.

Methoden der Psychologie

Teil 2: Versuchsplanung

SoSe 2016

Psychologisches Institut der
Johann-Gutenberg-Universität Mainz

Uwe Mortensen

24. 05. 2016

Inhaltsverzeichnis

1 Grundlegende Begriffe

Es gibt viele Gründe, sich zu fragen, worin die Unterschiede zwischen Menschen bestehen: warum sind manche Menschen erfolgreich und andere nicht, warum begehen manche Menschen Verbrechen und andere nicht, wodurch unterschieden sich religiöse von nicht religiösen Menschen, wodurch unterscheiden sich Frauen von Männern, etc. Ein Personalchef möchte gerne Bewerber, die "erfolgreich" sind, von Bewerbern, die "nicht erfolgreich" sind unterscheiden, während ein Sozialpolitiker daran interessiert ist, "Defizite" von "nicht erfolgreichen" Menschen auszugleichen, damit größere Chancengleichheit von Menschen hergestellt wird, – beide werden von unterschiedlichen Motivationen aus, um zu erklären, was sie unter "erfolgreich" verstehen, und diese Motivationen können eine Rolle spielen, wenn sie untersuchen, was erfolgreiche Menschen kennzeichnet.

Man kann versuchen, sich die jeweiligen Unterschiede am Schreibtisch oder im Lehnstuhl klar zu machen. Dazu stellt man sich vor, was z.B. Begriffe wie 'erfolgreich', 'religiös' , 'kriminell' etc eigentlich bedeuten und leitet daraus ab, welche Eigenschaften ein erfolgreicher, ein religiöser oder kriminellenr Mensch haben muß. Man kann dann wie etwa Eduard Spranger (1882 – 1963) zu einer Typologie gelangen, die durch "Lebensformen" charakterisiert werden: es gibt seiner Ansicht nach den religiösen, den ästhetischen, den sozialen, den politischen, den theoretischen und den ökonomischen Menschen. Spranger ist (wie Albert Wellek (1904 – 1972)) von Grundprinzipien wie dem der Polarität überzeugt und findet, dass Frauen im Unterschied zu Männern generell für das Gefühl und das ganzheitliche Erleben prädestiniert seien, während die Männer sich auf dem Gegenpol des "Geistigen" befänden, der den Frauen nicht zugänglich sei. Derartige Theorien seien auch von der Erfahrung bestimmt, aber es ist eine Erfahrung, die einerseits anekdotisch ist und sich andererseits an sprachlichen Stereotypen orientiert ohne darauf zu achten, ob diese Stereotypen zur Realität korrespondieren oder umgekehrt die Realität sogar mitbestimmen, – ein Fall von *self fulfilling prophecy*. Für jede Theorie kann man eine Anekdote, also ein Beispiel finden, die oder das der Theorie entspricht und sie "bestätigt".

1.1 Explorative Untersuchungen

Es ist also nötig, objektiv und ohne vorgefasste Meinungen nach Charakterisierungen von Menschen zu suchen. Sowohl der Personalchef wie auch der Sozialpolitiker könnten auf die Idee kommen, Gruppen von erfolgreichen und nicht erfolgreichen Menschen auf ihre jeweiligen Merkmale hin zu untersuchen. Dazu können beide auf dieselbe Idee kommen: man stellt eine Liste von Persönlichkeitsmerkmalen zusammen und schaut nach, welche Merkmale bei der einen Gruppe ausgeprägt sind und welche bei der anderen Gruppe. Sie werden feststellen, dass viele dieser Merkmale "miteinander einhergehen", was heißen soll, dass sie in einem statistischen Sinne miteinander assoziiert sind, und dass sich diese Assoziationen in den beiden Gruppen

möglicherweise unterscheiden. In derselben Weise kann man vorgehen, wenn man den Unterschied zwischen kriminellen und nicht kriminellen, zwischen religiösen und nicht religiösen etc Menschen bestimmen will, aus was für Gründen auch immer.

Ein derartiges Vorgehen ist explorativ. Bei explorativen Untersuchungen geht man im Allgemeinen nicht von bestimmten Hypothesen aus, sondern man versucht, Daten zu sammeln, aus denen sich Hypothesen ergeben. Ob diese Hypothesen korrekt sind, muß dann geprüft werden. Untersuchungen, die auf den Test von Hypothesen zielen, sind inferentiell: man möchte aus den Daten Schlüsse über die Gültigkeit der betrachteten Hypothesen ziehen. Inferentielle Untersuchungen werden im nächsten Abschnitt besprochen.

1.2 Inferentielle Untersuchungen

Die Unterscheidung zwischen explorativen und inferentiellen Untersuchungen gilt natürlich nicht nur für Fragen nach unterschiedllichen Persönlichkeiten, sie gilt im Prinzip für alle Fragen, die sich in der Forschung ergeben. Wie explorative Untersuchungen im Detail aussehen hängt dabei von der jeweiligen Fragestellung ab. Inferentielle Untersuchungen können sich auch aus vorangegangenen inferentiellen Untersuchungen ergeben, ebenso wie sich explorative Untersuchungen aus inferentiellen Studien ableiten lassen.

Sowohl in explorativen wie auch in inferentiellen Untersuchungen kann die Statistik eine Rolle spielen. Untersuchungen, in denen auf den Einsatz von Statistik verzichtet wird, heißen qualitative Untersuchungen, – sie sind in erster Linie explorativ. In explorativen Studien werden deskriptive Verfahren angewandt: man bestimmt elementare Statistiken wie Mittelwerte und Varianzen, stellt Häufigkeitsverteilungen auf oder berechnet Korrelationen (Maße für Assoziationen zwischen gemessenen Größen). Weiterführende Verfahren können dann unter Umständen aus Kovarianzen oder Korrelationen Informationen über "latente Variablen", also nicht direkt beobachtbare Variablen gewinnen, die die beobachteten statistischen Assoziationen zwischen verschiedenen Größen "erklären". Inferenzstatistische Verfahren prüfen dann, ob z.B. Mittelwertsunterschiede "signifikant", d.h. nicht nur zufällig sind, ob Korrelationen sich signifikant von Null unterscheiden, der beobachtete Zusammenhang also nicht nur zufällig ist, etc.

1.3 Protokollsätze, Induktion, und Deduktion

Die bisherige Darstellung kann den Eindruck erwecken, als würde man in der Wissenschaft von explorativen Beobachtungen induktiv zu Hypothesen gelangen, oder, alternativ dazu, Hypothesen deduktiv aus Theorien ableiten. Umgangssprachlich bedeutet Induktion so viel wie von besonderen Beobachtungen bzw. von Aussagen über Beobachtungen zu allgemeinen Aussagen über bestimmte Mengen von elementaren Aussagen, den Protokollsätzen, zu allgemeinen Aussagen zu gelangen, die zunächst

den Status von Hypohesen haben. Diese Ansicht wurde bereits im 19-ten Jahrhundert formuliert und wurde dann insbesondere von einigen Philosophen des *Wiener Kreises*, namentlich von Rudodf Carnap (1891 – 1970) vertreten. Der Wiener Kreis war eine Gruppe von ungefähr 20 Philosophen, von denen die meisten Physiker oder Mathematiker waren. Die Gruppe bestand zwischen den Jahren 1922 bis 1936; zum Wiener Kreis im engeren Sinne gehörte man dann, wenn man zu den wöchentlichen privaten Sitzungen des Philosophen und Physikers Moritz Schlick (1882 – 1936), dem Begründer des Wiener Kreises, eingeladen wurde. Man wollte eine im Sinne des Physikers und ebenfalls Philosophen Ernst Mach (1838 – 1916) positivistische, d.h. metaphysikfreie und nur auf objektiven Erfahrungen beruhende Philosophie begründen, wobei die *neue Logik* des britischen Mathematikers und Philosophen Bertrand Russell (1872 – 1970) eine zentrale Rolle spielen sollte: Russell hatte, zusammen mit Alfred North Whithead, die formale Logik weit über die seit Aristoteles existierende Syllogistik hinaus entwickelt, um Grundlagenprobleme in der Mathematik lösen zu können. Deshalb nannten sich die Mitglieder des Wiener Kreises auch logische Empiristen, oder Neopositivisten. Diskussionen zwischen R. Carnap und einem anderen Mitglied des Wiener Kreises, Otto Neurath (1882 – 1945), der ausnahmsweise kein Physiker oder Mathematiker, sondern Ökonom war, zeigten aber, dass "metaphysikfreie" Protokollsätze kaum jemals existieren. Die Details dieser Debatte können hier nicht dargestellt werden[1], – der wichtige Punkt hier ist, dass Protokollsätze, die in explorativen Forschungen aufgestellt weden, im Allgemeinen bereits theoretische und eventuell auch 'metaphysische' Komponenten enthalten, womit nicht etwa theologische Komponenten gemeint sind, sondern implizite Annahmen etwa über die Struktur des Raumes und der Zeit – nach Newton ist der Raum "absolut" und die Zeit "verfließt" unabhängig vom Raum gleichmäßig. Es ist eine reine ad hoc-Annahme, sie folgt nicht aus einem bereits bewiesenen Prinzip, vereinfacht aber die Gleichungen. Der Philosoph Immanuel Kant (1724 – 1804) hielt diese Annahmen für a priori wahr, nicht weiter hinterfragbar und empirisch nicht beweisbar, und insofern für metaphysisch. Spätestens seit Einstein (1905)[2] haben wir Grund zur Annahme; dass Kant hier falsch lag. Der französische Mathematiker, theoretische Physiker und Wissenschaftsphilosoph Henri Poincaré (1854 – 1912) hatte bereits ähnliche Betrachtungen angestellt, die in ein ähnliche Richtung gingen. Trotzdem gehen wir in vielen Untersuchungen und im täglichen Leben vom newtonschen Modell aus, weil es einfacher als das einsteinsche Modell ist und die Rechnungen drastisch vereinfacht, – und wir machen damit eine metaphysische Annahme.

Der ebenfalls aus Wien stammende Philosoph Karl R. Popper (1902 – 1994) kritisierte ebenfalls den Carnapschen Ansatz, induktiv über Protokollsätze zu theoretischen Aussagen zu gelangen und elaborierte die schon von dem britischen Philosophen David Hume (1711 – 1776) formulierte Kritik am Begriff der Induktion; Induktion könne logisch nicht gerechtfertigt werden, weshalb jede Hypothesenbildung

[1] Eine relativ knappe Darstellung finden man in Mortensen, Wissenschaftstheorie III (1), Abschnitt 6, s. http://www.uwe-mortensen.de

[2] Publikation der Speziellen Relativitätstheorie

ein deduktiver Akt sei. Popper sagte von sich, er sei es gewesen, der dem Positivismus den Todesstoß gegeben habe; seine deutschen Widersacher Th. W. Adorno (1903 – 1969) und Jürgen Habermas (1929 –) haben ihn gleichwohl unverdrossen einen Positivisten geschimpft, sie setzen Empirismus und Positivismus gleich, was nicht korrekt ist. Beide Philosophen behaupten, dass kritische, "philosophische" Reflexion mehr Einsicht bringe als eine "öde" (Dilthey) Empirie[3]. Ein Grund für diese Ansicht ist, dass sie sich der hegelschen Philosophie verpflichtet sehen. Hegel hatte in seiner *Phänomenologie des Geistes* behauptet, der Wahrheit durch die *Arbeit der Begriffe* näher zu kommen, denn "das Ganze ist das Wahre", und dem Ganzen, auch Totalität genannt, nähere man sich durch die dialektische Methode. Die geisteswissenschaftlichen Psychologen beziehen sich nicht notwendig auf Hegel, wenn sie der Ansicht sind, durch Introspektion und "Verstehen" der psychischen Prozesse allgemeingültige Einsicht in eben diese Prozesse bekommen zu können. Sie übersehen, dass unsere Introspektion uns kein objektives Wissen über uns selbst verschafft und per Analogschluß auch das Verstehen anderer nicht notwendig ermöglicht. Nisbett & Wilson (1977) und Nisbett & Ross (1980) haben eine Reihe von Experimenten durchgeführt, deren Ergebnisse Zimmer (1986) knapp mit der Bemerkung zusammenfasste: "wir *glauben* immer nur, zu wissen, warum wir etwas tun oder meinen." Zimmer (1986), p. 248. Die Details der Ergebnisse sollen hier nicht in aller Länge dargestellt werden, bis auf einige Zitate:

- "Das Beweismaterial [d.h. die experimentellen Ergebnisse] rechtfertigt also den größten Pessimismus hinsichtlich der menschlichen Fähigkeit, die eigenen Denkprozesse zutreffen zu beschreiben." (Nisbett & Wilson (1977), p. 247)

- Wenn Vpn[4] andere Personen beurteilen sollen, greifen sie nicht auf eigene Urteile zurück, sondern auf gängige Pop-Theorien, die gerade im Schwange sind: "Wie andere Menschen in Alltagssituationen haben wahrscheinlich auch die Probanden in diesen Experimenten nicht einmal versucht, ihre Erinnerungen an ihre eigenen Denkprozesse zu Rate zu ziehen." (Nisbett & Wilson, p. 249)

- "Es ist beängstigend, dass man nichts Gewisseres über die Arbeitsweise des eigenen Geistes weiß als ein Außenstehender, der die Lebensgeschichte kennt und weiß, was einen beeinflußt haben könnte, als man sich sein Urteil bildete." (Nisbett& Wilson, p. 257)

[3]Empirie kann durchaus öde oder auch blöde sein: So werden jährlich "IgNobel Preise" für die trivialste, dümmste etc Forschung aus allen Wissenschaftsbereichen verliehen (www.improbable.com/ig/ig-pastwinners.html), – der IgNobelpreis 2001 ging an Chittaranjan Andrade und B.S. Srihari von National Institute of Mental Health and Neurosciences in Bangalore: A Preliminary Survey of Rhinotillexomania in an Adolescent Sample, in *Journal of Clinical Psychiatry*, 62, 2001, 426-431. Dort wird festgestellt, dass das Nasebohren eine weitverbreitete Aktivität bei Adoleszenten ist. Der Ig Nobelpreis im Jahr 2000 wurde an Deavid Dunning und Justin Kreuger von der University of Illinois verliehen für ihre Arbeit: Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 77, 1999, 1121-1134.

[4]Versuchspersonen

Grundsätzlichen Kritikern der empirischen Forschung sei in Erinnerung gerufen, dass auch die Hermeneutik sich auf Erfahrung beruft, denn letztlich geht es um die Interpretation von Aussagen über die Welt. Empiriker versuchen nur, mögliche Interpretationen objektiv zu überprüfen. Geisteswissenschaftler habe eine notorisch schlechte Beziehung zum Begriff der Kausalität: damit seien mechanische Abhängigkeiten gemeint, die "im Geistigen" nicht existierten. Der Philosoph J. Habermas (wie auch Th. W. Adorno in seinen diversen Schriften) führt[5] bei seiner Preisung der Psychoanalyse den Begriff der "Kausalität des Schicksals" ein. Diese sei nicht die Kausalität der Natur, sondern eine Eigentümlichkeit des Geistes, mit deren "symbolischen Mitteln" sie (die istKausalität des Schicksals) herrsche, die aber durch "die Kraft der Reflexion bezwungen" werden könne. Woher er das weiß, läßt Habermas offen, er betreibt, was Popper *Offenbarungsphilosophie* bezeichnet: Habermas schreibt seine Thesen auf, als habe er sie nicht aus irgendwelchen Grundannahmen logisch abgeleitet, sondern als seien sie ihm offenbart worden, und der Leser muß sich den habermaschen Behauptungen anheimgeben, die – ironischerweise – Teil einer "Kritischen Theorie" sein sollen. Die schicksalhaften Ursachen des Geistes lassen sich, durch Reflexion, "aufheben", – wie dies geschieht, bleibt ebenfalls Habermas' Geheimnis, vermutlich bezieht er sich auf den hegelschen Begriff des Aufhebens. Dies ist ein Prozess, der nach Hegel vom Widerspruch zwischen These und Antithese zur Synthese führt[6]. Jedenfalls erscheint "das Geistige" als etwas über dem Natürlichen Schwebendes. Auch hier wird nur behauptet und nichts bewiesen oder auch nur hypothetisch formuliert. Das Gerede etwa von der Psychoanalyse als einer "anderen Wissenschaft" ist leer.

Definitionen, insbesondere operationale: Was aber an all diesem philosophischen Gerangel für Empiriker relevant ist, ist die Einsicht, dass die in einer empirischen Untersuchung verwendeten Begriffe so zu definieren sind, dass ihre Relation zu empirischen Beobachtungen deutlich wird. Denn durch das assoziative Schweben zwischen verschiedenen semantischen Konnotationen wird in der Hermeneutik alles irgendwie möglich, – oder auch nicht möglich, ganz, wie man es gerade haben möchte. Für eine nachprüfbare Empirie sollen also Indiktoren für einen Begriff gefunden werden, die z.B. den Begriff einer Messung zugänglich machen. Diese Art der Definition wurde von Bridgman (1927)[7] *operationale Definition* genannt. Will man etwa den Begriff 'geistige Versandung' bei der Charaktierisierung von Alkoholikern verwenden, so muß man angeben, wie man denn geistige Versandung feststellen, d.h. im allgemeinen Sinne messen will, denn wir haben alle eine intuitive Idee, was mit geistiger Versandung gemeint sein könnte, können aber im Gebrauch dieses Begriffes im Einzelfall voneinander abweichen, wenn es um die Feststellung von geistiger Versandung bei einer bestimmten Person geht. Es ist durchaus möglich, dass verschiedene ForscherInnen den Versandungsbegriff auf verschiedene Weise operational definie-

[5] In *Erkenntnis und Interesse* (1969)

[6] Den Dreischritt These → Antithese → Synthese hat Hegel in dieser Form nicht beschrieben, es handelt sich um eine Kurzfassung der hegelschen Theorie der Dialektik, die Moritz Chalybäus (1796 – 1862) formulierte.

[7] Percy Williams Bridgman (1882 – 1961), Physiker, Nobelpreis 1946

ren, aber die Tatsache, dass derartige Definitionen explizit genannt werden müssen, macht dann klar, wie es zu unterschiedlichen Bewertungen eines Phänomens (etwa: Alkoholismus) kommen kann.

1.4 Begriffe, Variablen, Konstrukte

1.4.1 Variablentypen

Eng damit zusammen hängt damit auch die Definition von Größen, die beobachtet werden sollen. Man spricht dabei von *Variablen*, womit angedeutet werden soll, dass die entsprechenden Größen von einer Beobachtung variieren können. Wichtig ist dabei insbesondere die Unterscheidung zwischen *abhängigen* und *unabhängigen* Variablen. Unabhängige Variable sind solche, die vom Experimentator kontrolliert werden können und von deren Wert die der Wert der abhängigen Variablen eben abhängt. So kann man die Versandung als abhängige Variable betrachten, deren Ausmaß von den unabhängigen Variablen (i) Intensität des Alkoholgenusses, und (ii) Dauer des (übermäßigen) Alkoholgenusses abhängt. Bei diesen unabhängigen Variablen ist einigermaßen klar, wie ihr Wert bestimmt werden kann, bei der abhängigen Variablen müssen aber noch die entsprechenden Indikatoren für Versandung geannt werden, so dass u.U. mehrere, möglicherweise wieder voneinander abhängende abhängige Variablen spezifiziert werden. In diesem Zusammenhang ergeben sich zwei weitere Begriffe, die bei der Definition von Variablen eine Rolle spielen:

1. *Reliabilität:* Eine Variable soll möglichst reliabel sein; dies ist der Fall, wenn ihr Wert von einer Messung zur nächsten bei gleicher Ausprägung der gemessenen Größe möglichst wenig variiert, d.h. wenn möglichst wenige oder gar keine zufälligen Effekte in die Bestimmung des Werts der Variablen eingehen.

2. *Validität:* Die Variable ist valide (gültig), wenn sie tatsächlich auch die Größe erfasst, die sie erfassen soll. Reliabilität und Validität sind zumindest im Prinzip empirisch bestimmbare Größen; wie sie empirisch erfasst werden, wird insbesondere in der Klassischen Testtheorie diskutiert.

Variablentypen: Für die Versuchsplanung sind die folgenden Variablentypen von Bedeutung:

1. *Intervenierende Variablen*: Personen zeigen in verschiedenen Situationen eine ähnliche Verhaltensweisen, z.B. in als eng empfundenen Räumen, seien es kleine Zimmer, Fahrstühle, Flugzeuge mit eng gestellten Sitzen etc erhöht sich der Blutdruck. Man ”erklärt” sich den Zusammenhang durch Einführung einer hypothetischen Variablen, die zwischen der Situation und der Reaktion (erhöhter Blutdruck) vermittelt.

2. *konfundierende Variable:*[8] In einem Experiment werden unabhängige und abhängige Variablen definiert. Die unabhängigen Variablen definieren experimentelle Bedingungen, die auf die abhängigen Variablen einwirken oder auch nicht einwirken – welche Fall zutrifft, soll in der Untersuchung herausgefunden werden. Eine nicht explizit kontrollierte Variable kann aber störend auf die abhängigen Variablen einwirken (*Störvariable*). In einem Experiment zu Arten des Vergessens werden Vpn angeworben, ohne auf die Zugehörigkeit zu einer bestimmten Altersgruppe zu achten. Die Art des Vergessens kann aber vom Alter einer Vp abhängen, so dass Effekte auftreten, die die Daten verzerren. Diese Effekte werden gewissermaßen mit der Wirkung der unabhängigen Variablen verwechselt (konfundiert). Das folgende Beispiel wurde von Nachtigall et al. (2000) konstruiert.

   Bei der Überprüfung der Wirksamkeit von Therapien werden oft nur geringe Effekte der Therapie gefunden, manche liegen auf dem Niveau von Placebos. Derartige Daten können allerdings täuschen, da konfundierende Variable nicht berücksichtigt werden. Tabelle 1 zeigt einen solchen Fall. Es sind zwei Gruppen

Tabelle 1: Wirkung einer Therapie, aggregierte Daten

	Heilung		
Therapie	ja	nein	Σ
ja	30	20	50
nein	30	20	50

   von Personen betrachtet worden: eine Gruppe, die sich einer Therapie unterzogen hat, und eine gleichgroße Kontrollgruppe. Auch bei der Kontrollgruppe trat eine Heilung auf, – es kommt hier nicht darauf an, zu spezifizieren, was "Heilung" bedeutet, da hier nur die Wirkung einer konfundierenden Variablen illustriert werden soll. Insgesamt haben 100 Personen an der Studie teilgenommen (die Daten sind fingiert!). Die bedingten Wahrscheinlichkeiten einer Heilung, gegeben eine Teilnahme oder Nicht-Teilnahme an der Therapie, ist in beiden Fällen

$$P(\text{Heilung}|\text{Th. ja})] = \frac{30}{50} = .6, \quad P(\text{Heilung}|\text{Th. nein}) = \frac{30}{50} = .6.$$

   Demnach kann man folgern, dass der Therapieerfolg nur einem Placebo-Effekt entspricht.

   Es ist aber eine weitergehende Analyse der Daten möglich, da man zu Beginn der Studie die Motivation der Personen, sich überhaupt einer Therapie zu unterziehen, gemessen hatte. 60 Personen waren motiviert, 40 nicht. Man kann nun die Daten für diese beiden separat betrachten. Man erhält nun Die

[8] von lat. confundere = verwechseln

Tabelle 2: Daten für motivierte und nicht motivierte Personen

motivierte Patienten				nicht motivierte Patienten			
	Heilung				Heilung		
Therapie	ja	nein	Σ	Therapie	ja	nein	Σ
ja	12	6	18	ja	18	14	32
nein	26	16	42	nein	4	4	8

Tabelle 3: Heilungserfolge bei motivierten und nicht motivierten Patienten

motivierte Patienten	nicht motivierte Patienten
$P(\text{Heil.}\vert\text{Therapie}) = \frac{12}{18} = .67$	$P(\text{Heil.}\vert\text{Therapie}) = \frac{18}{32} = .56$
$P(\text{Heil.}\vert\text{keine Therapie}) = \frac{26}{42} = .62$	$P(\text{Heil.}\vert\text{keine Therapie}) = \frac{4}{8} = .5$

Heilungserfolge sind bei den motivierten Patienten durchweg höher, – auch die Spontanheilungen bei nicht therapierten Patienten.

Addiert man die beiden Tabellen in Tabelle 2, so erhält man wieder die Tabelle 1. Die aggregierte Tabelle liefert ein anderes Ergebnis als die beiden Tabellen, die aggregiert wurden. Dieses Phänomen ist als *Simpsons Paradox* bekannt. Das Simpsonsche Paradox[9] ist ein Effekt konfundierender Variablen. Auf Seite 58 wird noch einmal darauf zurückgekommen.

3. *Moderatorvariablen:* Hierbei handelt es sich um Variablen, die systematischen Einfluß auf die abhängigen Variablen haben. So kann sich mathematische Begabung auf die Lernleistung im Mathematikunterricht so auswirken, dass mathematisch begabte Schüler unabhängig vom Unterrichtstil des Lehrers gute Leistungen erbringen, während die Leistungen von weniger begabten Schülern stark vom Stil eines Lehrers abhängen. Das Geschlecht kann eine Moderatorvariable sein: Mädchen habe eine höhere Wahrscheinlichkeit, in der Schule bessere Noten zu bekommen als Jungen.

1.4.2 Operationale Definition von Begriffen bzw. Variablen

Es sei zunächst an den Begriff der Gesetzmäßigkeit erinnert. Dies sind Aussagen, in denen gewisse Bedingungen und ihre Konsequenzen festgestellt werden. In der Psychologie haben sie im Allgemeinen eine probabilistische Form: Wenn Bedingung B gegeben ist, dann tritt das Ereignis Ereignis E ein, $B \Rightarrow E$. Freud hat im Rahmen seiner Triebtheorie die Behauptung aufgestellt, dass Frustration – erzeugt durch das Versagen einer "Triebabfuhr" – stets Aggression erzeuge, und Dollard et a. (1939)

[9] Allgemeine Diskussion: Zalta, in http://plato.stanford.edu/entries/paradox-simpson/

haben diese These weiter elaboriert Es zeigt sich aber, dass Frustration nicht, wie von Freud behauptet, *stets* zu Aggresion führt, so dass man allenfalls

$$F \overset{\text{stoch}}{\longrightarrow} A = P(A|F). \tag{1.1}$$

schreiben kann. Bekanntlich Freud hat gerne und viel behauptet, ohne sich wirklich um einen Beweis seiner Behauptungen zu kümmern, es genügte ihm, zu beteuern, dass er von seiner jeweiligen Behauptung fest überzeugt sei. Die Frage ist also, wie man Freuds *Hypothesen* testen kann. Dazu muß man auf jeden Fall F und A *messen*. Dies bedeutet, A und F in Termen von Variablen, für die Skalen existieren zu definieren. Die umgangssprachlichen Definitionen sind nicht hinreichend: Diese Definitionen von Frustration und Aggression erweisen sich für Zwecke dees messens als nicht hinreichend. So wird man auf den Begriff der *operationalen* Definition geführt. Nach Dollard et al. könnte man die Frustration mit der Definition "Frustration ist eine Emotion, die entsteht, wenn eine Person daran gehindert wird, ein angestrebtes Ziel zu erreichen" charakterisieren. Man könnte demnach Frustration experimentell erzeugen, indem man eine Versuchsperson dazu bringt, einerseits ein bestimmtes Ziel erreichen zu wollen, z.B. eine Aufgabe zu lösen, und sie dann anderseits davon abhält, sie tatsächlich zu lösen. Man muß nun noch eine Reaktion der Vp definieren, die anzeigen soll, in welchem Ausmaß die Person frustriert ist. Wenn man sagt, dass diese Reaktion eben eine aggressive Handlung sein soll, weil nach Hypothese Frustration ja Aggression erzeugt, so begeht man die berüchtigte *petitio principii*, die "Erschleichung des Beweises", denn man setzt damit implizit voraus, was erst noch gezeigt werden soll, nämlich die angebliche Kopplung der Aggression an die Frustration. Man muß also eine Reaktion finden, die das Ausmaß der erlebten Frustration reflektiert, die aber noch keine aggressive Reaktion ist. Hier wird deutlich, dass man erst einmal herausfinden muß, worin eine Reaktion, die Frustration anzeigt, überhaupt besteht. In einem zweiten Schritt muß man Reaktionen definieren, die Agression anzeigen, unabhängig davon, ob sie nun durch Frustration erzeugt wurde oder nicht. Mit Dollard et al könnte man Aggression operational so definieren: aggressiv ist "any sequence of behavior, the goal response to which is the injury of the person toward whom it is directed." Aber Verhalten kann doch "aggressiv" erscheinen, ohne dass gleich jemand verletzt wird. Hat man in den USA einen anderen Begriff von Aggression gehabt als man ihn hier hat? Die Schwierigkeit der operationalen Definition scheint u.a. darin zu bestehen, dass man zunächst eine Begriffsexegese durchführen muß, deren Adäquatheit dann bestimmt werden muß – ein komplexer Prozess, der nicht notwendig alle zufrieden stellen wird – bevor man operational definieren kann. Selbst wenn tatsächlich eine zur "Agression" korrespondierende Definition gefunden wird, wird die Vp nicht notwendig ein zu dieser Definition korrespondierendes Verhalten zeigen, und man weiß nicht, ob die Operationalisierung falsch ist oder die Hypothese nicht stimmt. Andererseits kommt man an einer Operationalisierung nicht vorbei: irgendwie muß man ja die vermutete Aggression messen. Man sieht, das sich Zusammenhänge zwischen psychischen Reaktionen leichter behaupten als nachweisen lassen.

