\boldsymbol{Z} = \boldsymbol{F}\boldsymbol{A}$ und unter Verwendung des Fundamentaltheorems der Faktorenanalyse wurde die Faktorladungsmatrix $\boldsymbol{A}$ berechnet, mit deren Hilfe die Stärke der Beziehungen zwischen Observablen und Faktoren ermittelt und eine inhaltliche Beschreibung der Faktoren durchgeführt werden konnte. Die damit erreichte Strukturierung und Reduzierung der Observablen ist in den meisten Fällen bereits das eigentliche Ziel der Anwendung von Faktorenanalysen. Prinzipiell ist es jetzt aber auch möglich, die Werte der Matrix $\boldsymbol{F}$, d.h. die *Faktorwerte* der theoretischen Variablen zu ermitteln. Diese explizite Bestimmung von Faktorwerten ist in folgenden Fällen sinvoll:

- Vereinfachte Beschreibung einer Analyseeinheit durch einen Faktorwert anstelle von vielen Observablenwerten.
- Konstruktvalidierung, d.h. Durchführung von Tests, ob eine Observable den Bedeutungsgehalt eines Faktors mißt.
- Weiterverwendung von Faktorwerten anstatt der Observablenwerte in nachfolgenden Analysen (z.B. als linear unabhängige Variablen in einer Regressionsanalyse).

Grundsätzlich ergeben sich für die Ermittlung von $\boldsymbol{F}$ folgende Möglichkeiten:

1. *Direkte Berechnung*

   Im Fall der Hauptkomponenten-/Hauptachsenanalyse, bei der genausoviele Faktoren wie Observablen erzeugt werden, ist $\boldsymbol{A}$ quadratisch. Falls $\boldsymbol{A}$ zusätzlich invertierbar ist, läßt sich aus der Ausgangsgleichung eine Bestimmungsgleichung für $\boldsymbol{F}$ ableiten:

   $$\begin{aligned} \boldsymbol{Z} &= \boldsymbol{F}\boldsymbol{A} \\ \boldsymbol{Z}\boldsymbol{A}^{-1} &= \boldsymbol{F}\underbrace{\boldsymbol{A}\boldsymbol{A}^{-1}}_{=\boldsymbol{E_n}} \end{aligned}$$

   Damit gilt:

   $$\boldsymbol{F} = \boldsymbol{Z}\boldsymbol{A}^{-1}$$

2. *Schätzung* (z.B. durch multiple Regression)

   Im Fall von r gemeinsamen Faktoren $(r < m)$ lassen sich die Faktorenwerte, d.h. die Elemente einer Matrix $\hat{\boldsymbol{F}}$, schätzen:

   $$\hat{\boldsymbol{F}} = \boldsymbol{Z}\boldsymbol{B}$$

   Gesucht wird $\boldsymbol{B}$ als $(m \times r)$-dimensionale Matrix von standardisierten Regressionskoeffizienten für r Faktoren. Nach den Ausführungen zur multiplen Regressionsanalyse (vgl. Kapitel 6.2.2) gilt für die Berechnung von Regressionskoeffizienten:

   $$\boldsymbol{b} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

   Daraus folgt für $\boldsymbol{B}$:

   $$\begin{aligned} \boldsymbol{B} &= (\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{F} \\ &= m(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\frac{1}{m}\boldsymbol{Z}^T\boldsymbol{F} \\ &= \underbrace{(\frac{1}{m}\boldsymbol{Z}^T\boldsymbol{Z})^{-1}}_{=\boldsymbol{R}^{-1}}\ \underbrace{\frac{1}{m}\boldsymbol{Z}^T\boldsymbol{F}}_{=\boldsymbol{R}_{ZF}} \end{aligned}$$

Somit ergibt sich insgesamt:

$$\begin{aligned}\hat{\boldsymbol{F}} &= \boldsymbol{ZB} \\ &= \boldsymbol{ZR}^{-1}\boldsymbol{R}_{ZF}\end{aligned}$$

$\boldsymbol{R}_{ZF}$ bezeichnet dabei die Korrelationsmatrix zwischen den Observablen z_j und den Faktoren f_r, für die gilt:

- $\boldsymbol{R}_{ZF} = \boldsymbol{A}$ $(= \boldsymbol{V}_{fp}$ (*factor pattern*)),
  falls die Faktoren unkorreliert sind
- $\boldsymbol{R}_{ZF} = \boldsymbol{V}_{fp} \cdot \boldsymbol{C}_f$ $(= \boldsymbol{V}_{fs}$ (*factor structure*)),
  falls die Faktoren untereinander korreliert sind (z.B. nach einer schiefwinkligen Rotation) mit $\boldsymbol{C}_f$ als Korrelationsmatrix zwischen den Faktoren

Nachfolgende SPSS-Ausgabe enthält für das bekannte Beispiel die Matrix der Regressionskoeffizienten $\boldsymbol{B}$, die, multipliziert mit den standardisierten Observablen $\boldsymbol{Z}$, die Faktorwerte für jede Analyseeinheit ergibt ($\boldsymbol{F} = \boldsymbol{ZB}$).

```
* Auszug aus: FACTOR      /VARIABLES V222 TO V231
*                         /PRINT EXTRACTION INITIAL ROTATION FSCORE
*                         /PLOT ROTATION(1,2).

Factor Score Coefficient Matrix:

              FACTOR  1     FACTOR  2

V222            .27067        .12456
V223            .06049        .32004
V224            .25276        .11784
V225            .18216       -.06497
V226           -.01343        .18120
V227            .08784        .34107
V228            .20154       -.01030
V229            .17406       -.04825
V230            .13770        .36234
V231            .24549        .07156
```

6.4 Diskriminanzanalyse

6.4.1 Problemstellung

Aufgabe der Diskriminanzanalyse ist es, verschiedene Aussagen über Beziehungen zwischen einer nominal skalierten und einer oder mehreren intervallskalierten Variablen zu treffen. Aussagen dieser Art antworten auf eine der folgenden Fragen:

- Wie genau läßt sich aus der Kenntnis mehrerer quantitativer (metrischer) Merkmale auf ein qualitatives Merkmal schließen? Mit welcher Sicherheit läßt sich aus der Kenntnis metrischer Merkmale einer Untersuchungseinheit auf die Zugehörigkeit zu einer Gruppe von Untersuchungseinheiten schließen?
- In welchem Ausmaß tragen einzelne quantitative Merkmale zur Genauigkeit der Vorhersage der Gruppenzugehörigkeit bei?
- Angenommen, ein Teil der Untersuchungseinheiten sei sowohl hinsichtlich seiner Gruppenzugehörigkeit (seines qualitativen Merkmals) als auch hinsichtlich seiner quantitativen Merkmale bestimmt: zu welcher Gruppe gehört eine Untersuchungseinheit, von der nur die quantitativen Merkmale bekannt sind?

Am Beispiel von zwei Gruppen und zwei intervallskalierten Variablen sei zunächst das Problem graphisch veranschaulicht (Abb. 6.17):[9]

Offensichtlich gibt es in Abb. 6.17 — einem Streuungsdiagramm, das von den Skalometern für SPD und FDP aufgespannt wird — ein Gebiet, in dem Angehörige der Gruppe 1 (CDU-Wähler) vorherrschen, und eines, in dem Angehörige der Gruppe 2 (SPD-Wähler) vorherrschen.

[9] Abbildung und Berechnung der Mittelwerte (im Beispiel auf Seite 190) beruhen auf einer Skala mit dem Wertebereich 1 bis 11, obwohl die Skalometerfragen im Bereich -5 bis $+5$ zu beantworten waren.

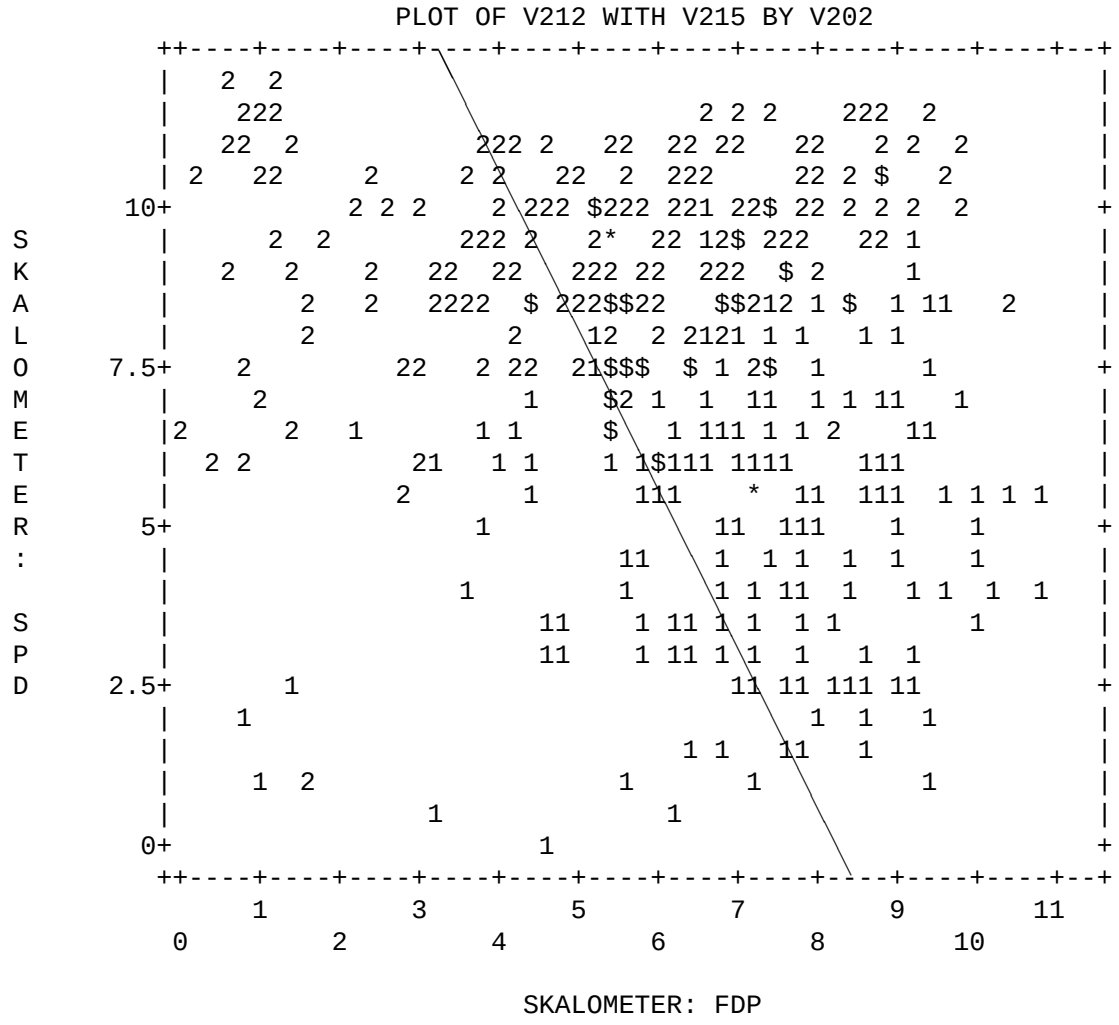


Abb. 6.17: CDU- und SPD-Wähler im Streuungsdiagramm aus FDP- und SPD-Sympathie. CDU-Wähler sind mit einer 1, SPD-Wähler sind mit einer 2 markiert, * steht für die beiden Mittelwerte, $ für Positionen, die von CDU- und SPD-Wählern besetzt sind. Die Verbindungslinie zwischen den beiden Mittelwerten gibt die erste Linearkombination an.

6.4.2 Verfahren

Zur Untersuchung des Problems geht die klassische Diskriminanzanalyse ([Klecka 1980], [Nie/Hull/u.a. 1975, S. 434–467]) davon aus, daß die beiden Gruppen multinormalverteilt mit identischer Kovarianzmatrix, aber verschiedenen Mittelwertvektoren sind, auch wenn dieses Modell in der Regel von den Daten nicht vollständig erfüllt ist.

Das folgende Beispiel nutzt alle verfügbaren Skalometerfragen und alle Befragten, die angegeben haben, CDU, SPD oder FDP wählen zu wollen, zur Beantwortung der Frage, wie genau die Kenntnis der Sympathieskalometer die Antwort auf die Sonntagsfrage vorhersagt:

```
* SPSS-Aufruf: DSCRIMINANT /GROUPS V202 (1,3)
*                          /VARIABLES V212 TO V231
*                          /ANALYSIS V212 TO V231
*                          /METHOD DIRECT
*                          /STATISTICS 1 2 10 13 15 16.

Number of Cases by Group

               Number of Cases
  V202      Unweighted     Weighted  Label
         1         176        176.0  CDU
         2         203        203.0  SPD
         3          24         24.0  FDP
     Total         403        403.0

Group means

  V202           V212 SPD       V213 CDU       V214 CSU       V215 FDP
         1        5.42045        9.47727        8.82386        7.21023
         2        9.39901        5.56158        4.55665        5.40887
         3        6.45833        7.45833        6.33333        8.33333
     Total        7.48635        7.38462        6.52605        6.36973

  V202           V220           V221           V222           V223 Brandt
         1        5.39205        3.44318        6.54545        4.35227
         2        8.53202        6.02463        5.10837        8.02463
         3        6.12500        4.16667        7.37500        5.37500
     Total        7.01737        4.78660        5.87097        6.26303

  V202           V224 Genscher  V225 Kohl      V226           V227
         1        7.86932        9.21591        3.32386        5.36932
         2        6.38916        4.83744        5.45813        9.16749
         3        8.70833        6.79167        4.50000        6.16667
     Total        7.17370        6.86600        4.46898        7.33002
```

Die erste Phase der Diskriminanzanalyse besteht darin, den Satz von unabhängigen Variablen — ähnlich wie in der Faktorenanalyse — durch eine ausreichende Anzahl neuer, untereinander nicht korrelierter Variablen, von sogenannten kanonischen Diskriminanzfunktionen, zu ersetzen. Aus der Anzahl der Ausprägungen der abhängigen, nominal skalierten Variablen ergibt sich die Anzahl der kanonischen Diskriminanzfunktionen: zwei Gruppen können durch eine Funktion, g Gruppen durch $g-1$ Funktionen getrennt werden. "Trennung" bedeutet hier die Varianz-Maximierung der Gruppenmittelwerte der Diskriminanzfunktionen unter der Nebenbedingung, daß alle Diskriminanzfunktionen Linearkombinationen aus den unabhängigen Variablen und untereinander unkorreliert sind. Für den Zwei-Gruppen-Fall kann "Trennung" auch verstanden werden als der Abstand zwischen den Mittelwerten der beiden Gruppen; die erste kanonische Diskriminanzfunktion entspricht damit der Richtung, die durch die beiden in der obigen Abbildung mit * angegebenen Gruppenmittelpunkte bestimmt ist. Für den Mehr-Gruppen-Fall reicht diese Definition offenbar nicht aus, an die Stelle des Mittelpunktabstands tritt daher die Varianz der Gruppenmittelwerte in der jeweiligen kanonischen Diskriminanzfunktion.

