# Vorbemerkungen **Empirische Forschung und Statistik**

Statistik ist ein wichtiger Bestandteil empirisch-wissenschaftlichen Arbeitens. Statistik beschränkt sich nicht nur auf die Zusammenfassung und Darstellung von Daten (dies ist Aufgabe der ***deskriptiven Statistik***, die im ersten Kapitel behandelt wird), sondern sie ermöglicht empirischen Wissenschaften objektive Entscheidungen über die Brauchbarkeit der überprüften Hypothesen. Dieser Teilaspekt der Statistik, der sich mit der Überprüfung von Hypothesen befasst, wird häufig als ***analytische Statistik*** oder ***Inferenz- (schließende) Statistik*** bezeichnet.

Wissenschaftliches Arbeiten zielt auf die Verdichtung von Einzelinformationen und Beobachtungen zu allgemein gültigen theoretischen Aussagen ab. Hierbei leitet die deskriptive Statistik zu einer übersichtlichen und anschaulichen Informationsaufbereitung an, und die Inferenzstatistik ermöglicht eine Überprüfung von Hypothesen an der beobachteten Realität.

Wenn beispielsweise das Sprachverhalten von Unterschichtkindern interessiert, könnten wir eine Schülerstichprobe beobachten und für verschiedene Sprachmerkmale Häufigkeitsverteilungen erstellen bzw. graphische Darstellungen anfertigen. Das erhobene Material wird in quantitativer Form so aufbereitet, dass man sich schnell einen Überblick über die in der untersuchten Stichprobe angetroffenen Merkmalsverteilungen verschaffen kann. Verallgemeinernde Interpretationen dieser deskriptiven statistischen Analyse, die über das erhobene Material hinausgehen, sind jedoch spekulativ.

Lassen sich theoretisch Erwartungen hinsichtlich der Häufigkeit des Auftretens bestimmter Sprachmerkmale begründen, wird eine allgemeingültige Hypothese formuliert, die sich nicht nur auf einige zufällig ausgewählte Kinder, sondern auf alle Kinder dieser Schicht bezieht. Die Tauglichkeit dieser Hypothese wird anhand der empirischen Daten getestet. Verfahren, die dies leisten und die verallgemeinerte, über die jeweils untersuchten Personen hinausgehende Interpretationen zulassen, bezeichnen wir als inferenzstatistische Verfahren.

**Die Inferenzstatistik ermöglicht im Unterschied zur deskriptiven Statistik die Überprüfung von Hypothesen.**

Hat man keine Theorie bzw. Erkennisse, die eine Hypothese begründen könnten, bezeichnen wir die Untersuchung als ein *Erkundungsexperiment,* das dazu dient, erste Hypothesen über einen bestimmten, noch nicht erforschten Gegenstand zu formulieren. Bevor diese Hypothesen akzeptiert und zu einer allgemeingültigen Theorie verdichtet werden können, bedarf es weiterer Untersuchungen, in denen mit inferenzstatistischen Verfahren die Gültigkeit der „erkundeten" Hypothesen gesichert wird.

Bereits an dieser Stelle sei nachdrücklich auf einen Missbrauch der Inferenzstatistik hingewiesen: das statistische Überprüfen einer Hypothese anhand derselben Daten, die die Formulierung der Hypothese veranlasst haben. Forschungsarbeiten, in denen dasselbe Material zur Formulierung und Überprüfung von Hypothesen herangezogen wird, sind unwissenschaftlich. Dies gilt selbstverständlich in verstärktem Maße für Arbeiten, in denen Hypothesen erst nach der statistischen Auswertung aufgestellt werden. Eine Forschungsarbeit, die ein gefundenes Untersuchungsergebnis im Nachhinein so darstellt, als sei dies die zu prüfende Hypothese gewesen, kann nur mehr oder weniger zufällige Ergebnisse bestätigen, die untereinander häufig widersprüchlich sind und sich deshalb eher hemmend als fördernd auf den Forschungsprozess auswirken.

Dies bedeutet natürlich nicht, dass Hypothesen grundsätzlich nur vor und niemals nach einer

empirischen Untersuchung formuliert werden dürfen. Falls in einer Untersuchung angesichts der erhobenen Daten neue Hypothesen aufgestellt werden, ist diese Untersuchung jedoch explizit als Erkundungsexperiment oder explorative Studie zu kennzeichnen. Diese Hypothesen sind dann Gegenstand weiterführender, Hypothesen prüfender Untersuchungen.