Eine andere Hypothese Freuds (von ihm allerdings nicht 'Hypothese', sondern 'Einsicht' genannt) ist, dass eine 'Verdrängung' unangenehmer Erlebnisse existiert in dem Sinne, dass sie automatisch, d.h. nicht bewußt gewollt, in das Unbewußte abgeschoben werden, dort als Gedächtnisinhalte "überleben" und dabei ihr Unwesen treiben, so dass neurotische Störungen entstehen. Aus der Sicht der empirischen Gedächtnispsychologie ist das Postulat der Verdrängung schon deswegen unplaubsibel, war nicht klar ist, warum unbewußte Gedächtnisinhalte nicht auch vergessen werden können, – wie Vokabeln, mathematische Formeln, die man lange nicht gebraucht hat, oder wie Namen alter Bekannter, die man jahrelang nicht mehr gesehen hat, oder wie Erlebnisse, die man nicht verdrängt hat, die Erinnerung an sie im Laufe der Zeit aber immer ungenauer wird. Die Frage ist dann, wie man die Freudsche (Hypo-)These überprüft. Dazu muß man einen kontrollierbaren Mechanismus defineren, der beschreibt, wie Verdrängung überhaupt geschehen kann, und einen weiteren, mit dem man die Unveränderlichkeit dieser Erinnerungen im Unbewußten überprüft. Die Beantwortung dieser Frage erweist sich als schwierig, zumal Psychoanalytiker nicht müde werden, 'Verdrängung' sei "eigentlich" anders definiert, als sie im jeweiligen Experiment definiert wurde (diese Aussage kommt im Allgemeinen mit dem Zusatz, ein derart "naturwissenschaftlicher" Ansatz der Überprüfung sei sowieso nicht angemessen, da sich die Psychoanalyse einer "positivistischen" Überprüfung entziehe, man müsse hier hermeneutisch vorgehen). Die Hermeneutik wird zu einem Werkzeug der Teflonisierung lieb und teuer gewordener Thesen: keine Kritik bleibt haften.

So bleibt zunächst nur der Hinweis, dass es allerdings in der Tat experimentell nachweisbar ist, dass Gedächtnisinhalte durch bestimmte Ereignissen implantiert werden können, die in der Realität nie stattgefunden haben, die Vpn aber nicht entscheiden können, ob es sich um reale oder implantierte Erinnerungen handelt. Mit einem solchen Argument zeigt man aber nur, dass von Erinnerungen nicht notwendig auf das Erleben einer entsprechenden Realität geschlossen werden kann: $P \to Q$ und nun Q erlaubt ja nicht, von Q zwingend auf P zu schließen.

Ein analoges Problem stellt die These, die Deutschen seien wegen der Naziverbrechen "intrinsically vicious" (zitiert nach Dawes (2001)), oder hätten einen "autoritären Charakter" (Adorno et al. (1950)), weshalb sie Hitler gefolgt seien und seine Befehle ausgeführt hätten. Den Versuch durchzuführen, eine operationale Definition des Begriffs "intrinsisch bösartig" zu finden um zu zeigen, dass tatsächlich alle Deutschen, die während des Nazi-Regimes eineschlägig tätig geworden sind diese Eigenschaften haben, sei der/dem LeserIn als Übung empfohlen. Adorno et al. haben tatsächlich versucht, eine operationale Definition für den autoritären Charakter zu geben, indem sie eine F-Skala konstruierten (F für Faschismus oder faschistoid), mit der man nach Ansicht der Autoren das Merkmal "autoritär" erfassen kann. Die F-Skala ist dem Anspruch allerdings schon deswegen nicht gerecht geworden, weil sie als eindimensional konzipiert wurde, der Begriff des Autoritarismus aber mehrdimensional ist, aber Adorno et al.s Ansatz hat zu über zweitausend Publikationen geführt, in denen dies These diskutiert wurde, und insofern war bzw. ist sie eine fruchtbare These.

Im Zusammenhang mit der der Einführung des Begriffs des Konstrukts wird noch einmal auf die operationale Definition zurückgekommen.

1.5 Skalen und Konstrukte

1.5.1 Skalentypen

Eine Variable ist eine Größe, die verschiedene Werte annehmen kann. Variable sind unbestimmten Zusammenhängen zufällige Veränderliche, z.B. wenn sie als abhängige Variable definiert sind, denn Messungen sind i.A. mit Messfehlern behaftet. Variable können auf verschiedenen Skalenniveaus gemessen werden; die gängigen Skalen sind

1. *Nominalskala:* Man kann den Objekten Zahlenwerte zuordnen, aber die Zahlen haben nur die Funktion von Namen. Jedem Objekt kann nur eine Zahl zugeordnet werden, aber die die Relationen zwischen den Zahlen haben eine Entsprechung zu irgendwelchen Relationen zwischen den Objekten.

   Im einfachsten Fall werden nur die Zahlen 0 oder 1 zugeordnet; die Variable ist dann eine *Indikatorvariable*: Ist $\omega \in \Omega$ ein Objekt aus der Menge Ω, so bedeutet für $X(\omega) = 1$, dass ω ein bestimmtes Merkmal M hat, und $X(\omega) = 0$ heißt, dass es das Merkmal M nicht hat. Derartige Variablen heißen auch *Bernoulli-Variable*[10] Ein anderes Beispiel sind die Körperbautypen nach Kretschmer: einem Menschen wird der Skalenwert $X = 1$ zugeordnet, wenn er einen pyknischen Körperbau hat, $X = 2$, wenn ein einen leptosomen, $X = 3$ wenn er einen athletischen, $X = 4$ wenn er einen dysplastischen und $X = 5$ wenn er einen atypischen Körperbau hat. Jede andere Zuordnung von X-Werten und Körperbautypen ist zulässig, so lange die Zuordnung nur eindeutig ist.

2. *Ordinalskala:* Die "Objekte" $\omega \in \Omega$ können in Bezug auf die Ausprägung eines Merkmals M in eine Rangreihe gebracht werden:

   $$\omega_1 \succeq \omega_2 \succeq \cdots \succeq \omega_n;$$

   Das Zeichen $\succeq$ bedeutet, dass das in Frage stehende Mekrmal bei ω_j stärker ausgeprägt ist als bei ω_k; das Zeichen $\geq$ bezieht sich auf numerische Werte. Die Indices $1, 2, \ldots, n$ können als Skalenwerte einer Rang- oder Ordinalskala betrachtet werden. Tatsächlich können auch andere Zahlenwerte gewählt werden, so lange $j \geq k$ für die entsprechenden Objekte ω und ω' gilt, denn der Abstand zwischen den Zahlen reflektiert *nicht* die Größe des Unterschieds der Ausprägung von M zwischen den Objekten.

3. *Intervallskala:* Das klassische Beispiel für eine Intervallskala ist eine Temperaturskala vom Typ der Celsius- oder der Fahrenheitskala. Bei einer solchen

[10] Nach dem schweizer Mathematiker Jakob I. Bernoulli (1654 – 1705), der den Ausdruck aber nicht einführte.

Skala wird der Nullpunkt willkürlich gewählt: bei der Celsius-Skala ist die Tempratur gleich 0 Grad, wenn Wasser zu frieren beginnt. Die Einheit wird festgelegt, indem ein zweiter Wert fixiert wird. Bei der Celsius-Skala ist dies der Siedepunkt des Wassers, – er wird gleich 100 Grad gesetzt. Verschiedene Intervallskalen können durch eine lineare Transformation der Art $y = ax + b$ in einander überführt werden, wobei die x-Werte die Skalenwerte von Skala 1 und die y-Werte die Skalenwerte der Skala 2 sind.

4. *Verhältnis- oder Ratio-Skala:* Ratio-Skalen sind Skalen vom Typ der Längen oder Gewichtsskalen: hier liegt der Nullpunkt fest: gilt für ein Objekt der Skalenwert $x = 0$ so heißt dies, dass das gemessene Merkmal die Ausprägung 0 hat. Verschiedene Ratio-Skalen können durch eine Transformation $y = ax$ ineinander überführt werden; a ist ein Faktor, durch den die Maßeinheit in eine andere überführt wird: ist ein Objekt 10 cm lang, so bedeutet $y = ax$ mit $a = 10$ den Übergang zur Millimeterskala.

5. *Absolutskala:* Bei Skalenwerten einer Absolutskala ist keinerlei Transformation mehr möglich. Ein Beispiel sind beobachtete Häufigkeiten, mit der Objekte in einer Untersuchung aufgetreten sind. Eine Multiplikation der Zahlen mit einer Konstanten $a \neq 1$ hieße, die Beobachtungen zu verändern, was keinen Sinn macht.

Die Nominal- und die Ordinalskala sind *diskret*, denn es können nur Skalenwerte aus einer endlichen Menge von Zahlen verwendet werden, und diese Skalenwerte können durchnummeriert werden: $x_1, x_2, \ldots$. Bei Ratio- und Intervallskalen sind Skalenwerte aus einem Intervall der reellen Zahlen zugelassen, etwa alle Zahlen zwischen 1 und 6, oder zwischen 0 und 1. *Alle* Zahlen heißt, dass *jede* reelle Zahl aus dem betrachteten Intervall ein möglicher Skalenwerte sein kann. Man sagt, die Skalenwerte seien auf einem Kontinuum definiert. Dieser Sachverhalt ist von Bedeutung, weil Skalenwerte Merkmalsausprägungen repräsentieren, die zufällig in einem bestimmten Bereich variieren können. Skalenwerte sind in diesem Sinne stetig variierende zufällige Veränderliche. So macht man bei der Anwendung von Thurstones *Law of Comparative Judgment* von der Annahme Gebrauch, dass die Differenz $X_j - X_k$ der wahrgenommenen Merkmalsausprägungen Gauß-verteilt und damit auf einem Kontinuum verteilt ist.

1.5.2 Rating-Skalen:

In vielen Studien möchte man Reaktionen von (Versuchs-)Personen auf Stimuli im allgemeinen Sinn des Wortes erfassen: ein Stimulus kann ein visuelles Muster oder ein Klang in einem psychophysischen Experiment sein, oder ein Begriff (z.B. Liebe oder Hass: ist "Liebe" eher "hoch" oder eher "tief", eher "krank" oder eher "gesund", etc – dies sind Befragungen im Rahmen eines Polaritätsprofils, auch Semantisches Differential genannt), oder ein(e) PolitikerIn (ist A. Merkel eher "vertrauenswürdig"

oder nicht?). Dazu bittet man die Befragungspersonen (Bpn), ihre Einschätzung auf einer Skala, eben einer Rating-Skala[11], anzukreuzen. Eine typische Rating-Skala hat die Form

-3 – -2 – -1 – 0 – 1 – 2 – 3

d.h. man kann Werte von -3 bis + 3 ankreuzen. Die 0 repräsentiert dann entweder eine mittlere Ausprägung des einzuschätzenden Merkmals, oder eine Ausprägung von Null; welche Bedeutung die Skalenwerte haben, geht aus der Art der Befragung hervor. Führt man eine Befragung im Rahmen des semantischen Differentials (Polaritätsprofil) durch und gibt man z.B. die Skala "hoch" vor, so haben negative Skalenwerte die Bedeutung von "tief" oder "eher tief". Eine andere Art von Rating-Skala hat die Form

1 – 2 – 3 – 4 – 5 – 6 – 7

Dabei hat "1" die Bedeutung von "sehr ausgeprägt" (im Sinne von "sehr gut" bei Schulnoten, die ja auch Werte auf einer Rating-Skala sind), oder von "sehr wenig oder gar nicht ausgeprägt", so dass die "7" die Bedeutung von "sehr ausgeprägt" hat. Die Anzahl der vorgegebenen Skalenwerte kann im Prinzip frei gewählt werden; im einfachsten Fall sind es nur zwei: 0 oder 1, ein Merkmal ist "nicht vorhanden" oder "vorhanden", oder man hat 11 oder sogar 12 vorgegebene Skalenwerte, je nach gewünschter Feinheit oder Auflösung der Merkmalsunterschiede. Untersuchungen über die optimale Anzahl von Skalenwerten zeigen, dass Auflösungen mit mehr als 10 bis Skalenwerten kaum größere Genauigkeit der Schätzungen liefern.

Eine wichtige Frage bei Rating-Skalen bezieht sich auf das Messniveau. Bei Skalen, die nicht nur die Werte -1 , +1 oder 0 oder 1 vorgeben, hätte man gerne ein Intervallskalenniveau, weil dann die Bildung von Mittelwerten Sinn macht. Natürlich kann man auch von Skalenwerten mit nur Ordinalniveau Mittelwerte bilden, aber sie haben nicht dieselbe Aussagekraft wie Werte mit Intervallskalenniveau. Man macht sich leicht klar, dass Schulnoten weniger Aussagekraft haben, wenn sie nur Ordinalskalenniveau haben, als wenn sie Werte auf einer Intervallskala haben. Die Frage ist dann, wie man feststellt, ob Bpn in der Lage sind, ein subjektives Kontinuum in gleiche Abschnitte aufzuteilen und diese Abschnitte auf eine Rating-Skala abzubilden. Für Skalen mit Intervallskalenniveau sind eindeutig bis auf lineare Transformationen der Form $y = ax + b$, wobei y ein Skalenwert auf Skala 1 und x ein Skalenwert auf Skala 2 ist (man denke an die Celsius- und die Fahrenheit-Skala). Man läßt Objekte (i) auf einer Skale mit vorgegebenen Werten von 1 bis 5, und (ii) auf einer Skala von 1 bis 7 oder von 1 bis 10 beurteilen und prüft dann, ob sich die geschätzten Werte auf einer Skala per linearer Regression auf die Skalenwerte auf der zweiten Skala vorhersagen lassen oder nicht. Überraschenderweise gelingt eine derartige Vorhersage in vielen Fällen.

[11] Von engl. *to rate* = einschätzen, beurteilen.

Es gibt bestimmte Typen von Rating-Skalen:

1. *Häufigkeit*: Es wird gefragt, wie häufig bestimmte Ereignisse auftreten. Eltern werden z.B. gefragt, wie häufig ihr Kind Kopfschmerzen hat, in einer Marketingstudie wird gefragt, wie häufig die Bpn während eines Monats im Restaurant essen, etc. Statt numerischer Werte gibt man beschreibende Kategorien vor wie nie, selten, gelegentlich, immer , etc.
2. *Intensität*: Hier werden etwa Studierende gefragt, wie zufrieden sie mit einer Vorlesung sind. Man gibt ebenfalls oft verbale Beschreibungen vor, wie gar nicht, kaum, mittelmäßig, ziemlich, sehr bzw. außerordentlich.
3. *Wahrscheinlichkeit*: hier werden Wahrscheinlichkeitsaussagen gewünscht, etwa: Für wie wahrschenllich halten Sie es, dass die AfD Teil einer Bundesregierung wird? Mögliche Kategorien sind gar nicht oder Null, wahrscheinlich nicht, vielleicht, ziemlich wahrscheinlich, ganz sicher.
4. *Bewertung*: Es wird eine Aussage vorgegeben, der man zustimmen kann oder nicht, etwa: "Die Psychologie ist eine Naturwissenschaft." Mögliche Kategorien sind völlig falsch, ziemlich falsch, unentschieden, ziemlich richtig, völlig richtig.

Den Kategorien kann man nachträglich Zahlen zuordnen, um Mittelwerte und Varianzen und andere statistische Größen, etwa Korrelationen zwischen Skalen berechnen zu können. Dies Zahlen sind eindeutig bis auf die zugelassenen Transformationen.

Aber es gibt eine Reihe von Fragen, die sich im Zusammenhang mit Rating-Skalen stellen. So stellt die Neigung vieler Bpn, sich nicht deutlich für eine Bewertung zu entscheiden, ein Problem dar, denn diese Neigung impliziert einen Trend, die eine mittlere Kategorie zu wählen; man spricht von *zentraler Tendenz* (central tendency). Ein möglicher Ausweg besteht darin, gar keine mittlere Kategorie zuzulassen, so dass die Bp gezwungen wird, sich für eine eher "positive" oder eher "negative" Einschätzung zu entscheiden.

Ein weiteres Problem ist der *Halo-Effekt*[12]. Dies ist ein zuerst von Thorndike (1920) beschriebener Effekt, der darin besteht, dass sich ein globaler Eindruck auf die Einzelurteile auswirkt. Obwohl bestimmte Merkmale unabhängig voneinander ausgeprägt sein können, werden durch den globalen Eindruck korrelierte Einschätzungen erzeugt. Andere Autoren sprechen in diesem Zusammenhang auch von einem "logischen Fehler" (Newcomb (1931)). Halo-Effekte treten insbesondere dann auf, wenn die einzuschätzenden Merkmale diffus defniert sind. Eine Möglichkeit, Halo-Effekte zu minimieren, besteht darin, alle Objekte zunächst auf einer Skala beurteilen zu lassen, anschließend alle Objekte auf der zweiten Skala, etc.

Der *Milde-Härte-Fehler* (leniency-severity error) tritt oft auf, wenn Personen zu beurteilen sind (z.B. in Prüfungen). Hier werden systematisch zu nachsichtige oder zu harte Urteile gefällt. Ein Prüfer möchte etwa vermeiden, dass die Prüflinge berufliche Nachteile durch eine zu strenge Beurteilung haben, weil er das Fach, das er prüft,

[12]Von engl. halo = Heiligen- oder Glorienschein

für nicht so wichtig für die berufliche Arbeit hält, oder er oder sie ist umgekehrt der Ansicht, dass das geprüfte Fach sehr wichtig ist, so dass eine strenge Beurteilung angezeigt sei.

Weiter ist die Rater-Ratee-Interaktion zu beachten. Eine Bp hat ihrer Eigenwahrnehmung zufolge eine bestimmte Position auf einer Merkmalsskala und beurteilt andere Personen nach Maßgabe der eigenen Position. Es gibt dann einen "Ähnlichkeitsfehler", wobei die Bp die anderen Personen in Richtung auf die eigene Position verschätzen. Dieser Effekt tritt besonders dann auf, wenn die eigene Position der Bp eher extrem ist (das Merkmal wenig oder stark ausgeprägt ist), und einen "Kontrastfehler", bei dem die Bp den wahrgenommen Kontrast zur eigenen Position erhöht, andere Personen als z.B. als deutlich weniger fähig oder deutlich fähiger einschätzt.

Schließlich ist noch der *Primacy-recency-Effekt* zu berücksitigen. Hier hängen die Beurteilungen von den zuvor gemachten Beurteilungen ab, insbesondere wenn die ersten Beurteilungen eher extremer Natur waren.

1.5.3 Konstrukte

Die Notwendigkeit, Begriffe (Verdrängung, Aggression, Intelligenz, etc) so zu definieren, dass sie durch beobachtbare Größen gewissermaßen dingfest gemacht werden, führt zum Begriff des Konstrukts. So "weiß" umgangssprachlich jeder, was mit dem Wort Intelligenz gemeint ist, weshalb der in der Psychologie verwendete Intelligenzbegriff bei psychologischen Laien oft empörte Ablehnung auslöst: "Für mich ist Intelligenz aber etwas ganz anderes!". Der Hinweis, der Begriff der Intelligenz sei jahrzehntelang hinsichtlich aller möglicher Aspekte von 'Intelligenz' diskutiert worden und man habe sich auf den durch einen Test definierte Intelligenzbegriff geeinigt, hilft dann wenig: denn der Volkspsychologie zufolge sind professionelle, insbesondere akademische Psychologen nicht ganz richtig im Kopf und haben keine Ahnung von der "richtigen" Psychologie. Der Intelligenzbegriff der Psychologie ist ein *Konstrukt* insofern, als einerseits bestimmte inhaltliche Aspekte zur Definition ausgewählt wurden, die andererseits durch (Sub-)Tests erfasst werden können, deren "Scores" (Punktwerte) wiederum in einer Gleichung zusammengefasst werden:

$$IQ = b_1 X_1 + b_2 X_2 + \cdots + b_p X_p + e. \tag{1.2}$$

Dabei sind die X_j die Scores in Untertests, die bestimmte kognitive Fähigkeiten repräsentieren, und die b_j sind "Gewichte", mit denen die Scores in den IQ-Wert eingehen. Die Scores sind im Prinzip Skalenwerte, die im Modell (1.2) Nominal- oder Intervallskalenniveau haben können; dieser Sachverhalt wird weiter unten noch ausführlich erläutert. Der IQ-Test wird anhand gewisser Kriterien *geeicht*, d.h. die Übereinstimmung der IQ-Werte mit Leistungen, die als Intelligenzleistungen angesehen werden, wird bestimmt; in diesem Zusammenhang werden auch die Koeffizienten b_j bestimmt. "Intelligenz ist, was der Intelligenztest misst", sagte bereits 1923 der Experimentalpsychologe Edwin Boring (1886 – 1968). Solche Aussagen erregen den

Laien, aber man muß bedenken, dass *jede* Bestimmung der Intelligenz einer Person eine bestimmte, wenn auch implizite Definition von Intelligenz voraussetzt, – über die man dann trefflich streiten kann. Auf jeden Fall illustriert die Gleichung (1.2), was man unter einem Konstrukt versteht, nämlich eine bestimmte Definition, derzufolge ein Begriff, der nicht unmittelbar gemessen wird (hier: Intelligenz) als aus messbaren Größen zusammengesetzt definiert wird. Das Wesentliche an dieser Definition ist, dass nicht auf ein nicht weiter spezifiertes "Wesen" der so definierten Größe rekurriert wird ("Das Wesen der Intelligenz besteht doch darin, Zusammenhänge verstehen zu können, – kopfrechnen zu können, hat nichts mit Intelligenz zu tun!"). Konstrukte können sich im Forschungsprozess verändern, – es gibt keinen a priori "wahren" Begriff von Intelligenz (bzw. von dem, was auch immer als Konstrukt definiert wird).

> Der Begriff des Konstrukts ist kein Ausdruck "positivistischer" Willkür. Er drückt nur aus, dass man sich auf eine Definition einigen muß, wenn man Sachverhalte diskutieren will. Insofern sind Definitionen keine "wahren" Aussagen, denn es ist in vielen Fällen denkbar, einen umgangssprachlich gegebenen Begriff auch durch ein anderes Konstrukt zu spezifizieren. Die Diskussion von Daten in Bezug auf ein Konstrukt kann sich demnach auf verschiedene Aspekte des umgangssprachlich gegebenen Begriffs beziehen. In manchen Fällen ist aber die Beziehung zwischen Definition und Gesetz nicht eindeutig. Ein Beispiel ist der Begriff der Kraft, der lange diskutiert wurde: nach Newton ist Kraft = Masse × Beschleunigung, nach Leibniz aber soll Kraft = Masse × Geschwindigkeit gelten. Aus guten Gründen, die hier nicht ausführlich diskutiert werden können, hat sich die newtonsche Definition durchgesetzt; die von Leibniz gegebene Definition entspricht heute dem, was man unter 'Impuls' versteht. In der Psychologie hat sich z.B. die Definition (1.2) für den Begriff *Intelligenz* durchgesetzt.

Die Beziehung zwischen Konstrukten und Empirie ist nicht eindeutig. Freuds Begriff des Unbewußten oder des Unterbewußtseins ist sicherlich ein Konstrukt: er meint damit mehr als das, was schon vor ihm 'unbewußt' genannt worden war. Es ist ein Ort mit eigener Dynamik, in dem Gedächtnisinhalte gewissermaßen unverändert gebunkert werden können und aus dem heraus sie in Interaktion mit Trieben das "Seelenleben" eines Menschen beeinflussen können. Ob das Unbewußte in dieser Form existiert ist eine andere Frage. Durch Introspektion und Reflexion scheint man sich dieser Frage nicht nähern zu können, wir können wohl unsere Empfindungen und Gedanken registieren, haben aber keinen Zugang zu ihrer Genese, wir *konstruieren* eher die Genese als dass wir ihr zusehen (vergl. die Befunde von Nisbett & Wilson (1977), Seite 7). Man spricht von *phänomenologischer Blindheit*. Konstrukte entstehen zum Teil durch einen Prozess, den man als assoziatives Ausmelken von Metaphern bezeichnen könnte. Dazu sei noch einmal Freuds Begriff der psychischen Energie bemüht. Den Wunsch, etwas bestimmtes zu tun oder zu erleben, kann man als Antrieb erleben und vor allen Dingen metaphorisch so bezeichnen. Wie ein Motor das Auto "antreibt" wirkt dann ein "Trieb" in uns, der einer Energie entspricht. Eine weitere "Erklärung" ergibt sich nun durch Analogiebildung: wir assoziieren zu diesem Begriff, was wir aus der Physik über den Energiebegriff wissen, und übertra-

gen diese Assoziationen auf den Begriff der psychischen Energie. Da Energie in der Physik erhalten bleibt, schlußfolgern wir, dass dieser Sachverhalt auch für psychische Energie gelten muß, so dass wir weiter folgern, dass Triebenergie "abgeführt" oder "verschoben" werden muß. Wenn wir unserer sexuellen "Energie" nicht freien Lauf lassen können, müssen wir sie "sublimieren" und in "geistige Energie" verwandeln oder transformieren (eine Art der Verschiebung), – und unversehens haben wir eine Theorie, die das Entstehen von Zivilisationen erklärt. Die Psychoanalyse wird zu einem aus Konstrukten zusammengebastelten Konstrukt, dessen anscheinende Stimmigkeit den Anhängern der Theorie gleichzeitig auch als ihre Bestätigung gilt. Aber scheinbare Konsistenz ist nicht gleichbedeutend mit Wahrheit, die Frage bleibt, ob ein Konstrukt auch zur Realität korrespondiert, d.h. ob es auch *valide* ist.

Man kann der Ansicht sein, dass wir unsere Theorien von der Welt stets konstruieren, und das keineswegs nur in der Psychologie. Die entsprechende wissenschaftstheoretische Richtung heißt *Konstruktionismus*. In der Psychologie hat insbesondere der Philosoph, Soziologe und Psychotherapeut Paul Watzlawik (1921 – 2007) einen anregenden Beitrag zum Konstruktivismus in der Psychologie mit seinem 1983 zuerst erschienenen Buch *Anleitung zum Unglücklichsein* geleistet.

Konstruktvalidität: Die Frage nach der Validität eines Konstrukts ist oft nicht leicht zu beantworten. Ob der Begriff der Intelligenz, so, wie er in Intelligenztests zugrunde gelegt wird, valide ist, läßt sich schlecht fragen, – ein Konstrukt ist ja eine Definition. Was mit dem Begriff der Konstruktvalidität gemeint ist bezieht sich auf die Frage, ob die Indikatoren für das Konstrukt valide sind, also ob zum Beispiel die Fähigkeit zum Kopfrechnen oder zum Analogienfinden tatsächlich Merkmale erfasst, die im *Konstrukt* 'Intelligenz' zusammengefasst werden; man könnte ja der Ansicht sein, dass das Kopfrechnen heute für intelligentes Verhalten nicht mehr gebraucht wird, da jeder auf seinem Taschentelefon auch einen kleinen Rechner hat. Der Begriff der Konstruktvalidität bezieht sich also mehr auf eine Analyse des Konzepts, das in einem Konstrukt spezifiziert wird. Im Unterschied dazu ist mit der *Kriteriumsvalidität* die Validität eines Tests gemeint, nämlich das Ausmaß, in dem ein Test das Merkmal, das er messen soll, uach tatsächlich mißt.

Führt man ein Experiment durch, so wird im Allgemeinen die Wirkung von unabhängigen Variablen auf bestimmte abhängige Variablen untersucht. Natürlich hätte man dann gerne, dass diese Einwirkung, falls vorhanden, auch psychologisch eindeutig ist. Eine Untersuchung ist *konstruktvalide* wenn diese Eindeutigkeit gegeben ist. So hat Milgram ein Experiment durchgeführt (Milgram (1963), in dem die Bereitschaft von Personen untersucht wird, Anweisungen zu folgen auch dann, wenn dadurch anderen Menschen Schaden zugefügt wird. Es wurde dabei ein (simuliertes) Lernexperiment durchgeführt. Es sollte die Lernleistung untersucht werden, wenn Fehler bestraft wurden. Die Bestrafung bestand in elektrischen Schocks, die um so stärker sein sollten, je häufiger ein Fehler gemacht wurde. Dazu mußten die eigentichen Vpn als Versuchsleiter agieren, denen ein Wissenschaftler sagte, wie intensiv die Schocks sein sollten. Die angeblich lernende Vpn war aber ein *stooge*, d.h. ein Schauspieler, der die Reaktionen auf die Schocks nur simulierte, – was die Vp, die

als Versuchsleiter fungierte, aber nicht wußte. Sie war nicht gezwungen, den Anweisungen des "Wissenschaftlers" zu folgen. Die meisten Vpn waren aber bereit, dessen Anweisungen zu folgen, auch wenn die geschockte Person anscheinend sehr unter den Schocks litt. Dem Experiment wird wegen seiner Lebensnähe eine hohe Konstruktvalidität zugeschrieben. Das Experiment wurde im Zusammenhang mit der These, wegen der Naziverbrechen seien Deutsche irgendwie anders als andere Menschen durchgeführt. Das Experiment wurde in vielen anderen Ländern durchgeführt und lieferte stets die gleichen Ergebnisse. Ob diese Konstruktvalidität auch hinreichend groß ist, eine Erklärung für die Verbrechen der Deutschen während der Nazi-Zeit zu liefern, ist aber noch eine andere Frage: Millionen Menschen in Gaskammern oder in oder vor Gräben zu treiben, wo man sie dann erschießt, hat doch eine andere Quzaltiät, als einem Wissenschaftler in einem weißen Kittel zu gehorchen, zumal viele Täter sich freiwillig zu den Mordaktionen gemeldet haben.