Zur Auffindung der ersten kanonischen Diskriminanzfunktion ist daher zunächst hypothetisch eine Linearkombination aller Variablen zu unterstellen und sodann in Abhängigkeit von den Koeffizienten dieser Linearkombination die Varianz ihrer Gruppenmittelwerte zu maximieren. Hierfür seien folgende Bezeichnungen eingeführt:

x_{ikm} : Wert der i-ten Variablen ($i = 1, \ldots, p$), gemessen an der m-ten Untersuchungseinheit ($m = 1, \ldots, n_k$) in der k-ten Gruppe ($k = 1, \ldots, g$), wobei $\sum_k n_k = n_\bullet$

$x_{ik\bullet}$: Mittelwert der i-ten Variablen in der k-ten Gruppe

$x_{i\bullet\bullet}$: Mittelwert der i-ten Variablen über alle Untersuchungseinheiten, d.h. ohne Ansehen ihrer Gruppenzugehörigkeiten

Für die Linearkombinationen wird der Buchstabe y verwendet; der (erste) Index j ($j = 1, \ldots, q$; anstelle von i) bezeichnet dann die j-te Linearkombination. Die Bedeutungen von y_{jkm}, $y_{jk\bullet}$ und $y_{j\bullet\bullet}$ sind den entsprechenden für x analog. Die Koeffizienten der j-ten Linearkombination werden mit u_{ji} bezeichnet. Dann gilt für den Wert der j-ten Linearkombination des m-ten Falles in der k-ten Gruppe:

$$y_{jkm} = u_{j0} + u_{j1}x_{1km} + u_{j2}x_{2km} + \ldots + u_{jp}x_{pkm}$$

Zur Vereinfachung der Ableitung wird von den Linearkombinationen (nicht aber von den Variablen) gefordert, daß sie wie folgt standardisiert seien:

$$\begin{aligned} y_{j\bullet\bullet} &= 0 \\ \frac{1}{n_\bullet - g} \sum_{k=1}^{g} \sum_{m=1}^{n_k} (y_{jkm} - y_{jk\bullet})^2 &= 1 \end{aligned}$$

Jede Linearkombination soll also über alle Gruppen den Mittelwert Null und innerhalb der Gruppen die Varianz Eins haben.

$$\begin{aligned} y_{j\bullet\bullet} = 0 &= \frac{1}{n} \sum_{k=1}^{g} \sum_{m=1}^{n_k} y_{jkm} \\ 0 &= \frac{1}{n} \sum_{k=1}^{g} \sum_{m=1}^{n_k} (u_{j0} + u_{j1}x_{1km} + u_{j2}x_{2km} + \ldots + u_{jp}x_{pkm}) \\ &= \frac{1}{n} \sum_{k=1}^{g} \sum_{m=1}^{n_k} \left(u_{j0} + \sum_{i=1}^{p} u_{ji}x_{ikm} \right) \\ \frac{1}{n} \sum_{k=1}^{g} \sum_{m=1}^{n_k} u_{j0} &= -\sum_{i=1}^{p} u_{ji} \left(\frac{1}{n} \sum_{k=1}^{g} \sum_{m=1}^{n_k} x_{ikm} \right) \\ u_{j0} &= -\sum_{i=1}^{p} u_{ji}x_{i\bullet\bullet} \\ y_{jkm} &= \sum_{i=1}^{p} u_{ji}(x_{ikm} - x_{i\bullet\bullet}) \end{aligned}$$

Damit ist

$$y_{jk\bullet} = \sum_{i=1}^{p} (x_{ik\bullet} - x_{i\bullet\bullet})u_{ji}$$

Die zu maximierenden Varianzen der Gruppenmittelwerte der kanonischen Diskriminanzfunktionen (die mit den Gruppengrößen gewichtet werden) betragen

$$s^2_{y_{jk\bullet}} = \frac{1}{n_\bullet - g} \sum_{k=1}^{g} n_k (y_{jk\bullet} - y_{j\bullet\bullet})^2$$

Wegen $y_{j\bullet\bullet} = 0$ (siehe oben) ist also folgende Funktion der Koeffizienten der Linearkombination zu maximieren:

$$V(\boldsymbol{u}) = \sum_{k=1}^{g} n_k y^2_{1k\bullet} \stackrel{!}{=} \textit{Maximum}$$

Unter Verwendung der Matrizenschreibweise ist dies gleichbedeutend mit

$$V = \boldsymbol{u}^T(\boldsymbol{X}_k - \boldsymbol{X}_\bullet^T)(\boldsymbol{X}_k - \boldsymbol{X}_\bullet)\boldsymbol{u}$$

Darin ist

$\boldsymbol{X}_k$	eine Matrix, die in jeder Zeile für einen Fall anstelle seiner Variablenwerte die zugehörigen *Gruppen*mittelwerte enthält, und
$\boldsymbol{X}_\bullet$	eine Matrix, die in jeder Zeile für einen Fall anstelle seiner Variablenwerte die zugehörigen *Gesamt*mittelwerte enthält;
$\boldsymbol{X}_k - \boldsymbol{X}_\bullet$	enthält also für jeden Fall die Differenz zwischen Gruppen- und Gesamtmittelwert.

Die Matrix $(\boldsymbol{X}_k - \boldsymbol{X}_\bullet)^T(\boldsymbol{X}_k - \boldsymbol{X}_\bullet) = \boldsymbol{B}$ wird als *between-groups sums of squares and crossproducts matrix* bezeichnet. Zu maximieren ist also

$$V = \boldsymbol{u}^T\boldsymbol{B}\boldsymbol{u}$$

Die Diskriminanzfunktionen sollen (vgl. Vorseite) standardisiert werden, so daß ihre intra-Gruppen-Varianz gleich 1 ist, d.h. es soll gelten:

$$\sum_{k=1}^{g}\sum_{m=1}^{n_k}(y_{jkm} - y_{jk\bullet})^2 = n_\bullet - g$$

Mit den letzten Ergebnissen für y_{jkm} und $y_{jk\bullet}$ läßt sich der Klammerausdruck dieser Gleichung als

$$y_{jkm} - y_{jk\bullet} = \sum_{i=1}^{p}(x_{ikm} - x_{ik\bullet})u_{ji}$$

und die Summe der Quadrate somit in Matrizenschreibweise als

$$\sum_{k=1}^{g}\sum_{m=1}^{n_k}(y_{jkm} - y_{jk\bullet})^2 = \boldsymbol{u}^T(\boldsymbol{X} - \boldsymbol{X}_k)^T(\boldsymbol{X} - \boldsymbol{X}_k)\boldsymbol{u}$$

darstellen.

Darin ist $\boldsymbol{X}$ die ursprüngliche Datenmatrix. Die Matrix $(\boldsymbol{X} - \boldsymbol{X}_k)^T(\boldsymbol{X} - \boldsymbol{X}_k) = \boldsymbol{W}$ wird als *within-groups sums of squares and crossproducts matrix* bezeichnet. Damit lautet die Nebenbedingung

$$n_\bullet - g - \boldsymbol{u}^T\boldsymbol{W}\boldsymbol{u} = 0$$

$\boldsymbol{u}$ errechnet sich aus dem Gleichungssystem

$$\frac{\partial H}{\partial \boldsymbol{u}} = 0$$

mit

$$H = \boldsymbol{u}^T \boldsymbol{B} \boldsymbol{u} + \lambda(n_\bullet - g - \boldsymbol{u}^T \boldsymbol{W} \boldsymbol{u})$$

Eingesetzt ergibt dies:

$$\frac{\partial H}{\partial \boldsymbol{u}} = 2\boldsymbol{B}\boldsymbol{u} - 2\lambda \boldsymbol{W}\boldsymbol{u} = 0$$

Nach Vereinfachung zu

$$\boldsymbol{B}\boldsymbol{u} = \lambda \boldsymbol{W}\boldsymbol{u}$$

stellt sich wiederum ein Eigenwertproblem, das mit Standardverfahren lösbar ist. Die letzte Gleichung hat $q = \min(p, (g-1))$ — in der Regel voneinander verschiedene — Lösungen. Dabei sind $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_q$ die zueinander orthogonalen Vektoren der Koeffizienten der kanonischen Diskriminanzfunktionen und $\lambda_1, \lambda_2, \ldots, \lambda_q$ die zugehörigen Eigenwerte. Aus den Koeffizientenvektoren ergibt sich (allerdings erst nach Standardisierung bzw. gerade dann, wenn die Datenmatrix $\boldsymbol{X}$ standardisierte Variablen enthält), welche Variablen zu welchen kanonischen Diskriminanzfunktionen beitragen, während aus den Eigenwerten die Diskriminationskapazität der zugrundeliegenden Variablen folgt.

Damit lassen sich für jeden Fall die Werte in den Linearkombinationen berechnen und insbesondere auch graphisch darstellen (vgl. Abb. 6.18 bis 6.22). Außerdem können die Gruppenmittelwerte der Linearkombinationen ausgerechnet werden:

```
Canonical Discriminant Functions evaluated at Group Means(Group Centroids)

   Group      FUNC   1    FUNC   2
       1       1.62852      .13854
       2      -1.49193      .05272
       3        .67675    -1.46188
```

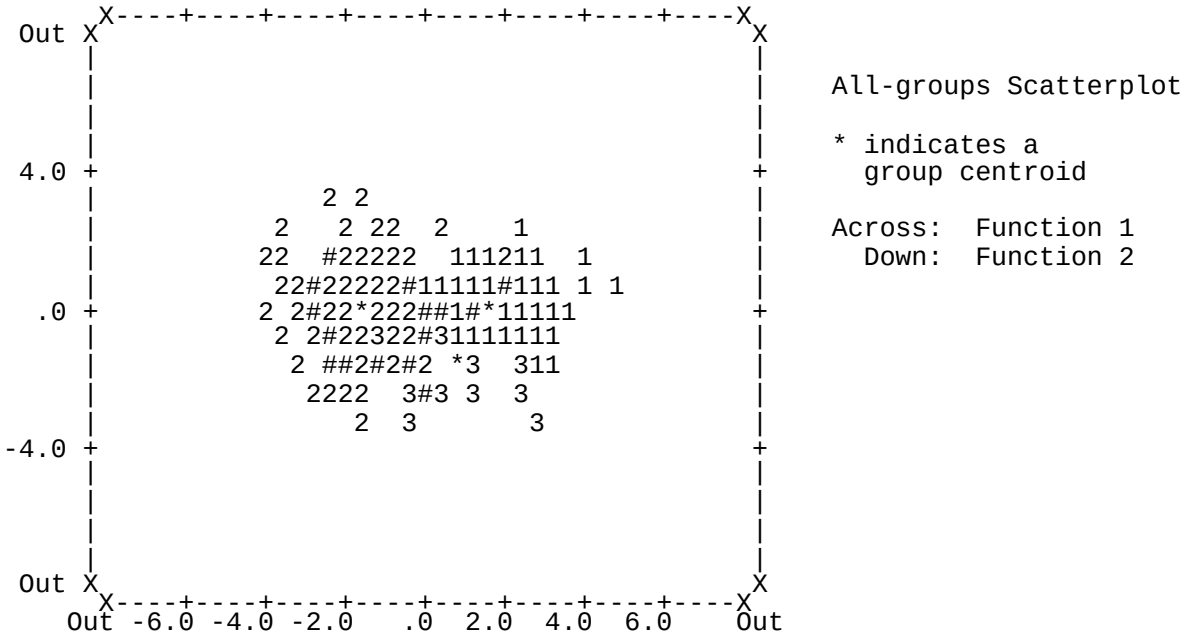


Abb. 6.18: “All-groups Scatterplot”

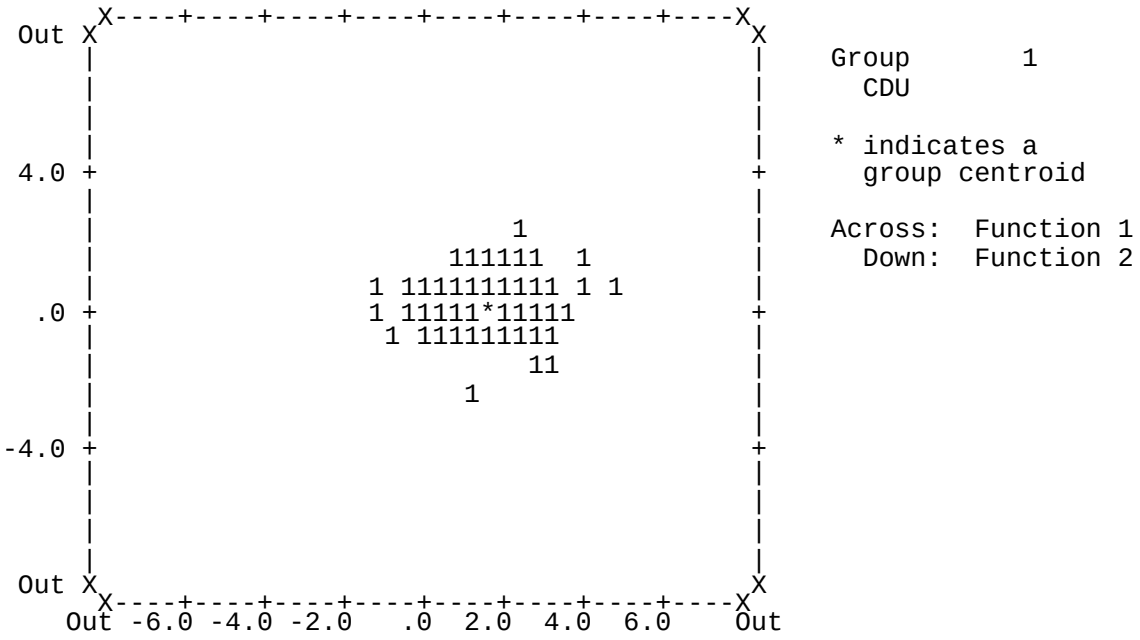


Abb. 6.19: Streuungsdiagramm für die CDU-Wähler

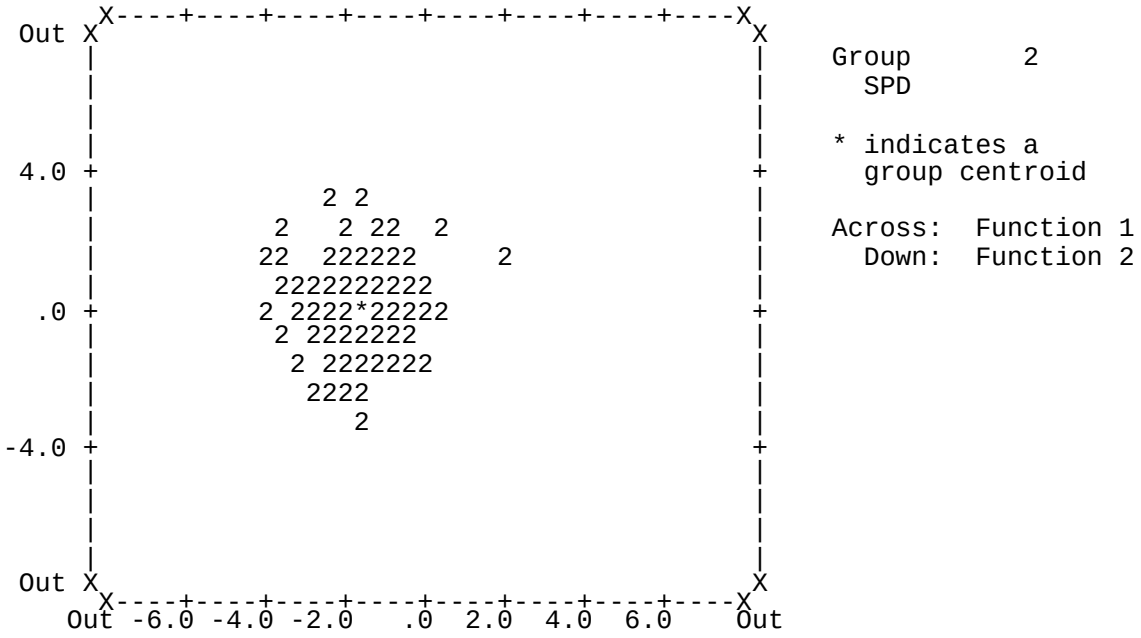


Abb. 6.20: Streuungsdiagramm für die SPD-Wähler

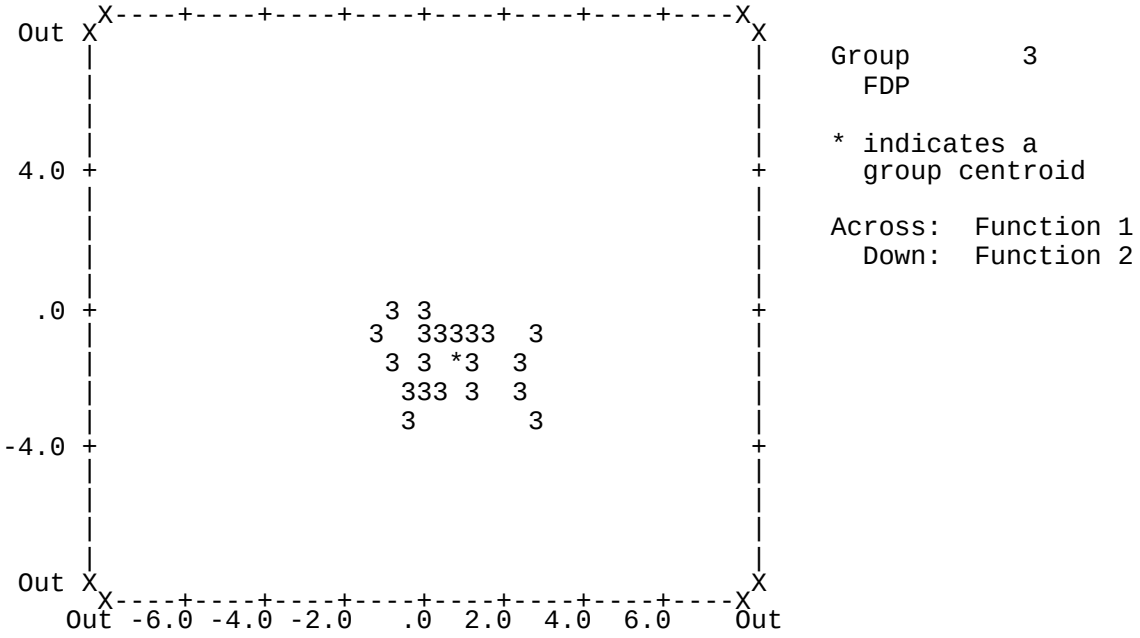


Abb. 6.21: Streuungsdiagramm für die FDP-Wähler

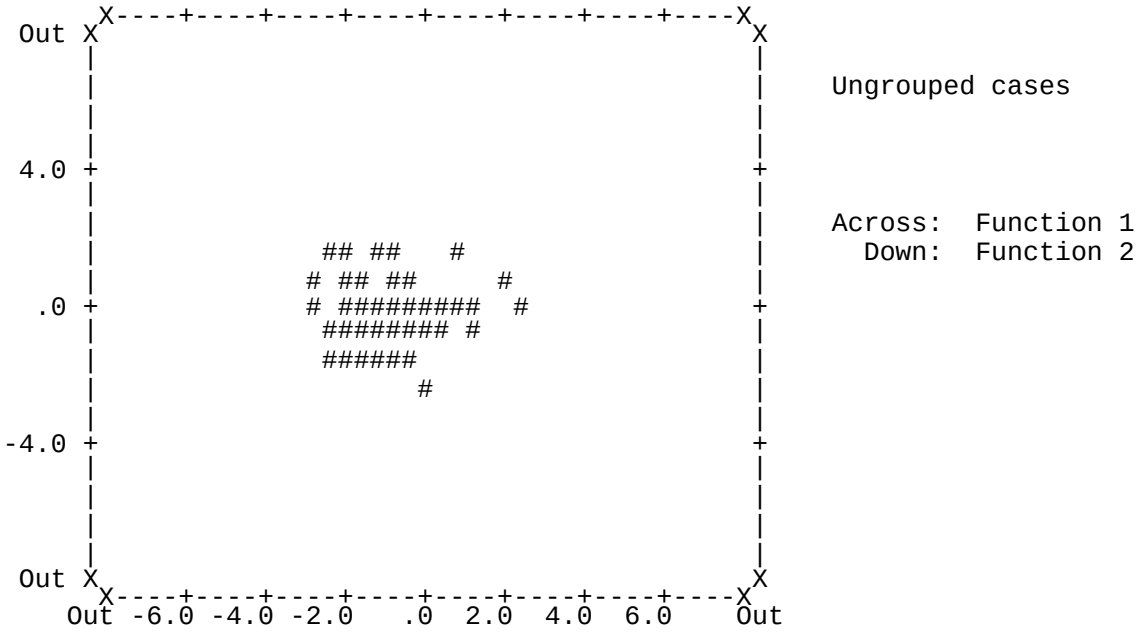


Abb. 6.22: Streuungsdiagramm für die Befragten mit unbekannter Wahlentscheidung

Zur Darstellung der Diskriminationskapazität wird *Wilks' Lambda* verwendet, das wie folgt definiert ist:

$$\Lambda_k = \prod_{i=k+1}^{q} \frac{1}{1+\lambda_i} \qquad 0 < \Lambda_k < 1$$

Dabei ist k die Anzahl der bereits (in der Reihenfolge der Größe ihrer Eigenwerte) berechneten kanonischen Diskriminanzfunktionen und Lambda ein Maß für die in den Variablen noch verbleibende Diskriminationskapazität. Insbesondere für $k = 0$ zeigt Λ die gesamte Diskriminationskapazität, und zwar ist Lambda umso größer (näher an 1.0), je geringer die Diskriminationskapazität ist, und umso kleiner (näher an 0.0), je größer sie ist. Damit ist die erste der drei in der Problemstellung (Seite 188) genannten Fragen beantwortet: es ist mit Wilks' Lambda ein skalares Maß für die Diskriminationskapazität gefunden (ein weiteres, anschaulicheres läßt sich nach Beantwortung der dritten Frage gewinnen).

Canonical Discriminant Functions

Fcn	Eigenvalue	Pct of Variance	Cum Pct	Canonical Corr	After Fcn	Wilks' Lambda	Chisquare	DF	Sig
					0	.2643	519.567	40	.0000
1*	2.3240	94.39	94.39	.8362	1	.8787	50.509	19	.0001
2*	.1381	5.61	100.00	.3483					

* marks the 2 canonical discriminant functions remaining in the analysis.

6.4.3 Beurteilung der einzelnen Variablen

Für die Beantwortung der zweiten Frage läßt sich die gleiche Analyse für jede einzelne Variable durchführen; auf diese Weise läßt sich jeder einzelnen Variablen ein Wilks' Lambda zuordnen; umgekehrt kann auch Wilks' Lambda für alle Variablen mit Wilks' Lambda für alle bis auf eine Variable verglichen werden, um so den marginalen Beitrag jeder einzelnen Variablen zur Gruppentrennung zu ermitteln.

Außerdem lassen sich die Korrelationen zwischen den Variablen und den Linearkombinationen berechnen:

```
Structure Matrix:

Pooled-within-groups correlations between discriminating variables
                                    and canonical discriminant functions
(Variables ordered by size of correlation within function)

             FUNC  1    FUNC  2
V212         -.69376*    .06056
V225          .61725*    .30842
V213          .60967*    .25745
V217          .59263*    .15546
V214          .58623*    .32877
V218          .58034*    .29889
V227         -.58021*    .10879
V223         -.53125*    .02788
V220         -.52181*    .07375
V230         -.44385*    .06111
V229          .42570*    .21021
V228          .38915*   -.17019
V221         -.34519*    .01691
V231          .30409*   -.25461
V216         -.27871*   -.06339
V226         -.26596*   -.13148
V215          .29337    -.51391*
V224          .24441    -.40322*
V219          .36946    -.39122*
V222          .21834    -.36431*
```

Die vorstehende Tabelle enthält die Korrelationen zwischen den Linearkombinationen und den Skalometern. Sie sind unerläßlich, um zu ermitteln, worin sich die Gruppen unterscheiden. Die erste — in den Bildern 6.18–6.22 waagerecht dargestellte — Linearkombination korreliert auffällig hoch negativ mit dem SPD-Skalometer (V212) und positiv mit dem Kohl-Skalometer (V225); hier scheint es

sich also um die “links”-“rechts”-Dimension zu handeln. Die zweite — in diesen Bildern senkrecht dargestellte — Linearkombination korreliert negativ mit den Skalometern für die FDP (V215) und für Genscher (V224); diese Dimension wäre vorläufig als “Anti-FDP-Dimension” zu bezeichnen.

6.4.4 Klassifikation

Für die dritte Frage muß zusätzlich zur Ableitung der kanonischen Diskriminanzfunktionen noch ein Klassifikationsverfahren beschrieben werden. Unterstellt war, daß jede einzelne Gruppe multinormalverteilt mit identischer Kovarianzmatrix und gruppenspezifischem Mittelwertvektor ist. Damit läßt sich (auch ohne kanonische Diskriminanzfunktionen) die Distanz jeder einzelnen Untersuchungseinheit zu allen Gruppenmittelpunkten (Zentroiden) ermitteln; unter Zuhilfenahme der kanonischen Diskriminanzfunktionen läßt sich diese Distanz auch unabhängig von den Skaleneinheiten der einzelnen Originalvariablen berechnen.

Im einfachsten Fall wird jede Untersuchungseinheit derjenigen Gruppe zugeordnet, zu der sie die geringste Distanz besitzt. Da nach Voraussetzung die kanonischen Diskriminanzfunktionen so berechnet worden sind, daß für alle Gruppen ihre Kovarianzmatrizen mit der Einheitsmatrix übereinstimmen, also sich in jeder Gruppe Varianzen auf Eins und Kovarianzen (Korrelationen) auf Null belaufen, sind für jede Gruppe die Distanzen der einzelnen Untersuchungseinheiten vom Gruppenmittelpunkt χ^2-verteilt mit q Freiheitsgraden. Für jede Untersuchungseinheit kann also mit Bezug auf jeden Gruppenmittelpunkt angegeben werden, wie groß die Wahrscheinlichkeit ist, Untersuchungseinheiten noch weiter entfernt anzutreffen. Diese Wahrscheinlichkeit wird üblicherweise verstanden als die Wahrscheinlichkeit, daß diese Untersuchungseinheit zur jeweiligen Gruppe gehört.

Der Gang der Rechnung wird anhand der ersten beiden Fälle des Datensatzes verfolgt,

```
Case   Mis        Actual   Highest Probability    2nd Highest     Discrim
Number  Val  Sel  Group    Group P(D/G) P(G/D)   Group P(G/D)     Scores
     1               1        1  .3599  .9774       3  .0168     1.6720
                                                                  1.5676
     2               2        2  .2161  .9910       3  .0089    -2.8357
                                                                 -1.0692
... ... ...
```

für die als Werte der beiden Linearkombinationen (“discriminant scores”)

Fall	Func 1	Func 2
1	1.6720	1.5676
2	−2.8357	−1.0692

berechnet wurden. Das Quadrat der Entfernung von den Gruppenmittelwerten beträgt

$$D_{km}^2 = \sum_{j=1}^{q} (y_{jkm} - y_{jk\bullet})^2$$

Die Entfernungen des Falles 1 zu den Mittelwerten der drei Gruppen betragen

Gruppe	D_{k1}	D_{k1}^2	$P(\chi^2 > D_{k1}^2)$	$P(G\|D)$
Gruppe 1	1.42972	2.04410	0.35986	0.9774
Gruppe 2	3.50789	12.30529	0.00213	0.0021
Gruppe 3	3.18877	10.16825	0.00619	0.0168
Summe			0.36818	1.0000

so daß — schon nach der Spalte D_{k1} — Fall 1 am ehesten zu Gruppe 1 zu gehören scheint (was auch tatsächlich nach dem Wahlverhalten laut Datensatz der Fall ist).

D_{km}^2 ist um den Gruppenmittelpunkt $(y_{1k\bullet}, \ldots, y_{qk\bullet})$ χ^2-verteilt mit q Freiheitsgraden, denn es handelt sich um — vgl. obige Gleichung — eine Summe von q Quadraten normalverteilter Zufallszahlen (vgl. hierzu auch Kapitel 5.2.1 und 5.2.2). Damit läßt sich die Wahrscheinlichkeit $P(\chi^2 > D^2)$ berechnen, daß Fälle gefunden werden, die noch weiter vom Gruppenmittelpunkt entfernt liegen als D — diese Wahrscheinlichkeit sei identifiziert mit der Wahrscheinlichkeit $P(D|G)$, der Wahrscheinlichkeit, bei dieser Distanz zur Gruppe zu gehören. Im allgemeinen werden sich diese Wahrscheinlichkeiten nicht zu 1 addieren; daher werden sie durch Division durch die Summe dieser Wahrscheinlichkeiten in die Wahrscheinlichkeiten $P(G|D)$ — die als Wahrscheinlichkeiten ansehbar sind, daß Fall 1 zu einer der drei Gruppen gehört — umgerechnet.

Die Zuweisung zu derjenigen Gruppe, bei der für eine Untersuchungseinheit diese Wahrscheinlichkeit am höchsten ist, stimmt mit einer auf Distanzbasis erfolgenden überein; die Wahrscheinlichkeitsklassifizierung kann jedoch zusätzlich noch an im voraus bekannte Zugehörigkeitswahrscheinlichkeiten angepaßt werden. Dazu werden die ursprünglich abgeleiteten Gruppenzugehörigkeitswahrscheinlichkeiten $P(D|G)$ mit der *a-priori*-Wahrscheinlichkeit, zu der betreffenden Gruppe zu gehören, multipliziert. Die Zuweisung erfolgt dann zu der Gruppe, bei der dieses Produkt am höchsten ist. Bei der Auswertung von *a-priori*-Wahrscheinlichkeiten wird die Zugehörigkeit zu großen Gruppen besser, die zu kleinen schlechter prädiziert.

Auf die gleiche Weise können Fälle, deren Gruppenzugehörigkeit nicht bekannt ist oder bei der Berechnung von Mittelwerten und kanonischen Diskriminanzfunktionen nicht berücksichtigt wurde, ebenfalls klassifiziert werden.

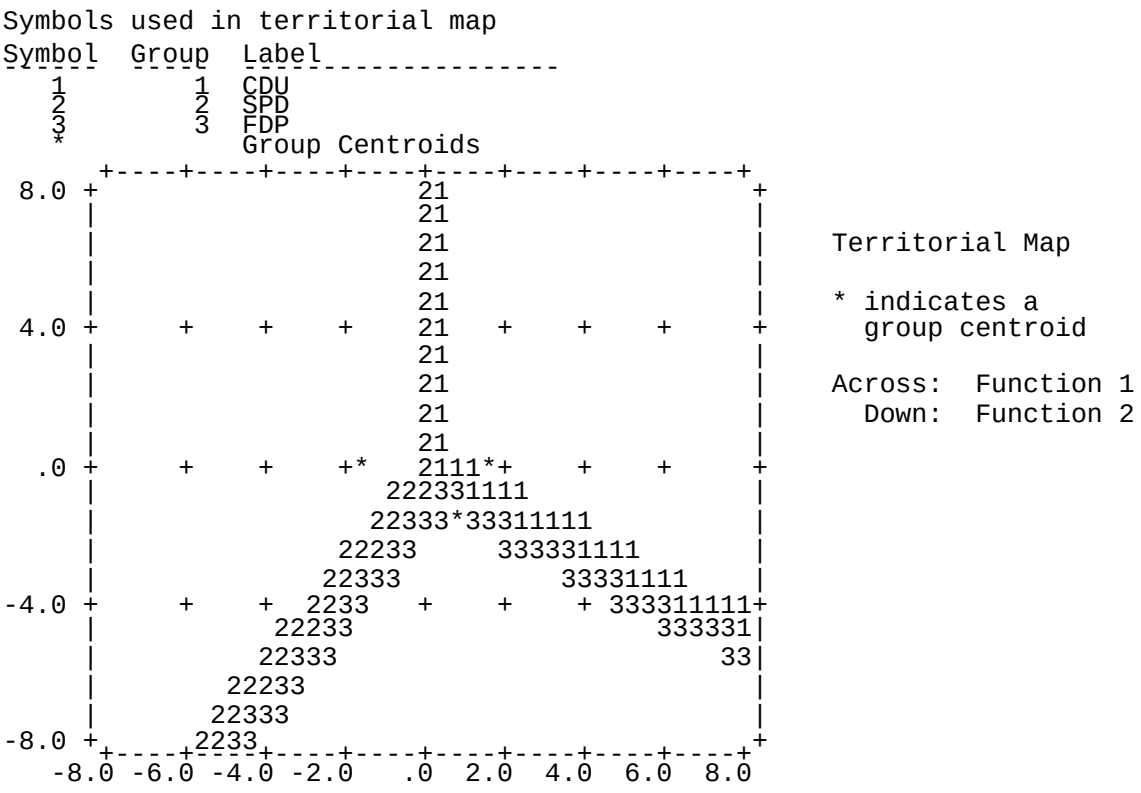


Abb. 6.23: “Territorial map”

Darüber hinaus ist es möglich, den von den kanonischen Diskriminanzfunktionen oder den von den Originalvariablen aufgespannten Raum in Gruppenterritorien zu gliedern (vgl. Abb. 6.23). Im Zwei-Variablen-Fall handelt es sich dabei um die Zerlegung der Ebene durch g Geraden, wobei wieder identische Kovarianzmatrizen der Originalvariablen in den einzelnen Gruppen unterstellt sind.

Auf der Basis der Klassifikation läßt sich ein weiteres Maß für die Güte der Gruppentrennung ableiten, das die Anzahl der Fehlklassifikationen zur Grundlage hat. Bei bekannten Gruppengrößen wird bei zufälliger “Klassifikation” ein Teil der Fälle richtig zugeordnet (dieser Anteil ergibt sich als Summe der Quadrate der *a-priori*-Wahrscheinlichkeiten). Wird die Zahl der durch Raten richtig zugeordneten Fälle sowohl von der Zahl der durch die Diskriminanzanalyse richtig klassifizierten als auch von der Zahl aller Fälle subtrahiert, und werden die beiden

Differenzen dividiert, so ist das Ergebnis das Assoziationsmaß τ von Goodman und Kruskal [Liebetrau 1983, S. 24–31], das gerade Null ist, wenn die Diskriminanzanalyse ein Ergebnis gebracht hat, welches ebenso schlecht ist, wie das Ergebnis schlichten Ratens, und das Eins ist, wenn die Diskriminanzanalyse alle Fälle richtig klassifiziert hat.

```
Classification Results -

                    No. of    Predicted Group Membership
   Actual Group      Cases          1          2          3
--------------------  ------   --------   --------   --------
Group       1           176        149         14         13
CDU                                84.7%       8.0%       7.4%

Group       2           203          4        188         11
SPD                                 2.0%      92.6%       5.4%

Group       3            24          3          3         18
FDP                                12.5%      12.5%      75.0%

Ungrouped Cases          70          8         46         16
                                   11.4%      65.7%      22.9%

Percent of "grouped" cases correctly classified:  88.09%
```

Im vorliegenden Fall wäre ohne Kenntnis der Skalometer die Einschätzung der Antwort der 403 Befragten auf die Sonntagsfrage

$$176 * \frac{176}{403} + 203 * \frac{203}{403} + 24 * \frac{24}{403} = 77 + 102 + 1 = 180$$

d.h. in 44.8% der Fälle, richtig gewesen. Mit Kenntnis der Skalometer und nach Durchführung der Diskriminanzanalyse hat sich die Anzahl der richtigen Einschätzungen auf $149 + 188 + 18$, d.h. auf 88.1% verbessert. Die anteilige Reduktion des Fehlers (PRE, vgl. Kapitel 5.2.3) beträgt also 0.784:

Prozentualer Fehler ohne Kenntnis	0.552
Prozentualer Fehler mit Kenntnis	0.119
Fehlerreduktion	0.433
Anteilige Fehlerreduktion	0.784

Die letzte Zeile der ungruppierten Fälle in der Klassifikationstabelle gibt an, wie sich die Befragten mit unbekannter Parteipräferenz vermutlich entscheiden

würden, wenn sie ihre Partei- und Politikersympathien auf die gleiche Weise in ihre Stimmabgabe umsetzen wie die Befragten mit bekannter Parteipräferenz.