Ethische Fragen Milgram wurde kritisiert, weil er seine Vpn (die im Experiment als "Versuchsleiter" dienten) "schwer traumatisiert" habe. Andererseits hat er gezeigt, dass "ganz normale" Menschen dazu gebracht werden können, anderen Menschen schweren Schaden zuzufügen, wenn sie unter geeigneten Bedingungen dazu aufgefordert werden. Nach Milgram sützen diese Resultate Hannah Ahrendts These von der Banalität des Bösen, die sie im Zusammenhang mit den Eichmann-Prozess eingeführt hatte: man muß nicht "anders" sein, um verbrecherischen Befehlen zu folgen. Diese Einsicht mag deprimierend sein, ist andererseits aber wichtig für die Einsicht in die menschliche Natur, weshalb die Beurteilung des Experiments als unethisch nicht von allen Wissenschaftlern geteilt wird.

2 Qualitative versus quantitative Untersuchungen

Diese beiden Typen von Untersuchungen entsprechen (grob) dem geisteswissenschaftlichen (qualitativ) und dem naturwissenschaftlichen (quantitativ) Ansatz der Psychologie. Qualitative Untersuchungen können wichtige Vorinformatonen liefern und zur Bildung von testbaren Hypothesen führen, die dann im Rahmen von quantitativen Untersuchungen getestet werden.

Gleichzeitig gibt es einen alten Streit zwischen Gruppen von entweder qualitativ oder quantitativ arbeitenden ForscherInnen. Dies sind Gruppen die grundsätzlich die Psychologie als Geisteswissenschaft oder grundsätzlich zumindest in methodischer Hinsicht als Naturwissenschaft sehen. Die geisteswissenschaftlich motivierten Psychologen/innen argumentieren, die einzig sinnvolle Methode sei das Verstehen im Sinne Diltheys, weshalb ein ganzheitlicher (holistischer) Ansatz notwendig sei, es ginge um Einsicht und nicht um mechanisches Erklären, dass das Ziel der "Variablenpsychologie" sei, deren Ergebnisse, sofern es überhaupt welche gebe, zu eng und zu trivial seien, um den Aufwand der Untersuchung zu rechtfertigen. Beim qualitativen Vorgehen würde man nicht von vorgefassten Theorien über den Forschungsgegen-

stand ausgehen und Schlußfolgerungen würden sich induktiv ergeben. Es käme auf die *emergente Flexibilität* an, denn der *teilnehmende Beobachter* könne flexibel auf aufscheinende (emergente) Phänomene reagieren, wohingehend quantitativ arbeitende Forscher an ihren fixen Versuchsplan gebunden seien und auf kausale Erklärungen fokussieren (Hussy et al., p. 185).

Die "naturwissenschaftlich" arbeitenden ForscherInnen dagegen stellen einige der Grundannahmen der geisteswissenschaftlichen Gruppe in Frage. Gerade im berühmten Protokollsatzstreit sei gezeigt worden, dass es eben keine voraussetzungsfreien Beobachtungen gebe, der Anspruch, ganzheitlich vorzugehen, sei einerseits leer, weil auch der quantitative Ansatz ganzheitlich sei, – sofern dies möglich sei, nur würde man explizite Modelle (z.B. den Ansatz der multiplen Regression, das Allgemeine Lineare Modell) verwenden, um spezielle Aspekte testen zu können, wohingegen der "verstehende" Ansatz der Geisteswissenschaftler akzidentelle (zufällige) Phänomene nicht von systematisch auftretenden trennen können. Aus den Beobachtungen abgeleitete Aussagen seien nur sinnvoll, wenn ihre statistische Signifikanz nachweisbar sei.

Ein derartiger Streit ist unfruchtbar. Zum einen läßt sich argumentieren, dass die Droysen-Diltheysche Unterscheidung zwischen Erklären und Verstehen artifiziell ist; der Philosoph Wolfgang Stegmüller (1923 – 1991) hat die Unterscheidung extensiv untersucht; eine Kurzfassung der Argumentation findet man in Stegmüller (1971)[13], wo er feststellt "... die Diltheysche Gegenüberstellung [ist] *die mit Abstand unfruchtbarste*" (p. 66, Stegmüllers Kursivsetzung). Andererseits ist der Fokus auf den Signifikanztest innerhalb der Statistik Gegenstand langer und intensiver Diskussionen, so dass eine Absolutsetzung des Signifikanztests nicht gerechtfertigt ist[14]. Eine ausführliche Darstellung dieser Diskussion sowie der stegmüllerschen Argumente ist hier nicht möglich, festzuhalten ist aber, dass es produktiver ist, qualitative und quantitative Forschung als verchiedene, sich aber komplettierende Seiten ein und derselben Münze zu sehen.

2.1 Qualitative Forschung

Es können hier nur einige Grundzüge qualitativer Forschung dargestellt werden, Hussy et al (2010) geben einen ausführlicheren Überblick, ebenso Breuer (1996).

Der ursprüngliche Ansatz besteht darin, dass die Forschung zum Beispiel über menschliches Zusammenleben in der natürlichen Umgebung der Menschen durchgeführt wird. Im Unterschied zum klassischen Experiment werden also Randbedingungen nicht systematisch variiert; das Verhalten von Mitgliedern jugendlicher Banden wird nur innerhalb der Bande untersucht, nicht aber in anderen Umgebungen. Der

[13] Der sogenannte Zirkel des Verstehens. In: Stegmüller (1971)

[14] Eine Diskussion der verschiedenen Ansätze findet man in Mortensen (2010/2014): http://www.uwe-mortensen.de/Inferenzstatistikd.pdf; es wird dort allerdings eine gewissen Bekanntschaft mit den Grundlagen der Statistik vorausgesetzt.

Vorteil dieser Art von Untersuchung ist offenbar, dass die Jugendlichen in einer gewohnten Umgebung handeln und sich deshalb "natürlich" verhalten. Der Nachteil wiederum ist, dass nicht deutlich wird, welche Wirkung andere Umgebungen auf sie haben. Darüber hinaus ist es oft schwierig akzidentelle (zufällige) Effekte von systematischen Effekten zu trennen.

In den späten zwanziger und beginnenden dreißiger Jahren des 20-ten Jahrhunderts waren in Deutschland und Österreich viele Menschen arbeitslos. Ein berühmtes Beispiel qualitativer Forschung, die im Übringen von zu quantitativer Forschung sehr wohl befähigten ForscherInnen durchgeführt wurde, hatte die Frage zum Gegenstand, wie sich längerfristige Arbeitslosigkeit auf die Einstellungen der Betroffenen auswirkt. Marxistisch motivierten Theorien zufolge könnte man die Entstehung revolutionärer Einstellungen beobachten, als Alternativhypothese wurde die Entstehung von Resignation und Hoffnungslosigkeit "vorhergesagt".

Die empirisch arbeitenden Marie Jahoda (Sozialpsychologin), der Soziologe und Mathematiker Paul Lazarsfeld und Jurist und Statistiker Hans Zeisel beschlossen, eine Studie zur mentalen Situation der Arbeitslosen zu unternehmen, die 1933 unter dem Titel *Die Arbeitslosen von Marienthal. Ein soziographischer Versuch über die Wirkungen langandauernder Arbeitslosigkeit* erschien und die zu einem Klassiker der Forschung auf diesem Gebiet wurde. Das Hauptergebnis war, dass Arbeitslosigkeit statt zur Revolution eher zu Apathie und Resignation führt. Jahoda et al. fanden, dass man sich der Fragestellung zunächst qualitativ nähern sollte, d.h. es wurden spezielle Interviewtechniken entwickelt, mit denen die für die Einstellungen der Arbeitslosen charakteristische Variablen bestimmt werden sollten. Auf der Basis einer derartigen qualtitativen Arbeit können dann z.B. Fragebögen entwickelt werden, die für eine quantitative Auswertung geeignet sind.

Alle drei Wissenschaftler mußten wegen der Nationalsozialisten emigrieren und forschten in England und in den USA weiter. P. Lazarsfeld hat sich u.a. in der Psychologie wegen seiner Entwicklung mathematischer Methoden zur Auffindung "latenter Variablen", also Variablen, die nicht direkt beobachtet werden können, einen Namen gemacht.

Fallstudien: Fallstudien sind ein Spezialfall qualitativer Forschung. Hussy et al. (p. 193) stellen sie als *holistische Forschungsmethode* dar, d.h. die Fälle sollen ganzheitlich untersucht werden. Ein typisches Beispiel ist die Fallgeschichte der Anna O. von Freud und Breuer (1895); Anna O. war die erste Patientin, die einer Psychoanalyse unterzogen wurde[15]. Fallstudien müssen nicht auf einzelne Personen beschränkt werden, es ist möglich, ganze Institutionen zu untersuchen, z.B. ein Krankenhaus, oder eine Abteilung eines Krankenhauses. Typisch für Fallstudien ist, dass verschiedene Erhebungsmethoden bzw. Beobachtungsmethoden verwendet werden, deren Resul-

[15]Bei Anna O. handelte es sich um Bertha Pappenheim (1895 – 1936), ursprünglich Patientin von Josef Breuer. Sie wurde als hysterisch diagnostiziert. Sie sagte, ihre Seele würde durch das Aussprechen ihrer Probleme entlastet, woraufhin insbesondere Freud die Katharsis-These entwickelte, die zu einem wesentlichen Element der Psychonanalyse wurde.

tate bei der Interpretation der Ergebnisse integriert werden.

Theorienbildung: Eine Frage, die im Zusammenhang mit qualitativer Forschung zu stellen ist, ist die nach der Formulierung von Theorien. Glaser und Strauss (1965) haben im Zusammenhang mit der Erforschung der Ineraktion von Klinikpersonal und Todkranken die *gegenstandbezogene Theorienbildung* vorgeschlagen (sie wird von ihnen als *grounded theory* bezeichnet. Der Punkt bei dieser Art von Theorienbildung ist, dass sie sich von dem hypothesenprüfenden Vorgehen der quantitativen Forschung unterscheiden soll. Glaser et al.s Idee war, dass sich die Theorie direkt auf die Daten beziehen soll. Die Stichproben werden sukzessive im Verlauf der Untersuchung gebildet. Die Methoden können beliebig gewählt werden, die =. Hauptsache ist, dass man über sie Informationen von den betroffenen Personen erhält. Es gibt drei Arten der Auswertung; (i) das *offene*, (ii) das *axiale* und (iii) das *selektive Kodieren*. Das offene Kodieren besteht in einer erste Herausbildung von Konzepten zur Charakterisierung der Befragten; diese Konzepte sind die *Codes*. Beim axialen Kodieren wird von den offenen Kodierungen abstrahiert, d.h. es werden theoretische *Codes* entwickelt; dies sind die *Kategorien*, die der Strukturierung des Datenmaterials dienen. Beim selektiven Kodieren werden die axialen *Codes* zueinander in Beziehung gesetzt, so dass ein Gesamtmodell bzw. eine Gesamttheorie entsteht. Dies ist die eigentliche Theoriebildung.

Die verwendeten bzw. entwickelten Kategorien werden während der Arbeit zu einer Theorie verknüpft. Die Untersuchung gilt als abgeschlossen, wenn neue Fälle keine neue Information liefern; dies ist der Zustand der *theoretischen Sättigung*. Die Datenerhebung und die Theoriebildung bilden simultane, interagierende Prozesse.

Natürlich stellt sich die Frage, wie man eine derart zustandegekommene Theorie in Bezug auf ihre Adäquatheit überprüft. Hussy et al. gehen auf diese Frage nicht weiter ein. Da der Ansatz sich bewußt vom quantitativen Ansatz unterscheidet kann man vermuten, dass ein Test der Theorie nicht weiter vorgesehen ist, zumal ja davon ausgegangen wird, dass sie sich unmittelbar aus den Daten ergibt. Gleichwohl ergeben sich einige Fragen: selbst wenn eine gegenstandbezogene Theorie in Teamarbeit aufgestellt wird, kann es doch sein, dass es zu teamspezifischen Kategorienbildungen kommt, die von den Voreinstellungen der jeweiligen ForscherInnen abhängen. In gewisser Weise haben Breuer und Freud (1895) mit der Psychoanalyse eine "gegenstandsbezogene Theorie" formuliert, – sie haben anhand ihrer Beobachtungen Verhaltenskategorien gebildet und diese aufeinander bezogen, und natürlich hingen diese Kategorienbildungen von impliziten theoretischen Vortellungen ab. Die Kritik an den Protokollsätzen, die in der frühen Phase des Wiener Kreises den ihm angehörenden Philosophen als die Basis für zu konstruierende Theorien galten, greift ja auch hier: Aussagen über unmittelbar gegebene Sachverhalte sind bereits "theoriegeladen". Bei Breuer und Freud war dieser theoretische Hintergrund durch die diffuse Vorstellung von Trieben gegeben, von denen sie nun annahmen, dass sie aus dem Unterbewußtsein heraus wirken; das "Triebhafte" stellte ja in der Wiener Gesellschaft der Jahrhundertwende ein tabuiertes Thema dar, insbesondere wenn es um den "Sexualtrieb" ging. Ein Evidenzerlebnis von Adäquatheit stellt aber noch lange

keinen Wahrheitsbeweis dar. Darüber hinaus stellt sich die Frage, ob das Verhalten des Klinikpersonals, das von Glaser et al untersucht wurde, dem Verhalten in irgend einer anderen Klinik entspricht oder ob man eine sehr klinikspezifische Theorie erhält. Schließlich stellt sich die Frage nach den akzidentellen Aspekten des Datenmaterials, das hier "ganzheitlich" in die Theorie integriert wird. Auf diese Frage wird bei der Diskussion der Frage nach den Prädiktoren in einem multiplen Regressionsansatz noch einmal eingegangen. Andererseits sind die so gewonnenen theoretischen Vorstellungen durchaus nützlich für weitgehende, quantitavive kreuzvalidierende[16] Untersuchungen.

Deskriptive Feldforschung: Diese ist eine Erfindung der Ethnologen: ForscherInnen leben in einer bestimmten Kultur, um diese gewissermaßen von innen heraus zu verstehen. Die Datenerhebungstechnik ist die *teilnehmende Beobachtung*. Die Technik ist offenbar auch für Psychologen interessant, die verstehen wollen, wie ein bestimmter Kulturkreis funktioniert. Natürlich kann man sich einem solchen Kreis (es müssen ja nicht die Bewohner Western Samoas oder Papua Neuguineas sein) mit einem Arsenal von Variablen nähern, die sich ordentlich messen lassen, aber es ist möglich, dass diese Variablen die relevanten Aspekte der betrachteten sozialen Gruppe (man denke an Obdachlose, die sich bestimmter bürgerlicher Normen entledigt haben) gar nicht erfassen. Die Analyse der Daten entspricht im Wesentlichen dem Vorgehen bei der Formulierung gegenstandsbezogener Theorien. Auch hier gilt, dass eine Theoriebildung anhand unmittelbar gegebener Beobachtungen keineswegs notwendig der Realität adäquat sein muß. Ein Beispiel hierfür ist das Bild, dass die berühmte Anthropologin Margaret Mead (1928) von der polynesischen Gesellschaft auf Western Samoa geliefert hat. Sie berichtet, dass das Zusammenleben von Jugendlichen und ihren Eltern frei von Spannungen sei, – Samoa als Paradies. Sie hatte übersehen, dass Polynesier der Ansicht sind, dass man Fremden keineswegs alles erzählen darf, da diese sonst Macht über sie bekämen. Die samoanische Gesellschaft ist keineswegs so konfliktfrei, wie sie sich M. Mead dargestellt hat (vergl. Pinker (1997)).

2.2 Quantitative Forschung

Forschung ist "quantitativ", wenn sie in einem sehr allgemeinen Sinn messbare Variablen definiert. Es muß gleich gesagt werden, dass die Grenze zwischen qualitiativen und quantitativen Methoden oft fließend ist: die Betrachtung von Häufigkeiten kann rein deskriptiven Zwecken dienen und auf rein qualitative Folgerungen zielen. Häufig wird bei der quantitativen Forschung neben der Messung von Variablen noch der *statistische* Test von Hypothesen vorgenommen. Der wesentliche Unterschied zwischen quantitativen und qualitativen Untersuchungen scheint darin zu bestehen, dass die zu erhebenden Variablen vor der eigentlichen Untersuchung definiert werden, so dass gewisse Vorkenntnisse über die zu betrachtenden Variablen vorhanden sind. Andererseits kann man auch Untersuchungen durchführen, bei denen einfach

[16] Auf den Begriff der Kreuzvalidierung wird weiter unten noch näher eingegangen.

Hypothesen über den Effekt bestimmter Variablen getestet werden, ohne dass einer solchen Untersuchung Forschung der qualitativen Art vorangegangen sein muß, in der sich die betrachtete Variable als relevant gezeigt hat. Eine Reihe von Beispielen mag hilfreich sein, um den Begriff der quantitativen Forschung zu verdeutlichen.

Effekte unabhängiger Variablen: Ein zweiter Typ von quantitativer Forschung ergibt sich, wenn man gar nicht auf erklärende Modelle der betrachteten Prozesse abzielt, sondern nur den Effekt bestimmter Variablen erfassen will. So kann man fragen, ob die Vergessenskurven von Variablen wie Alter, Geschlecht (dies ist eine nominal skalierte Variable), Intelligenz etc abhängen. Die getesteten Hypothesen sind dann *Nullhypothesen*: jüngere und ältere Menschen unterscheiden sich nicht hinsichtlich ihrer Fähigkeit, sinnlose Silben zu memorisieren, Männer und Frauen unterscheiden sich nicht hinsichtlich dieser Fähigkeit, etc. Die Alternativhypothesen sind dann einfach die Negationen der Nullhypothesen. Interessant sind bei dieser Art von Experiment die möglichen *Wechselwirkungen* zwischen den den unabhängigen Variablen. So kann es sein, dass sich *im Mittel* Männer und Frauen nicht in ihren Gedächtnisleistungen unterscheiden, dass aber das Alter einen Einfluß auf die Gedächtnisleistungen hat, wobei die Art des Einflusses für Männer und Frauen verschieden ist.

Zusammenhänge zwischen Variablen: Man misst eine Reihe von Variablen bei einer Anzahl von Personen, etwa V_1 die Fähigkeit, sinnlose Silben zu behalten, die Reaktionszeit V_2 auf einfache Stimuli (innerhalb eines Zeitintervalles wird zu einem zufällig gewählten Zeitpunkt ein Ton erzeugt, und die Versuchsperson (VP) muß so schnell wie möglich eine Taste drücken; gemessen wird die Zeit (Reaktionszeit), die zwischen der Darbietung des Tons und dem Drücken der Taste verstreicht), V_3 der Blutdruck der Vpn während des Reaktionszeitexperiments, etc. Insgesamt werden bei jeder Vp n Variable $V_1, \ldots, V_n$ gemessen. Der Messwert bei der i-ten Vp ($i = 1, 2, \ldots, m$) und der j-ten Variable ($j = 1, 2, \ldots, n$) sei v_{ij}. Untersucht wird die Frage, ob sich die v_{ij} als Kombinationen von hypothetischen "latenten Variablen" oder "Dimensionen" darstellen lassen. Derartige Untersuchungen werden etwa in der Persönlichkeitspsychologie vorgenommen, wo man an Persönlichkeitsdimensionen ("Persönlichkeitsfaktoren") interessiert ist. Hypothesen werden nicht notwendig getestet, die Datenanalysen sind zunächst rein deskriptiv.

Meinungsumfragen können von dieser Art sein. Man ist etwa daran interessiert, in Bezug auf welche Eigenschaften sich die Wähler einer bestimmten Partei von den Wählern anderer Parteien unterscheiden, d.h. etwa welche Wähler, die bislang die SPD, die CDU, Die Grünen oder die FDP gewählt haben, wählen nun die AfD? Man hat Verfahren entwickelt, die es gestatten, aus einer Zahl von Eigenschaften diejenigen herauszufiltern, die vorauszusagen gestatten, wer zur AfD wechselt, wobei die Vorhersagefehler minimiert werden. Auch hier kann entweder rein deskriptiv oder hypothesenprüfend vorgegangen werden.

Prozessmodelle: Ebbinghaus-Kurven Hermann Ebbinghaus hat erste systematische Untersuchungen zum Lernen und Vergessen gemacht und damit die moderne

Gedächtnispsychologie begründet. Um den Effekt Assoziationen zwischen Wörtern zu vermeiden, lernte er Listen sinnloser Silben auswendig und prüfte, wieviele Silben er nach bestimmten Zeitspannen (Stunden, Tage, Wochen, Monate)noch reproduzieren konnte. Es ergaben sich stets ähnliche Kurven (Ebbinghaus-Kurven oder Lernkurven), die auf den ersten Blick Exponentialfunktionen ähnlich sind, aber dann doch systematisch von Exponentialfunktionen abweichen. Währen die Lernkurven Exponentialkurven, so würden sie nahelegen, dass das Vergessen ein völlig zufällig wirkender Prozess ist, d.h. die Wahrscheinlichkeit, eine Silbe zu vergessen, würde nicht von der Zeitdauer bis zum Moment des Vergessens abhängen. Es zeigt sich, dass scheinbar sinnlose Silben gar nicht so sinnlos sind, weil Assoziationen zu Wörtern möglich sind, und die verschiedenen Assoziationen ein rein zufälliges Vergessen verhindern. Man kann mathematische Modelle über den Vergessensprozess entwickeln, mit denen die Vergessenskurve vorausgesagt werden kann. Diese Modelle liefern dann Einsicht in mögliche Prozesse des Vergessens.

Der Schwerpunkt wird im Folgenden auf quantitativen Studien liegen.

3 Versuchsplanung

3.1 Generelle Aspekte der Planung einer Untersuchung

3.1.1 Die Rolle der Planung vor der Datenerhebung

Quantitative Studien zielen nicht notwendig auf den Test von Hypothesen, sondern können rein explorativ angelegt sein. Dies kann dazu verleiten, zunächst einmal mit dem Sammeln von Daten zu beginnen in der Hoffnung, dass sich dann schon irgendwelche statistischen Methoden finden werden, mit denen sich interessante Aussagen aus den Daten herausdestillieren lassen.

So sei man zum Beispiel an der Frage interessiert, ob bestimmte Persönlichkeitstypen mit größerer Wahrscheinlichkeit AlkoholikerInnen werden als andere. Man ist fast dazu gezwungen, sich auf Patienten/innen zu konzentrieren, die stationär in einer Klinik behandelt werden, da "frei lebende" Alkoholiker selten bereits sind, sich derartigen Befragungen auszusetzen. Der Vorteil einer derartigen Eingrenzung der Stichporbe ist, das bereits Material aus qualitativen Beobachtungen vorliegt, die die Befragung leiten können. Man beginnt mit einer Befragung, ohne schon einen Versuchsplan festgelegt zu haben. Nach zwei Jahren möchte man die Befragungen auswerten, und wenn man Glück hat, findet man ein geeignetes Verfahren, mit dem sich betimmte Hypothesen testen lassen. Es zeigt sich aber, dass bestimmte formale Voraussetzungen der Anwendung verletzt sind, oder interessierende Variablen konfundiert wurden. Bei rechtzeitiger Planung hätten sich diese Probleme vermeiden lassen. Man kann man die Ergebnisse nicht publizieren, weil die Gutachter sie wegen der Fehler in der Datenerhebung ablehnen, – zwei Jahre sind vertan.

Die Hoffnung, dass sich in einer großen Datenmenge stets etwas Interessantes fin-

den läßt, kann sehr trügerisch sein, denn sie übersieht die Rolle zufälliger Effekte[17]. So kann es sein, dass sich während des Zeitraums der Untersuchung Patienten in der Klinik befinden, deren kommunikatives Verhalten gestört zu sein scheint. Man kommt zu der Vermutung, dass Störungen im kommunikativen Verhalten eine wesentliche Komponente in der Entstehung von Alkoholismus sind. Nach Erhebung entsprechender Daten findet man aber, dass derartige Störungen in der Tat eine Rolle bei der Entstehung von Alkoholismus spielen können, aber nur in Interaktion mit speziellen sozialen Situationen. Aber die hat man bei der Erhebung nicht berücksichtigt, und es existiert kein statistisches Verfahren, das es erlauben würde, den Effekt der sozialen Situation nachträglich zu berücksichtigen.

Die Fehler liegen jedesmal in der mangelnden Versuchsplanung. Man muß sich *vor* Beginn der Datenerhebung darüber im Klaren sein, mit welchen Methoden man die Daten analysieren kann. Es kann sinnvoll sein, sich von vornherein auf bestimmte Hypothesen zu konzentrieren und dementsprechend die Daten so zu erheben, dass sichergestellt ist, dass die Daten anschließend in Bezug auf die Hypothesen getestet werden können.

Die zufällige Variation, also die Varianz in den Daten ist ein Problem bei allen Untersuchungen. Man spricht auch von "Rauschen", in Anlehnung an das Rauschen im Telefon oder im Radio, das den Empfang der eigentlichen Signale – Texte, Bilder, etc – durch Überlagerung stört. Sucht man in einer explorativen Studie nach Strukturen, so werden diese durch das Rauschen verwischt, und in einer hypothesentestenden Studie werden die interessierenden Effekte der verwendeten unabhängigen Variablen undeutlich. Diese zufällige Variation besteht zum Teil aus den Effekten, die nicht kontrollierte Variable erzeugen. Man wird also versuchen wollen, möglichst viele dieser Variablen ebenfalls als unabhängige Variable zu berücksichtigen, d.h. zu kontrollieren. Ein Teilaspekt der zufälligen Variation ist der schon erwähnte Effekt von konfundierenden Variablen.

3.1.2 Validität

Es werden werden zwei Arten von Validität voneinander unterschieden: (i) die interne und die externe. Ein Experiment ist *intern valide*, wenn die Messwerte der abhängigen Variablen eindeutig auf die im Experiment definierten unabhängigen Variablen zurückgeführt werden können. Diese Eindeutigkeit wird als notwendig angesehen, eine kausale Beziehung zwischen abhängigen und unabhängigen Variablen postulieren zu können. Man möchte sicher stellen, dass keine alternativen – also im Experiment nicht betrachteten – Variablen dieselben Effekte auf die abhängigen Variablen

[17] Ein beliebtes Argument der Kritiker statistischer Verfahren besteht in der Behauptung, es gebe keine zufälligen Effekte, was bedeuten kann, dass es irgendwelche nicht weiter spezifizierten geistigen Kräfte gebe, die die Geschicke jedes Einzlenen leiten, oder dass alles mit allem in irgendeiner Weise zusammenhänge. Darüber kann man diskutieren, aber man muß es im gegebenen Zusasmmenhang nicht, weil hier "zufällig" nichts anderes als den Effekt von Einflüssen bedeutet, die im Rahmen der Untersuchung nicht weiter kontrolliert werden können.

haben können. Dies bedeutet, dass keine Störvariablen (konfundierende Variablen) einen Effekt auf die abhängigen Variablen haben. Kann z. B. keine Randomisierung (s. unten) vorgenommen werden, können sich Artefakte ergeben. nach Campbell & Stanley (1966) (zitiert nach Sarris (1990), p. 216) sind typische Artefakte:

1. *Zeiteinflüsse:* Dies sind Einflüsse, die nicht auf die experimentellen Bedingungen zurückgehen, sondern auf Ereignisse, die zwischen experimentellen Behandlungen wirken.

2. *Reifungseffekte:* Zwischen zwei oder mehr Messungen können Veränderungen der Probanden durch Reifungsprozesse stattfinden; dieser Fall kann insbesondere bei entwicklungspsychologischen Untersuchungen eintreten.

3. *Testeffekte:* In einer Untersuchung werden psychometrische Tests verwendet. Wird unter verschiedenen Bedingungen wiederholt getestet, so können Probanden spezielle, testspezifische Reaktionsweisen entwickeln. Dies sind die Testeffekte.

4. *Instrumentierungseffekte:* Auch der Versuchsleiter kann sein Verhalten im Laufe der Untersuchung verändern, und diese Veränderungen können einen Einfluß auf die Ergebnisse haben. Dies gilt möglicherweise auch für die Instrumente, mit denen Messungen durchgeführt werden, – deshalb ist von Instrumentierungseffekten die Rede.

5. *Statistische Regressionseffekte:* Hat man Personen mit extremen Ausprägungen der untersuchten Merkmale, so kann es zum Phänomen der *Regression zu Mitte* kommen. Dieser Effekt wurde zuerst von Francis Galton (1822 – 1911) einem Cousin von Charles Darwin, beobachtet: Eltern mit hoher Intelligenz haben oft Kinder, die weniger intelligent sind, und ebenso haben Eltern mit unterdurchschnittlicher Intelligenz Kinder, deren Intelligenz näher an der durchschnittlichen Intelligenz liegt. Der Effekt geht auf rein statistische Fluktuationen, also zufällige Effekte zurück und findet sich immer dann, wenn insbesondere extreme Merkmalsausprägungen beobachtet werden. So sind Sportler, die eine extreme Leistung vollbracht haben, in folgenden Wettkämpfen oft weniger herausragend.

6. *Auswahleffekte:* In verschiedenen Experimentalbedingungen werden verschiedene Gruppen untersucht. Bei mangelnder Randomisierung kann es vorkommen, dass sich zwei oder mehrere solcher Gruppen systematisch durch ein Merkmal voneinander unterscheiden, das einen Effekt auf abhängigen Variablen hat. Dann sind die Unterschiede in den Messwerten der abhängigen Variablen eben nicht nur auf die verschiedenen Versuchsbedingungen, sondern auch auf dieses Merkmal zurückzuführen. Dieser Sachverhalt wird als Auswahleffekt bezeichnet.

Eine Untersuchung ist *extern valide*, wenn es möglich ist, von der Laborsituation (allgemein Untersuchungssituation) auf die natürlichen Situationen zu schließen. Ein wichtiger Aspekt der externen Validität ist die Repräsentativität der Vpn-Stichprobe: bekanntlich dienen in vielen psychologischen Untersuchungen Studierende der Psychologie als Vpn. Aber die Population dieser Studierenden ist nur ein Teil der Gesamtpopulation und deswegen nicht notwendig repräsantativ für die Gesamtpopulation.

3.1.3 Experimente und Quasi-Experimente

Um Hypothesen testen zu können, müssen i.A. mindesten zwei Gruppen von Versuchspersonen gebildet werden. Es wird üblicherweise zwischen Experimenten und Quasi-Experimenten unerschieden. Bei Experimenten ist gewährleistet, dass jede Veränderung einer unabhängigen Variablen sich in einer korrespondierenden Änderung der abhängigen Variablen niederschlägt; diese Forderung impliziert, dass Vpn *randomisiert* (von englisch random = zufällig, randomise = zufällig anordnen) auf alle Bedingungen verteilt werden, d.h. eine möglichst homogene Stichprobe von Personen wird per Zufall auf die Experimental- und die Kontrollgruppe aufgeteilt. Experimental- und Kontrollgruppe sollen sich möglichst wenig voneinander unterscheiden, so dass etwaige Unterschiede im Verhalten nur auf die Wirkung der unabhängigen Variablen zurückgeführt werden können.

Bei Quasi-Experimenten ist die Randomisierung nicht mehr gegeben, was implizieren kann, dass die Variation der abhängigen Variablen nicht mehr notwendig auf die verschiednen Versuchsbedingungen, d.h. auf die unabhängigen Variablen zurückgeführt werden kann. Darüber hinaus kann die Variierbarkeit der unabhängigen Bedingungen eingeschränkt sein, wodurch die interne Validität reduziert wird. Eine wichtige Klasse von Untersuchungen wie Therapie-Verlaufsstudien fallen unter die Kategorie der Quasi-Experimente: es ist häufig nicht möglich, Patienten randomisiert auf verschiedene Gruppen aufzuteilen und die abhängigen Variablen von nicht weiter kontrollierbaren Einflußgrößen frei zu halten. In Abschnitt 3.4 wird auf derartige Studien ausführlicher eingegangen.

3.1.4 Die Bildung von Stichproben

Ein generelles Ziel von Forschung ist es, von speziellen Beobachtungen (Aussagen) zu allgemeinen Aussagen zu kommen. Dieses Ziel bedeutet, geeignete Stichproben zu finden, die hinreichend repräsentativ sind, um Schlüsse auf die entsprechende Population zu erlauben, also von den Arbeitslosen von Marienthal auf alle Arbeitslosen zu schließen, die für längere Zeit keine Arbeit gefunden haben. In der qualitativen Forschng werden insbesondere zwei Arten der Stichprobenbildung betrachtet: die *bottom-up* Strategie und die *top-down*. Bei der erst Genannten werden die Auswahlkriterien erst während der Untersuchung gebildet, während sie bei der letzteren vorher festgelegt werden. Beim Bottom-up-Verfahren werden einerseits Fälle mit

maximaler Ähnlichkeit bezüglich des interessierenden Merkmals betrachtet, und in einem zweiten Schritt werden Fälle nach dem Prinzip der *maximalen Differenz* betrachtet, die in Bezug auf die interessierende Variable eine andere Ausprägung als die Fälle mit maximaler Ähnlichkeit aufweisen. Hat man den Eindruck, dass es keine weiteren Fälle gibt, die weitere Informationen liefern können, so gilt die Stichprobe als *gesättigt*. Der Punkt bei dieser Stichprobenbildung ist, dass man Personen mit möglichst großen Unterschieden bezüglich der interessierenden Variablen findet. Dieses Prinzip gilt auch für die Top-down-Stichproben, für die die Auswahlkriterien vorher festgelegt wurden.

Der Begriff der Population muß spezifiziert werden. Er kann die Weltbevölkerung bedeuten, oder die Bewohner einer Stadt, oder die Population aller Mitglieder von Turn- oder Trachtenvereinen. Führt man Experimente zur Ebbinghausschen Vergessenskurve durch, so nimmt man implizit an, dass die Ergebnisse *im Prinzip* für alle Menschen gleichermaßen gelten und die eventuell existierenden Unterschiede sich nur auf die Parameter der Vergessenskurven beziehen (z.B. höhere oder niedrigere Rate des Vergessens), – vielleicht gibt es Völkerschaften, die generell besser memorisieren können als andere, aber die *Struktur* des Vergessens ist bei allen dieselbe, weil die neuronalen Mechanismen des Vergessens für alle Menschen dieselben sind. Gleichzeitig kann es innerhalb der Weltbevölkerung Teilpopulationen geben, die sich systematisch voneinander unterscheiden, – z.B. jüngere und ältere Menschen. Darüber hinaus unterscheiden sich Menschen hinsichtlich ihrer Intelligenz. Führt man die Vergessensexperimente nur mit Studierenden der Psychologie aus, so betrachtet man eine Teilpopulation von jüngeren und im Allgemeinen überdurchschnittlich intelligenten Menschen, bei denen die Vergessensprozesse möglicherweise anderen Mechanismen folgen als bei älteren, nicht akademisch gebildeten Menschen, – hier könnten kognitive Strategien zum Memorisieren die Unterschiede bewirken. Will man die Fähigkeit, logische Aussagen auf ihre Korrektheit zu überprüfen testen, so können sich sogar innerhalb einer studentischen Population systematische Unterschiede zeigen. Studierende der Mathematik und Physik haben hier mehr Training, mehr Motivation und *möglicherweise* auch mehr angeborene Fähigkeiten, derartige Aufgaben unter Zeitbegrenzung zu lösen. Untersucht man nur Mathematik-/Physikstudenten, oder *nur* Literaturstudenten, so erhält man vermutlich kein repäsentatives Bild der Fähigkeiten der Population aller Studierenden.

Im Prinzip gleicht die Bildung einer für eine bestimmte Untersuchung repräsentative Stichprobe dem Ziehen von Kugeln aus einer Urne. Im einfachen Fall enthalte die Urne n Kugeln. Die Kugeln mögen sich in Bezug auf k Merkmale M_j voneiander unterscheidenb, und es gebe n_1 Kugeln mit dem Merkmal M_1, n_2 Kugeln mit dem Merkmal M_2 und schließlich n_k Kugeln mit dem Merkmal M_k. Die Stichprobe ist repräsentativ für die Gesamtpopulation der Kugeln, wenn der Anteil der Kugeln mit dem Merkmal M_j in guter Näherung gleich n_j/n ist, für $j = 1, \ldots, n$, d.h. für alle Teilpopulationen in der Urne. Ist m der Umfang der Stichprobe, so ist intuitiv klar, dass die Anteile $\hat{n}_j/n$ um so näher bei den tatsächlichen n_j/n liegen werden,

je größer der Umfang m der Stichprobe ist. Generell wird man sagen können, dass

$$\lim_{m \to n} \frac{\hat{n}_j}{n} = \frac{n_j}{n} \tag{3.1}$$

gilt. Das Problem ist nun, dass oft aus praktischen und aus Kostengründen der Wert von m klein im Verhältnis zu n ist. Um so mehr muß man dann auf die Repräsentativität achten. Für die Urne bedeutet dies, dass die Kugeln gut durchmischt sein sollen, wenn man zufällig Kugeln zieht. Liegen die Kugeln zum Beispiel geschichtet in der Urne, so dass die Kugeln mit M_1 am Boden liegen, darauf die Kugeln mit M_2, etc., so kann die Stichprobe in einer Weise zusammengesetzt sein, die sich deutlich von der Zusammensetzung der Population unterscheidet. Will man eine Stichprobe aus den Bewohnern einer Stadt bilden, so muß man darauf achten, dass die Personen nicht alle aus einem bestimmten Stadtteil kommen, denn die Bewohner verschiedener Stadtteile können sich deutlich hinsichtlich sozio-ökonomischer Variablen voneinander unterscheiden. Eine Möglichkeit ist, Stichproben aus den Populationen verschiedener Stadtteile zu bilden, die anteilsmäßig den jeweiligen Bevölkerungsanteilen entsprechen, und die Gesamtstichprobe aus diesen Teilstichproben zusammenzusetzen.

Die oben betrachtete Aufteilung nach Merkmalsträgern ist natürlich nur eine Vereinfachung, um das Prinzip der Repräsentativität zu verdeutlichen. Es wurde angenommen, dass ein Merkmal M_j, $1 \leq j \leq n$ entweder vorhanden ist oder nicht. In der Realität wird man diesen einfachen Fall nicht immer haben: Merkmale sind oft unterschiedlich ausgeprägt. Die Auswahl einer Stichprobe wird dann selten perfekt sein in dem Sinne, dass auch die verschiedenen Ausprägungen repräsentativ in der Stichprobe vertreten sind. Sofern die Werte der abhängigen Variablen von diesen Merkmalsausprägungen abhängen, werden diese unterschiedlichen Ausprägungen die Variabilität (d.h. die Fehlervarianz) der Werte der abhängigen Variablen erhöhen und damit die Effekte, die die interessierenden unabhängigen Variablen erzeugen in einem gewissen Ausmaß camouflieren. Generell wird man also daran interessiert sein, die Fehlervarianz so klein wie nur irgend möglich zu halten. Die Aufgabe, die Fehlervarianz zu minimieren ist ein wesentlicher Teil einer guten Versuchsplanung.

3.1.5 Signifikanz und Effektstärke

Hypothesen werden geprüft, indem man eine bestimmte zufällige Veränderliche T definiert, deren Wahrscheinlichkeitsverteilung von der Gültigkeit der Hypothesen abhängt. Üblicherweise werden zwei Hypothesen betrachtet; H_0 und $H_1 = \neg H_0$. Man möchte etwa die Wirksamkeit einer Behandlung prüfen. Dazu betrachtet man die Werte einer abhängigen Variablen X, die den Effekt der Behandlung bzw. der Nichtbehandlung repräsentiert. Unter H_0 habe X den Erwartungswert $\mu_0 = \mathbb{E}(X|H_0)$, und unter H_1 habe X den Erwartungswert $\mu_2 = \mathbb{E}(X|H_1)$[18]. Der Einfachheit halber

[18] Die Erwartungswerte $\mathbb{E}(X|H_0)$ und $\mathbb{E}(X|H_1)$ sind *bedingte Erwartungswerte* – die Verteilung und damit die Parameter der Verteilung hängen von den Bedingungen ab, die hier durch H_0 bzw. H_1 spezifiziert werden.

habe X unter beiden Bedingungen dieselbe Varianz σ^2. Man kann nun die zufällige Veränderliche $T = X_1 - X_0$ definieren. Der Erwartungswert von T ist $\mathbb{E}(T) = \mu_1 - \mu_0$ (der Erwartungswert einer Differenz ist stets gleich der Differenz der Erwartungswerte). Es werde nun angenommen, die Behandlung habe keinen Effekt, so dass $\mu_1 = \mu_0$ gilt. In diesem Fall gilt H_0 und es ist $\mathbb{E}(T) = \mu_1 - \mu_0 = 0$, so dass T um den Erwartungswert 0 verteilt ist. Gilt dagegen H_1, so ist $\mathbb{E}(T) \neq 0$. Da μ_1 und μ_2 nicht bekannt sind (es handelt sich ja um Populationsparameter) läßt sich in diesem Fall nichts über den Wert von $\mathbb{E}(T)$ sagen. Gegeben sind aber die Mittelwerte $\bar{x}_0$ und $\bar{x}_0$ der Werte aus der Kontrollgrupppe und aus der Experimentalgruppe. Schon wegen der Stichprobenfehler findet man $\Delta\bar{x} = \bar{x}_0 - \bar{x}_1 \neq 0$ auch dann, wenn H_0 wahr ist[19]. Der Test von H_0 besteht also darin, zu prüfen, ob die Differenz $\Delta\bar{x}$ *hinreichend* von Null verschieden ist. Ist X normalverteilt, so ist auch die Differenz $\Delta\bar{x}$ normalverteilt, und $\sigma^2_{\Delta\bar{x}} = \sigma^2/n$, n der Stichprobenumfang. Dann ist

$$z = \frac{\bar{x}_1 - \bar{x}_0}{\sigma^2_{\Delta\bar{x}}}$$

standardnormalverteilt. Findet man $|z| > 1.96$ so hat man einen Wert, der unter H_0 insgesamt nur in 5% der Fälle vorkommt. Das ist ziemlich selten, so dass man folgert, dass H_1 die wahrscheinlichere Alternative ist – man entscheidet sich für H_1 und sagt, die gefundene Mittelwertsdifferenz sei *signifikant*. Im Normalfall ist σ^2 aber nicht bekannt und man muß mit der Schätzung s^2 Vorlieb nehmen, mit $s^2_{\Delta\bar{x}} = s^2/n$ als Schätzung der Varianz der Differenzen. s^2 ist ebenfalls zufällig verteilt (es handelt sich um die χ^2-Verteilung), so dass

$$t_{n-1} = \frac{\bar{x}_1 - \bar{x}_0}{s_{\Delta\bar{x}}} \tag{3.2}$$

nicht mehr normal, sondern t-verteilt mit $n-1$ Freiheitsgraden ist; die Verteilung von t ist als STUDENTsche t-Verteilung bekannt; unter H_0 gilt $\mathbb{E}(t_{n-1}) = 0$. Die Logik der Entscheidung für oder gegen H_0 bleibt dieselbe: ist $|t_{n-1}| > t_{crit}$ so entscheidet man sich für H_1, andernfalls behält man H_0 bei. t_{crit} hängt vom Stichprobenumfang n ab. Die Frage ist nun, wie man t_{crit} bestimmt. Dazu macht man von der Tatsache Gebrauch, dass

$$p(x) = \int_{-\infty}^{x} f(x)dx, \quad 1 - p(x) = \int_{x}^{\infty} f(x)dx$$

gilt, wobei im Falle des t-Tests $x = t$ und f die Dichtefunktion von t ist. Kennt man p, so kann man auf den Wert von $x = t$ schließen. Deshalb legt man einen Wert zum Beispiel von $1 - p$ fest und schließt für gegbebenen Wert von n auf den zugehörigen $t = t_{crit}$-Wert. Im obigen Beispiel wurde ein zweiseitiger Test betrachtet; diesen Fall hat man, wenn man die Hypothese $|\mu_1 - \mu_0| \neq 0$ testen will. In diesem Fall legt man eine Wahrscheinlichkeit α fest und berechnet den Wert von t_{crit} für $\alpha/2$.

[19] Δ ist der griechische Buchstabe Delta (großes Delta), δ ist das kleine Delta. Δ steht in der Mathematik oft für Differenz; *Delta*$\bar{x}$ bezeichnet die Differenz zweier Mittelwerte.

Alternativ kann man auch die Wahrscheínlichkeit $p = P(t > \Delta\bar{x})$ oder $p = P(t \leq \Delta\bar{x})$ oder $P(|\Delta_{\bar{x}}| > 0)$ berechnen und entscheidet nach Maßgabe dieses p-Werts. Ist er hinreichend klein, etwa $p < .05$, entscheidet man sich für H_1, sonst für H_0. Analoge Betrachtungen gelten für die Signifikanz von Korrelationen und anderen Statistiken.

Es gibt eine Reihe von Problemen bei der Entscheidung über Hypothesen nach Maßgabe von p-Werten, auf die hier nicht im Einzelnen eingegangen werden kann. Aber eines der unangenehmeren Probleme ist, dass man auch für sehr kleine Mittelwertsdifferenzen eine Signifikanz erreichen kann, wenn man nur den Stichprobenumfang n hinreichend groß werden läßt. Es läßt sich leicht zeigen, dass die Varianz des Stichprobenmittelwerts duch $\sigma^2_{\bar{x}} = \sigma^2/n$ gegeben, ist, wobei σ^2 die Varianz der X-Werte in der Population ist und n der Stichprobenumfang; offenbar folgt

$$\sigma^2_{\bar{x}} = \frac{\sigma^2}{n} \to 0 \text{ für } n \to 0,$$

womit gesagt werden soll, dass man für "hinreichend" großen Wert von n eine beliebig kleine Varianz des Stichprobenmittelwerts erhält, – d.h. $\bar{x}$ liegt in einer beliebig kleinen Nachbarschaft von μ. Dieselbe Betrachtung gilt natürlich auch auch für Mittelwertsdifferenzen: die Varianz der Differenz von Mittelwerten wird beliebig klein für hinreichend großen Stichprobenumfang. Dies bedeutet, dass t_n hinreichend groß wird, um größer als der kritische Wert (für gegebenen α-Wert) zu werden. Dieser Sachverhalt kann im Extrem bedeuten, dass man die Nullhypothese verwirft, obwohl in Wirklichkeit $\mu_0 = \mu_1$ gilt. Tatsächlich ist H_1 aber nur interessant, wenn sich μ_0 und μ_1 substantiell unterscheiden, – wobei zugegebenermaßen der Begriff 'substantiell' nicht sehr scharf definiert ist.

So ist es eine altbekannte These, dass die visuelle Vorstellungskraft von Frauen weniger ausgeprägt sei als die von Männern. Dieser Sachverhalt sei evolutionstheoretisch gesehen plausibel: die Männer mußten in der Savanne oder im Wald jagen und anschließend zur heimatlichen Höhle zurückfinden, um dort Frau und Kind mit Wildbret zu versorgen, während die Frau Feuer und Kinder hütete, wozu sie keine visuelle Vorstellungskraft benötigte, wohingegen der Mann verlorenginge, hätte er keine derartige Vorstellungskraft. Derartige "Erklärungen" haben den gleichen Informationswert wie psychoanalytische Erklärungen aggressiven Verhaltens männlicher Jugendlicher wegen des Ödipuskomplexes, – nämlich keinen. Sorgfältige Untersuchungen (d.h. adäquate, repräsentative Stichprobenbildungen) zu dieser Behauptung zeigen aber, dass die Differenzen $\bar{x}_w - \bar{x}_m$ so klein sind (falls sie überhaupt existieren), dass es nicht besonders sinnvoll ist, $\Delta\mu = \mu_w - \mu_m \neq 0$ anzunehmen. Denn selbst wenn tatsächlich $\Delta\mu < 0$ wäre, so wird $\Delta\mu$ so klein sein, dass immer noch knapp 50% der Frauen ein besseres räumlich Vorstellungsvermögen haben als knapp unter 50% der Männer; dies folgt aus der Überlappung der meistens symmetrischen Verteilungen der Scores für visuelle Vorstellungskraft. Es macht also keinen Sinn, Frauen *generell* als weniger geeignet für den Architektenberuf anzusehen als Männer. Analoge Aussagen gelten für andere Fähigkeiten.

Es erscheint demnach sinnvoller, nicht auf die Signifikanz von Unterschieden, sondern auf die Effektstärke

$$\varepsilon = \frac{\bar{x}_1 - \bar{x}_0}{s} \tag{3.3}$$

zu fokussieren. Hat man Vorinformationen über diese Größen, kann man den Stichprobenumfang bestimmen, der nötig ist, diese Effektgröße auch zu finden (bei zu kleinen Stichprobenumfängen kann ein Effekt im Rauschen verloren gehen.

3.2 Korrelationsstudien

3.2.1 Multiple Regression

In vielen Studien – zum Beispiel im Rahmen der Persönlichgkeitspsychologie – ist man an der Beziehung zwischen einer Reihe von Variablen interessiert. IM einfachsten Fall hat man nur zwei Variable, und man kann die Regressionen

$$y = b_{yx}x + a_{yx} + e_y \tag{3.4}$$

$$x = b_{xy}y + a_{xy} + e_x \tag{3.5}$$

betrachten,. In Abschnitt 2.2.1 von Methoden der Psychologie, Teil 1, wurde gezeigt, dass die Methode der Kleinsten Quadrate die Schätzungen

$$b_{yx} = \frac{Kov(x,y)}{s_x^2} \tag{3.6}$$

$$b_{xy} = \frac{Kov(x,y)}{s_y^2} \tag{3.7}$$

und

$$a_{yx} = \bar{y} - b_{yx}\bar{x} \tag{3.8}$$

$$a_{xy} = \bar{x} - b_{xy}\bar{y} \tag{3.9}$$

liefert. Dabei ist

$$Kov(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \tag{3.10}$$

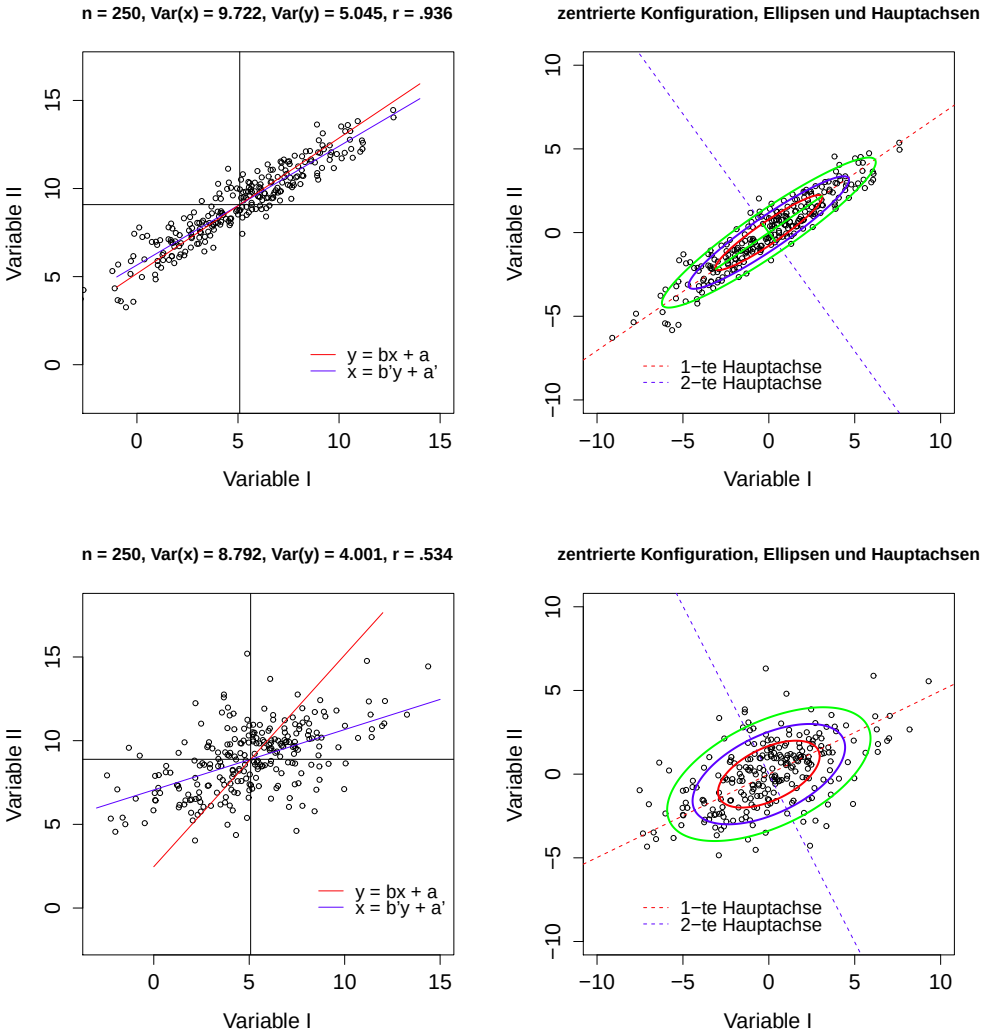
die Kovarianz zwischen X und Y. Abbildung 1 links illustriert eine einfache Regression. Auf die Abbildung rechts wird weiter unten zurückgekommen.

Multipliziert man b_{yx} mit dem Quotienten s_x/s_y, so erhält man

$$b_{yx}\frac{s_x}{s_y} = \frac{Kov(x,y)s_x}{s_x^2 s_y} = \frac{Kov(x,y)}{s_x s_y} = r_{xy}, \tag{3.11}$$

r_{xy} die Produkt-Moment-Korrelation, im Folgenden einfach Korrelation genannt. Analog dazu findet man $b_{xy}s_x/s_y = r_{xy}$. Es ist $-1 \leq r_{xy} \leq 1$, und für $r_{xy} = 0$ kann man die stochastische Unabhängigkeit von X und Y annehmen, wenn X

Abbildung 1: Links: Regressionsgeraden, rechts: dieselbe Konfiguration, allerdings zentriert, mit Ellipsen und zugehörigen Hauptachsen.



und Y bivariat normalverteilt sind (es lassen sich Fälle konstruieren, bei denen $Kov(x, y) = 0$ und damit $r_{xy} = 0$, obwohl eine perfekte Abhängigkeit zwischen X und Y existiert, beim Schluß von $r_{xy} = 0$ auf stochastische Unabhängigkeit ist also Vorsicht geboten!). Weiter gilt

$$r_{xy}^2 = \frac{s_{\hat{y}}^2}{s_y^2} = 1 - \frac{s_e^2}{s_y^2} \tag{3.12}$$

r_{xy}^2 heißt *Determinationskoeffizient* und gibt den Anteil der vorgesagten Varianz ($s_{\hat{y}}^2$) an der Gesamtvarianz s_y^2 der Y-Werte an. Natürlich ist $r_{xy}^2 \leq 1$, so dass $s_e^2/s_y^2 \leq 1$,

d.h. $s_e^2 \leq s_y^2$. Für $s_e^2 \to s_y^2$ fogt $r_{xy}^2 \to 0$; dann hängt die Variation der y-Werte gar nicht mehr von der Variation der x-Werte ab, d.h. es gibt keine Beziehung zwischen x und y. Umgekehrt gilt $s_e^2 \to 0 \Rightarrow r_{xy}^2 \to 1$; je kleiner die Fehlervarianz, desto höher ist die Korrelation und desto besser ist die Vorhersage von y auf der Basis von x.

Will man also die Vorhersage verbessern, so muß man die Fehlervarianz verkleinern. Dies kann man durch Hinzunahme von weiteren Prädiktoren X_j erreichen. Für p Prädiktoren ergibt sich

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_p X_p + e \tag{3.13}$$

Die Betrachtung zur Interpretation von r^2 überträgt sich auf diesen Fall, mit $\hat{y} = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_p X_p$. Hier sind die b_j, $0 \leq j \leq p$ zunächst unbekannte Parameter, die wieder mit der Methode der Kleinsten Quadrate geschätzt werden können. Auf die Details dieser Schätzungen soll hier nicht weiter eingegangen werden, es genügt zunächst, festzustellen, dass die Vorhersage der Y-Werte um so besser wird, je mehr Prädiktoren X_j man hat. Die Beziehung (3.13) ist *linear*, d.h. die X_j gehen nur additiv in die Vorhersage von Y ein. Die Gleichung kann aber verallgemeinert werden, indem man Terme der allgemeinen Form $f(x_{j_1}, \ldots, X_{j_r})$ hinzufügt. Die f repräsentieren irgendwelche nichtlinearen (= nicht additiven) Beziehungen zwischen Teilmengen der Prädiktoren. Die können aus einzelnen Elementen der Form X_j^r, $r \neq 1$ bestehen, oder aus Produkten, etwa $X_1 X_2$, die Wechselwirkungen abbilden, etc.

Vom Standpunkt der Versuchsplanung aus gesehen gibt es aber ein paar Fragen in Bezug auf den Ansatz (3.13). Zunächst einmal sollten die Prädiktoren stochastisch unabhängig voneinander sein. Denn die Schätzungen $\hat{b}_j$ der Parameter b_j hängen von der Stichprobe der Messwerte ab und sind damit selbst zufällige Veränderliche. Die Varianz der Schätzungen $\hat{b}_j$ wird aber um so größer, je größer die Korrelationen zwischen den Prädiktoren sind. Darüber hinaus werden die $\hat{b}_j$ schwer zu interpretieren. Sind die Korrelationen zwischen den Prädiktoren zu hoch, ergibt sich für die Schätzungen der b_j ein oszillierender Effekt: ist b_1 klein, so wird b_2 groß, und b_3 wird wieder klein, etc. Für b_1 groß wird b_2 klein, etc. Dieser Effekt hat nichts mit der Wichtigkeit oder Unwichtigkeit der Prädiktoren zu tun, sondern ist ein rein numerischer Effekt, der die inhaltliche Interpretation sehr erschwert. Für den Fall, dass die Korrelationen zwischen den Prädiktoren alle gleich Null sind haben aber die $\hat{b}_j$ eine einfache Interpretation: sie können als Gewichte betrachtet werden, mit denen die einzelnen Prädiktoren in die Vorhersage der Y-Werte eingehen.