Bei Nutzung der Information über die Gruppengrößen (*a-priori*-Wahrscheinlichkeiten) ergeben sich bei CDU und SPD 159 ($+10$) bzw. 200 ($+12$) richtige Zuordnungen, bei der FDP nur noch 7 (-11).

6.4.5 Kanonische Korrelation

Abschließend sei erwähnt, daß die Diskriminanzanalyse formal äquivalent ist mit einer kanonischen Korrelationsanalyse ([Levine 1977], [Gaensslen/Schubö 1976, S. 165–198]), bei der der eine Variablensatz die Menge der diskriminierenden Variablen, also wiederum der Variablen x_i ($i = 1, \ldots, p$) und der andere ein Satz von $(g-1)$ Dummy-Variablen $z_1, z_2, \ldots, z_{g-1}$ ist. Für den Fall m aus der Gruppe k ist $z_{kkm} = 1$ und sind alle $z_{lkm} = 0$ für $l \neq k$. Die z-Werte eines Falles aus der Gruppe g (der Gruppe mit der höchsten Gruppennummer) sind sämtlich Null, so daß mit den Werten z_{lkm} ($l = 1, \ldots, (g-1)$) von denen höchstens einer gleich Eins, alle anderen aber gleich Null sind, die Gruppenzugehörigkeit genau beschrieben ist. Aufgabe der kanonischen Korrelationsanalyse ist es nun, Linearkombinationen

$$s_{jkm} = \sum_{i=1}^{p} (z_{ikm} - z_{i\bullet\bullet}) v_i$$

und

$$t_{jkm} = \sum_{i=1}^{p} (x_{ikm} - x_{i\bullet\bullet}) u_i$$

so zu bestimmen, daß

- die Korrelation von s_1 und t_1 maximal wird und
- die partielle Korrelation von s_j und $t_j, j > 1$, unter rechnerischer Konstanthaltung aller $s_{j'}$ und $t_{j'}, j' < j$, maximal wird, wobei gleichzeitig alle Korrelationen zwischen s_j und $s_{j''}$, t_j und $t_{j''}$, s_j und $t_{j''}$ sowie $s_{j''}$ und $t_j, j'' < j$, verschwinden.

Danach ergibt sich als Korrelationsmatrix der kanonischen Variablen s_j und t_j jeweils miteinander die Einheitsmatrix und als Korrelationsmatrix der s_j mit den t_j eine Diagonalmatrix mit von links oben nach rechts unten fallenden sogenannten kanonischen Korrelationen, die die gesamte Information über die linearen Abhängigkeiten zwischen den beiden Originalvariablensätzen enthalten. Übliches

Maß für die Stärke des (linearen) Zusammenhangs insgesamt zwischen den beiden Variablensätzen ist wiederum Wilks' Lambda, das auf der Basis der kanonischen Korrelationen definiert ist als

$$\Lambda = \prod_{j=1}^{q}(1 - r_{cj}^2)$$

wobei r_{cj} für eine der q kanonischen Korrelationen steht und q die kleinere der beiden Variablenanzahlen in den beiden Variablensätzen ist.

Werden — wie oben beschrieben — z_i ($i = 1, \ldots, (g-1)$) als Dummy-Variablen zur Beschreibung der Gruppenzugehörigkeit gebildet, so sind die Linearkombinationen t_j ($j = 1, \ldots, q$) der kanonischen Analyse mit den kanonischen Diskriminanzfunktionen y_j der Diskriminanzanalyse identisch.

6.5 Clusteranalyse

6.5.1 Problemstellung

Die Clusteranalyse (in ihren verschiedenen Ausprägungen) hat zur Aufgabe, die Fälle (Objekte, Befragten) zu möglichst homogenen Gruppen (Clustern) zusammenzufassen. Hierzu muß zwischen den Fällen ein Ähnlichkeitsmaß (oder, in umgekehrter Betrachtungsweise: ein Distanzmaß) definiert sein, das es erlaubt, ähnliche Fälle bzw. Fälle mit geringer Distanz zu gruppieren, mit dem Ziel, daß die Gruppen untereinander möglichst unähnlich sind oder möglichst große Distanzen aufweisen.

6.5.2 Verfahren

Im wesentlichen werden hierarchische agglomerative und partitionierende Verfahren unterschieden.

Hierarchische agglomerative Verfahren beginnen bei den einzelnen Fällen und fassen zunächst die zwei ähnlichsten Fälle zusammen (agglomerativ) und fügen dann alle weiteren Fälle nach dem Maße ihrer Ähnlichkeit zu dem entstehenden Cluster hinzu, wobei eine Hierarchie der Zusammenfassungen entsteht, die sich in einer baumartigen Struktur (Dendrogramm) veranschaulichen läßt. Je nach dem, wie das Anfügen im einzelnen geschieht (“sorting strategy” [Aldenderfer/Blashfield 1984, S. 38]), wird differenziert:

- *single linkage:*
  Ein neuer Fall wird einem Cluster angefügt, wenn wenigstens ein Fall aus diesem Cluster auf demselben Ähnlichkeitsniveau steht wie der neu einzufügende Fall. Für die Beurteilung der Ähnlichkeit zwischen neuem Fall und Cluster wird das Cluster also durch denjenigen seiner Fälle repräsentiert, der dem einzufügenden Fall am ähnlichsten ist.

- *complete linkage:*
  Ein neuer Fall wird einem Cluster angefügt, wenn alle Fälle aus diesem Cluster auf demselben Ähnlichkeitsniveau stehen wie der neu einzufügende Fall. Für die Beurteilung der Ähnlichkeit zwischen neuem Fall und Cluster wird das Cluster also durch denjenigen seiner Fälle repräsentiert, der dem einzufügenden Fall am wenigsten ähnlich ist.

- *average linkage:*
  Ein neuer Fall wird einem Cluster angefügt, wenn die Fälle aus diesem Cluster im Mittel auf demselben Ähnlichkeitsniveau stehen wie der neu einzufügende Fall. Für die Beurteilung der Ähnlichkeit zwischen neuem Fall und Cluster wird das Cluster also gewissermaßen (genau gilt das nur für die Zentroid-Methode — nicht zu verwechseln mit der veralteten Zentroidmethode bei der Faktorenanalyse —, die allerdings metrische Merkmale voraussetzt) durch seinen Mittelpunkt repräsentiert.

- *Ward's method:*
  Auch dieses Verfahren setzt — wie die Zentroid-Methode — metrische Merkmale voraus. Sie beurteilt die Summe der quadrierten Entfernungen der Fälle eines jeden Clusters vom Mittelpunkt ihres jeweiligen Clusters (Fehlerquadratsumme, "error sum of squares") und führt jene Clusterzusammenfügung aus, die den geringsten Zuwachs in dieser Fehlerquadratsumme zur Folge hat.

Für alle Varianten der agglomerativen Clusteranalyse ist es erforderlich, erst einmal eine Matrix der Distanzen (oder Ähnlichkeiten) zwischen allen Fällen (und Clustern) zu berechnen; sie sind daher sehr speicheraufwendig. Distanzen oder Ähnlichkeiten über metrische Variable werden bezeichnet als:

- Euklidischer Abstand

$$d_{ij} = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}$$

- Quadratischer euklidischer Abstand

$$d_{ij} = \sum_{k=1}^{m} (x_{ik} - x_{jk})^2$$

- City-Block-Abstand

$$d_{ij} = \sum_{k=1}^{m} |x_{ik} - x_{jk}|$$

- Chebyshev-Abstand

$$d_{ij} = \max_{k=1}^{m} |x_{ik} - x_{jk}|$$

- Verallgemeinerter Abstand

$$d_{ij} = \sqrt[r]{\sum_{k=1}^{m} |x_{ik} - x_{jk}|^p}$$

($p = 2, r = 2$ ergibt den euklidischen Abstand, $p = 2, r = 1$ den quadrierten euklidischen Abstand, $p = 1, r = 1$ den City-Block-Abstand und $p = \infty, r = \infty$ den Chebyshev-Abstand)

Daneben eignen sich — vor allem wenn die Blickrichtung umgekehrt wird und eine Gruppierung ähnlicher Merkmale zu Merkmals-Clustern beabsichtigt ist — alle Korrelations-, Assoziations- und PRE-Maße zur Ähnlichkeitsmessung.

Schließlich kann die Clusteranalyse — ohne Auswertung von Merkmalen bei Fällen — gleich auf einer Ähnlichkeits- oder Distanzmatrix aufsetzen.

Partitionierende Verfahren beginnen mit einer beliebigen Anfangsaufteilung der Fälle in Cluster und gehen dann folgendermaßen iterativ vor:

1. Sie berechnen die Mittelpunkte der Cluster (hierzu bedarf es wiederum metrischer Merkmale),

2. weisen dann jeden Fall demjenigen Cluster zu, dessen Mittelpunkt am nächsten liegt und

3. berechnen die Mittelpunkte der Cluster neu.

4. Die Schritte 2 und 3 werden solange ausgeführt, bis in einem Durchgang kein Fall mehr seine Cluster-Zugehörigkeit ändert.

In SPSS zum Beispiel sind die beschriebenen Varianten der agglomerativen Verfahren in der Methode `CLUSTER` realisiert, während eine Variante der partitionierenden Verfahren in der Methode `QUICK CLUSTER` durchgeführt wird.

6.5.3 Partitionierende Verfahren: Ein Beispiel

Anhand einer weiter verkleinerten Anzahl von Fällen aus dem Wahl-Datensatz kann gezeigt werden, daß mit den auch zur Illustration der Diskriminanzanalyse verwendeten Skalometerfragen eine Bildung von Gruppen recht gut möglich ist. Anders als bei der Diskriminanzanalyse lautet die Fragestellung hier nicht:

> Lassen sich die Wähler unterschiedlicher Parteien nach ihrer Sympathie für Parteien und Politiker voneinander unterscheiden?

sondern:

> Lassen sich die Wähler in homogene Gruppen aufteilen, die sich in den Sympathieeinstufungen ihrer Mitglieder unterscheiden?

An diese Frage läßt sich hernach die zweite anschließen: Wodurch lassen sich die gefundenen Gruppen, außer durch ihre Sympathieeinstufungen, noch charakterisieren?

Das zunächst vorzustellende partitionierende Verfahren in `QUICK CLUSTER` vermag noch alle Fälle zu verarbeiten. Zur Clusterbildung werden alle Sympathieskalometer benutzt (`V212 TO V231`), die Anfangsaufteilung beginnt mit den am weitesten voneinander entfernten (unähnlichsten) Fällen (`/INITIAL SELECT`) in zwei Clustern (`/CRITERIA CLUSTERS (2)`). Am Schluß der Auswertung werden die Abstände von den endgültigen Clustermittelpunkten ausgegeben (`/PRINT DISTANCE`), außerdem wird eine Varianzanalyse (`ANOVA`) durchgeführt. Die endgültige Clusterzugehörigkeit wird in der Variablen `SKQC2` abgespeichert (`/SAVE CLUSTER (SKQC2)`).

```
* SPSS-Aufruf:   QUICK CLUSTER V212 TO V231
*                     /INITIAL SELECT /CRITERIA CLUSTERS (2)
*                     /PRINT DISTANCE ANOVA  /SAVE CLUSTER (SKQC2).
```

Der Algorithmus sucht nun bei einem ersten Durchgang durch die Daten die beiden am weitesten voneinander entfernten Fälle. Zunächst werden die beiden ersten Fälle genommen; ihre Variablenwerte und ihr Abstand werden gespeichert. Dann wird der dritte Fall genommen und daraufhin überprüft, ob die kürzeste Distanz von ihm zu einem der ersten beiden Fälle (d.h. den vorläufigen Clustermittelpunkten) größer ist als die Distanz zu den beiden ersten Fällen. Wenn das der Fall ist, ersetzt er denjenigen vorläufigen Clustermittelpunkt, dem er am nächsten liegt. Mit allen folgenden Fällen wird ebenso verfahren, so daß immer nur die Positionen der

vorläufigen Clustermittelpunkte und ihre Distanzen gespeichert werden müssen. (Diese Positionen und Distanzen werden von SPSS nur ausgegeben, wenn die Option `/CRITERIA ... NOUPDATE` eingesetzt wird. In diesem Fall werden allerdings für die endgültige Berechnung der Clusterzugehörigkeiten auch nur die im ersten Durchgang gefundenen am weitesten voneinander entfernten Fälle als Clustermittelpunkte verwendet.)

Ohne die Option `/CRITERIA ... NOUPDATE` wird sodann eine erste Zuweisung aller Fälle zu den vorläufigen Clustern vorgenommen. Deren Mittelpunkte werden als *Classification Cluster Centers* ausgegeben:

```
Classification Cluster Centers.

 Cluster       V212          V213          V214          V215

       1         4.5816       10.3732        9.6905        3.0913
       2        10.1201        3.2893        4.9435        7.6474

 Cluster       V216          V217          V218          V219

       1         7.4773       10.3162       10.3580        9.6900
       2         9.7784        2.3445        2.3422        2.3572

 Cluster       V220          V221          V222          V223

       1         2.7116        1.8687        8.0241        3.8151
       2        10.0341       10.0610        2.0119       10.3927

 Cluster       V224          V225          V226          V227

       1        10.0111       10.2255        1.7929        3.0347
       2         2.1816        1.6317        9.9457       10.4019

 Cluster       V228          V229          V230          V231

       1        10.4107       10.3033        2.0559        9.8184
       2         2.2248        1.4955       10.1247        1.4907
```

In einem zweiten Durchgang durch die Daten werden nun diese *Classification Cluster Centers* als Clustermittelpunkte für die endgültige Zuweisung der Fälle zu den Clustern verwendet: Jeder Fall gehört zu dem Cluster, dessen Clustermittelpunkt er am nächsten liegt. Nach dieser endgültigen Zuweisung der Fälle zu Clustern werden die endgültigen Clustermittelpunkte berechnet:

```
Final Cluster Centers.
```

```
 Cluster       V212          V213          V214          V215

       1         6.2819        8.9614        8.2239        7.1236
       2         8.8364        4.5093        3.6215        4.8832

 Cluster       V216          V217          V218          V219

       1         3.5444        8.6023        8.8224        7.4093
       2         6.4626        4.4626        4.6262        4.6682

 Cluster       V220          V221          V222          V223

       1         6.0193        3.8417        6.6911        5.1081
       2         8.1168        6.7243        4.3364        7.9486

 Cluster       V224          V225          V226          V227

       1         7.8571        8.6795        3.6486        6.0386
       2         5.7290        3.7196        6.4299        8.7570

 Cluster       V228          V229          V230          V231

       1         8.6062        8.4015        5.6023        6.7027
       2         5.8925        3.9813        7.3505        3.8224
```

Außerdem werden die Entfernungen zwischen den Clustermittelpunkten ausgegeben:

```
Distances between Final Cluster centers.
 Cluster              1             2
       1          .0000
       2        14.5655         .0000
```

Eine Varianzanalyse prüft, ob sich die Cluster in den einzelnen zur Analyse benutzten Variablen signifikant unterscheiden:

```
Analysis of Variance.
 Variable Cluster MS   DF  Error MS        DF           F   Prob
 V212       764.7097    1    4.9803     471.0     153.5483   .000
 V213      2322.5834    1    3.4673     471.0     669.8549   .000
 V214      2482.1568    1    4.5761     471.0     542.4151   .000
 ...
 V229      2289.5140    1    5.6139     471.0     407.8268   .000
 V230       358.1048    1    5.4751     471.0      65.4065   .000
 V231       972.1195    1    4.2916     471.0     226.5149   .000
```

Wie ersichtlich, unterscheiden sich die beiden Cluster auf allen Variablen höchst signifikant.