Eine große Zahl von Prädiktoren erfordert eine entsprechend große Stichprobe von Messungen, und eine große Stichprobe kann kosten: auf jeden Fall Zeit, aber oft auch finanzielle Kosten. Abgesehen von den Kosten ergibt sich ein anderes Problem: zwar kann mit mit einer großen Zahl von Prädiktoren *für einen gegebenen Datensatz* eine gute Anpassung der vorhergesagten Y-Werte $\hat{y}$ an die tatsächlich gemessenen Werte Y erreichen, aber sobald man die $\hat{b}_j$-Werte für die Vorhersage mittels anderer X_j Werte verwendet (dies ist bei psychometrischen Tests der Fall, wobei die X_j Testscores in Untertests sind), ergeben sich schlechte Vorhersage. Der Grund

dafür ist, dass einige der Prädiktoren nur zufällige Effekte in der ersten Stichprobe abbilden, die in einer zweiten Stichprobe nicht vorhanden sind, aber durch andere zufällige Effekte ersetzt werden. Die gute Anpassung leifert nur eine Pseudoerklärung der Y-Werte anhand der Prädiktoren. Die Hinzunahme nichtlinearer Terme wie X_j^r, X_jX_k etc bedeutet notgedrungen Korrelationen zwischen diesen Termen und den übrigen Prädiktoren und damti instabile Schätzungen der dazu korrespondierenden Parameter, so dass nichtlineare Terme oft nur eine Pseudoverbesserung der Vorhersage bedeuten. Es sei noch einmal an das Zitat eines Hermeneutikers aus Zimmer (1986) erinnert:

> "Gerade jene vom einheitswissenschaftlichen Standpunkt aus bemängelte Unterbestimmtheit und Vagheit der Freudschen Formulierungen ist es, die sie für das hermeneutische Feld so fruchtbar macht, weil dadurch Sinnantizipationen durchgehalten werden können, die noch dem kleinsten Traumfragment eine integrierbare Deutung abzuringen vermögen."
> Stuber, in Mertens (1983), p. 97 (zitiert nach Zimmer (1986), p. 231)

Hier gehen die "kleinsten Traumfragmente" gewissermaßen als zusätzliche Prädiktoren in die Betrachtungen des Hermeneutikers ein und scheinen einen sinnstiftenden Effekt auf die Interpretation zu haben. Übersehen wird die Möglichkeit, dass Träume auch zufällige Komponenten enthalten, Ausschmückungen von "Tagesresten" ohne tiefere Bedeutung. Für den Hermeneutiker machen sie aber das Bild "stimmiger". Bekanntlich kann man auch in Wolkenformationen Gesichter erkennen, und ein gut interpretierter Kaffeesatz kann durchaus die oft dunklen Wege des Lebens erhellen.

Die weitere Diskussion der Problematik der multiplen Regression kann hier nicht geführt werden, aber man sollte sich schon einmal merken, dass es sie gibt, und dass sie Ausdruck einer allgemeinen, nicht nur auf eine Formel bezogenen Problematik ist. Es muß andererseits festgehalten werden, dass die multiple Regression eine Art kanonisches Modell für viele Datenanalysen darstellt, was in den folgenden Abschnitten noch erläutert werden wird.

3.2.2 Analyse von Korrelationen

In vielen Untersuchungen werden für jede Person mehrere Variablen $X_1, \ldots, X_p$ gemessen, und man ist an den Kovariationen bzw. Korrelationen zwischen den X_j interessiert. Die X_j können Fragen in einem Persönlichkeitsfragebogen sein, oder man läßt zum Beispiel in einer Umfrage Politiker hinsichtlich einer Reihe von Eigenschaften bewerten, und die X_j sind Skalen, die diese Eigenschaften repräsentieren, oder man läßt im Rahmen von Marktforschungsuntersuchungen Produkte auf den Skalen X_j bewerten. Man erhält dann insgesamt

$$\binom{p}{2} = \frac{p(p-1)}{2}$$

Korrelationen, – wegen der Symmetrie $r_{jk} = r_{kj}$ müssen nicht alle p^2 Korrelationen betrachtet werden. Für $p = 5$ sind 10 Korrelationen zu betrachten, für $p = 10$ sind es schon 45, und für $p = 20$ hat man bereits 380 Korrelationen zu interpretieren. Schaut man ohne weitere Analyse auf diese Korrelationen erschließt sich ihre Bedeutung nur schwer.

Latente Variable: Korrelationen bedeuten nicht notwendig, dass kausale Beziehungen zwischen den entsprechenden Variablen bestehen. So kann es sein, dass die Korrelation r_{jk} zwischen zwei Variablen X_j und X_k dadurch zu erklären ist, dass beide Variablen durch eine dritte Variable L_1 beeinflußt werden, so dass

$$X_j = b_{jL}L_1 + a_j + e_j \tag{3.14}$$
$$X_k = b_{kL}L_1 + a_k + e_k \tag{3.15}$$

gilt. Darüber hinaus könnte L_1 auch weitere Variable mitbestimmen. L ist eine *latente Variable*. Darüber hinaus könnte man die Korrelationen zwischen den Variablen durch eine zweite latente Variable L_2 erklären. Am Ende könnte es sein, dass mit nur wenigen latenten Variablen $L_1, \ldots, L_q$, $q < p$ alle gemessenen Variablen X_j "erklärt" werden können, und damit auch die Korrelationen zwischen den X_j. Wenn darüber hinaus die $L_1, \ldots, L_q$ auch noch paarweise unabhängig voneinander wären, so müßte man eigentlich nur diese latenten Variablen betrachten, da sie die gesamte Information in den X_j enthalten. Die Frage ist nur, wie man diese unbekannten Variablen findet.

Die Lösung dieses Problems wird in Abbildung 1, Seite 36 angedeutet. Die Punkte in den Koordinatensystemen sind Fälle (Personen), deren Koordinaten die Messwerte (x_{i1}, x_{i2}) auf den Skalen X_1 und X_2 sind. Es läßt sich zeigen, dass die Korrelation r_{12} eine Schar von achsenparallelen Ellipsen definiert; jeder Punkt liegt auf einer Ellipse[20]. Eine Ellipse ist durch ihre Hauptachsen bestimmt; die sind in den Ellipsen in Abb. 1 eingezeichnet worden. Es läßt sich weiter zeigen, dass die Hauptachsen die Orientierung der latenten Variablen angeben. Man findet diese Hauptachsen schlicht durch eine Drehung der gesamten Punktekonfiguration, bzw. durch eine Transformation (Drehung) der ursprünglichen Koordinatenachsen derart, dass sie mit den Hauptachsen der Ellipsen zusammenfallen. Man bestimmt nun die Projektionen der Punkte auf diese neuen Koordinatenachsen. Diese Projektionen sind die Ausprägungen der latenten Merkmale – denn es handelt sich bei den latenten Variablen um Merkmale, auch wenn man sie nicht sofort identifizieren kann – bei den einzelnen Punkten, d.h. Personen. Ist die Fehlervarianz bei einer Regression klein, so genügt offenbar nur eine latente Variable, um die Konfiguration zu beschreiben, ist sie groß, so müssen zwei latente Variable betrachtet werden, d.h. die "Fehler" enthalten Komponenten, die eine *systematisch* wirkende zweite latente Variablen zurückgehen. Bei der Interpretation von Korrelationen zwischen mehr als zwei Variablen kann es mehr als zwei latente Variablen geben.

[20]Die Herleitung dieses Sachverhalts ergibt sich leicht mittels der Vektor- und Matrixrechnung, die in der Veranstaltung Multivariate Methoden eingeführt wird.

Hat man mehr als zwei Variable, so bestimmen die r_{jk} ein Ellipsoid. Die Hauptachsen dieses Ellipsoids sind mögliche latente Variablen. Da Hauptachsen von Ellipsen bzw. Ellipsoiden stets senkrecht aufeinander stehen, können sie als stochastisch unabhängige latente Variable betrachtet werden (dieser Sachverhalt kann bewiesen werden, was hier nicht geschehen soll). Man gelangt also zu latenten Variablen durch Drehung des ursprünglichen Koordinatensystems; man spricht von einer *Hauptachsentransformation*. Es gibt stets so viele Hauptachsen (= latente Variable) wie es Variable X_j gibt, aber nicht alle müssen interpretiert werden, oft genügen zwei oder drei, um alle Korrelationen zwischen den X_j zu "erklären".

Das Semantische Differential: Eine spezielle Form von Zusammenhangsanalyse liefert die Technik des *semantischen Differentials*, auch *Polaritätsprofil*(Osgood et al (1957)). Es wird eine Menge von Eigenschaftspaaren wie hoch-tief, krank-gesund, warm-kalt, etc gewählt; sie bilden jeweils die Enden einer polaren Skala, die etwa von -3 über 0 bis zu +3 läuft: Die Eigenschaften sollen eine repräsentative Stichprobe

Tabelle 4: Polaritätsprofil

hoch	-3	- 2	-1	0	1	2	3	tief
krank	-3	- 2	-1	0	1	2	3	gesund
klein	-3	- 2	-1	0	1	2	3	groß
⋮				⋮				⋮
kalt	-3	- 2	-1	0	1	2	3	warm

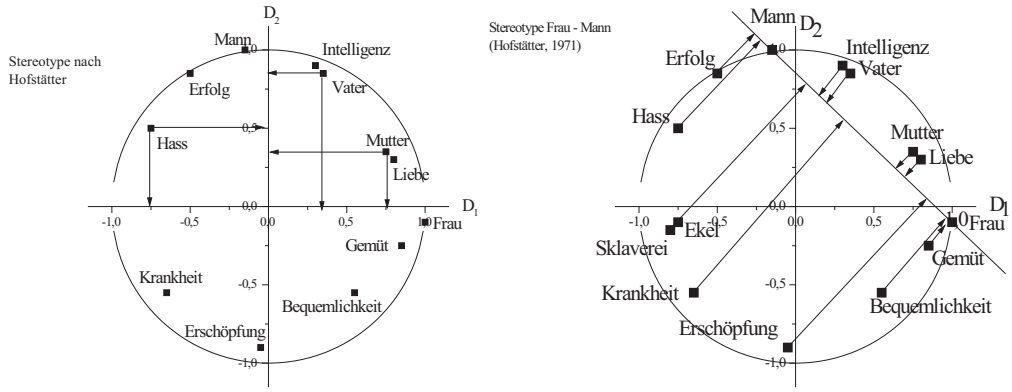
aus der Menge von Eigenschaften sein, die für eine Menge von Objekten (Personen) charakteristisch sind. Jedes Objekt oder jede Person aus einer bestimmten Menge von Objekten/Personen wird nun auf jeder dieser Skalen eingeschätzt. Trägt man diese Schätzungen für ein Objekt auf den Skalen ein, so entsteht ein Profil von Einschätzungen für dieses Objekt. Sind zwei Objekte sich ähnlich, so werden die Profile ähnlich sein. Einem gängigen Ansatz zufolge können diese Ähnlichkeiten durch Korrelationen ausgedrückt werden: man betrachtet dazu die Einschätzungen (x_i, y_i), $i = 1, 2, \ldots, m$, m die Anzahl der Eigenschaftspaare ("Polaritäten"), als Messwertpaare, für die die Kovarianz

$$Kov(\omega_1, \omega_2) = \frac{1}{m} \sum_{i=1}^{m} (x_i - \bar{x})(y_i - \bar{y})$$

und damit dann auch die Korrelation zwischen den Profilen für die Objekte ω_1 und ω_2 berechnet werden kann. Diese Korrelationen können dann wieder in Bezug auf mögliche latente Variablen analysiert werden.

Das Problem ist, dass die auf diese Weise berechneten Korrelationen keinen Sinn machen. Denn die Anordnung der Eigenschaften ist beliebig: es gibt keine Regel, nach der links "klein" und rechts "groß" stehen müßte, ebensogut könnte links "groß"

Abbildung 2: Stereotypen zur "Polarität" weiblich - männlich, I



stehen und rechts "klein". Damit hängen die numerischen Werte der Schätzungen der Ausprägungen der jeweiligen Eigenschaften von der beliebig gewählten Anordnung ab, und folglich sind die Korrelationen beliebig, denn jede Skala kann ebensogut in der inversen Form präsentiert werden: mn kann einerseits die Skala

$$\text{hoch} \quad -3 \quad -2 \quad -1 \quad 0 \quad 1 \quad 2 \quad 3 \quad \text{tief}$$

präsentieren, oder die Skala

$$\text{tief} \quad -3 \quad -2 \quad -1 \quad 0 \quad 1 \quad 2 \quad 3 \quad \text{hoch}$$

Die Skalenwerte -3, -2, bis 2, 3 haben in beiden Fällen eine unterschiedliche Bedeutung, weshalb sich die Korrelationen für die beiden Positionen voneinander unterscheiden, – durch den Faktor -1, d.h. die Korrelation für die Skala ist gerade die negative Korrelation für die andere; in vielen Darstellungen des semantischen Differentials wird dieses Problem übersehen.

Der Punkt ist, dass eine Korrelation oder Kovarianz stets für *zwei Variablen* berechnet wird, die beliebig gewählten Positionen der Eigenschaften in der obigen Liste repräsentieren aber keine Werte auf bestimmten Variablen. Trotzdem kann der Anssatz gerettet werden. Man betrachtet als Skala jeweils nur einen Pol eines Eigenschaftspaares, also etwa nur "klein" statt "klein versus groß", etc. Es ist gleichgültig, welche der beiden Eigenschaften (klein oder groß) man wählt, man muß sich nur auf eine Eigenschaft festlegen. Man korreliert dann jeweils zwei Eigenschaften miteinander; für die Kovarianz zwischen den Skalen S_j und S_k erhält man dann

$$Kov(S_j, S_k) = \frac{1}{n} \sum_{i=1}^{n} (s_{ij} - \bar{s}_j)(s_{ik} - \bar{s}_k),$$

wobei n die Anzahl der beurteilten Objekte ist. Ein Beispiel für die Anwendung des semantischen Differentials ist die Stereotypforschung. Man gibt z.B. Begriffe

wie Vater, Mutter, Diktator, Erschöpfung, Intelligenz, etc vor und läßt sie auf den Skalen S_1 (hoch), S_2 (krank), S_3 (klein) etc von p Personen bewerten. Für jeden dieser Begriffe und jede Skala kann dann eine mittlere Einschätzung (arithmetisches Mittel über die Schätzungen der befragten Personen) berechnen; $\bar{x}_{ij}$ ist dann die mittlere Einschätzung des i-ten Begriffs auf der j-ten Skala. Man kann dann die Korrelationen zwischen den Skalen berechnen; die Kovarianzen sind

$$Kov(S_j, S_k) = \frac{1}{n} \sum_{i=1}^{n} (\bar{x}_{ij} - \bar{x}_{\cdot j})(\bar{x}_{ik} - \bar{x}_{\cdot k})$$

Für dies Kovarianzen oder die entsprechenden Korrelationen kann dann die oben genannte Hauptachsentransformation auf latente Variablen durchgeführt werden. Ein Beispiel liefert Abbildung 2. Sie beruht auf Daten aus den späten 50-er Jahren des 20-ten Jahrhunderts und bei einer Wiederholung des Versuchs heute wird man warscheinlich ein anderes Bild erhalten. Interessant ist gleichwohl, dass "männlich" und "weiblich" eben *nicht* als Polaritäten in den Stereotypen auftauchen, sondern als voneinander unabhängige Konzepte. Diese Aussage folgt aus der Tatsache, dass "Mann" und "Frau" nahezu perfekt (bis auf Stichprobenfehler) als Pole auf Dimensionen auftreten, deren Rechtwinkligkeit (Orthogonalität im Jargon dieser Analyse) eben die stochastische Unabhängigkeit von Variationen hinsichtlich der Merkmalsausprägungen der Variable 'Geschlecht' bedeutet. Es folgt auch aus den Abbildungen, dass die Konzepte 'Erfolg', 'Intelligenz', 'Vater' als typisch männliche Eigenschaften gelten, – aber die Orthogonalität bedeutet, dass sie nur bei männlichen Wesen auftreten können. Typisch 'weiblich' sind auch die Konzepte 'Mutter', 'Liebe' und 'Gemüt', aber diese Merkmale können durchaus auch bei Männern auftreten. Das Konzepot 'Erschöpfung' erscheint als typisch unmännlich, während 'Hass', 'Sklaverei' und 'Ekel' typisch unweiblich sind.

Gleichwohl kann man die von Philosophen[21], Dichtern (z.B. Goethe) und manchen Psychologen vertretene Polaritätsthese in diesen Resultaten rekonstruiert werden. Sie entspricht zwar nicht den Daten, aber man sehen, wie die Idee entstehen konnte. Dazu betrachte man die Gerade zwischen 'Frau' und 'Mann', auf der diese beiden Konzepte als Gegensatzpaar erscheinen, wenn man die Projektionen der anderen Punkte auf diese Gerade betrachtet. Diese Gerade entspricht ziemlich gut den Konstruktionen, die man von Nietzsche, Wellek und anderen Denkern kennt. Stereotype sind vereinfachte und verzerrte Bilder, die nur wenig oder gar nicht mit der Realität übereinstimmen müssen, deren Abbild zu sein sie vorgeben.

Untersuchungen dieser Art sind einerseits quantitativ, andererseits exploratorisch, denn es werden i.A. keinerlei Hypothesentests durchgeführt. Eine detaillierte Darstellung der Hauptachsentransformation setzt Kenntnisse der Vektor- und Matrizenrechnung (lineare Algebra) voraus, die in der Vorlesung *Multivariate Methoden* vermittelt werden.

[21] Die Idee der Polarität von 'männlich' und 'weiblich' geht bis in die Antike zurück.

3.3 Varianzanalytische Versuchspläne

Die Varianzanalyse ist eine Methode, mit der Mittelwertsvergleiche durchgeführt werden. Warum man dennoch von 'Varianzanalyse' spricht, wird gleich verdeutlicht werden.

Bei der multiplen Regression wurde der Einfluß von p Prädiktorvariablen auf eine abhängige Variable untersucht, wobei die Prädiktoren kontinuierliche Variablen sind. Wenn man statt kontinuierlicher Prädiktoren *Indikatorvariablen* verwendet, führt man de facto Mittelwertsvergleiche durch: die Regressionskoeffizienten ergeben sich dann als Mittelwerte. Dass dieser Vergleich von Mittelwerten dann auch noch Varianzanalyse heißt, mag auf den ersten Blick verwirrend sein, ergibt sich aber leicht, wenn man von der Beschreibung der Varianzanalyse ausgeht.

Als Beispiel werde der Vergleich verschiedener Therapien betrachtet. Gegegen seien etwa p verschiedene Therapien, etwa $p = 5$, und es soll entschieden werden, ob sie sich hinsichtlich ihrer Effektivität voneinander unterscheiden. Dazu könnte man paarweise vorgehen: man testet alle möglichen Paare mittels eines t-Tests. Man hat dann

$$\binom{5}{2} = \frac{5 \times 4}{s} = 10$$

t-Tests durchzuführen. Ein solches Vorgehen ist allerdings problematisch. Bekanntlich ist bei Gültigkeit der Nullhypothese die Wahrscheinlichkeit, dass H_0 irrtümlich abgelehnt wird, gleich α, sagen wir also $\alpha = .05$. Dass wir bei 10 t-Tests mindestens eine derartige Fehlentscheidung treffen, ist gleich 1 minus die Wahrscheinlichkeit, keine Fehlentescheidung zu treffen. Gilt also generell die Nullhypothese, so ist die Wahrscheinlichkeit *keiner* Fehlentscheidung gleich $(1 - \alpha)^5$, und somit ist die Wahrscheinlichkeit mindestens einer Fehlentscheidung

$$P(\text{mindestens eine Fehlentsch.}) = 1 - (1 - \alpha)^5. \tag{3.16}$$

Diese Wahrscheinlichkeit ist größer als α. Denn sicherlich ist[22]

$$1 - \alpha > (1 - \alpha)^p, \quad p > 1,$$

und somit folgt $1 - (1 - \alpha)^p > \alpha$. Je mehr Therapien man miteinander vergleichen möchte, desto größer wird die Wahrscheinlichkeit, mindestens einmal irrtümlich die Nullhypothese zurückzuweisen: man postuliert Unterschiede, wo möglicherweise keine sind.

Die Idee ist nun, die vielen t-Tests zunächst einmal durch einen einzigen Test zu ersetzen, mit dem man herausfindet, ob es überhaupt ein Paar von Therapien gibt, die sich signifikant in ihrer Effektivität unterscheiden. Dazu geht man von einem linearen Modell aus:

$$y_{ij} = \mu_i + e_{ij}, \quad i = 1, \ldots, p \tag{3.17}$$

[22] Es sei a eine Zahl zwischen 0 und 1, $0 < a < 1$. Dann ist eine ganzahlige Potenz von a stets kleiner als a: $.5^2 = .25$, $.7^3 = .343$, etc. Da $0 < 1 - \alpha < 1$, folgt die Behauptung.

Dabei ist μ_i die "wahre" Effektivität der i-ten Therapie, und e_{ij} ist ein Fehler (der Effekt aller nicht kontrollierten Variablen) bei der Messung der j-ten Person in der i-ten Therapie. Die generelle Nullhypothese ist dann

$$H_0\text{: } \mu_1 = \mu_2 = \cdots = \mu_p = \mu \tag{3.18}$$

Die im Folgenden dargestellte Idee, H_0 durch eine Zerlegung der Gesamtvarianz der Daten zu testen, geht auf R.A. Fisher[23] zurück. Dazu werde zunächst angenommen, dass H_0 wahr sei. Weiter werde angenommen, dass die Fehler e_{ij} von den jeweiligen μ_i-Werten unabhängig seien, so dass für die Varianz

$$Var(y_{ij} = Var(\mu_i + e_{ij}) = Var(e_{ij}) = \text{konstant für alle } i \tag{3.19}$$

gilt; denn μ_i ist eine Konstante, die nichts zur Varianz der y_{ij} beiträgt, zumal unter H_0, da dann ja $\mu_i = \mu$ für alle i gilt. Für $Var(e_{ij})$ wird auch σ_e^2 geschrieben, also $\sigma_e^2 = Var(y_{ij})$. Natürlich hat man für die Varianzen nur Schätzungen $s^2_{y_{ij}}$, s_e^2, $s^2(\mu_i)$, und für die μ_i hat man nur die Schätzungen $\bar{y}_i$, die Stichprobenmittelwerte für die einzelnen Gruppen. Für jede Experimentalgruppe kann man nun σ_e^2 durch

$$s_i^2 = \frac{1}{n}\sum_{j=1}^{n}(y_{ij} - \bar{y}_i)^2 \tag{3.20}$$

schätzen[24]. Wegen der allgegenwärtigen Stichprobenfehler werden die s_i^2 nicht alle identisch sein, auch wenn die Annahme $\sigma_e^2 =$ konstant für alle i korrekt ist. Eine bessere Schätzung wird man bekommen, wenn man die s_i^2 noch einmal mittelt, – wie man das macht, wird gleich deutlich werden. Unter H_0 werden die Schätzungen $\bar{y}_i$ um μ herum zufällig verteilt sein, und für ihre Varianz gilt $Var(\bar{y}_i) = \sigma_e^2/n$. Das heißt aber, dass $\sigma_e^2 = nVar(\bar{y}_i)$ ist, so dass man aus der Varianz der Mittelwerte ebenfalls eine Schätzung für σ_e^2 gewinnen kann. Gilt H_0 *nicht*, so werden sich also die μ_i ebenfalls eine von Null verschiedene Varianz haben und die Schätzung von σ_e^2 anhand der Formel $\sigma_e^2 = nVar(\bar{y}_i)$ wird zu groß ausfallen.

Um von diesen Überlegungen zu einem wirklichen Test zu gelangen betrachte man nun die Gesamtvarianz der Daten. Dazu sei $\bar{y}$ der Mittwert aller Messwerte. Weiter sei y_{ij} die gemessene Effektivität bei der j-ten Person in der i-ten Therapie. Dann ist

$$s^2 = \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2 \tag{3.21}$$

die Gesamtvarianz der Daten, und n_i ist die Anzahl der Personen, die die i-te Therapie gewählt haben. Unter der Bedingung, dass die Nullhypothese (3.18) gilt, werden die Mittelwerte $\bar{y}_i$ für die verschiedenen Gruppen zufällig um μ variieren. Die

[23] Ronald Aylmer Fisher (1890 – 1962), britischer Mathematiker und Statistiker, Biologe/Genetiker, Evolutionsforscher, etc, der eine Reihe von statistischen Verfahren erfand und die heute übliche Inferenzstatistik begründete.

[24] Bei kleineren Stichproben teilt man durch $n-1$ statt durcn n, um eine systematische Unterschätzung zu vermeiden. Da hier am Ende nur die Quadratsummen interessieren, muß hierauf nicht wetier eingegangen werden.

Gesamtvarianz s^2 kann nun in Varianzkomponenten aufgebrochen werden. Denn sicherlich gilt

$$(y_{ij} - \bar{y})^2 = (y_{ij} - \bar{y}_i + \bar{y}_i - \bar{y})^2.$$

Nun kann man aber

$$a_{ij} = y_{ij} - \bar{y}_i \quad b_i = \bar{y}_i - \bar{y}$$

setzen, so dass

$$(y_{ij} - \bar{y})^2 = (a_{ij} + b_i)^2 = a_{ij}^2 + b_i^2 + 2a_{ij}b_i$$

gilt, d.h.

$$(y_{ij} - \bar{y})^2 = (y_{ij} - \bar{y}_i)^2 + (\bar{y}_i - \bar{y})^2 + 2(y_{ij} - \bar{y}_i)(\bar{y}_i - \bar{y}). \tag{3.22}$$

Summiert man jetzt noch über i und j, erhält man

$$ns^2 = QS_{ges} = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2 = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^{p}\sum_{j=1}^{n_i}(\bar{y}_i - \bar{y})^2 \tag{3.23}$$

denn

$$\sum_{i=1}^{p}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)(\bar{y}_i - \bar{y}) = 0, \tag{3.24}$$

wovon man sich durch nachrechnen überzeugen kann. Diese Gleichung besagt aber, dass die Variablen $y_{ij} - \bar{y}_i$ und $\bar{y}_i - \bar{y}$) voneinander statistisch unabhängig sind, denn ihre Kovarianz ist offenbar stets gleich Null[25]. Damit hat man gezeigt, dass die Quadratsumme, die für die Berechung der Gesamtvarianz benötigt wird (QS_{ges}) in zwei voneinander unabhängige, additive Komponenten zerlegen läßt: in

$$QS_{inn} = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2 \tag{3.25}$$

mit QS_{inn} für *Quadratsumme innerhalb* der Gruppen, und in

$$QS_{zwisch} = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(\bar{y}_i - \bar{y})^2, \tag{3.26}$$

mit QS_{zwisch} für *Quadratsumme zwischen* den Gruppen, so dass man symbolisch

$$QS_{ges} = QS_{inn} + QS_{zwisch}. \tag{3.27}$$

schreiben kann. Die zu $2ab$ korrespondierende Summe (3.24) entspricht der Kovarianz zwischen den $(y_{ij} - \bar{y}_i)$, den Abweichungen der Messwerte y_{ij} von ihrem Gruppenmittelwert, und den $(\bar{y}_i - \bar{y})$, den Abweichungen der Gruppenmittelwerte vom

[25] Dieser Schluß gilt allerdings nicht uneingeschränkt, denn es läßt zeigen, dass zwei Variablen eine Kovarianz gleich Null haben können, aber dennoch statistisch abhängig voneinander sind. Sind aber die y_{ij} alle Gauß-verteilt, so gilt dieser Schluß, so dass bei der Varianzanalyse noch die zusätzliche Annahme, dass die y_{ij} Gauß-verteilt sind, gemacht werden muß!

Gesamtmittel, und diese Kovarianz ist nach (3.24) gleich Null, so dass die a_{ij} und b_i als stochastisch unabhängige Größen betrachtet werden können – zumindest, wenn die Daten normalverteilt sind. Dann folgt aber auch, dass QS_{inn} und QS_{zwisch} stochastisch unabhängige Variablen sind. Beide Größen können als Schätzungen für die Fehlervarianz σ^2 betrachtet werden. QS_{inn} repräsentiert einen Mittelwert über die Schätzungen von σ_e^2 innerhalb der Gruppen, und QS_{zwisch} ist eine Schätzung von $s_{\bar{y}}^2$, der Varianz der Mittelwerte. Es ist aber bekannt, dass $s_{\bar{y}}^2 = s_e^2/n$, so dass $ns_{\bar{x}}^2 = s_e^2$. Man hat also zwei voneinander unabhängige Schätzungen der Fehlervarianz, – falls H_0 korrekt ist. Gilt dagegen H_1, die Hypothese, dass H_0 *nicht* korrekt ist, so wird die Schätzung von s_e^2 auf der Basis von $s_{\bar{y}}^2$ zu groß sein, um noch mit H_0 kompatibel zu sein. Man muß also nur zwei Varianzschätzungen auf Gleichheit testen. Ein solcher Test ist der F-Test. Die Details der Herleitung der Anwendung des F-Tests auf die Varianzkomponenten QS_{inn} und QS_{zwisch} können hier übergangen werden, es resultiert jedenfalls der Test

$$F_{p-1,n-p} = \frac{QS_{zwisch}/(p-1)}{QS_{inn}/(n-p)} \tag{3.28}$$

$n = \sum_i n_i$ ist die Gesamtzahl der Messungen, und $p-1$ und $n-p$ sind die Freiheitsgrade für den Quotienten. Für $F_{p-1,n-p}$ läßt sich für ein gewähltes α ein kritischer Wert F_{crit} finden derart, dass H_0 mit der Fehlerwahrscheinlichkeit α zurückgewisen werden kann, wenn $F_{p-1,n-p} > F_{crit}$. Tritt dieser Fall ein, kann davon ausgegangen werden, dass mindestens ein Paar von Mittelwerten $(\bar{y}_i, \bar{y}_k)$ überzufällig verschieden ist. Man kann dann anhand von a posteriori-Tests auf die Suche nach denjenigen Mittelwertspaaren gehen, die sich signifikant voneinander unterscheiden, worauf hier aber nicht eingegangen werden soll, da der Fokus auf grundlegende Aspekte der Versuchsplanung liegt.

Es sollte jetzt klar sein, warum der Ausdruck 'Varianzanalyse' gebraucht wird, wenn man an Mittelwertsunterschieden interessiert ist. Denn man zerlegt eine Varianz, um die möglichen Unterschiede zwischen Mittelwerten erkennen zu können. Für den Ausdruck Varianzanalyse ist auch das Acronym ANOVA üblich – von englisch Analysis Of Variance.

Im Beispiel ist die Frage betrachtet worden, ob sich p Therapien voneinander unterscheiden. In der Sprache der Varianzanalyse definieren die Therapien einen *Faktor*, und die einzelnen Therapien sind die *Stufen* (engl. levels) des Faktors. Dementsprechend heißt die bis jetzt besprochene Versuchsanordnung auch *Einfaktorielles Design.*

Natürlich kann es sein, dass σ_e^2 den Effekt von einer Reihe weiterer Variablen reflektiert, d.h. die Fehlervarianz kann reduziert werden, wenn man derartige Variablen explizit kontrolliert. Im Therapiebeispiel könnte das Geschlecht der Patienten, die an den Therapien teilnehmen, eine Rolle spielen. Man kann also zu einem 2-faktoriellen Design (Plan) übergehen, bei dem der Faktor 'Geschlecht' mit den Stufen 'weiblich' und 'männlich' explizit kontrolliert wird. Der Plan (das Design) hat dann die Form der Tabelle 5. Die μ_{wj} und μ_{mj} sind die Erwartungswerte der abhängigen Variablen

für die jeweilige Kombination von Faktorstufen. Für diese Erwartungswerte kann ein *Strukturmodell* angeschrieben werden:

$$\mu_{wj} = \mu + \mu_{w\cdot} + \mu_{\cdot j} + \mu_{w\times j}. \tag{3.29}$$

Darin ist μ ein unspezifischer Parameter, der nur Eigenschaften der verwendeten Skala für die abhängige Variable widerspiegelt. $\mu_{w\cdot}$ ist eine systematische Komponente (ein *Effekt*) der Stufe w des Faktors Geschlecht, die unabhängig von anderen Faktoren wirkt. $\mu_{\cdot j}$ ist eine systematische Komponente der Therapie T_j, die unabhängig von allen anderen Faktoren wirkt. $\mu_{w\times j}$ schließlich ist eine systematische Komponente, die durch eine Wechselwirkung (Interaktion) nur zwischen der Stufe w und der Stufe T_j existiert; die Wechselwirkung ist ein Effekt, der sich spezifisch aus dem Zusammentreffen dieser beiden Stufen ergibt. Analog zu (3.29) kann man

$$\mu_{mj} = \mu + \mu_{m\cdot} + \mu_{\cdot j} + \mu_{m\times j}. \tag{3.30}$$

anschreiben, oder ganz allgemein

$$\mu(G,T) = \mu + \mu_G + \mu_T + \mu_{G\times T}. \tag{3.31}$$

μ_G und μ_T sind die *Haupteffekte* der beiden Faktoren, und $\mu_{G\times T}$ sind die Wechselwirkungen zwischen den Faktoren. Für die Messwerte gilt dann z.B.

$$y_{wj} = \mu_{wj} + e = \mu + \mu_{w\cdot} + \mu_{\cdot j} + \mu_{w\times j} + e \tag{3.32}$$

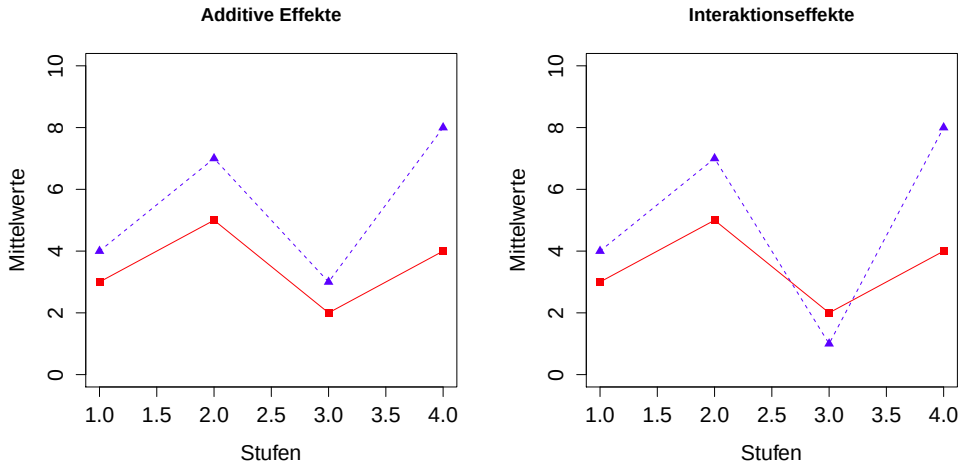
Die einzelnen Haupt- und Wechselwirkungseffekte gehen *additiv* ein, und insofern ist das Strukturmodell ein lineares Modell. Es läßt sich zeigen, dass es im Prinzip dem Ansatz der multiplen Regression entspricht; auf die Details soll hier nicht weiter eingegangen werden, damit nicht zu viele formale Komponenten in die Darstellung geraten. Man spricht deswegen vom *Allgemeinen Linearen Modell* (ALM). Im Rahmen des ALM lassen sich also Zusammenhangshypothesen und Unterschiedshypothesen testen.

Für das zweifaktorielle Design gibt es insgesamt drei Nullhypothesen: H_{01}: alle Effekte des ersten Faktors (etwa des Faktors 'Therapie') sind gleich Null, H_{02}: alle Effekte des zweiten Faktors (Geschlecht) sind gleich Null, und H_{03}: alle Wechselwirkungseffekte sind gleich Null. Die Hypothesen sind voneinander unabhängig und werden durch F-Tests geprüft. Die Details dieser Tests werden hier übergangen, es soll aber illustriert werden, wie sich Haupt- und Wechselwirkungseffekte auswirken. Den Erwartungswerten μ_{wj}, μ_{mj}, und $\mu_{w\times j}$ etc entsprechen Mittelwerte $\bar{y}_{wj}$, $\bar{y}_{mj}$, etc. Es werde angenommen, dass die Wechselwirkungseffekte alle gleich Null seien. Man kann dann gemäß (3.29) und (3.30) die Mittelwerte $\bar{y}_{\cdot j}$ für $j = 1, \ldots, p$ in ein Diagramm einzeichnen und ebenso die Mittelwerte $\bar{y}_{\cdot j}$. Es entstehen zwei Kurven, die parallel sein müssen, da sie sich nur durch die Konstanten $\bar{y}_{m\cdot}$ und $\bar{y}_{w\cdot}$ voneinander unterschieden. Umgekehrt kann man aus der Parallelität der Kurven schließen, dass keine Wechselwirkungseffekte existieren. Sind die Kurven überdies in guter Näherung Gerade, die parallel zur x-Achse liegen, so läßt sich vermuten,

Tabelle 5: Zweifaktorieller Versuchsplan

Geschlecht	Therapien			
	T_1	T_2	$\cdots$	T_p
w	μ_{w1}	μ_{w2}	$\cdots$	μ_{wp}
m	μ_{m1}	μ_{m2}	$\cdots$	μ_{mp}

Abbildung 3: Anova: Mittelwerte ohne und mit Wechselwirkungseffekte



dass die Haupteffekte gleich Null sind. Existieren dagegen Wechselwirkungseffekte, so werden sich die Kurven überschneiden.

Mehrdimensionale Designs: Im Prinzip können auf diese Weise Versuchspläne mit drei, vier oder mehr Faktoren aufgestellt werden. Je mehr Faktoren man kontrolliert, desto geringer wird die Fehlervarianz werden (die Fehlervarianz wird durch die Faktoren "aufgeklärt"). Es können Interaktionen zwischen Paaren von Faktoren, zwischen Tripeln von Faktoren etc getestet werden. Es müssen allerdings stets die beiden Annahmen:

1. Die Daten müssen normalverteilt sein,
2. Die Varianzen inden den verschiedenen Gruppen müssen homogen sein.

Der Ausdruck 'homogene Varianzen' heißt nichts weiter, als dass die Fehlervarianz σ_e^2 für alle Kombinationen von Stufen der Faktoren dieselbe sein soll. Diese Forderungen ergeben sich aus dem Ziel, die einzelnen Hypothesen durch F-Tests prüfen zu wollen, die voraussetzen, dass die Daten normalverteilt mit homogener Varianz sind. Insbesondere die Forderung nach Varianzhomogenität ist relativ restriktiv: bei

vielen Daten verändert sich mit den Mittelwerten auch die Varianz. Inhomgene Varianzen können zu Fehlentschiedungen über die Hypothesen führen. Die allgemeine Beliebtheit der ANOVA-Designs erklärt isch u.a. aber dadurch, dass sich die F-Tests als relativ robust gegen Verletzungen der Voraussetzungen erweisen. Die Robustheit der Tests ist ein Thema, das in Statistikveranstaltungen diskutiert werden muß.

Mehrdimensionale Versuchspläne haben den Vorteil, die Daten einigermaßen detailliert analysieren zu können. Hat man R Faktoren mit jeweils $r_1, r_2, \ldots, r_R$ Stufen, so hat man insgesamt $r_1 \times r_2 \times \cdots \times r_R$ Stichproben mit, sagen wir, jeweils n Fällen zu bilden. Dieser Sachverhalt ist als *curse of dimensionality* bekannt (statt des Ausdrucks 'Faktor' wird auch der Ausdruck 'Dimension' verwendet). Man rechnet leicht nach, dass der Aufwand, die entsprechenden Stichproben zu bilden, sehr schnell sehr groß wird. Der Vorteil, jede Kombination von Bedingungen untersuchen zu können, wird durch den Nachteil, entsprechend viele Stichproben bilden zu müssen, schnell wieder aufgehoben.

Messwiederholungen: Die Stichprobenproblematik kann abgeschwächt werden, wenn zumindest teilweise dieselben Personen unter den verschiedenen Bedingungen (Kombinationen von Stufen der Faktoren) teilnehmen. Man spricht dann von *Messwiederholumngen* (engl.: repeated measurements, im Unterschied zu replicated measurements, bei denen verschiedene Personen in den verschiedenen Gruppen vermessen werden). Dieser Fall wird um Teil durch die Fragestellung erzwungen, etwa, wenn der Effekt von Trainings-, Lern- oder Interventionsmaßnahmen (z.B. Therapien) untersucht werden soll. Außer der Ersparnis an Vpn (Versuchspesonen) kann man den Effekt einer Reduktion der Fehlervarianz haben: die Varianz, die durch Unterschiede zwischen den Vpn entsteht, entfällt zumindest teilweise. Der Preis dafür ist: die Messungen in den verschiedenen Gruppen sind nicht mehr notwendig unabhängig voneinander. In den klassischen Darstellungen der ANOVA (etwa Scheffé (1959) wurden recht artifiziell anmutende, restriktive Annahmen eingeführt, um eine Schätzung der F-Quotienten zu ermöglichen, z.B. die Annahme, dass die Korrelationen zwischen allen Vpn gleich groß sind, was im Allgemeinen sicher nicht der Fall ist. Grundsätzliche Lösungen, wie sie schon von Scheffé angedacht wurden, waren wegen des sehr hohen Rechenaufwands und mangelnder Computer nicht praktikabel. Heute bieten Programmpakete entsprechende Module an, vor allem das frei verfügbare Program R (The R Project for Statistical Computing, https://www.r-project.org/, https://cran.r-project.org/) bietet Module (packages) an, die verwendet werden können. Die Module werden in der internationalen Statistikergemeinde ständig weiterentwickelt und können frei heruntergeladen werden, – der Nachteil ist allerdings, dass man sich eine gewisse Zeit mit ihnen beschäftigen muß, um sie effektiv anwenden zu können[26].

[26]Speziell für Psychologen: Luhmann, M.: R für Einsteiger – Einführung in die Statistiksoftware für dieSozialwissenschaften. Weinheim 2011. Maike Luhmann ist Psychologin und hat das Buch speziell für Psychologen geschrieben, R enthält eine *package* mit nahezu allen statistischen Verfahren, die in der Psychologie angewendet werden. Der Zugang zu R wird dadurch sehr leicht, überdies ist die Grafik-Software von R sehr gut. Hinweis: man benutze R über R-Studio.

Fixed-, Random- und Mixed Models: Bisher sind die Stufen der Faktoren als feste Größen betrachtet worden, und die Aussagen der ANOVA beziehen sich dann auch nur auf diese Stufen. Gelegentlich repräsentiert aber ein Faktor eine kontinuierliche Variable. Ein Beispiel ist das Alter: will man den Effekt des Alters auf Therapie- oder Lern- bzw. Vergessensprozesse untersuchen, so ist das Alter der Vpn sicher eine sinnvolle unabhängige Variable. Aber wir altern jede Sekunde, ja jede Millisekunde, und es sind noch feinere Aufteilungen denkbar. Man wird, je nach Fragestellung, Alter aber wohl eher in Tagen, Wochen, Monaten oder Jahren erfassen wollen. Selbst wenn es um Jahre geht, wird man kaum 70 oder 80 Stufen betrachten wollen oder können, und deshalb größere Lebensabschnitte betrachten. So könnte man die Abschnitte 15 - 20, 25 - 30, 35 - 40 und 45 - 50 wählen, falls man damit der Fragestellung gerecht wird. Betrachtet man diese Auswahl als *fixed*, so kann man nur über genau diese Abschnitte Aussagen machen. Will man generell Aussagen über den Effekt des Alterns machen, so muß man sie als zufällige Auswahl (random factors bzw levels) ansehen, und dies schlägt sich in der Konstruktion der F-Tests nieder. In bestimmten Versuchsplänen mit Messwiederholungen kann man einen Vpn-Faktor betrachten, bei dem jede Vp eine Stufe dieses Faktors repräsentiert. Natürlich ist man nicht nur an diesen speziellen Vpn interessiert, sondern möchte generell etwas über Vpn aussagen. In diesem Fall ist dieser Faktor ein *Random Factor*.

Versuchspläne, in denen nur *fixed factors*, also feste Faktoren, vorkommen, heißen demnach *fixed factor designs*, Versuchsplane, in denen nur random Faktoren vorkommen, heißen Random Factor Designs und Pläne, in dennen sowohl fixed wie random Faktoren vorkommen, heißen *mixed factor designs*. Sie unterscheiden sich in der Art der Berechung der F-Quotienten. Es ist wichtig, diesen Sachverhalt zu berücksichtigen, um zu korrekten Interpretationen der Daten zu gelangen.

Hierarchische Versuchspläne: Bisher ist stillschweigend angenommen worden, dass für alle $r_1 \times r_2 \times \cdots r_p$ Kombinationen von Faktorstufen in einem p-faktoriellem (oder auch p-dimensionalem) Versuchsplan auch Stichproben erhoben werden können, was ein Maximum an möglichen Hypothesenprüfungen erlaubt. Man spricht von *vollständigen Versuchplänen* oder *completely crossed designs*. Es ist aber so, dass bestimmte Hypothesen, d.h. Effekte, gar nicht von besonderem Interesse sind. So sind Interaktionen höherer Ordnung (und je größer die Anzahl der Faktoren, also der unabhängigen Variablen sind, desto mehr Interaktionen höherer Ordnung gibt es) aber oft gar nicht von Interesse, schon weil sie zum Teil sehr schwer zu interpretieren sind. Man kann dann *geschachtelte*, auch *hierarchische* Versuchspläne betrachten (englisch: nested designs). Bei diesen Plänen wird nicht jede Stufe eines Faktors mit jeder Stufe der übrigen Faktoren miteinander verglichen.

So sei K eine Krankheit, für die es zwei mögliche Therapien gibt: diese definieren den Faktor T mit den zwei Stufen T_1 und T_2. Man möchte die Wirksamkeit der Therapien miteinander vergleichen, findet aber, dass die verschiedenen Kliniken – diese definieren den Faktor B - jeweils nur eine der beiden Therapien verwenden. So verwenden etwa die Kliniken B_1, B_2 die Therapie T_1, und die Kliniken B_3, B_4, B_5 die Therapie T_2. Man erhält dann den Versuchsplan Für die Kombinationen von

Tabelle 6: Zweifaktorieller hierarchischer Versuchsplan

	Therapie (T)	
Klinik (B)	T_1	T_2
B_1	Mess-	keine
B_2	werte	Messwerte
B_3	keine	Mess-
B_4	Mess-	werte
B_5	werte	

B_1 und B_2 und T_2 gibt es keine Messwerte, ebenso nicht für die Kombinationen von B_3, B_4, B_5 mit T_1. Die übliche Darstellung ist Da hier bestimmte Kombinationen von

Tabelle 7: Alternative Darstellung eines hierarchischen Designs

T_1		T_2		
B_1	B_2	B_3	B_4	B_5
M	e	s	s	–
w	e	r	t	e
		⋮		

Stufen nicht vorkommen können die entsprechenden Wechselwirkungen nicht getestet werden, – es ist denkbar, dass die Therapie T_1 in der Klinik B_3 besonders gut funktioniert hätte, weil die Bedingungen für T_1 dort besser gewesen wären, aber diese Information läßt sich nicht aus diesem Versuchsplan herausfiltern. Bei der Planung eines hierarchischen Designs muß man also darauf achten, auf welche Interaktionen man verzichten kann. Dies bedeutet, dass auch die Haupteffekte schwieriger zu interpretieren sind als in einem vollständigen Versuchsplan: ein möglicher Unterschied zwischen den Therapien T_1 und T_2 ist nur deduzierbar, wenn man den Effekt der Kliniken vernachlässigen kann, d.h. dass die Interaktionen zwischen Kliniken und Therapieform vernachlässigbar ist. Mit diesem Problem muß man aber leben, da es kaum möglich sein wird, Kliniken zu finden, die alle Therapieformen gleichermaßen anwenden.

Quadratische Versuchspläne: Es werde angenommen, dass man vier Therapieformen $T_1, \ldots, T_4$ habe und ebenso vier Kliniken, in denen jeweils alle Therapieformen angewendet werden. Der vollständige Versuchsplan kann dann durch ein Quadrat repräsentiert werden: es gibt vier Zeilen $B_1, \ldots, B_4$ und vier Spalten $T_1, \ldots, T_4$, und die Zelle (B_i, T_j) enthält die Daten für die Klinik B_i und die Therapie T_j. Man möchte aber noch das Alter der Patienten mitberücksichtigen, und man definiert vier Altersgruppen $A_1, \ldots, A_4$. Man erhält dann einen $4 \times 4 \times 4$ Kubus. Tatsächlich

läßt sich daraus aber wieder ein quadratisches Design machen, das *Lateinische Quadrat*. Das Design sieht dann so aus: Wie man sieht, kommt jede Stufe des Faktors

Tabelle 8: Lateinisches Quadrat

	T_1	T_2	T_3	T_4
B_1	A_1	A_2	A_3	A_4
B_2	A_2	A_3	A_4	A_1
B_3	A_3	A_4	A_1	A_2
B_4	A_4	A_1	A_2	A_3

'Alter' in Kombination mit jeder Stufe des Faktors 'Therapie' und mit jeder Stufe des Faktors 'Klinik' vor. Man sagt deshalb auch, dass der Versuchsplan in Bezug auf die Haupteffekte *ausbalanciert* sei.

Wie man der Tabelle 8 leicht entnehmen kann, ist der Versuchsplan in Bezug auf die Interaktionen *nicht* ausbalanciert. Dazu müßten alle Faktorstufen in Kombination auftreten, was hier aber offenbar nicht der Fall ist. Dieser Sachverhalt hat Rückwirkungen auf die Interpretation der Haupteffekte: sie sind nur interpretierbar, wenn die Interaktionen zwischen den Faktoren vernachlässigbar sind. Es gibt weitere Versuchspläne dieser Art, etwa *griechisch-lateinische Quadrate*, auf die aber in diesem Skript nicht weiter eingegangen werden soll.

3.4 Veränderungshypothesen

Die Untersuchung von Veränderungen von Einstellungen oder Verhaltensweisen ist ein wichtiges Forschungsgebiet. Die Evaluation des Verlaufs von Therapien ist zum Beispiel von großer praktischer Relevanz und steht in enger Beziehung zur Evaluation des Erfolgs von Therapien. Gleichzeitig sind sie im Allgmeinen vom Typ eines Quasi-Experiments, da sich oft Probleme mit der Randomisierung ergeben und die Bedingungen der Untersuchung oft nicht in dem Maße kontrolliert werden können, wie es in Laborexperimenten möglich ist. Man definiert mindestens eine abhängige Variable und erhebt deren Wert in aufeinander folgenden Zeitabschnitten. Man erwartet,dass ein Trend in der abhgängigen Variablen sichtbar bzw. nachweisbar wird, der idealerweise den Erfolg der Therapie reflektiert. Im Prinzip kann die Analyse varianzanalytisch erfolgen: ein Faktor ist die Zeit, und die Stufen dieses Faktors müssen so gewählt werden, dass einerseits die Kosten nicht zu groß werden, andererseits mögliche Veränbderungen sichtbar gemacht werden können. Im einfachsten Fall hat man nur zwei Messzeitpunkte: zu Beginn der Therapie und zum Ende der Therapie. Dieser Faktor ist ein Messwiederholungsfaktor und es gilt, den Effekt der Korrelation zwischen den Messwerten so in Rechnung zu stellen, dass ein Test der Nullhypothese (die Therapie hat keinen Effekt) möglich wird. Ist man nicht nur am Erfolg, sondern auch am Verlauf der Therapie interessiert, zum Beispiel um den

Prozess der Veränderung näher zu untersuchen, kann die Anzahl der Stufen des Zeitfaktors entsprechend erhöht werden.

Ein möglicher Aspekt, der eine derartige Untersuchung zum Quasi-Experiment macht, sind die möglicherwiese unterschiedlichen Verläufe der abhängigen Variablen bei den verschiedenen PatientenInnen. Mittelt man über diese Verläufe, so kann es passieren, dass der mittlere Verlauf anders als die individuellen Verläufe ist.

Zeitreihenversuchspläne: Zeitreihen sind Folgen von Messungen in zeitlich gleichem Abstand, also Messungen der Art

$$X_t, X_{t+1}, X_{t+2}, \ldots, X_{t+j}, \ldots$$

wobei 1, 2, ... Zeitintervalle von konstanter Dauer sind. Zeitreihen sind *stationär*, wenn der Erwartungswert $\mathbb{E}(X_{t+j} = \mu$ und die Varianz $Var(X_{t+j} = sigma^2$ für alle j konstant sind, d.h. die X_{t+j} sind zufällige Veränderliche der Form $X_{t+j} = \mu + \sigma^2$. Stationarität bedeutet, dass sich die gemessene Größe nicht systematisch in der Zeit verändert. Die statistische Theorie der Zeitreihen beschäftigt sich u.a. mit der Möglichkeit, Zeitreihen in Bezug auf ihre Stationarität (H_0) bzw. auf Trends (H_1) zu prüfen. Auf die Details kann hier nicht eingegangen werden, das Problem bei der Analyse von Zeitreihen sind jedenfalls die Korrelationen zwischen den X_{t+j}, die ja schon bei der ANOVA mit Messwiederholungen ein Problem darstellen. Gleichwohl kommt man auf die Analyse von Zeitreihen zumindest im Prinzip nicht herum, wenn man den Verlauf von Lernprozessen oder von Effekten von Trainingsprogrammen oder Therapien untersuchen möchte.

Ein relativ einfacher Versuchsplan ist der Ein-Gruppen-Plan mit mehreren Vorher-Nachher-Messungen. Solche Pläne heißen auch ABAB-Pläne: in einem ersten Messabschnitt A werden die Ausgangswerte von abhängigen Variablen gemessen (Kontrollmessungen). Im folgenden Zeitabschnitt B wird das jeweils interessierende Progamm (etwa eine Therapie) durchgeführt und es werden Messungen unter dem Einfluß der Intervention gemacht. Es folgt wieder ein Kontrollabschnitt A ohne Intervention, anschließend wieder eine Phase B mit Intervention, etc. Man hat nur eine Gruppe, die gewissermaßen simultan als Kontrool- und als Experimentalgruppe dient. Kann man mehrere Gruppen gleichzeitig untersuchen, so entsteht ein *Mehrgruppen-Zeitreihen-Design*. Die Gruppen sind durch interessierende Merkmale (weiblich-männlich etc) definiert.

Ereignisanalysen: Auf die statistischen Fragen, die bei diesen Designs auftreten, soll hier nicht weiter eingegangen werden, weil sie einen Exkurs in die Zeitreihenanalyse voraussetzen, der hier nicht gegeben werden kann. Es soll aber darauf hingewiesen werden, dass die Notgwendigkeit, die Messungen in gleichgroßen Zeitintervallen durchzuführen, die Datenerhebung oft sehr erschwert. Eine Alternative zu den Zeitreihendesigns ist die *Ereignisanalyse*. Der Name ist ein wenig irreführend, weil nicht irgendwelche Ereignisse analysiert werden, sondern die zeitlichen Abstände zwischen ihnen. Ein Beispiel hierfür sind Patienten, die an immer wieder auftretenden schizophrenen Schüben leiden. Die Zeitspanne zwischen zwei Schüben variiert

zufällig, und der Erfolg einer Therapie kann – zumindest im Prinzip – daran gemessen werden, dass diese Zwischenzeiten länger werden, im Estrem "unendlich lang", wenn also ein Patient gar keinen Schub mehr bekommt. Um die Zwichenzeiten zu analysieren, wird die *Hazard-Funktion* eingeführt. Diese ist die bedingte Wahrscheinlichkeit, dass das in Frage stehende Ereignis in einem Zeitintervall $[t, t+\Delta t)$ eintritt *unter der Bedingung, dass es bis zum Zeitpunkt t noch nicht eingetreten ist*:

$$h(t) = P(\tau \in [t, t + \Delta t) | \tau > t). \tag{3.33}$$

Die Hazard-Funktion kann im Prinzip für beliebige Ereignisse definiert werden. So kann man fragen, wie groß die Wahrscheinlichkeit ist, dass ein gerade geborenes Kind im 81-ten Lebensjahr stirbt. Diese Wahrscheinlichkeit hängt von den allgemeinen Lebensbedingungen ab; sie ist zum Beispiel in Norwegen größer als in Deutschland. Man kann aber auch fragen, wie groß die Wahrscheinlichkeit ist, im folgenden Lebensjahr zu sterben, wenn man gerade seinen 80-ten Geburtstag feiert. Diese Wahrscheinlichkeit ist (wenn man nicht bereits schwer krank ist) kleiner als die des Babies, im 81-ten Lebensjahr zu sterben, denn all die Gründe, deretwegen man im Laufe eines Lebens sterben kann (Unfälle, Erkrankungen etc), hat man im Alter von 80 jahren ja schon überlebt.

Dioe in (3.33) definierte Hazard-Funktion hängt von der Wahrscheinlichkeitsverteilung der *Wartezeit* τ ab. Ein Spezialfall hierfür ist die Exponentialverteilung $f(\tau) = \lambda e^{-\lambda\tau}$. Bestimmt man für diese Verteilung die zugehörige Hazard-Funktion, so erhält man

$$h(t) = \lambda. \tag{3.34}$$

d.h. $h(t)$ hängt gar nicht von t ab. Es ist, als würde man zu jedem Zeitpunkt die Münze werfen, um zu erfahren ob man das folgende Intervall $[t, t + \Delta t)$ überlebt. Insofern repräsentiert der Fall (3.34) die *reine Zufälligkeit* der Ereignisse und damit eine Art von Nullhypothese.

$h(t)$ kann als von unabhängigen Variablen abhängig definiert werden; da ja λ einen kleineren oder größeren Wert haben kann, gilt dies auch für (3.34), so dass man in diesem Fall

$$\lambda = \lambda(x_1, \ldots, x_n) \tag{3.35}$$

hat.

Die Hazard-Funktion kann auch verteilungsfrei geschätzt werden und erweist sich damit als ein Messinstrument von relativ großer allgemeiner Bedeutung, – wenn es nicht darauf ankommt, bestimmte Größen zu messen, wie etwa in psychophysiologischen Untersuchungen, sondern nur, zu konstatieren, dass ein bestimmtes Ereignis eingetreten ist oder nicht. Zsuammenfassend kann gesagt werden, dass Zeitreihen - und Ereignisanalysen typischerweise Quasi-Experimente sind.