Schließlich wird die Anzahl der Fälle in den einzelnen Clustern angegeben:

```
Number of Cases in each Cluster.
 Cluster      unweighted cases    weighted cases
       1                 259.0             259.0
       2                 214.0             214.0
 Missing                    27
 Total                   473.0             473.0
```

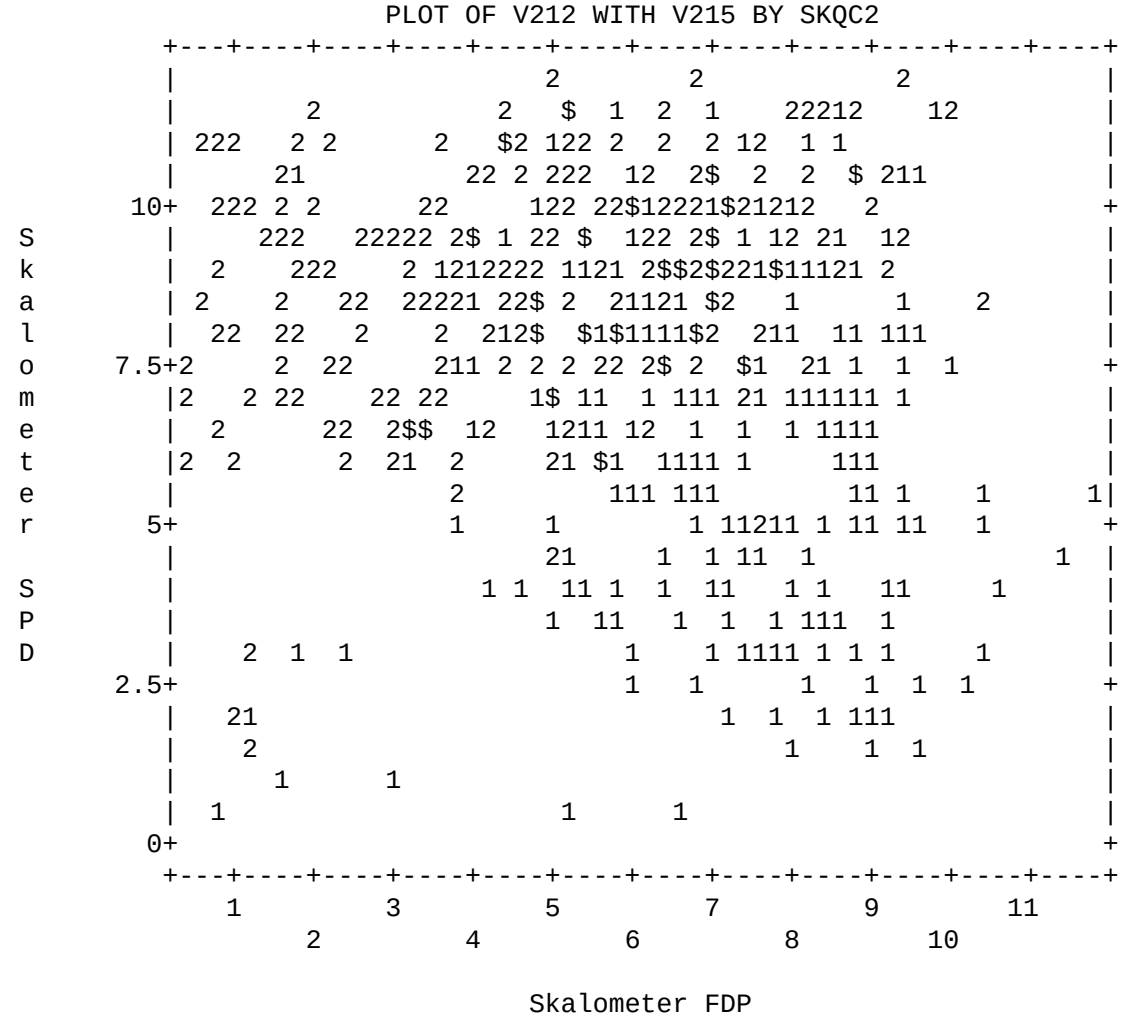


Abb. 6.24: Vermutete CDU- und SPD-Wähler im Streuungsdiagramm aus FDP- und SPD-Sympathie

Die Clusterzugehörigkeit läßt sich nun weiter untersuchen, beispielsweise indem — ähnlich wie bei der Einführung der Diskriminanzanalyse (vgl. Kapitel 6.4) — alle Fälle mit einer Kennung in ein von zwei Variablen aufgespanntes Streuungsdiagramm eingetragen werden (Abbildung 6.24).

Dieses Streuungsdiagramm ähnelt sehr demjenigen in Abbildung 6.17 (siehe Kapitel 6.4); es liegt also nahe, in Cluster 1 den größten Teil der CDU-Wähler und in Cluster 2 den größten Teil der SPD-Wähler zu vermuten. Tatsächlich erbringt eine Kreuztabellierung der Sonntagsfrage mit der Clustervariablen `SKQC2` das folgende Ergebnis:

	Cluster	
Sonntagsfrage	1	2
CDU	170	6
SPD	51	152
FDP	20	4
Grüne	2	39

Es liegt nun nahe, einen ähnlichen Versuch mit mehr als zwei Clustern vorzunehmen, um womöglich auch die Wählerschaften der kleineren Parteien aus den Skalometerfragen zu rekonstruieren. Das Ergebnis ist — bei vier Clustern — wiederum eine mit der Varianzanalyse nachgewiesene perfekte Trennung. Es ergeben sich vier Clustern mit 131, 215, 34 und 93 Fällen. Die Kreuztabellierung mit der Sonntagsfrage zeigt, daß zwei Drittel der CDU-Wähler dem Cluster 1 und der Rest dem Cluster 2 zugewiesen wurden, die SPD-Wähler verteilen sich auf die Cluster 2 und 4, die FDP-Wähler verteilen sich gleichmäßig auf die Cluster 1 und 2, während die Wähler der Grünen zur Hälfte im Cluster 3 und zu je einem Viertel in den Clustern 2 und 4 liegen. Die Übereinstimmung dieser Clusteraufteilung mit den Antworten auf die Sonntagsfrage ist also bei den beiden großen Parteianhängerschaften schlechter als im Zwei-Cluster-Fall (für die Anhängerschaften der kleineren Parteien gibt es naturgemäß keinen Vergleich). Die Frage, was mit den beiden Clustervariablen `SKQC2 und SKQC4` eigentlich gemessen wird, ist also nicht leicht zu beantworten. Wird vom Grundgedanken der Clusteranalyse ausgegangen, so unterstellen die beiden alternativen Ergebnisse der Zwei- und der Vier-Cluster-Analyse, daß es zwei bzw. vier (und mit noch mehr Versuchen wahrscheinlich auch drei, vielleicht gar fünf oder sechs) gut separierbare Gruppen gibt, sie weisen es aber nicht nach. Um das hieraus resultierende Problem zu veranschaulichen, kann ein Test durchgeführt werden, bei dem zwei simulierte, normalverteilte und unkorrelierte Variablen für eine Zwei-Cluster-Analyse Verwendung finden. Nach Konstruktion gibt es hier keine separierbaren Gruppen,

denn alle Fälle sind Realisierungen eines bivariat und unkorreliert normalverteilten Zufallsvektors — aber auch hier liefert QUICK CLUSTER zwei (etwa gleich große) Cluster, die (laut Varianzanalyse) höchst signifikant getrennt werden, obwohl die Wahrscheinlichkeitsverteilung ihre höchste Dichte an der Gruppengrenze hat (im Prinzip ergibt sich dasselbe für mehr als zwei Cluster): ein Ergebnis, das zur sorgfältigen Interpretation von QUICK CLUSTER-Ergebnissen veranlassen sollte. Genau genommen müßte immer geprüft werden, ob die gemeinsame Verteilung aller Fälle mehrgipflig ist; dies ist jedoch mit Standardmethoden prinzipiell nicht möglich (vgl. [Troitzsch 1990, Herlitzius 1990]).

6.5.4 Agglomerative Verfahren: Ein Beispiel

Agglomerative Verfahren haben gegenüber den partitionierenden den Vorteil, daß die Anzahl der entstehenden Cluster nicht im vorhinein festgelegt werden muß. Vielmehr wird im Laufe des Verfahrens die Anzahl der jeweils aktuellen Cluster immer weiter verkleinert, bis schließlich alle Fälle zu einem Cluster zusammengefaßt sind. Auf jeder Stufe läßt sich beurteilen, wie gut die Cluster separiert sind. Dank der großen Zahl von Distanzmaßen und Methoden der Zusammenfassung bleibt aber auch hier genug (wenn nicht zuviel) Entscheidungsspielraum für den Anwender, und auch hier ist eine sorgfältige Interpretation angebracht.

Wegen des hohen Speicherbedarfs lassen sich mit der PC-Version von SPSS nur ungefähr 260 Fälle bearbeiten.

```
* SPSS-Aufruf:   CLUSTER V212 TO V231
*                    /PRINT SCHEDULE /PLOT DENDROGRAM /MEASURE DEFAULT
*                    /SAVE CLUSTER (2,5) /METHOD COMPLETE (SKHCCO).
```

Die Clusteranalyse benutzt also auch hier sämtliche Skalometerfragen (V212 TO V231). Als Distanzmaß wird die Voreinstellung “quadratischer euklidischer Abstand” (/MEASURE DEFAULT) verwendet, als “sorting strategy” das “complete linkage” (/METHOD COMPLETE). Die Zugehörigkeit der Fälle zu den jeweiligen Clustern in den letzten fünf Agglomerationsschritten wird in den SPSS-Variablen SKHCCO2 bis SKHCCO5 gespeichert (/SAVE CLUSTER (2,5) ... (SKHCCO)). Zur Beurteilung der Qualität der Ergebnisse werden eine Übersicht über den Verlauf der Zusammenfassung (/PRINT SCHEDULE) und ein Dendrogramm ausgegeben (/PLOT DENDROGRAM).

Zunächst werden die Distanzen zwischen allen Fällen ausgerechnet, sodann werden die beiden Fälle mit der geringsten Distanz zu einem Cluster vereinigt:

Agglomeration Schedule using Complete Linkage

Stage	Clusters Cluster 1	Combined Cluster 2	Coefficient	Stage Cluster 1st Appears Cluster 1	Cluster 2	Next Stage
1	126	127	7.000000	0	0	17
2	168	169	8.000000	0	0	106
3	190	191	9.000000	0	0	60
4	242	249	11.000000	0	0	24
5	101	213	14.000000	0	0	53
6	74	100	14.000000	0	0	22
7	176	240	16.000000	0	0	60
8	146	223	16.000000	0	0	61
9	16	178	17.000000	0	0	41
10	12	55	17.000000	0	0	27
11	167	228	18.000000	0	0	40
12	98	131	18.000000	0	0	123
13	81	123	18.000000	0	0	33
14	134	162	20.000000	0	0	51
15	96	132	20.000000	0	0	26
16	13	17	20.000000	0	0	86
17	126	239	21.000000	1	0	81
...						
80	114	141	46.000000	0	40	161
81	16	126	46.000000	41	17	129
82	120	121	46.000000	0	0	123
...						
112	3	231	62.000000	0	0	199
113	124	149	62.000000	0	0	184
114	33	81	62.000000	70	98	192
...						
128	19	224	70.000000	0	0	202
129	16	150	70.000000	81	0	162
130	91	103	71.000000	0	0	186
...						
199	3	50	150.000000	112	0	244
...						
239	1	70	376.000000	227	230	246
240	11	72	414.000000	235	214	247
241	14	23	414.000000	232	233	243
242	4	8	447.000000	237	234	245
243	10	14	477.000000	236	241	248
244	2	3	499.000000	238	199	245
245	2	4	641.000000	244	242	247
246	1	5	647.000000	239	228	248
247	2	11	679.000000	245	240	249
248	1	10	824.000000	246	243	249
249	1	2	1397.000000	248	247	0

Der ersten Zeile ist zu entnehmen, daß die beiden Fälle mit der geringsten Distanz die Fälle 126 und 127 — Distanz 7 — sind. Sie werden zu einem Cluster vereinigt, das die Nummer 126 bekommt, wie — der Verweis findet sich in der letzten Spalte von Zeile 1 — in Zeile 17 zu sehen, denn dort wird das Cluster 126 mit dem Fall 239 vereinigt; die Distanz des Falles 239 zu dem am weitesten entfernten Mitglied des Clusters 126 beträgt 21. In den nächsten beiden Spalten erfährt man, daß Cluster 126 zuletzt in Zeile 1 ergänzt wurde, während Cluster 239 als Cluster noch gar nicht vorkam. Der Eintrag "81" in der letzten Spalte läßt erkennen, daß es in Zeile 81 mit der Ergänzung des Clusters 126 weitergeht.

Im einzelnen läßt sich dieses anhand der Originaldaten verfolgen:

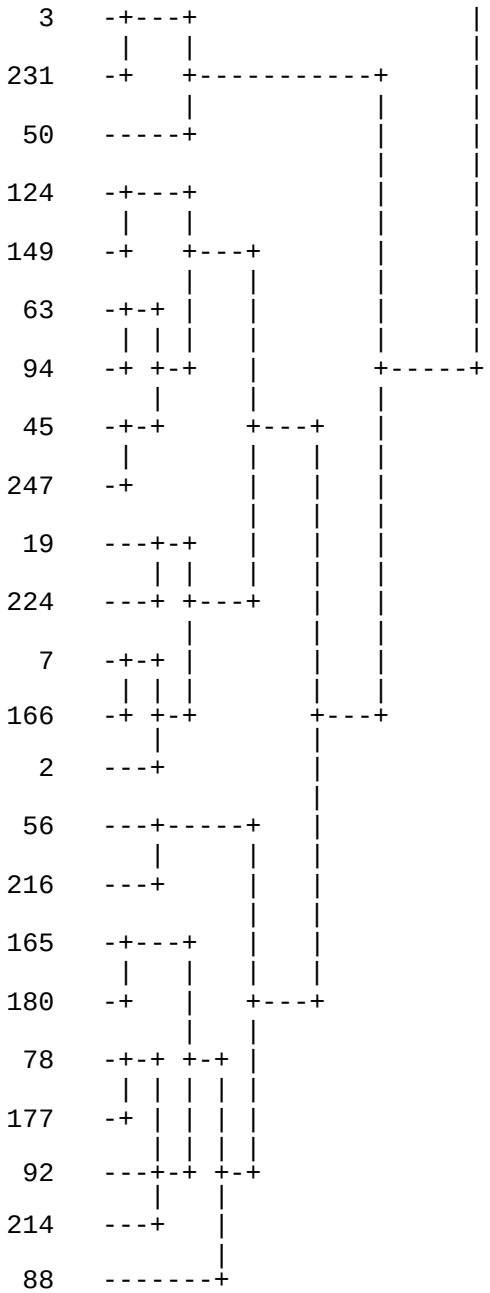
```
Nr   V212                                                          V231
126     8  5  5  6  7  4  4  4  7  7  6  7  6  5  7  9  6  6  8  6
                  !  !                    !  !                 !  !  !
127     8  5  6  5  7  4  4  4  7  6  5  7  6  5  7 10  7  7  8  6
                  !  !     !  !     !  !        !  !  !  !!        !!  !!
239     8  5  5  4  7  5  5  4  8  7  5  7  5  5  8  8  7  7  6  4
                     !!     !  !     !     !     !     !  !  !  !  !!  !!
126     8  5  5  6  7  4  4  4  7  7  6  7  6  5  7  9  6  6  8  6
```

Wie sofort ersichtlich ist, unterscheiden sich die Fälle 126 und 127 an sieben Positionen jeweils um 1, während sich der Fall 239 sowohl von Fall 126 als auch von Fall 127 an drei Positionen um 2 und an neun Positionen um 1 unterscheidet ($9 \times 1^2 + 3 \times 2^2 = 21$). In dieser Weise werden alle Eintragungen in der Matrix der Distanzen zwischen allen Fällen abgearbeitet.

In welcher Weise die Fälle zu Clustern zusammengefügt worden sind, läßt sich außer in der Übersicht auch noch in sogenannten "Eiszapfendarstellungen" ("icicle plots") — die aber nur für kleine Fallzahlen wirklich anschaulich sind — und im Dendrogramm sehen. Das Dendrogramm ist ein Baum, an dessen Blättern die Fälle erscheinen, während die inneren Knoten Clusterzusammenfügungen wiedergeben.

Die folgende Abbildung ist ein Ausschnitt aus jenem Dendrogramm, das zur Interpretation der Informationen aus der tabellarischen Übersicht von Seite 216 erzeugt wurde. Die waagerechten Entfernungen zwischen den Knoten (+) im Dendrogramm repräsentieren die Distanzen (`Coefficient` in der Übersicht). Kaum zu erkennen ist, daß jeder Knoten einem Schritt (`Stage`) entspricht. Mit diesen Erläuterungen läßt sich nun nachvollziehen, daß die Fälle 3 und 231 (im 112. Schritt) bei einer Distanz von 62 vereinigt werden. Dem so entstandenen Cluster wird (im 199. Schritt, `Next Stage`) der Fall 50 (Distanz 150) hinzugefügt. Dieses Cluster wird (im 244. Schritt) bei einer Distanz von 499 mit dem

Cluster 2 vereinigt (das zu diesem Zeitpunkt die übrigen 20 in dem Ausschnitt aufgezählten Fälle umfaßt). Dieses 23-Fälle-Cluster wird unter der Clusternummer 2 mit dem Cluster 4 vereinigt usw., bis schließlich für den letzten Schritt — im Dendrogramm-Ausschnitt nicht enthalten — nur noch zwei Cluster übrigbleiben, die eine Distanz von 1397 zueinander aufweisen.



Zur Beurteilung dieser Distanz gilt es zu vergegenwärtigen, daß 1397 sich zum Beispiel als Summe von 20 Quadratzahlen in der Größenordnung von 70 ergibt; zwei Fälle, die sich an allen 20 Positionen um 8 oder 9 Skaleneinheiten unterscheiden, werden also eine Distanz zwischen 1280 und 1620 haben, wobei hier noch einmal daran zu erinnern ist, daß die Distanz zwischen zwei Clustern bei der "complete linkage"-Methode die Distanz zwischen den am weitesten entfernten Fällen aus beiden Clustern ist.

Wie bei den partitionierenden Verfahren ist es auch bei den agglomerativen zweckmäßig, zur Interpretation der Cluster andere Variablen hinzuzuziehen. Die erste Prüfung mag der Frage gelten, ob die Cluster des "complete linkage"-Verfahrens mit denen des partitionierenden Verfahrens übereinstimmen. Bei den Zwei-Cluster-Lösungen ergibt sich eine gute Übereinstimmung: Mehr als 90% aller Fälle sind von beiden Methoden in die gleichen Cluster eingeordnet worden:

	Agglomerativ:	
Partitionierend:	Cluster 1	Cluster 2
Cluster 1	131	7
Cluster 2	14	98

Cluster 1 ist auch bei der agglomerativen Methode ein CDU-Wähler-Cluster: 93 von 98 CDU- und 11 von 14 FDP-Wählern finden sich hier, während sich in Cluster 2 die meisten (rund zwei Drittel) der SPD-Wähler und fast alle Wähler der Grünen befinden.

Der Versuch, mittels der Diskriminanzanalyse zu prüfen, wie gut sich aus den Skalometerfragen die Clusterzugehörigkeit rekonstruieren läßt, ergibt einen Prozentsatz richtiger Klassifikationen von 96% für das Zwei-Cluster-Ergebnis, für die Ergebnisse mit drei bis fünf Clustern, d.h. vor den abschließenden Zusammenfügungen, ergeben sich 92.8%, 89.6% bzw. 91.2%, was zugleich bedeutet, daß sich die Cluster auch hinsichtlich der 20 Variablen hochsignifikant unterscheiden.

Die Übersicht über die Schritte zur Agglomeration erlaubt eine vorsichtige Beantwortung der Frage, zu wievielen Clustern sich die Fälle zusammenfügen lassen. Immer dort, wo zwischen zwei Schritten die Distanz zwischen zwei zusammenzufügenden Clustern stark ansteigt, ist es angemessen, die Zusammenfügung nicht fortzusetzen, ist doch das neu zusammengefügte Cluster wesentlich weniger homogen als alle bisherigen. So spricht viel für eine Zwei-Cluster- und wenig für eine Ein-Cluster-Lösung, denn die Vereinigung der letzten beiden Cluster ergibt ein Cluster mit einer maximalen Distanz zwischen den Fällen von 1397, während die beiden letzten Cluster interne Distanzen von nur 679 bzw. 824 haben. Auch

schon die Vereinigung der Cluster 2 und 4 im Schritt 245 (und damit alle späteren Vereinigungen) wäre vielleicht unangemessen gewesen, denn während Cluster 2 und 4 interne Maximaldistanzen von 447 und 499 haben, bringt es das vereinigte Cluster auf eine anderthalb mal so hohe interne Maximaldistanz von 641. Es lägen dann allerdings sechs Cluster mit den Fallzahlen 53, 23, 62, 2, 90 und 20 vor, deren inhaltliche Interpretation nicht ganz unproblematisch ist:

	Cluster					
Partei	1	2	3	4	5	6
CDU	47	1	4	1	45	
SPD	2	18	36		34	17
FDP	4	1	1		7	1
Grüne		2	16	1		
???						1
Nichtw.		1	3		3	1
sysmiss			2		1	
	53	23	62	2	90	20

Cluster 1 besteht wiederum zum größten Teil aus CDU-Wählern, Cluster 2 und Cluster 6 jeweils zum größten Teil aus SPD-Wählern, das große Cluster 3 besteht sowohl aus SPD-Wählern als auch aus Wählern der Grünen. Cluster 5 umfaßt neben fast der Hälfte der CDU- und einem Drittel der SPD-Wähler auch die Hälfte der FDP-Wähler. Cluster 4 schließlich besteht aus einem CDU- und einem Grün-Wähler. Auch hier erweist eine Diskriminanzanalyse einen Klassifikationserfolg von 91.2%, daneben erlaubt sie aber auch eine zweidimensionale Darstellung der Fälle und Cluster (Abb. 6.25).

Alle Angehörigen des Clusters 1 finden sich ganz rechts, die Angehörigen der Cluster 2, 3 und 6 ganz links (2 und 6 links oben, 3 links unten), die Mitglieder des Clusters 5 versammeln sich in der Mitte und etwas rechts von der Mitte, während die beiden Mitglieder des Clusters 4 ihre Positionen etwas rechts von der Mitte und ganz weit unten einnehmen. Die erste der beiden Diskriminanzfunktionen ist also (vgl. S. 201) wieder die “links”-“rechts”-Dimension, die zweite dürfte eine Dimension der politischen Zufriedenheit (vgl. [Troitzsch 1990]) sein. Mit ein wenig Phantasie (und dem Rückgriff auf eine entsprechende Frage aus der Wahlstudie) lassen sich die Cluster als (nicht ganz sauber abgegrenzte) Gruppen von Befürwortern verschiedener Koalitionen auffassen: 1 CDU/CSU-FDP, 2 “rot-grün” (oder allenfalls sozialliberal), 3 “rot-grün” (oder allenfalls Große Koalition), 4 “schwarz-grün”, 5 auf jeden Fall Koalition, 6 Große Koalition oder sozialliberal.

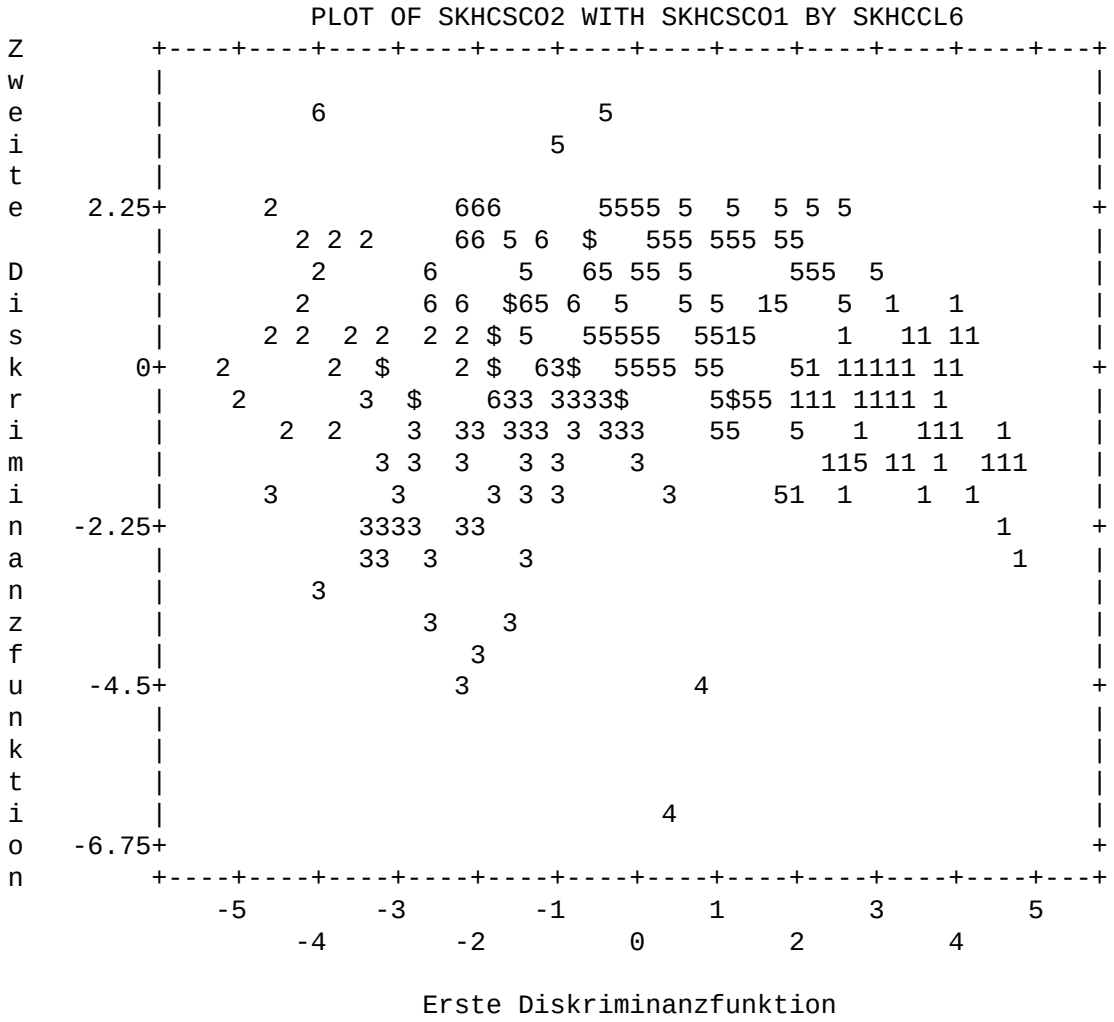


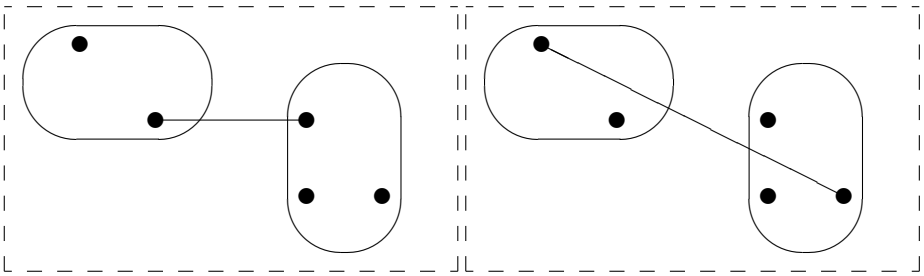
Abb. 6.25: Streuungsdiagramm aus Fällen und Clustern

	Cluster					
Koalitionspräferenz	1	2	3	4	5	6
CDU/CSU-FDP	42		2		37	
CDU/CSU-SPD	7	2	11		23	11
SPD-Grüne	1	11	36		11	2
SPD-FDP	1	8	6		10	5
CDU/CSU allein	1				1	
SPD allein		1				1
sonstiges			1	2	3	1

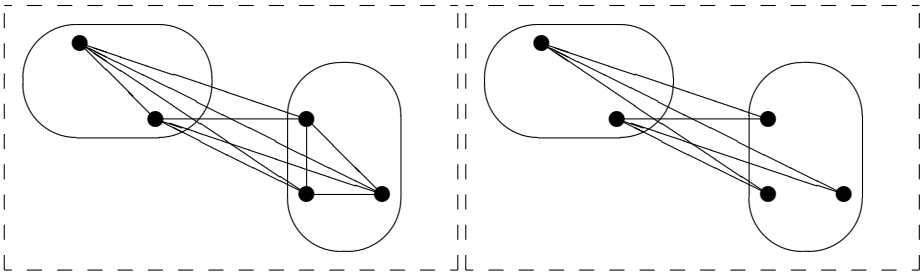
6.5.5 Vergleich verschiedener agglomerativer Verfahren

Abschließend soll anhand des Beispiels untersucht werden, welche Auswirkungen auf das Ergebnis der Clusteranalyse die Wahl eines anderen Distanzmaßes oder einer anderen Agglomerationsmethode hat. Der Vergleich der Agglomerationsmethoden beschränkt sich dabei auf folgende Methoden:

- `WAVERAGE`, eine der beiden "average linkage"-Methoden: bei ihr wird als Distanz zwischen zwei zu vereinigenden Clustern die mittlere Distanz aller Fälle des Vereinigungsclusters genommen.
- `BAVERAGE`, die andere "average linkage"-Methode: bei ihr wird als Distanz zwischen zwei zu vereinigenden Clustern der Mittelwert aller Distanzen von jedem Fall aus einem Cluster zu jedem Fall aus dem anderen Cluster genommen.
- `SINGLE`: als Distanz zwischen zwei Clustern gilt die kürzeste Distanz zwischen je einem Fall aus jedem Cluster.
- `COMPLETE`: als Distanz zwischen zwei Clustern gilt die längste Distanz zwischen je einem Fall aus jedem Cluster.



single linkage complete linkage



average linkage (`WAVERAGE`) average linkage (`BAVERAGE`)

Als Distanz zwischen den Clustern gelten also die Mittelwerte der in den vier Abbildungen oben eingezeichneten Entfernungen.

Im vorliegenden Beispiel sind die Ergebnisse mit der “single linkage”-Methode praktisch nicht brauchbar. In den letzten 35 Agglomerationsschritten werden 32 einzelne Fälle und drei ganz kleine Cluster mit dem bis dahin einzigen, großen Cluster vereinigt. Es zeigt sich hier der auch in der Literatur ([Aldenderfer/Blashfield 1984, S. 39]) diskutierte Effekt der Kettenbildung, der daraus folgt, daß es im allgemeinen genügend einzelne Fälle gibt, die nahe an der Grenze des ersten gebildeten Clusters liegen.

Die Ergebnisse der beiden “average linkage”-Methoden kommen den Ergebnissen des “complete linkage”, die hier als Referenzmethode gedient hat, ziemlich nahe, wenn auch im Falle `BAVERAGE` — ähnlich wie beim “single linkage” — ganz zum Schluß noch einige bis dahin nicht berücksichtigte einzelne Fälle angefügt werden, so daß Lösungen mit wenigen Clustern eine Reihe ganz kleiner Cluster enthalten. Demgegenüber haben die Methoden `WAVERAGE` und `COMPLETE` bei jeweils vier Clustern knapp 80% der Fälle gleich gruppiert (bei jeweils zwei Clustern beträgt die Übereinstimmung sogar 94%).

Unterschiedliche Distanzmaße können sich ebenfalls auf die Clusterbildung auswirken. Der quadratische euklidische Abstand und der euklidische Abstand haben bei “single linkage” und “complete linkage” das gleiche Ergebnis, da es ja nur auf die Ordnung der Distanzen ankommt; wo hingegen Mittelwerte der Distanzen gebildet werden, verändern sich die Ergebnisse naturgemäß, denn der Mittelwert der Quadratwurzeln ist nicht gleich der Quadratwurzel des Mittelwerts. Im vorliegenden Beispiel kommen die Ergebnisse der City-Block-Distanz dem (quadratischen) euklidischen Abstand, der hier als Referenz dient, näher als die Ergebnisse der Chebyshev-Distanz. City-Block- und Euklid-Ergebnisse stimmen bei einer Vier-Cluster-Lösung zu 72%, bei einer Drei-Cluster-Lösung zu 75.6% und bei einer Zwei-Cluster-Lösung zu 82% überein. Die entsprechenden Zahlen für die Chebyshev-Distanz lauten 45.2%, 52.8% und 65.2%. Da diese beiden Distanzmaße im vorliegenden Beispiel ohnehin unangemessen sind, lohnt es nicht, die Ergebnisse weiter zu diskutieren, der Vergleich lehrt aber, daß die Entscheidungen über das zu wählende Distanzmaß — ebenso wie über die zu wählende Agglomerationsmethode — nicht leichtfertig gefällt werden dürfen.

Anhang A

Exkurse zu ausgewählten Themen

A.1 Multivariate Modellbildung in der Meßtheorie (Indexbildung)

Unter dem Begriff "Index" wird ein Meßmodell der folgenden Art verstanden:[1]

$$I_k = f(x_1, x_2, \ldots, x_k)$$

Die Variablen $x_1, x_2, \ldots, x_k$ werden als *Indikatoren* der zu messenden Variablen X bezeichnet und stellen die Elemente des Indexes I_k dar. Ihre Anzahl k bestimmt auch die Freiheitsgrade des Indexes. Bezüglich der zu messenden Variablen X sind die x_i durch folgende Relation definiert:

$$x_i = g_i(X, Z_1, Z_2, \ldots, Z_n)$$

wobei Z_j $(j = 1, 2, \ldots, n)$ eine Reihe von Variablen darstellt, die die Messung von X durch I_k beeinflußt. Die Konstruktion eines Indexes I_k bedeutet demnach, die Funktionen f und g_i so zu definieren, daß

$$I_k = h(X) + e_{I_k}$$

wobei angenommen wird, daß

$$e_{I_k} < e_{I_{k-1}}$$

Das Symbol e_{I_k} kennzeichnet den Meßfehler, d.h. die Summe der zufallsbedingten Schwankungen, deren mathematische Erwartung $E(e)$ gleich Null ist.

Bemerkungen:

[1] vgl. [Besozzi/Zehnpfennig 1976, S. 12ff]

- Anzahl der Indikatoren des Meßmodells: $k \geq 2$

- Kriterium für die Bestimmung der Anzahl der Indikatoren ist die Größe des Meßfehlers innerhalb einer modellspezifischen Meßgenauigkeit.

- Die zu messende Variable X ist meistens nicht nur mit k, sondern vielmehr mit m (für $m > k$) beobachtbaren Variablen funktional verbunden. Die Funktion für I_k definiert dementsprechend nur eine mögliche Untermenge

  $$A_{I_k}(x_1, x_2, \ldots, x_k)$$

  aus der Menge

  $$B_{I_k}(x_1, x_2, \ldots, x_k \ldots, x_m)$$

  der m Variablen, die für die Messung von X aufgrund einer Theorie von X in Frage kommen (Meßtheorie). Die Auswahl von k Indikatoren zur Messung von X setzt die Entwicklung von Kriterien voraus, nach denen eine optimale Indikatorenmenge identifiziert werden kann.