3.5 Die Analyse von Häufigkeiten

3.5.1 Log-lineare Analysen

Oft liegen nur kategoriale Daten vor und man kann, um den Zusammenhang zwischen den entpsrechenden Variablen zu bestimmen, nur die Häufigkeiten erheben, mit denen Konjunktionen von Kategorien auftreten. Ein Beispiel ist die Tabelle 9; es gibt zwei Variablen, Körperbau und psychische Erkrankung (wobei die Epilepsie keine psychische Erkrankung im engeren Sinne ist), und für beide Variablen sind Kategorien gegeben. So ist ein Körperbau entweder pyknisch, oder athletisch etc, und bei einer Erkrankung handelt es sich entweder um eine manisch-depressive Störung, eine Schizophrenie oder eine Epilepsie. Ob eine Beschränkung auf derartige Kategorien sinnvoll ist, soll hier nicht weiter diskutiert werden; die Tabelle wurde von Westphal (1931) im Zusammenhang mit einer von dem Psychiater Kretschmer aufgestellten These, derzufolge Körperbau und Art der Erkrankung in einem bestimmten Zusammenhang stehen, zusammengestellt, indem der in den Landeskrankenhäusern für Psychiatrie Patienten inspizierte und gemäß der Kretschmerschen Theorie klassifizierte. Von der Anzahl (insgesamt 8099 Patienten) her ist das Material beachtlich,

Tabelle 9: Körperbau und psychische Erkrankung nach Kretschmer

Typ	man.dep.	Epilepsie	Schizophr.	Σ
pyknisch	879	83	717	1679
athletisch	91	435	884	1410
leptosom	261	378	2632	3271
dysplastisch	15	444	550	1009
atypisch	115	165	450	730
Σ	1361	1505	5233	8099

auf mögliche methodische Fehler bzw. Probleme wird weiter unten eingegangen.

Man kann die Tabelle als zweifaktoriellen Plan betrachten und dementsprechend zur Analyse der Daten an eine zweifaktorielle ANOVA denken, um etwaige Haupteffekte (systematische Unterschiede im Auftreten der Körperbautypen einerseits und der Erkrankungen andererseits) und Interaktionseffekte (bestimmte Körperbautypen gehen mit bestimmten Erkrankungen einher) zu identifizieren. Derartige Analysen von Häufigkeitstabellen werden gemacht und sind oft gut interpretierbar, aber die Ergebnisse sind mit einem Fragezeichen zu versehen. Denn Häufigkeiten sind, wenn man Glück hat, allenfalls approximativ normalverteilt und die Varianzen von Häufigkeitsverteilungen sind keinesfalls notwendig homogen. Es werden also die Annahmen der ANOVA verletzt und es ist nicht klar, ob die oft beschworene Robustheit der F-Tests auch für eine gegebene Tabelle postuliert werden kann.

Die Anwendung der ANOVA ist auch nicht notwendig, denn es gibt andere Mög-

lichkeiten der Analyse. Die Nullhypothese ist, dass Körperbau und psychische Erkrankung voneinander unabhängig sind. Diese Hypthese erlaubt, aus der Tabelle die Häufigkeiten zu errechnen, die man erwartet, wenn die Nullhypothese korrekt ist. Dazu betrachtet man bei einer zufällig gewählten Person das gemeinsame Auftreten eines bestimmten Körperbaus (A) und einer bestimmten Erkrankung (B) als zufälliges Ereignis $A \cap B$, und aus der Wahrscheinlichkeitstheorie ist bekannt, dass bei stochastischer Unabhängigkeit von A und B die Formel

$$P(A \cap B) = P(A)P(B) \tag{3.36}$$

gilt. $P(A)$ und $PB)$ lassen sich aber aus den Randhäufigkeiten der Tabelle schätzen. Ein Beispiel ist das gemeinsame Auftreten des Körperbaus 'athletisch' und der Erkrankung 'Epilepsie'. Personen mit athletischem Körperbau treten in der Stichprobe 1410 mal auf, die relative Häufigkeit ist 1410/8091505 = .174, so dass $\hat{P}$(athletisch) = .174 eine Schätzung von P(athletisch) ist. Personen mit der Erkrankung Epilepsie treten mit der Häufigkeit 1505 in der Stichprobe auf, und mn erhält die Schätzung $\hat{P}$(1505/8099) = .186. Demnach würde man

$$\hat{P}(A \cap B) = .174 \times .186 = .032$$

erwarten, was einer Häufigkeit von 262 entspricht. In der Tabelle 10 sind die tatsächlichen Häufigkeiten und die unter H_0 erwarteten zusammen aufgeführt worden. Der

Tabelle 10: Körperbau und psychische Erkrankung: beobachtete und erwartete Häufigkeiten

		Erkrankung			
Typ		man./dep.	Epilepsie	Schizophr.	Σ
pyknisch	n_{ij}	879	83	717	1679
erwartet	$\hat{n}_{ij}$	282	312	1085	1679
athletisch	n_{ij}	91	435	884	1410
erwartet	$\hat{n}_{ij}$	237	262	911	1410
leptosom	n_{ij}	261	378	2632	3271
erwartet	$\hat{n}_{ij}$	549	608	2114	3271
dysplastisch	n_{ij}	15	444	550	1009
erwartet	$\hat{n}_{ij}$	170	187	652	1009
atypisch	n_{ij}	115	165	450	730
erwartet	$\hat{n}_{ij}$	123	136	471	730
	Σ	1361	1505	5233	$N = 8099$

Test von H_0 besteht in einem Vergleich von tatsächlichen und unter H_0 erwarteten Häufigkeiten. Der Vergleich muß so geschehen, dass man eine Statistik erhält, deren Verteilung man kennt, damit man einen kritischen Wert und die zugehörige Wahrscheinlicheit unter H_0 berechnen kann. Dazu dient der χ^2-Test (Chi-Quadrat-Test,

χ ist der griechische Buchtabe chi). Er basiert auf der χ^2-Verteilung: gegeben seien n standardnormalverteilte, unabhängige zufällige Variablen $z_1, \ldots, z_n$. Gesucht ist die Verteilung der Summe der quadrierten z_i^2:

$$\chi_n^2 = z_1^2 + z_2^2 + \cdots + z_n^2. \tag{3.37}$$

Diese Verteilung ergibt sich, wenn man nach der Verteilung der Stichprobenvarianz $s^2 = \sum_{i=1}^n (x_i - \bar{x})^2/n$ bei normalverteilten x_i fragt, und diese Frage tritt wiederum bei inferenzstatistischen Fragen auf. Ist n_{ij} ddie Häufigkeit, mit der der i-te Körperbautyp mit der j-ten Art von Erkrankung auftritt, und sind n_{i+} die Häufigkeiten des i-ten Körperbautyps und n_{+j} die Häufigkeiten des j-ten Erkrankungstyps, so kann man statt der z_j die Größen

$$\frac{n_{ij} - \hat{n}_{ij}}{\sqrt{\hat{n}_{ij}}}, \quad \hat{n}_{ij} = n\frac{n_{i+}}{n}\frac{n_{+j}}{n} = \frac{n_{i+}n_{+j}}{n}$$

berechnen; die $\hat{n}_{ij}$ sind die unter H_0 erwarteten Häufigkeiten nach (3.36). Es läßt sich zeigen, dass dann

$$S^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}} \to \chi^2_{(I-1)(J-1)},\ n \to \infty \tag{3.38}$$

gilt, wobei I die Anzahl der Zeilen, J die Anzahl der Spalten der Tabelle ist, und $\chi^2_{(I-1)(J-1)}$ ist ein Wert der χ^2-Verteilung mit $(I-1)(J-1)$ Freiheitsgraden. Die Forderung $n \to \infty$ bedeutet, dass die Größe S^2 für n gegen unendlich gegen die χ^2-Verteilung strebt. Gemeint ist, das die Annäherung für "große" n gilt. Glücklicherweise zeigt sich, dass bei diesem Test $n = 5$ schon eine große Zahl ist; diue Häufigkeiten in den Zellen der Tabelle sollten größer als 5 sein, damit S^2 schon als χ^2-verteilt gelten kann.

Man wählt einen Wert für α, den Fehler erster Art, etwa $\alpha = .05$ und findet den dazu gehörigen kritischen Wert χ^2_{crit}; ist der berechnete Wert $\chi^2_{(I-1)(J-1)}$ größer als χ^2_{crit}, so kann man H_0 verwerfen.

Log-lineare Modelle: Dies ist die Standardbehandlung einer Tabelle, die auch auf mehr als 2-dimensionale Tabellen angewendet werden kann. Aber es wird nur die Frage beantwortet, ob es überhaupt irgendwelche Abhängigkeiten gibt, – um welche Abhängigkeiten es sich handelt, wird dabei nicht gesagt. Derartige Fragen lassen sich aber beantworten, wenn man die *log-linearen Modelle* für Häufigkeitstabellen anwendet. Eine vollständige Behandlung derartiger Modelle kann hier nicht gegeben werden, es kann nur das Prinzip angedeutet werden.

Unter H_0 sind die erwarteten Häufigkeiten durch

$$\hat{n}_{ij} = \frac{n_{i+}n_{+j}}{n}$$

gegeben. Logarithmiert man die $\hat{n}_{ij}$ so erhält man

$$\log \hat{n}_{ij} = \log n_{i+} + \log n_{+j} - \log n = \mu + \mu_i^A + \mu_j^B \tag{3.39}$$

mit $\mu = \log n, \mu_i^A = \log n_{i+}, \mu_j^B = \log n_{+j}$. Gilt H_0, so lassen sich alle $\hat{n}_{ij}$ in diese additive Darstellung bringen, und gilt H_0 nicht, so lassen sich eben nicht alle $\hat{n}_{ij}$ in diese Form bringen. Gibt es spezifische Abbhängigkeiten, so lassen sie sich über einen Interaktionsterm ausdrücken:

$$\log \hat{n}_{ij} = \mu + \mu_i^A + \mu_j^B + \mu_{ij}^{A\times B} \tag{3.40}$$

$\mu_{ij}^{A\times B}$ repräsentiert – wie bei einer ANOVA – alle Effekte in n_{ij}, die nicht durch μ_i^A und μ_j^B erfasst werden, d.h. sie repräsentieren spezifische Abhängigkeiten. Man beachte, dass (3.40) (und als Spezialfall (3.39)) analog zu den Strukturgleichung einer ANOVA aufgebaut sind, nur sind hier die "Haupteffekte" μ_i^A und μ_j^B von eher geringem Interesse, da sie ja nur die Randhäufigkeiten reflektieren. Von Interesse sind Schätzungen $\hat{\mu}_{ij}^{AB}$ (AB ist nur eine einfachere Schreibweise als $A \times B$.

Der Ansatz läßt sich auf höherdimensionale Tabellen verallgemeinern; dabei treten dann nicht nur Interaktionsterme $\mu_{ij}^{AB}, \mu_{ik}^{AC}, \mu_{jk}^{BC}$ auf, sondern noch Terme, die in diesem Fall, noch die Interaktion zwischen den Faktgoren A, B und C repräsentieren: μ_{ijk}^{ABC}. Tests bestimmter Hypothesen werden über entsprechend konstruierte χ^2-Tests durchgeführt. Natürlich könnte man versuchen, spezielle Abhängigkeiten durch prüfen der entsprechenden bedingten Wahrscheinlichkeiten zu schätzen, z.B.

$$P(\text{schiz}|\text{dyspl}) \stackrel{?}{\approx} P(\text{schiz})P(\text{dyspl}).$$

Es ergibt sich dabei aber wie bei vielen t-Tests das Problem der α-Inflation, anlog zur ANOVA, das durch die log-lineare Analyse umgangen wird. Wie bei der Analyse von Daten aus dem semantischen Differential ergibt sich auch bei Tabellen die Möglichkeit, die Daten auf "latente Variablen" hin zu untersuchen; das Ergbnis ist in Methoden, Teil 1, schon vorgestellt worden. Für mehr als 2-dimensionale Tabellen muß man versuchen, sie in 2-dimensionale Tabellen umzuschreiben. Details dieses Verfahrens werden in der Veranstaltung *Multivariate Verfahren* besprochen.

Simpsons Paradox Tabellen können Zusammenhänge suggerieren, die in Wirklichkeit nicht existieren. Man betrachte die folgende Tabelle (Rachelet (1981): Die

Tabelle 11: Verhängung der Todesstrafe in den USA

	Todesstrafe		
Angeklagte	ja	nein	Σ
weiß	19	141	160
schwarz	17	149	166
Σ	36	290	326

Wahrscheinlichkeit, nach einem Schuldspruch zum Tode verurteilt zu werden, unter der Bedingung, ein Weißer zu sein, ist $P(T|w) = 19/160 = .119$, während die Wahrscheinlichkeit, nach einem Schuldspruch zum Tode verurteilt zu werden, unter der

Bedingung, ein SChwarzer zu sein, beträgt $P(T|s) = 17/166 = .102$. Umgekehrt ist die bedingte Wahrscheinlichkeit, ein Weißer zu sein, wnn man zum Tode verurteilt wurde, $P(w|T) = 19/36 = .528$, und die bedingte Wahrscheinlichkeit, schwarz zu sein, wenn man zum Tode verurteilt wurde, ist $P(s|T) = 17/36 = .472$. Offenbar ist die Behauptung, dass Schwarze häufiger zum Tode verurteilt werden als Weiß, nicht mit den Daten kompatibel.

Tatsächlich aber ist die von Radelet publizierte Tabelle 3-dimensional; die Tabelle 12 ist die vollständige Tabelle. Die bedingten Wahrscheinlichkeiten sprechen hier eine

Tabelle 12: Verhängung der Todestrafe in den USA

		Todesstrafe		
Angeklagte	Opfer	ja	nein	Anteil (ja)
weiß	weiß	19	132	.126
	schwarz	0	9	.000
schwarz	weiß	11	52	.175
	schwarz	6	97	.058

andere Sprache als die der Tabelle 11. Die Wahrscheinlichkeit, zum Tode verurteilt zu werden, hängt offenbar nicht nur von der Hautfarbe des Täters ab, sondern auch von der des Opfers. Es stehe A für Angeklagter. Es ist $P(T|A = w, O = w) = .126$ und $P(T|A = s, O = w) = .175$. Ist das Opfer also weiß, so ist die Wahrscheinlichkeit, zum Tode verurteilt zu werden, höher, wenn man Schwarzer ist, als wenn man Weißer ist. Weiter ist die Wahrscheinlichkeit, als Weißer zum Tode verurteilt zu werden wenn das Opfer schwarz ist, gleich Null, während man als schwarzer Täter mit einem schwarzen Opfer immer noch mit der Wahrscheinlichkeit .058 zum Tode verurteilt wird.

Tabelle 11 ist eine über die Dimension "Opfer" *aggregierte* Tabelle. Während in den Teiltabellen für die Opfer ein deutlicher Trend aufscheint, dass Schwarze häufiger zum Tode verurteilt werden als Weiße, ist dieser Trend in der aggregierten Tabelle gerade umgekehrt. Dieses Phänomen ist als *Simpsons Paradox* bekannt. Erzeugt wird diese Umkehrung durch eine Interaktion zwischen Täter und Opfer: es ist schlimmer, einen Weißen umzubringen als einen Schwarzen. Die Opferfarbe ist eine Moderatorvariable, die, wenn sie nicht explizit berücksichtigt wird, einen konfundierenden Effekt hat.

Deskriptive Modelle Oft liefern Methoden einen Einblick in die Daten, die auf der Basis von bestimmten Berechnungen eine anschauliche Beschreibung der Daten liefern. Die Korrespondenzanalyse ist ein solches Verfahren, das bei unübersichtlichen Tabellen gelegentlch mehr Einsicht verschafft als ein inferenzstatistisches Verfahren und dabei so gut wie keine Vorannahmen macht. Als Illustration für eine unübersichtliche Tabelle ist die Tabelle 13, in der die Häufigkeiten von Doktorgraden in

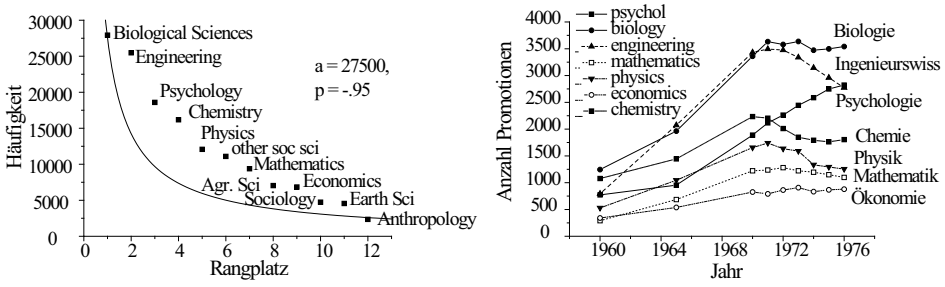
den verschiedenen Fächern in den Jahren zwischen 1960 und 1976 zusammengefasst wurden.

Tabelle 13: Trends bei Doktorgraden in den USA in den Jahren 1960 - 1976

	1960	1965	1970	1971	1972	1973	1974	1975	1976	Σ
Engineer.	794	2073	3432	3495	3475	3338	3144	2959	2773	25483
Mathem	291	685	1222	1236	1281	1222	1196	1149	1099	9381
Physics	530	1046	1655	1740	1635	1590	1334	1293	1254	12077
Chemistry	1078	1444	2234	2204	2011	1849	1792	1762	1804	16178
Earth Sci	253	375	511	550	580	577	570	556	584	4556
Biol. Sci	1245	1963	3360	3633	3580	3636	3473	3498	3541	27929
Agric. Sci	414	576	803	900	855	853	830	904	908	7043
Psychology	772	954	1888	2116	2262	2444	2587	2749	2822	18594
Sociology	162	239	504	583	638	599	645	680	687	4737
Economy	341	538	826	791	863	907	833	867	879	6845
Anthropology	69	82	217	240	260	324	381	385	394	2352
other soc sci	314	502	1079	1392	1500	1609	1531	1550	1616	11093
Σ	6263	10477	17731	18880	18940	18948	18316	18352	18361	146268

Man kann sich einen ersten Überblick verschaffen, indem man die Randhäufigkeiten betrachtet, vergl. Abbildung 4. Die eingezeichneten Kurve in der linken Abbildung ist eine sogenannte Pareto-Kurven, die hier nicht weiter diskutiert werden muß, da das Modell, auf dem die Kurve beruht, offensichtlich sowieso nicht passt. Für die zeitliche Entwicklung der Häufigkeiten gibt es für die meisten Fächer ein Maximum zwischen den Jahren 1970 bis 1972, bis auf die Fächer Psychologie, Ökonomie und vielleicht noch die Biologie. Die Korrespondenzanalyse repräsentiert die Fächer

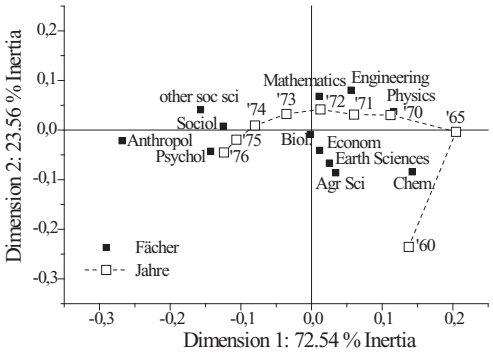
Abbildung 4: Verteilung der Häufigkeiten der Promotionsfächer



einerseits und die Jahreszahlen andererseits simultan in einem Koordinatensystem, das von einander unabhängige latente Variablen repräsentiert, über deren Bedeutung man sich hermeneutische)!) Gedanken machen kann. Eine solche Darstellung

heißt Biplot, eben weil zwei Variablensätze (Fächer und Jahreszahlen) simultan dargestellt werden. Das Resultat ist überraschend: die Entwicklung der Promotionen zeigt einen klaren Trend: Die erste latente Variable (Dimension I) erklärt 72% der

Abbildung 5: Biplot Doktorate - Jahre



Abhängigkeiten in den Daten, die zweite (Dimension II) ungefähr 24%. Die erste Dimnnsion reflektiert Prozesse, die den Übergang von 1965 bis 1976 reflektieren. Dieser ist durch eine Verschiebung von den naturwisswenschaftlichen Fächern hin zu den "weichen" sozialwissenschaftlichen Fächern. Die zweite Dimension spannt den Bogen zwischen Chemie, Geologie etc zu den eher mathematischen Fächern Mathematics, Phyics und Engineering. Weitgergehende Interpretationen wird man finden, indem man gesellschaftlch-politsche Entwicklungen in den USA in die Betrachtungen einbezieht.

Eine andere, einigermaßen unübersichtliche Tabelle (Tabelle 14) stellt Daten über Selbstmorde in Westdeutschland in den Jahren 1974 bis 1977 zusammen. Die Inspektion der Tabelle zeigt, dass es klare Unterschiede in der Wahl der Methoden gibt, – aber wie hängen sie mit dem Geschlecht einerseits und dem Alter andererseits zusammen? Die Korrespondenzanalyse kann zunächst einmal nur 2-dimensionale Tabellen analysieren, aber man kann aus 3-dimensionalen Tabellen 2-dimensionale basteln, indem man z.B. die Teiltabellen für männliche und weibliche Selbstmörder nebeneinander schreibt (es gibt andere Kombinationen, aber diese liefert die klarste Struktur. Zunächst inspiziere man die Häufigkeitsverteilung der Selbstmorde, getrennt nach Geschlechtern. Ofenbar ist die Neigung, Selbstmord zu begehen, bei Männern stärker ausgeprägt als bei Frauen, aber abgesehen davon ist der Verlauf, d.h. die Häufigkeiten als Funktion des Alters, bei beiden Geschlechtern im Altersabschnitt 15 – 60 Jahre unterschiedlich: während die Häufigkeiten bei den Frauen einigermaßen monoton ansteigen, haben die Männer eine Maximum bei ca 40 Jahren. In diesem Alter fallen Entscheidungen über Karrieren – man scheitert oder man scheitert nicht, gleichzeitig löst sich die Ehe auf oder auch nicht. Interessant ist der nahezu parallele Verlauf im Altersabschnitt 60 bis 90 Jahre, mit einem Maximum im Alter von 70 Jahren.

Abbildung 6: Häufigkeitsverteilung der Selbstmorde, aggregiert über Methoden

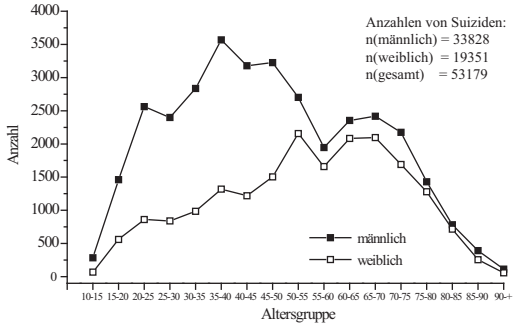
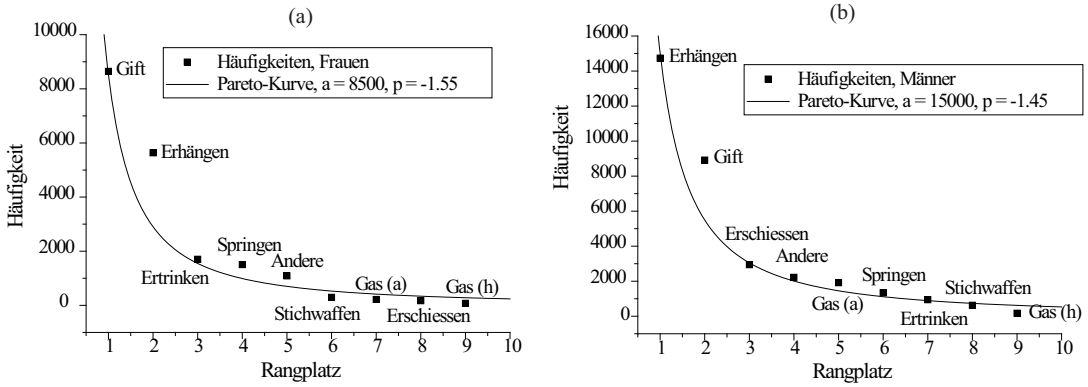


Abbildung 7: Ranggeordnete Häufigkeit der Methoden, (a) Frauen, (b) Männer.Zur Definition der Pareto-Kurve siehe Text.



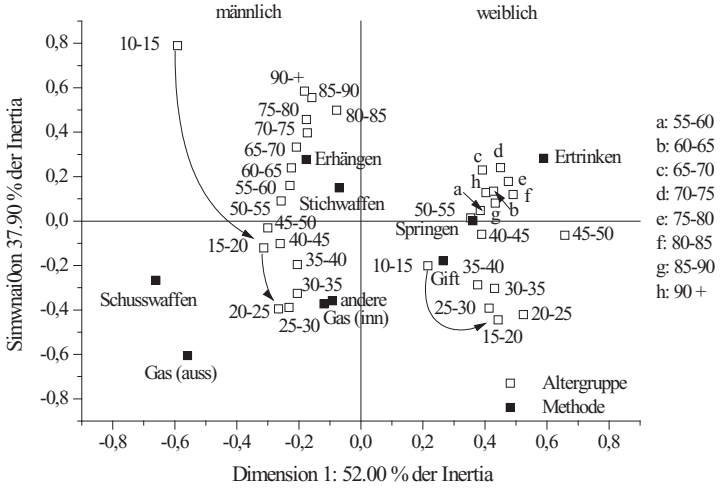
Darüber hinaus kann man die Häufigkeitsverteilungen für die Wahl der Methoden betrachten (Abbildung 7). Offenbar haben Frauen und Männer hier unterschiedliche Präferenzen. Aber die Interaktion zwischen den Kategorien Methode, Geschlecht und Alter wird erst deutlich, wenn man den Biplot in Abbildung 8 betrachtet. Offenbar unterscheiden sich Männer und Frauen deutlich hinsichtlich der Wahl der Methoden in Abhängigkeit vom Alter. Die erste Dimension erklärt ca 52% der Abhängigkeiten in der 3-dimensionalen Tabelle durch die Unterschiede zwichen den Geschlechtern, während die zweite Dimension 38% der Abhängigkeiten erklärt, die durch eine deutliche Ordnung der Altersgruppen bei den Männern entstehen, die im Übrigen mit einer dazu korrespondierenden Wahl der Waffen einhergeht: in jungen Jahren präferiert man Schusswaffen oder bringt sich durch Autoabgase oder Mißbrauch des häuslichen Gasherds um, in den ”besten Jahren” zwischen 50 und 60 neigt man eher

Tabelle 14: Selbstmorde in Westdeutschland 1974-1977

Alter/männl	Materie	Gas (h)	Gas (a)	Hängen	Ertrinken	Schußw.	Stichw.	Springen	Andere
10-15	4	0	0	247	1	17	1	6	9
15-20	348	7	67	578	22	179	11	74	175
20-25	808	32	229	699	44	316	35	109	289
25-30	789	26	243	648	52	268	38	109	226
30-35	916	17	257	825	74	291	52	123	281
35-40	1118	27	313	1278	87	293	49	134	268
40-45	926	13	250	1273	89	299	53	78	198
45-50	855	9	203	1381	71	347	68	103	190
50-55	684	14	136	1282	87	229	62	63	146
55-60	502	6	77	972	49	151	46	66	77
60-65	516	5	74	1249	83	162	52	92	122
65-70	513	8	31	1360	75	164	56	115	95
70-75	425	5	21	1268	90	121	44	119	82
75-80	266	4	9	866	63	78	30	79	34
80-85	159	2	2	479	39	18	18	46	19
85-90	70	1	0	259	16	10	9	18	10
90+	18	0	1	76	4	2	4	6	2
Alter/weibl	Materie	Gas (h)	Gas (a)	Hängen	Ertrinken	Schußw.	Stichw.	Springen	Andere
10-15w	28	0	3	20	0	1	0	10	6
15-20w	353	2	11	81	6	15	2	43	47
20-25w	540	4	20	111	24	9	9	78	67
25-30w	454	6	27	125	33	26	7	86	75
30-35w	530	2	29	178	42	14	20	92	78
35-40w	688	5	44	272	64	24	14	98	110
40-45w	566	4	2	343	76	18	22	103	86
45-50w	716	6	24	447	94	13	21	95	88
50-55w	942	7	26	691	184	21	27	129	131
55-60w	723	3	14	527	163	14	30	92	92
60-65w	820	8	8	702	245	11	35	140	114
65-70w	740	8	4	785	271	4	38	156	90
70-75w	624	6	4	610	244	1	27	129	46
75-80w	495	8	1	420	161	2	29	129	35
80-85w	292	3	2	223	78	0	10	84	23
85-90w	113	4	0	83	14	0	6	34	2
90+w	24	1	0	19	4	0	2	7	0

zu Stichwaffen, und die älteren Jahrgänge haben eine Neigung zum Strang. 'Bei den Frauen ist das Bild nicht so klar, aber es wird doch deutlich, dass man in jüngeren Jahren eher zum Gift einschließlich Schlaftabletten greift, in den mittleren dann eher den Sprung in den Abgrund vorzieht und im höheren Alter den Schritt ins Wasser als geeigntete Methode ansieht. Natürlich liefert der Biplot keine Begründung für diese Präferenzen, denn die Tabelle enthält ja nur die Kategorien Geschelcht, Alter und Methode und keine weiteren Kategorien, die auf kausale Zusammenhänge

Abbildung 8: Biplot: Selbstmorde: Methode, Altersgruppen und Geschlecht



zwischen diesen Kategorien hinweisen könnten. Andererseits werden systematische Unterschiede auf einen Blick deutlich, sie legen eventuell nahe, welche weiteren Informationen man für die Gründe dieses menschlichen Unglücks heranziehen muß.