Weitere Voraussetzung für die meßtheoretische Modellbildung:

1. Entwicklung einer Meßtheorie bzw. eines Modells, daß die Beziehungen zwischen manifesten Variablen (Indikatoren) und latenten Variablen (Index) erklärt (kausalanalytischer Ansatz der Indexbildung).

2. Spezifikation der Störvariablen (Z_j) und der Bedingungen, unter denen von ihrem Einfluß auf die Messung abgesehen werden kann.

3. Spezifikation des funktionalen Zusammenhangs zwischen der Indikatorvariablen und der zu messenden theoretischen Variablen. Bestimmung der Indikatorfunktion g_i.

4. Spezifikation der Art des funktionalen Zusammenhangs zwischen der zu messenden theoretischen Variablen und den Indikatorvariablen — Gewichtung der Einzelindikatoren und Bestimmung der Indexfunktion f.

Beispiel: Ansatz einer Meßtheorie für das Konstrukt “Parteisympathie”

Hypothese: Parteisympathie ist eine Projektion der persönlichen Sympathien und Antipathien gegenüber einzelnen Parteipolitikern auf eine Partei.

Erklärung: Inhalte spielen eine immer untergeordnetere Rolle in der Politik:

- kaum unterscheidbare Programmatik in wichtigen Politikfeldern: Frieden, Wohlstand, Sicherheit, Umwelt
- Medienvermarktung von Politik

X : Parteisympathie

x_i : Sympathieskalometervariablen zu Politikern

> “Und was halten Sie — so ganz allgemein — von dem Politiker <XXX>?
> Sagen Sie es bitte anhand einer Skala von $+5$ bis -5:
> $+5$ bedeutet, daß Sie sehr viel von <XXX> halten.
> -5 bedeutet, daß Sie überhaupt nichts von ihm halten.”

Z_j :

- Anwortvorgaben
- Frageformulierungseinflüsse
- situative Einflüsse (z.B. Anwesenheit Dritter, Interviewerverhalten)

g_i : Im Sinne der Definition von “Messen” (vgl. Kapitel 2) bezeichnen die g_i jeweils die Zuordnung der Sympathie für einen Politiker auf einer Skala von -5 bis $+5$.

f : Mit Hilfe einer Faktorenanalyse werden die manifesten Sympathieskalometervariablen x_i zu zwei Faktoren “zusammengefaßt” (vgl. Kapitel 6.3.6):

$$\begin{array}{lcl} f_1(=I_1) & : & \text{Sympathie Regierung} \\ f_2(=I_2) & : & \text{Sympathie Opposition} \end{array}$$

Grundlage der Faktorenanalyse ist dabei die Annahme einer *linearen* Beziehung zwischen manifesten und latenten Variablen, d.h. (vgl. Kapitel 6.3.7)

$$\begin{aligned} \boldsymbol{F} &= f(\boldsymbol{Z}) \\ &= \boldsymbol{ZB} \end{aligned}$$

mit:

$\boldsymbol{F}$: $(n \times r)$–Matrix der Faktorwerte
$\boldsymbol{Z}$: $(n \times m)$–Matrix der standardisierten Indikatorenwerte
$\boldsymbol{B}$: $(r \times m)$–Matrix der Gewichtungskoeffizienten

A.2 Wahrscheinlichkeitsfunktion, Verteilungsfunktion

Der Begriff der "Wahrscheinlichkeitsverteilung"[2] ergibt sich aus der Frage, wie sich die Wahrscheinlichkeiten bei einem Zufallsexperiment auf die verschiedenen Ereignisse (= Realisierungen einer Zufallsvariablen) verteilen. Entsprechend dem Skalenniveau der zugrunde liegenden Zufallsvariablen wird zwischen *stetigen* und *diskreten* Wahrscheinlichkeitsverteilungen unterschieden.

Eine Wahrscheinlichkeitsverteilung läßt sich allgemein durch zwei Funktionen beschreiben:

- *Wahrscheinlichkeitsfunktion* $f(X)$:

  Die Wahrscheinlichkeitsfunktion gibt an, mit welchen Wahrscheinlichkeiten die verschiedenen Ereignisse der Zufallsvariablen vorkommen können.

  Beispiel:

  - *diskret*: Wahrscheinlichkeit, eine bestimmte Zahl zu würfeln.
    Sie beträgt für jeden Wurf $1/6$, und es gilt

    $$\sum_{i=1}^{6} f(x_i) = 1$$

  - *stetig*: Normalverteilung
    Die Wahrscheinlichkeitsfunktion beschreibt hier eine Wahrscheinlichkeitsdichte, für die gilt

    $$\int_{-\infty}^{\infty} f(v)dv = 1$$

- *Verteilungsfunktion* $F(x)$:

  Die Verteilungsfunktion gibt die Wahrscheinlichkeit an, mit der die Zufallsvariable kleiner ist als x, d.h.

  $$F(x) = P(X \leq x)$$

[2] vgl. [Kreyszig 1979, S. 70ff]

Beispiel:

- *diskret*: Wahrscheinlichkeit, eine bestimmte Zahl oder eine kleinere zu würfeln

  Die Verteilung für die Zahl 3 beträgt $F(3) = P(X \leq 3) = 3/6$, und es gilt allgemein:

  $$F(x) = \sum_{x_i \leq x} f(x_i)$$

- *stetig*: Normalverteilung

  Allgemein gilt:

  $$F(x) = \int_{-\infty}^{x} f(v)dv$$

Weiterhin gilt folgende Beziehung zwischen Wahrscheinlichkeitsfunktion und der Verteilungsfunktion:

$$P(a < x \leq b) = F(b) - F(a) = \int_{a}^{b} f(v)dv$$

A.3 Konstruktion von Wahrscheinlichkeitsverteilungen

1. Konstruktion einer χ^2-verteilten Zufallsvariablen

Seien $z_1, \ldots, z_n$ stochastisch unabhängige, normalverteilte Zufallsvariablen:

$$z_i \sim N(0,1), \quad (i = 1, \ldots, n)$$

Dann ist die durch

$$\chi^2 = \sum_{i=1}^{n} z_i^2$$

definierte Zufallsvariable *zentral-χ^2-verteilt* mit n Freiheitsgraden.

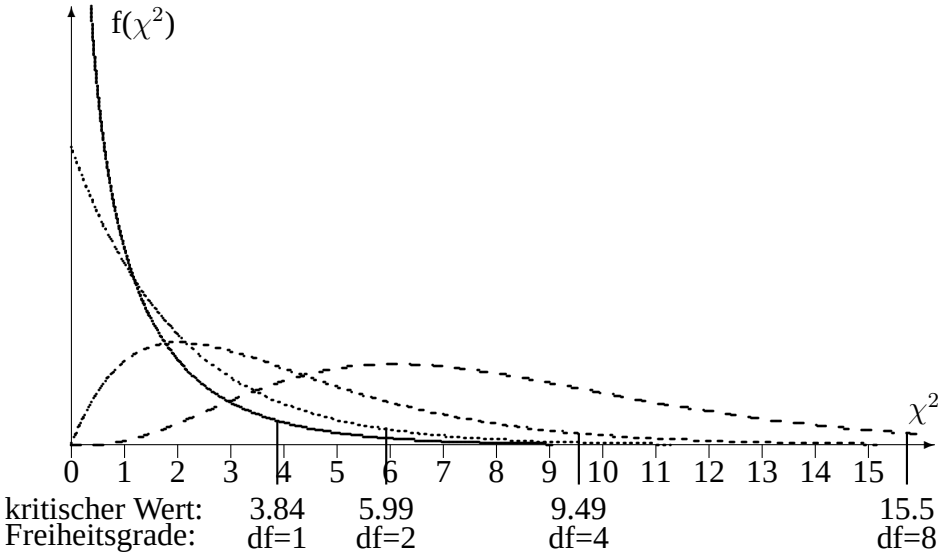


Abb. A.1: Graph der χ^2-Verteilung

2 Konstruktion einer t-verteilten Zufallsvariablen

Sei z eine standardnormalverteilte Zufallsvariable und v eine zentral-χ^2-verteilte Zufallsvariable mit n Freiheitsgraden. Dann ist die Zufallsvariable t

$$t = \frac{z}{\sqrt{\frac{v}{n}}}$$

eine *zentral-t-verteilte* Zufallsvariable mit n Freiheitsgraden.

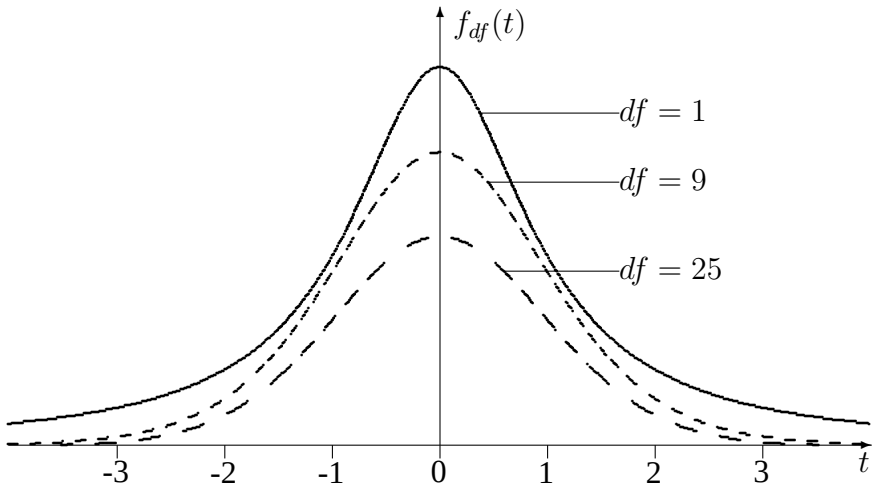


Abb. A.2: Graph der t-Verteilung

3. Konstruktion einer F-verteilten Zufallsvariablen

Seien v_1 und v_2 zwei stochastisch unabhängige, zentral-χ^2-verteilte Zufallsvariablen mit df_1 bzw. df_2 Freiheitsgraden. Dann ist die Zufallsvariable F

$$F = \frac{v_1/df_1}{v_2/df_2} = \frac{v_1 \cdot df_2}{v_2 \cdot df_1}$$

eine *zentral-F-verteilte* Zufallsvariable mit df_1 Zähler- und df_2 Nennerfreiheitsgraden.

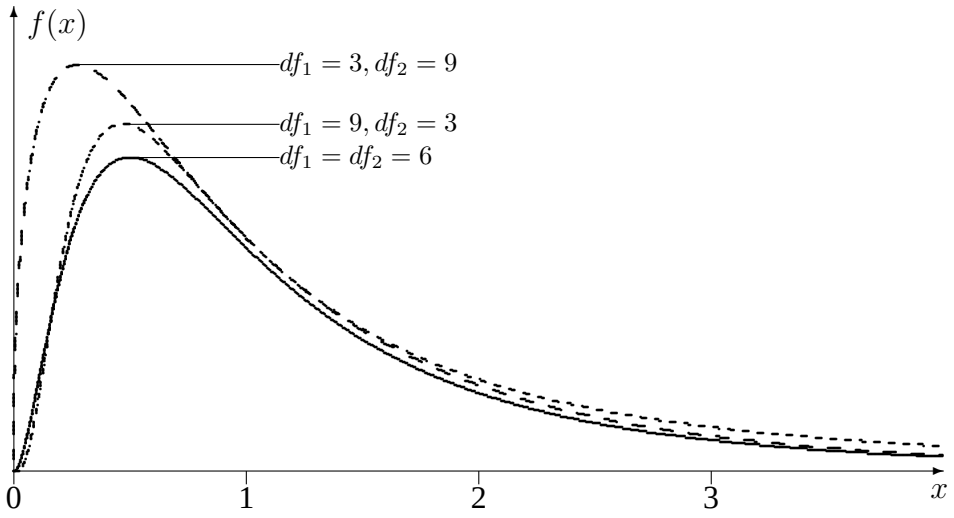


Abb. A.3: Graph der F-Verteilung

Ebenso wie für die Normalverteilung liegen auch die Werte der χ^2-, t- und F-Verteilung tabelliert vor, so daß anhand einer vorgegebenen (Vertrauens-) Wahrscheinlichkeit sowie den vorliegenden Freiheitsgraden die zugehörigen Intervallgrenzen ermittelbar sind.

A.4 Der maximale Kontingenzkoeffizient C_{max}

Bei $r \times r$-Tabellen mit maximaler Assoziation sind nur die Felder in der Diagonalen belegt. Jedes Diagonalfeld enthält den Wert $\frac{1}{r}N$.

$\frac{1}{r}N$			
	$\frac{1}{r}N$		
		$\ddots$	
			$\frac{1}{r}N$

N

Für die r Felder in der Diagonalen (von links oben nach rechts unten) gilt:

$$\begin{aligned}
h_{ij} &= \frac{1}{r}N \\
\hat{h}_{ij} &= \frac{1}{r^2}N \\
h_{ij} - \hat{h}_{ij} &= \frac{1}{r}N - \frac{1}{r^2}N \\
(h_{ij} - \hat{h}_{ij})^2 &= \frac{1}{r^2}N^2 - \frac{2}{r^3}N^2 + \frac{1}{r^4}N^2 \\
\frac{(h_{ij} - \hat{h}_{ij})^2}{\hat{h}_{ij}} &= \frac{N^2\ r^2}{r^2\ N} - \frac{2N^2\ r^2}{r^3\ N} + \frac{N^2\ r^2}{r^4\ N} = N - \frac{2N}{r} + \frac{N}{r^2}
\end{aligned}$$

Für die übrigen $r^2 - r$ Felder gilt:

$$\begin{aligned}
h_{ij} &= 0 \\
\hat{h}_{ij} &= \frac{1}{r^2}N \\
h_{ij} - \hat{h}_{ij} &= -\frac{1}{r^2}N \\
(h_{ij} - \hat{h}_{ij})^2 &= \frac{1}{r^4}N^2 \\
\frac{(h_{ij} - \hat{h}_{ij})^2}{\hat{h}_{ij}} &= \frac{N^2\ r^2}{r^4\ N} = \frac{1}{r^2}N
\end{aligned}$$

$\Longrightarrow$

$$\begin{aligned}
\chi^2_{max} &= r(N - \frac{2N}{r} + \frac{N}{r^2}) + (r^2 - r)\frac{1}{r^2}N \\
&= rN - 2N + \frac{N}{r} + N - \frac{N}{r} \\
&= (r - 2 + \frac{1}{r} + 1 - \frac{1}{r})N = (r-1)N \\
C_{max} &= \sqrt{\frac{(r-1)N}{(r-1)N + N}} = \sqrt{\frac{(r-1)N}{rN}} = \sqrt{\frac{r-1}{r}}
\end{aligned}$$

A.5 Die Kovarianz zweier Variablen

Die Kovarianz beschreibt die (gemeinsame) Varianz zweier Variablen und ist eine weitere charakteristische Maßzahl über den Zusammenhang zweier Variablen, sie hängt eng mit dem Produkt-Moment-Korrelationskoeffizienten r zusammen.

$$\text{cov}(x, y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

Für standardisierte Variablen ($s^2_{x^*} = s^2_{y^*} = 1$, $\bar{x}^* = \bar{y}^* = 0$) gilt:

$$\text{cov}(x^*, y^*) = r = b_1$$

Weitere Eigenschaften der Kovarianz:

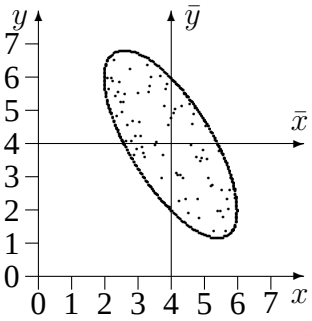
1. Das Vorzeichen der Kovarianz definiert die Richtung der Beziehung zwischen zwei Variablen.

2. Sind die beiden Variablen statistisch unabhängig, nähert sich der Wert der Kovarianz dem Betrag 0.

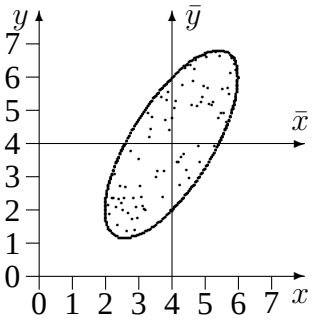
3. Der Wert der Kovarianz liegt für jede Gesamtheit in dem Intervall:

$$-\sqrt{\text{var}(x)}\sqrt{\text{var}(y)} \leq \text{cov}(x, y) \leq \sqrt{\text{var}(x)}\sqrt{\text{var}(y)}$$

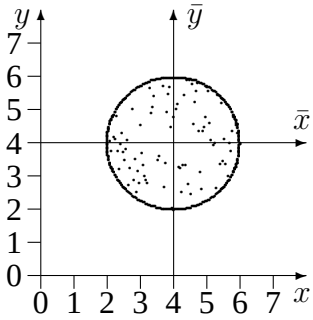
   Ober- und Untergrenze der Kovarianz variieren also von Grundgesamtheit zu Grundgesamtheit (von Stichprobe zu Stichprobe), sie sind aber für jede Grundgesamtheit (Stichprobe) fest.



negative Kovarianz



positive Kovarianz



Kovarianz = 0

Abb. A.4: Graphische Darstellung der Kovarianz

A.6 Varianzzerlegung von y (x,y metrisch skaliert)

$$
\begin{aligned}
(y_i - \bar{y}) &= (y_i - \hat{y}_i) &&+ (\hat{y}_i - \bar{y}) \\
(y_i - \bar{y})^2 &= (y_i - \hat{y}_i)^2 &&+ (\hat{y}_i - \bar{y})^2 \\
&&&+ 2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) \\
\sum_{i=1}^{n}(y_i - \bar{y})^2 &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 &&+ \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \\
&&&+ 2\underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})}_{=0,\text{ s. Folgeseite}} \\
&= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 &&+ \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \\
\textit{Gesamtvariation} &= \textit{nicht erklärte Variation} &&+ \textit{(durch x) erklärte Variation} \\
1 &= \frac{\textit{nicht erklärte Variation}}{\textit{Gesamtvariation}} &&+ \frac{\textit{erklärte Variation}}{\textit{Gesamtvariation}} \\
1 &= \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} &&+ \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \\
1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} &= \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \\
R^2 &= \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}
\end{aligned}
$$

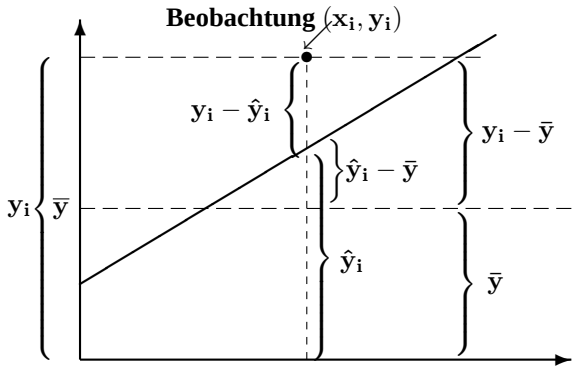


Abb. A.5: Graphische Erläuterung der Varianzzerlegung

Herleitung : $0 = \sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$

$$= \sum_{i=1}^{n}(y_i - [\bar{y} - b_1\bar{x} + b_1x_i])([\bar{y} - b_1\bar{x} + b_1x_i] - \bar{y})$$

$$= \sum_{i=1}^{n}((y_i - \bar{y}) - b_1(x_i - \bar{x}))(b_1(x_i - \bar{x}))$$

$$= \sum_{i=1}^{n}[(b_1(x_i - \bar{x})(y_i - \bar{y}) - (b_1^2(x_i - \bar{x})^2)]$$

$$= b_1\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) - b_1^2\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) - \left(\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)^2\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$= \frac{(\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}))^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} - \frac{(\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}))^2}{(\sum_{i=1}^{n}(x_i - \bar{x})^2)^2}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$= \frac{(\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}))^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} - \frac{(\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}))^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$= 0$$

Literaturverzeichnis

[Hinweis] Die Literaturangaben sind nach Kapiteln gruppiert, um die thematische Suche nach Quellen zu erleichtern.

[Kapitel 1]

[Besozzi/Zehnpfennig 1976] C. Besozzi; H. Zehnpfennig. *Methodologische Probleme der Index-Bildung*. In: Koolwijk, J. v. (Hrsg.); Wieken-Mayser, M. (Hrsg.). Techniken der empirischen Sozialforschung. Bd. 5: Testen und Messen. München: Oldenbourg, 1976. S. 9–55.

[Engel/Möhring 1994] A. Engel; M. Möhring. *Der Beitrag der Sozialwissenschaftlichen Informatik zur Sozialwissenschaftlichen Modellbildung und Simulation*. In: Donhauser, K.; Irrgang, B.; Klawitter, J. (Hrsg.). Forum interdisziplinäre Forschung. Dettelbach: Röll, 1994 (in Vorb.).

[Falter 1973] J.W. Falter. *Faktoren der Wahlentscheidung*. Köln: Heymanns, 1973.

[Frey/Kunz/Lüschen 1990] J. Frey; G. Kunz; G. Lüschen. *Telefonumfragen in der Sozialforschung*. Opladen: Westdeutscher Verlag, 1990.

[Friedrichs 1973] J. Friedrichs. *Methoden empirischer Sozialforschung*. Reinbek: Rowohlt, 1973.

[Haag/Haux/Kieser 1992] U. Haag; R. Haux; M. Kieser. *Statistische Auswertungssysteme*. Stuttgart: Fischer, 1992.

[Hoffmeyer-Zlotnik 1993] J.H.P. Hoffmeyer-Zlotnik. *Operationalisierung von "Beruf" als zentrale Variable zur Messung von sozio-ökonomischen Status*. in : ZUMA (Hrsg.). ZUMA-Nachrichten. 32(Mai 1993). S. 135–141.

[Joop/De Bier/De Leeuw 1990] J. Joop; S.E. De Bie; E.D. Leeuw. *Computer-assisted Telephone Interviewing — A Review*. In: J. Gladitz; K.G. Troitzsch (Hrsg.). Computer Aided Sociological Research: Proceedings of a Workshop Organized by the Research Committee 33 of the International Sociological Association and the Academy of Sciences of the GDR Held at Holzhau, GDr, October 2nd to 6th, 1989. Berlin: Akademie-Verlag, 1990, p. 305–317.

[Kerlinger 1978b] F. Kerlinger. *Grundlagen der Sozialwissenschaften, Bd. 2*. Weinheim: Beltz, 1978, 2. Aufl.

[Leiner 1989] B. Leiner. *Stichprobentheorie: Grundlagen, Theorie und Technik*. München: Oldenbourg, 1989.

[Porst 1985] R. Porst. *Praxis der Umfrageforschung*. Stuttgart: Teubner, 1985. (Studienskripten zur Soziologie).

[Roth/Heidenreich 1984] E. Roth (Hrsg.); K. Heidenreich (Mitarb.). *Sozialwissenschaftliche Methoden*. München: Oldenbourg, 1984.

[Schneid 1991] M. Schneid. *Einsatz computergestützter Befragungssysteme in der Bundesrepublik Deutschland*. Mannheim: ZUMA, 1991. ZUMA-Arbeitsbericht 91/20.

[Schnell/Hill/Esser 1988] R. Schnell; P.B. Hill; E. Esser. *Methoden der empirischen Sozialforschung*. München: Oldenbourg, 1988.

[SPSS 1993] SPSS. *SPSS für Windows: Anwenderhandbuch für das Basis System*. München: SPSs, 1993.

[Wahlstudie 1987] Zentralarchiv für Empirische Sozialforschung (Hrsg.). *Wahlstudie 1987*. Köln, 1987.

[**Kapitel 2]**

[Besozzi/Zehnpfennig 1976] siehe [**Kapitel 1**]

[Friedrichs 1973] siehe [**Kapitel 1**]

[Schnell/Hill/Esser 1988] siehe [**Kapitel 1**]

[**Kapitel 3]**

[Benninghaus 1985] H. Benninghaus. *Deskriptive Statistik*. Stuttgart: Teubner, 1985. 2. Aufl. (Statistik für Soziologen 1).

[Böker 1993] F. Böker. *Statistik lernen am PC.* Göttingen: Vandenhoeck & Ruprecht, 1993.

[Krämer 1992] W. Krämer. *Wie lügt man mit Statistik?* Frankfurt: Campus, 1992.

[Linder/Berchthold 1979] A. Linder; W. Berchthold. *Elementare statistische Methoden*. Basel: Birkhäuser, 1979. (Uni-Taschenbücher, 796).

[Patzelt 1985] W. Patzelt. *Einführung in die sozialwissenschaftliche Statistik*. München: Oldenbourg, 1985.

[Uehlinger/Hermann/Huebner/Benke 1992] H.-M. Uehlinger; D. Hermann; M. Huebner; M. Benke. *SPSS/PC+ Benutzerhandbuch, Band 1*. Stuttgart: Fischer, 1992.

[**Kapitel 4]**

[Kreyszig 1979] E. Kreyszig. *Statistische Methoden und ihre Anwendungen*. Göttingen: Vandenhoeck & Ruprecht, 1979, 7. Aufl.

[Patzelt 1985] siehe [**Kapitel 3**]

[Uehlinger/Hermann/Huebner/Benke 1992] siehe [**Kapitel 3**]

[**Kapitel 5]**

[Achen 1978] C. Achen. *Interpreting and Using Regression*. Beverly Hills: Sage, 1978. (Quantitative Applications in the Social Sciences, 11).

[Benninghaus 1985] siehe [**Kapitel 3**]

[Gaensslen/Schubö 1976] H. Gaensslen; W. Schubö. *Einfache und komplexe statistische Analyse*. München: Reinhardt, 1976, 2. Aufl. (Uni-Taschenbücher, 274).

[Gruber 1981] J. Gruber. *Regressionsanalyse I: Einführung in die multiple Regression und Ökonometrie*. Hagen: Fernuniversität, 1981.

[Hildebrand/Laing/Rosenthal 1977] D.K. Hildebrand; J.D. Laing; H. Rosenthal. *Analysis of Ordinal Data*. Beverly Hills: Sage, 1977. (Quantitative Applications in the Social Sciences, 8).

[Hummell 1986] H.J. Hummell. *Grundzüge der Regressions- und Korrelationsanalyse*. In: Koolwijk, J. v. (Hrsg.); Wieken-Mayser, M. (Hrsg.). Techniken der empirischen Sozialforschung. Bd. 8: Kausalanalyse. München: Oldenbourg, 1986. S. 9–76.

[Lewis-Beck 1980] M. Lewis-Beck. *Applied Regression*. Beverly Hills: Sage, 1980. (Quantitative Applications in the Social Sciences, 20).

[Liebetrau 1983] A.M. Liebetrau. *Measures of Association*. Beverly Hills: Sage, 1983. (Quantitative Applications in the Social Sciences, 32).

[Linder/Berchthold 1979] siehe [**Kapitel 3**]

[Linder/Berchthold 1982] A. Linder; W. Berchthold. *Statistische Methoden II: Varianzanalyse, Regressionsrechnung*. Basel: Birkhäuser, 1982. (Uni-Taschenbücher, 1110).

[Patzelt 1985] siehe [**Kapitel 3**]

[Reynolds 1977] H.T. Reynolds. *Analysis of Nominal Data*. Beverly Hills: Sage, 1977. (Quantitative Applications in the Social Sciences, 7).

[Schroeder/Sjoquist/Stephan 1986] L. Schroeder; D. Sjoquist; P. Stephan. *Understanding Regression Analysis*. Beverly Hills: Sage, 1986. (Quantitative Applications in the Social Sciences, 57).

[Schubö/Hagen/Oberhofer 1983] W. Schubö; K. Hagen; W. Oberhofer. *Regressions- und kanonische Analyse*. In: [Bredenkamp/Feger 1983, S. 207–292].

[Uehlinger/Hermann/Huebner/Benke 1992] siehe [**Kapitel 3**]

[Urban 1982] D. Urban. *Regressionstheorie und Regressionstechnik*. Stuttgart: Teubner, 1982. (Studienskripten zur Soziologie, 36).

[**Kapitel 6]**

[Achen 1978] siehe [**Kapitel 5**]

[Aldenderfer/Blashfield 1984] M.S. Aldenderfer; R.K. Blashfield. *Cluster Analysis*. Beverly Hills: Sage, 1984. (Quantitative Applications in the Social Sciences, 44).

[Arminger 1979] G. Arminger. *Faktorenanalyse*. Stuttgart: Teubner, 1979.

[Berry/Feldman 1985] W.D. Berry; S. Feldman. *Multiple Regression in Practice*. Beverly Hills: Sage, 1985. (Quantitative Applications in the Social Science).

[Bredenkamp/Feger 1983] J. Bredenkamp; H. Feger (Hrsg.). *Strukturierung und Reduzierung von Daten*. Göttingen: Hogrefe, 1983. (Enzyklopädie der Psychologie, Themenbereich B: Methodologie und Methoden, Serie I, Bd. 4).

[Fahrmeir/Hamerle 1984] L. Fahrmeir; A. Hamerle (Hrsg.). *Multivariate statistische Verfahren*. Berlin, New York: De Gruyter, 1984.

[Flury/Riedwyl 1983] B. Flury; H. Riedwyl. *Angewandte multivariate Statistik: computerunterstützte Analyse mehrdimensionaler Daten*. Stuttgart: Fischer, 1983.

[Gaensslen/Schubö 1976] siehe [**Kapitel 5**]

[Hartung/Elpelt 1984] J. Hartung; B. Elpelt. *Multivariate Statistik*. München: Oldenbourg, 1984.

[Herlitzius 1990] L. Herlitzius. *Schätzung nicht-normaler Wahrscheinlichkeitsdichtefunktionen*. In: Gladitz, J. (Hrsg.); Troitzsch, K.G. (Hrsg.). Computer Aided Sociological Research. Proceedings of the Workshop “Computer Aided Sociological Research” (CASOR’89), Holzhau/DDr, October 2nd–6th, 1989. Berlin: Akademie-Verlag, 1990. S. 379–396.

[Holm 1982] K. Holm. *Die Befragung 3: Die Faktorenanalyse*. München: Francke, 1982, 2. Aufl. (Uni-Taschenbücher, 372).

[Holm 1977] K. Holm. *Die Befragung 5: Pfadanalyse, Coleman-Verfahren*. München: Francke, 1977. (Uni-Taschenbücher, 435).

[Holm 1979] K. Holm. *Die Befragung 6: Das allgemeine lineare Modell*. München: Francke, 1979. (Uni-Taschenbücher, 436).

[Jae-On/Mueller 1978a] K. Jae-On; C. Mueller. *Introduction to Factor Analysis*. Beverly Hills: Sage, 1978. (Quantitative Applications in the Social Sciences, 13).

[Jae-On/Mueller 1978b] K. Jae-On; C. Mueller. *Factor Analysis*. Beverly Hills: Sage, 1978. (Quantitative Applications in the Social Sciences, 12).

[Klecka 1980] W.R. Klecka. *Discriminant Analysis*. Beverly Hills: Sage, 1980. (Quantitative Applications in the Social Sciences, 19).

[Küchler 1979] M. Küchler. *Multivariate Analyseverfahren*. Stuttgart: Teubner, 1979. (Studienskripten zur Soziologie, 35).

[Lenk 1983] W. Lenk. *Faktorenanalyse ein Mythos?* Weinheim: Beltz, 1983. (Beltz Forschungsberichte).

[Levine 1977] M.S. Levine. *Canonical Analysis and Factor Comparison*. Beverly Hills: Sage, 1977. (Quantitative Applications in the Social Sciences, 6).

[Lewis-Beck 1980] siehe [**Kapitel 5**]

[Liebetrau 1983] siehe [**Kapitel 5**]

[Linder/Berchthold 1982a] A. Linder; W. Berchthold. *Statistische Methoden III: Multivariate Methoden*. Basel: Birkhäuser, 1982. (Uni-Taschenbücher, 1189).

[Newbold/Bos 1985] P. Newbold; T. Bos. *Stochastic Parameter Regression Models*. Beverly Hills: Sage, 1985. (Quantitative Applications in the Social Sciences).

[Nie/Hull/u.a. 1975] N.H. Nie; C.H. Hull; J.G. Jenkins; K. Steinbrenner; D.H. Bent. *SPSS — Statistical Package for the Social Sciences*. New York,...: McGraw Hill, 1975, 2. Aufl.

[Schroeder/Sjoquist/Stephan 1986] siehe [**Kapitel 5**]

[Schuchard-Ficher 1985] C. Schuchard-Ficher; u.a. *Multivariate Analysemethoden*. Berlin: Springer, 1985, 3. Aufl.

[Scott Long 1986] J. Scott Long. *Confirmatory Factor Analysis*. Beverly Hills: Sage, 1986. (Quantitative Applications in the Social Sciences, 33).

[Steyer 1983] R. Steyer. *Modelle zur kausalen Erklärung statistischer Zusammenhänge*. In: [Bredenkamp/Feger 1983, S. 59–153].

[Sturm/Vajna 1976] M. Sturm; T. Vajna. *Grundzüge der Faktorenanalyse*. In: Koolwijk, J. v. (Hrsg.); Wieken-Mayser, M. (Hrsg.). Techniken der empirischen Sozialforschung. Bd. 5: Testen und Messen. München: Oldenbourg, 1976. S. 184–216.

[Thurstone 1947] L. Thurstone. *Multiple Factor Analysis*. Chicago: University of Chicago Press, 1947.

[Troitzsch 1990] K.G. Troitzsch. *Self-organization in social systems*. In: Gladitz, J. (Hrsg.); Troitzsch, K.G. (Hrsg.). Computer Aided Sociological Research. Proceedings of the Workshop "Computer Aided Sociological Research" (CASOR'89), Holzhau/DDr, October 2nd–6th, 1989. Berlin: Akademie-Verlag, 1990. S. 353–377.

[Überla 1977] K. Überla. *Faktorenanalyse*. Berlin: Springer, 1977, 2. Aufl.

[Uehlinger/Hermann/Huebner/Benke 1992] siehe [**Kapitel 3**]

[Urban 1982] siehe [**Kapitel 5**]

Index

F

G

H

I

K

L

M

N

O

P

Q

R

S

T

U

V

W

Z