3.5.2 Logistische Modelle

Oft steht man vor der Frage. die Wahrscheinlichkeit bestimmter Ereignisse auf der Basis von Prädiktorvariablen abzuschätzen bzw. vorherzusagen. Ein typisches Beispiel ist die Vorhersage der möglichen Rückfälligkeit (Sucht, kriminelles Verhalten, etc) unter gegebenen Randbedingungen, die Gefahr von Infektionen bei Operationen, etc. Der allgemeine Ansatz besteht darin, vom Begriff der bedingten Wahrscheinlichkeit auszugehen: es sei H die Hypothese, dass das in Frage stehende Ereignis eintritt, und $x = (x_1, x_2, \ldots, x_n)$ repräsentiere einen Satz von Prädiktorvariablen. Man hat dann

$$P(H|x) = P(x|H)\frac{P(H)}{P(x)}. \tag{3.41}$$

$P(x)$ kann üüber den Satz der Totalen Wahrscheinlichkeit berechnet werden:

$$P(x) = P(x|H)P(H) + p(x|\neg H)P(\neg H) \tag{3.42}$$

wobei $\neg H$ für "das Ereignis tritt nicht ein" steht. Aus (3.41) folgt dann

$$P(H|x) = \frac{P(x|H)P(H)}{P(x|H)P(H) + p(x|\neg H)P(\neg H)}. \tag{3.43}$$

Dividiert man auf der rechten Seite Zähler und Nenner durch $P(x|H)P(H)$, so ergibt sich

$$P(H|x) = \frac{1}{1 + \frac{P(x|\neg H)P(\neg H)}{P(x|H)P(H)}} \tag{3.44}$$

Nun sei

$$Q = \frac{P(x|\neg H)P(\neg H)}{P(x|H)P(H)}$$

Dann ist

$$Q = e^{\log Q},$$

wobei $\log = \log_e$ der *natürliche Logarithmus* ist, also der zur Basis e (dies ist die Eulersche Zahl $e = 2.71828\ldots$ Diese Beziehung ist zunächst einmal völlig trivial, weil für $y = e^x$ der Logarithmus einfach die Umkehrfunktion ist, also $\log y = x$. Andererseits ist

$$\log Q = \log P(x|\neg H) - \log P(x|H) + \log(P(\neg H)/P(H)) \tag{3.45}$$

$P(H)$ hängt nicht von x ab, so dass $\log(P(\neg H)/P(H)) = k$ eine Kontante gesetzt werden kann. Für den Moment sei $x = x_1$, und x sei normalverteilt, unter H mit dem Erwartungswert μ und unter $H_0 = \neg H$ mit dem Erwartungswert μ_0. Dann hat man

$$P(x|H_0) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu_0)^2/2\sigma^2}, \quad P(x|H) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma}$$

wobei zur Vereinfachung sowohl für H_0 als auch für H dieselbe Varianz σ^2 angenommen wurde. Dann ist aber

$$\log P(x|H_0) = \frac{(x-\mu_o)^2}{2\sigma^2} + c, \quad \log P(x|H) = \frac{(x-\mu)^2}{2\sigma^2} + c,$$

Multipliziert man die quadratischen Ausdrücke aus und berücksichtigt (3.45), so sieht man nach ein wenig Rechnerei die Gleichung (3.44) in

$$P(H|x) = \frac{1}{1 + e^{-(ax+b)}} \tag{3.46}$$

übergeht (in die Parameter A und B sind die ebenfalls unbekannten Parameter μ_0, μ und σ "absorbiert" worden). Dies ist die *logistische Funktion*[27] Die Betrachtung kann für den Fall $x = (x_1, \ldots, x_n)$ verallgemeinert werden, und man erhält

$$P(H|x) = \frac{1}{1 + e^{-(b_0 + b_1 x_1 + \cdots b_n x_n x)}} \tag{3.47}$$

[27]Der Name 'logistische Funktion' leitet sich vom französischen Wort logis für Wohung ab; der belgische Mathematiker Pierre-François Verhulst (1804 – 1849) hat diese Funktion in Zusammenhang mit einem Modell für den Bedarf an Wohnungen in Paris abgeleitet, allerdings auf eine ganz andere Weise.

Der Exponent von e hat Form, die auch in der multiplen Regresson vorkommt. Man kann die Gleichung nach $x = b_0 + b_1x_1 + \cdots b_nx_nx$ auflösen und bekommt dann aus $p = 1/(1 + e^{-x})$ den Ausdruck $e^{-x} = p/(1 - p)$, d.h.

$$\log \frac{P(H|x)}{1 - P(H|x)} = b_0 + b_1x_1 + \cdots + b_nx_nx \tag{3.48}$$

Die linke Seite heißt *Logit-Funktion* und ist nichts anderes als der Logaritmus der Wettchance für das Ereignis H.

(3.48) impliziert, dass für die Wettchance gilt

$$\frac{P(H|x)}{1 - P(H|x)} = e^{b_0+b_1x_1+\cdots+b_nx_n} \tag{3.49}$$

$$= e^{b_0}e^{b_1x_1} \cdots e^{b_nx_n} \tag{3.50}$$

Die Gleichung (3.50) erleichtert die Interpretation der Ergebnisse, verg. Beispiel 3.1. Im diesem Beispiel wird von der Vektorschreibweise Gebrauch gemacht; ein *Vektor* ist einfach eine Anordnung von $p > 1$ Zahlen $x_1, \ldots, x_p$; man schreibt dafür $\vec{x} = (x_1, \ldots, x_p)$. Die $b_0, b_1, \ldots, b_n$ sind "freie" Parameter, d.h. ihre Werte sind unbekannt und es gibt keine weiteren Einschränkungen, was die Werte angeht. Sie müssen aus den Daten geschätzt werden. Die dazu verwendete Methode ist *nicht* die Methode der Kleinsten Quadrate (es gibt keinen Fehlerterm in (3.49)), sondern die Maximum-Likelihood-Methode, auf die hier nicht weiter eingegangen werden muß.

Beispiel 3.1 Infektionsrisiko bei Kaiserschnittgeburten Es soll das Risiko einer Infektion bei einer Geburt mit Kaiserschnitt berechnet werden. Der Kaiserschnitt kann geplant oder nicht geplant durchgeführt werden, und die Patientin kann Risikofaktoren haben oder nicht. Man hatte die folgenden Daten zur Verfügung (gepl = geplant, RF = Risikofaktor):

Tabelle 15: Risiko einer Infektion beim Kaiserschnitt (KS); – Kaiserschnitt geplant, - nicht geplant

		KS geplant		KS nicht geplant	
		Infektion		Infektion	
Antibiotika	Risikofaktor (RF)	ja	nein	ja	nein
gegeben	vorhanden	1	17	11	87
	nicht vorhanden	0	2	0	0
nicht gegeben	vorhanden	28	30	23	3
	nicht vorhanden	8	32	0	9

$$x_1 = \begin{cases} 1 & \text{nicht gepl} \\ 0 & \text{gepl} \end{cases}, \; x_2 = \begin{cases} 1 & \text{RF} \\ 0 & \text{kein RF} \end{cases}, \; x_3 = \begin{cases} 1 & \text{AB} \\ 0 & \text{kein AB} \end{cases} \tag{3.51}$$

Das Modell (3.52) ist das *Haupteffektmodell*, da keinerlei Interaktionen zwischen den drei unabhängigen Variablen angenommen werden.

$$\log \frac{p(\text{Infektion}|\vec{x})}{p(\text{keine Infektion}|\vec{x})} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3. \tag{3.52}$$

Daraus folgt

$$\frac{p(\text{Infektion}|\vec{x})}{p(\text{keine Infektion}|\vec{x})} = \exp(\beta_0)\exp(\beta_1 x_1)\exp(\beta_2 x_2)\exp(\beta_3 x_3) \tag{3.53}$$

Für einen speziellen Vektor $\vec{x}_i$ wird die abgekürzte Schreibweise

$$\rho_i = \frac{p(\text{Infektion}|\vec{x}_i)}{p(\text{keine Infektion}|\vec{x}_i)} = \frac{p(\text{I}|\vec{x}_i)}{p(\neg\text{I}|\vec{x}_i)} \tag{3.54}$$

eingeführt. Die Daten werden in der Tabelle 15 gezeigt, und die geschätzten Parameterwerte findet man in Tabelle 16. Ein nicht geplanter Kaiserschnitt erhöht

Tabelle 16: Parameterwerte

	Gewicht/Prädiktor			
	1	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
Wert	-1.89	1.07	2.03	-3.25
$\sqrt{\text{Var}(\hat{\beta})}$	.41	.43	.46	.48
t	-4.61	2.49	4.41	-6.77

nach (3.53) das Infektionsrisiko um den Faktor $\exp(\beta_1 x_1) = \exp(1.07) = 2.92$, ein vorhandener Risikofaktor erhöht das Infektionsrisiko um den Faktor $\exp(\beta_2 x_2) = \exp(2.03) = 7.6$, und ein Antibiotikum erniedrigt das Risiko um den Faktor $\exp(\beta_3 x_3) = \exp(-3.25) = .0388$, d.h. hat man

$$\frac{p(\text{Infektion})}{p(\text{keine Infektion})} = 1$$

im Falle keiner Antibiotikagabe, so wird das Risko auf

$$\frac{p(\text{Infektion})}{p(\text{keine Infektion})} = .0388$$

bei Antibiotikagabe gesenkt. Es sind weitere Abschätzungen für verschiedene Vektoren $\vec{x}_i = (x_1, x_2, x_3)'$ möglich, vergl. Tabelle 17. Die x_1, x_2, x_3 nehmen Werte an, die durch den i-ten Fall angezeigt sind. Man kann jetzt noch Vergleiche von ρ_i mit ρ_j berechnen. So kann man die Wirkung von Antibiotika für den Fall bestimmen, dass der Kaiserschnitt nicht geplant war und Risikofaktoren vorliegen. Man findet

$$\frac{\rho_1}{\rho_4} = \frac{1.300}{33.509} = .0388. \tag{3.55}$$

Tabelle 17: Mögliche Fälle; $\rho_i = p(I|\vec{x}_i)/p(\neg I|\vec{x}_i)$, $1 \leq i \leq 8$

Fall	x_1		x_2		x_3		ρ_i
1	1	npl	1	RF	1	AB	1.300
2	0	pl	1	RF	1	AB	.445
3	1	npl	0	kRF	1	AB	.171
4	1	npl	1	RF	0	kAB	33.509
5	1	npl	0	kRF	0	kAB	.445
6	0	pl	1	RF	0	kAB	11.476
7	0	pl	0	kRF	1	AB	.056
8	0	pl	0	kRF	0	kAB	1.510

Bei einem nicht geplanten Kaiserschnitt im Fall von Risikofaktoren erhöht sich das Risiko einer Infektion fast um das 26-fache, wenn *keine* Antibiotika gegeben werden, bzw. reduziert sich auf das .04-fache, wenn Antibiotika gegeben werden.

Die bisherigen Betrachtungen illustrieren die Art und Weise, in der Parameter für das gegebene Modell interpretiert werden. Es ist natürlich möglich, zu testen, ob das Modell überhaupt mit den Daten verträglich ist. Hier soll nur angemerkt werden, dass die *Deviance* den Wert 10.997 hat, was bedeutet, dass mit Bezug auf $\alpha = .05$ das Modell *extrem schlecht passt*! Die in Tabelle 16 angegebenen t-Werte sind hochsignifikant, aber das besagt ja noch nicht, dass das Modell insgesamt mit den Daten kompatibel ist. Man kann nun eine Erweiterung des Modells testen. Diese Erweiterung ist durch

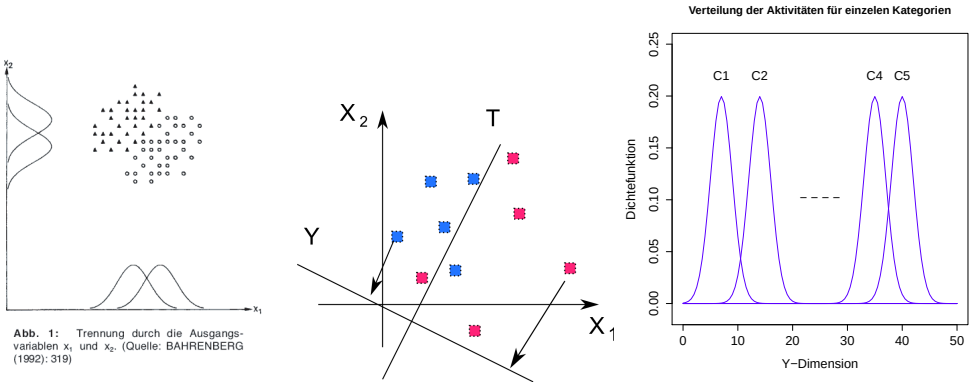
$$\log \frac{p(\text{Infektion}|\vec{x})}{p(\text{keine Infektion}|\vec{x})} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 \tag{3.56}$$

gegeben. Es wird also noch eine Wechselwirkung zwischen der Planung des Kaiserschnitts und dem Risikofaktor angenommen. Es zeigt sich aber, dass die Parameterschätzungen für dieses Modell eine starke Verzerrung haben, da die Tabelle 15 in den für die Schätzung wichtigen Zellen Nullen enthält, so dass nicht weiter auf dieses Modell eingegangen werden kann. □

3.6 Klassifikationen

Objekte, Personen, psychische Zustände etc müssen bzw sollen für viele Zwecke klassifiziert werden (Typ der Erkrankung, Eignung für Studienfächer, ...). Dazu seien die Prädiktorvariablen $X_1, \ldots, X_p$ gegeben, aus denen für das i-te Objekt (Person, Zustand, etc) die Klassen- oder Kategorienzugehörigkeit C_j, $j = 1, \ldots, K$ berechnet werden soll, wobei K die Anzahl der möglichen Kategorien ist. Ronald A. Fisher hat 1937 ein Verfahren gefunden, Klassifikationen durchzuführen, das in vielen prakti-

Abbildung 9: Fishersche Lineare Diskriminanzanalyse



schen Situationen sehr gute Dienste leistet. Der Ansatz besteht darin, die Prädiktoren *linear zu kombinieren*:

$$Y = u_1 X_1 + u_2 X_2 + \cdots + u_p X_p. \tag{3.57}$$

Y ist ein Wert auf einer Skala, die (i) Fälle, die zu verschiedenen Kategorien gehören, maximal trennt und (ii) Fälle, die zu einer Kategorie gehören, möglichst eng zusammenfasst. Damit soll sichergestellt werden, dass die Verteilungen der Y-Werte für die verschiedenen Kategorien sich so wenig wie möglich überlappen und damit die Wahrscheinlichkeit einer Fehlklassifikation so klein wie möglich wird. Die Aufgabe besteht darin, die Gewichte $u_1, \ldots, u_p$ zu finden, die eine Klassifikation in diesem Sinne ermöglichen. Abbildung 9 zeigt das Prinzip des Fisherschen Ansatzes für den Fall von nur zwei Prädiktorvariablen X_1 und X_2. Die linke Abbildung zeigt eine Konfiguration von Punkten in den zwei Koordinatenachsen X_1 und X_2, die Fälle aus zwei Kategorien enthalten (Dreiecke und Quadrate. Die Dicheverteilungen auch den beiden Achsen zeigen relativ große Überlappungen, d.h. in Bezug auf diese beiden Variablen sind die beiden Gruppen nicht gut auseinander zu halten. Die mittlere Abbildung zeigt das Fishersche Prinzip: man finde eine Achse Y in diesem 2-dimensionalen Raum derart, dass die Projektionen der Punkte auf diese Achse für die beiden Kategorien maximal getrennt sind. Senkrecht auf der Y-Achse steht eine andere Gerade T, die Trennlinie oder Trennebene (eine Gerade ist eine 1-dimensionale "Ebene"). Die rechte Abbildung zeigt die Verteilungen dieser Projektionen auf die Y-Gerade für fünf Kategorien $C_1, \ldots, C_5$. Wenn sich diese Verteilungen gar nicht übnerlappen, kann fehlerfrei kategorisiert werden. Im Normalfall ergeben sich Überlappungen, so dass es zu Fehlkategorisierungen kommt. Y wird allerdings so gewählt, dass die Überlappungen minimal sind.

Man muß nun spezifizieren, was "minimal" bedeuten soll. Wenn sich schon die Punktekonfigurationen im (X_1, X_2)-Raum nicht überlappen, gelingt ene überlappungsfreie Repräsentation auf einer Y-Dimension, andernfalls nicht. Man kann für

jede der zu einer Kaegorie korrespondierenden Punktekonfiguration einen Schwerpunkt berechnen; dies ist die mittlere Position der Fälle einer Kategorie. Die Fälle innerhalb einr Kategorie variieren unm diesen Schwerpunkt – dies ist die Variation innerhalb der Kategorie. Die Schwerpunkte sind im Allgemeinen verschieden voneinander und variieren ebenfalls, – dies ist die Variation zwischen den Kategorien. Analog dazu kann man in Bezug auf die Y-Dimension von einer Varianz "innerhalb" der Kategorien und einer Varianz "zwischen" den Kategorien sprechen. Diese Größen sind wie in der Varianzanalyse definiert.

Die tatsächlichen Überlappungen hängen Vom Verhältnis der Varianz zwischen den Gruppen und der Varianz innerhalb der Gruppen ab. Wie bei der ANOVA kann man dazu die Zerlegung der Gesamtvarianz bzw. der Quadratsumme, die die Gesamtvarianz definiert, in eine Quadratsumme zwischen und eine Qudratsumme innerhalb der Gruppen zerlegen:

$$QS_{ges} = \sum_{k=1}^{K}\sum_{i=1}^{n_k}(Y_{ik} - \bar{y})^2 = \sum_{i=1}^{n_k}\sum_{j=1}^{K}((Y_{ij} - \bar{y}_j) - (\bar{y}_j - \bar{y}))^2,$$

und man erhält

$$QS_{ges} = \sum_{i=1}^{n_k}\sum_{j=1}^{K}(Y_{ij} - \bar{y}_j)^2 + \sum_{k=1}^{K} n_k(\bar{y}_k - \bar{y})^2 = QS_{inn} + QSzwisch. \tag{3.58}$$

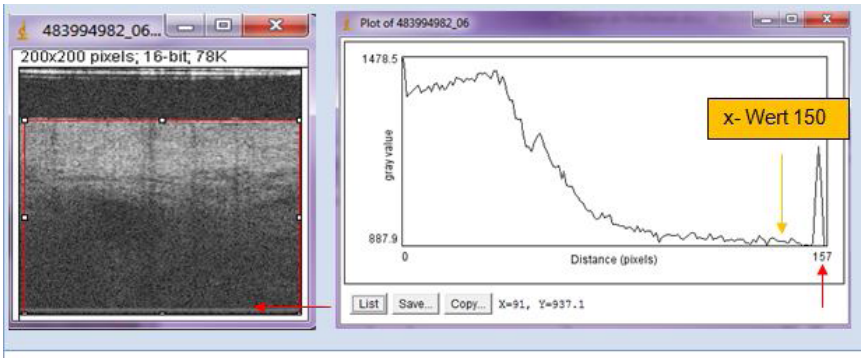
Das Fisher-Kriterium für die gesuchte Dimension Y dann

$$\lambda(u_1, \ldots, u_p) = \frac{QS_{zwisch}}{QS_{inn}} \stackrel{!}{=} \max \tag{3.59}$$

Die Y-Werte hängen aber wegen (3.57) von den X_j und damit auch von den u_j ab, so dass der Quotient λ eine Funktion der u_j ist. Die u_j werden so gewählt, dass der Quotient maximal ist. Die Details dieser Rechnung müssen hier übergangen werden, da sie ohne die Vektor- und Matrixrechnung nicht gut dargestellt werden können. Es zeigt sich aber, dass es im Allgemeinen mindestens eine Lösung $\mathbf{u} = (u_1, \ldots, u_p)$ gibt, für die die y-Werte gemäß (3.57) berechnet werden können; es ist durchaus möglich, dass sich mehr als ein Satz von u_j-Werten ergibt. Die sich ergebenden X-Skalen sind statistisch unabhängig und werden als zueinander orthogonale Koordinaten für die Fälle benützt. (Tatsächlich ist es so, dass die Bestimmung der neuen Y-Koordinaten aus mathematischer Sicht eine Rotation der ursprünglichen Koordinaten entspricht, die das Kriterium (3.59) erfüllt, – in diesem Sinne sind die Y-Koordinaten latente Variablen), wie sie schon in (3.14) und (3.15) explizit betrachtet wurden, wenn auch deren Berechnung ein anderes Kriterium unterliegt.

Die Diskriminanzanalyse kann an einem Beispiel aus der Medizin illustriert werden. Mit der Optical Coherence Tomography (OTC) können Bilder von Gewebeteilen gemacht werden, ohne dass diese Teile operativ entfernt werden müssen; damit können im Prinzip Biopsien umgangen werden, da die Bilder zu histologischen Untersuchungen verwendet werden können. Die Frage ist, ob es möglich ist, anhand

Abbildung 10: Gewebeprobe und OCT-Profil



der Bilder korrekte Kategorisierungen bezüglich möglicher Erkrankungen vorzunehmen. Insbesondere sollte geprüft werden, ob nicht die gesamte Information über die Kategorienzugehörigkeit im Helligkeitsverlauf vom äußeren Epithel bis zu einem gewissen Punkt innerhalb des Gewebes enthalten ist. Dieser Verlauf kann durch ein Helligkeitsprofil repräsentiert werden (s. Abbildung 10). Pro Kategorie wurden über hundert derartige Profile bestimmt. Die Länge eines Profils kann durch die Anzahl der Bildpunkte (Pixel) eines OCT-Bildes bestimmt werden, und diese Pixel können als Prädiktorvariablen für eine Kategorisierung verwendet werden. Abbildung 11 zeigt die mittleren Profile sowie die Standardabweichungen. Offenbar unterscheiden sich die Profile insbesondere in den ersten fünzig Pixeln; in diesem Bereich sind auch die Standardabweichungen am größten. Die Differenzierungs- und Kategorisierungsleistungen müssen in diesem Bereich erfolgen. Abbildung 12 zeigt die Ergebnisse der Diskriminanzanalyse. Es gibt drei Y-Skalen, die einen 3-dimensionalen Raum definieren, in dem die einzelnen Profile als Punkte erscheinen (Abbildung oben links). Wie man sieht, erscheinen die Klassen als gut separierte Punktwolken. Die übrigen Abbildungen zeigen die Konfigurationen der Punkte als Projektionen auf die drei Ebenen, die jeweil durch ein Paar von Koordinatenachsen (Y-Skalen) definiert werden: I und II, I und III und schließlich II und III. Die Analyse zeigt, dass die für die Kategorisierung wichtige Information in einem 3-dimensionalen Teilraum des 150-dimensionalen Prädiktorraums komprimiert ist! Der Ansatz (3.57) erinnert an die multiple Regression, der Unterschied ist allerdings, dass Y keine gemessene, sondern eine latente Variable ist. Aber wie schon bei der multiplen Regression können auch bei der Diskriminanzanalyse Korrelationen zwischen de Prädiktoren zu einem Problem werden. Sind diese Korrelationen zu hoch, kann man oft für die gegebene Stichprobe eine gute Trennung von Gruppen erreichen, die aber zusammenbricht, sobald neue Fälle anhand der Gewichte $u_1, \ldots, u_p$ kategorisiert werden sollen, die anhand der alten Stichprobe geschätzt wurden. Darüber hinaus ist die Diskriminanzanalyse auf *lineare* Trennungen zwischen den Gruppen beschränkt; dies bedeutet, dass die Fälle (Objekte, Personen, etc), die zu einer Kategorie korrespondieren,

Abbildung 11: Mittlere Profile mit Standardabweichungen

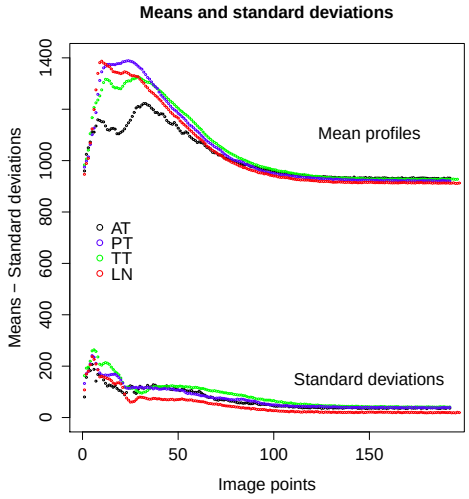
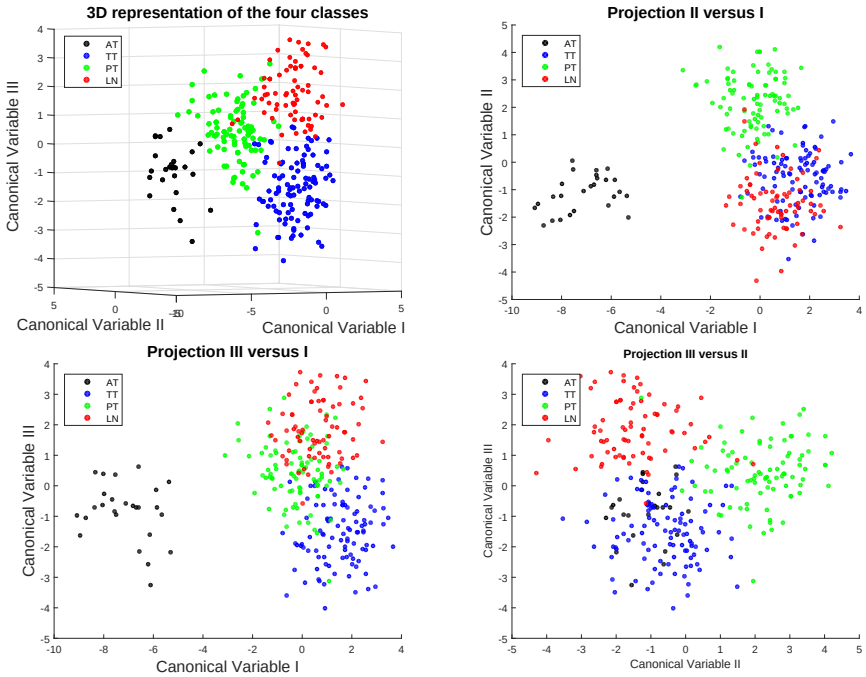


Abbildung 12: Ergebnisse der Diskriminanzanalyse



stets *konvexe* Punktmengen bilden müssen, d.h. für irgendzwei Fälle in einer solchen Menge müssen die Fälle, die auf einer Verbindungsgeraden zwischen diesen Fällen

liegen, ebenfalls zu der Kategorie gehören. Es gibt Verfahren, die auch im Falle nicht konvexer Punktmengen bzw. im Fall korrelierender Prädiktoren angewendet werden können (sogenannte regularisierte Diskriminanzanalysen und Support Vector Machines), die aber kaum ohne Rückgriff auf die Vektor- und Matrixalgebra dargestellt werden können.

Literatur

[1] Adorno, Th. W., Brunswik, E.F., Levinson, D. J., Sanford, N. R.: The Authoritarian Personality. Harper und Brothers, New York 1950.

[2] Bridgman, P. W.: The Logic of Modern Physics. MacMillian, New York 1927

[3] Breuer, F.: Qualitative Psychologie. Grundlagen, Methoden und Anwendugnen eines Forschungsstils. Opladen 1996

[4] Dawes, R.M.: Everyday Irrationality. How Pseudo-Scientists, Lunatics, and the rest of us systematically fail to think ratinoally. Westview Press, Boulder (Colorado) 2001

[5] Dollard, J., Miller, N.E., Doob, L. W. , Mowrer, O. H., Sears, R. R.: Frustration and aggression. New Haven, CT, US: Yale University Press (1939).

[6] Ebbinghaus, H.: Über das Gedächtnis. Untersuchungen zur experimentellen Psychologie. Leipzig 1885.

[7] Hussy, W., Schreier, M., Echterhoff, G.: Forschungsmethoden in Psychologie und Sozialwissenschaften für Bachelor. Berlin Heidelberg 2010

[8] Luhmann, M.: R für Einsteiger – Einführung in die Statistiksoftware für dieSozialwissenschaften. Weinheim 2011

[9] Mead, M.: Coming of age in Samoa (1928)

[10] Milgram,S. (1963) Behavioral Study of Obedience. *Journal of Abnormal and Social Psychology*, 67, 371–378

[11] Osgood, C.E., Suci, G., Tannenbaum, P.: The measurement of meaning. Urbana, IL: University of Illinois Press, 1957

[12] Nachtigall, C., Suhl, U., Steyr,R.: Einführung in die Konfundierungsanalyse. methvalreport 2(1), Jena 2000

[13] Nisbett, R.E., Wilson, T.D.: Telling mor than we can know – verbal reports on mental processes. *Psychological Revies*, 84(3), 231–259

[14] Nisbett, R.E., Ross, L.: Human inference – strategies and shortcomings of social judgment. Prentice-Hall,Engle Woode Cliffs 1980

[15] Popper, K.R.: Unended Quest. An intellectual Autobiography, London 1974; deutsch: Ausgangspunkte, Hamburg 1979

[16] Pinker, S.: How the mind works. New York 1997.

[17] Rachelet, M. (1981) Racial characteristics and imposition of the death penalty. Amer. Sociol. Review **46**, 918-927

[18] Sarris, V.: Methodologische Grundlagen der Psychologie, Band 1: Erkenntnisgewinnung und Methodik der experimentellen Psychologie. München Basel 1990

[19] Scheffé, H.: The Analysis of Variance, New York 1959

[20] Simpson, E.H. (1951) The Interpretation of Interaction in Contingency Tables. *Journal of the Royal Statistical Society. Series B.*, Vol. 13(2), 238–241

[21] Stegmuller, W.: Das Problem der Induktion: Humes Herausforderung und moderne Antworten. Wissenschaftliche Buchgesellschaft Darmstadt 1971

[22] Watzlawik, P.: Anleitung zum Unglücklichsein. München Zürich 2009

[23] Zimmer, D.E.: Tiefenschwindel – Die endlose und beendbare Psychoanalyse. Reinbeck bei Hamburg 1986

Index