**Für den sinnvollen Einsatz der Inferenzstatistik ist es erforderlich, dass vor Untersuchungsbeginn eine theoretisch gut begründete Hypothese oder Fragestellung formuliert wurde.**

Der sinnvolle Einsatz statistischer Verfahren, der über die reine Deskription des Untersuchungsmaterials hinausgeht, setzt also gründliche, theoretisch-inhaltliche Vorarbeit voraus. So gesehen kann der Wert einer konkreten statistischen Analyse immer nur im Kontext einer vollständigen Untersuchungsanlage erkannt werden, für die theoretische Vorarbeit, Hypothesenformulierung und eine genaue Untersuchungsplanung essentiell sind.

### Phasen der empirischen Forschung

Wegen der engen Verknüpfung statistischer Methoden mit inhaltlichen und untersuchungsplanerischen Fragen soll vor der eigentlichen Behandlung statistischer Techniken deren Funktion im Kontext empirischer Untersuchungen genauer verortet werden. Bei dieser Gelegenheit sind auch einige Fachbegriffe einzuführen, die in der empirischen Forschung gebräuchlich sind.

Wir unterteilen den empirischen Forschungsprozess in sechs verschiedene Phasen (vgl. Abb. 1), die im Folgenden kurz beschrieben werden. Ausführlichere Hinweise zur Planung und Durchführung empirischer Untersuchungen sowie weiterführende Literatur zu diesem Thema findet man z. B. bei Bortz u. Döring (2002), Campbell u. Stanley (1963), Czienskowski (1996), Hager (1987), Hussy u. Jain (2002), Lüer (1987), Rogge (1995), Sarris (1990, 1992) und Selg et al. (1992). Wissenschaftstheoretische Aspekte empirischer Forschung werden z. B. bei Chalmers (1986), Schnell et al. (1999, Kap. 3) und Westermann (2000) erörtert. Für eine grundlegende Orientierung sei die Enzyklopädie über „Methodische Grundlagen der Psychologie“ von Herrmann u. Tack (1994) empfohlen.

#### Erkundungsphase

Zur Erkundungsphase zählen die Sichtung der für das Problem einschlägigen Literatur, Kontaktaufnahmen mit Personen, die am gleichen Problem arbeiten, erste Erkundungsuntersuchungen, Informationsgespräche mit Praktikern, die in ihrer Tätigkeit mit dem zu untersuchenden Problem häufig konfrontiert werden, und ähnliche, zur Problemkonkretisierung beitragende Tätigkeiten. Ziel dieser Erkundungsphase ist es, die eigene Fragestellung in einen theoretischen Rahmen einzuordnen bzw. den wissenschaftlichen Status der Untersuchung – Hypothesen prüfend oder Hypothesen erkundend – festzulegen. Manche Forschungsthemen knüpfen direkt an bewährte Theorien an, aus denen sich für ein Untersuchungsvorhaben gezielte Hypothesen ableiten lassen. Andere hingegen betreten wissenschaftliches Neuland und machen zunächst die Entwicklung eines theoretischen Ansatzes erforderlich. Systematisch erhobene und objektiv beschriebene empirische Fakten müssen in einen gemeinsamen widerspruchsfreien Sinnzusammenhang gestellt werden, der geeignet ist, die bekannten empirischen Fakten zu erklären bzw. zukünftige Entwicklungen oder Konsequenzen zu prognostizieren. (Ausführliche Informationen zur Bedeutung und Entwicklung von Theorien und weitere Literatur hierzu findet man bei Bortz u. Döring 2002, Kap. 6.)

Die Erkundungsphase ist – wie empirische Wissenschaft überhaupt – gekennzeichnet durch ein Wechselspiel zwischen Theorie und Empirie bzw. zwischen induktiver Verarbeitung einzelner Beobachtungen und Erfahrungen zu allgemeinen Vermutungen oder Erkenntnissen und deduktivem Überprüfen der gewonnenen Einsichten an der konkreten Realität.

Hält man die „vorwissenschaftliche“ Erkundungsphase für abgeschlossen, empfiehlt sich eine logische und begriffliche Überprüfung des theoretischen Ansatzes.

#### Theoretische Phase

Bevor man eine Hypothese empirisch überprüft, sollte man sich vergewissern, dass die Hypothese

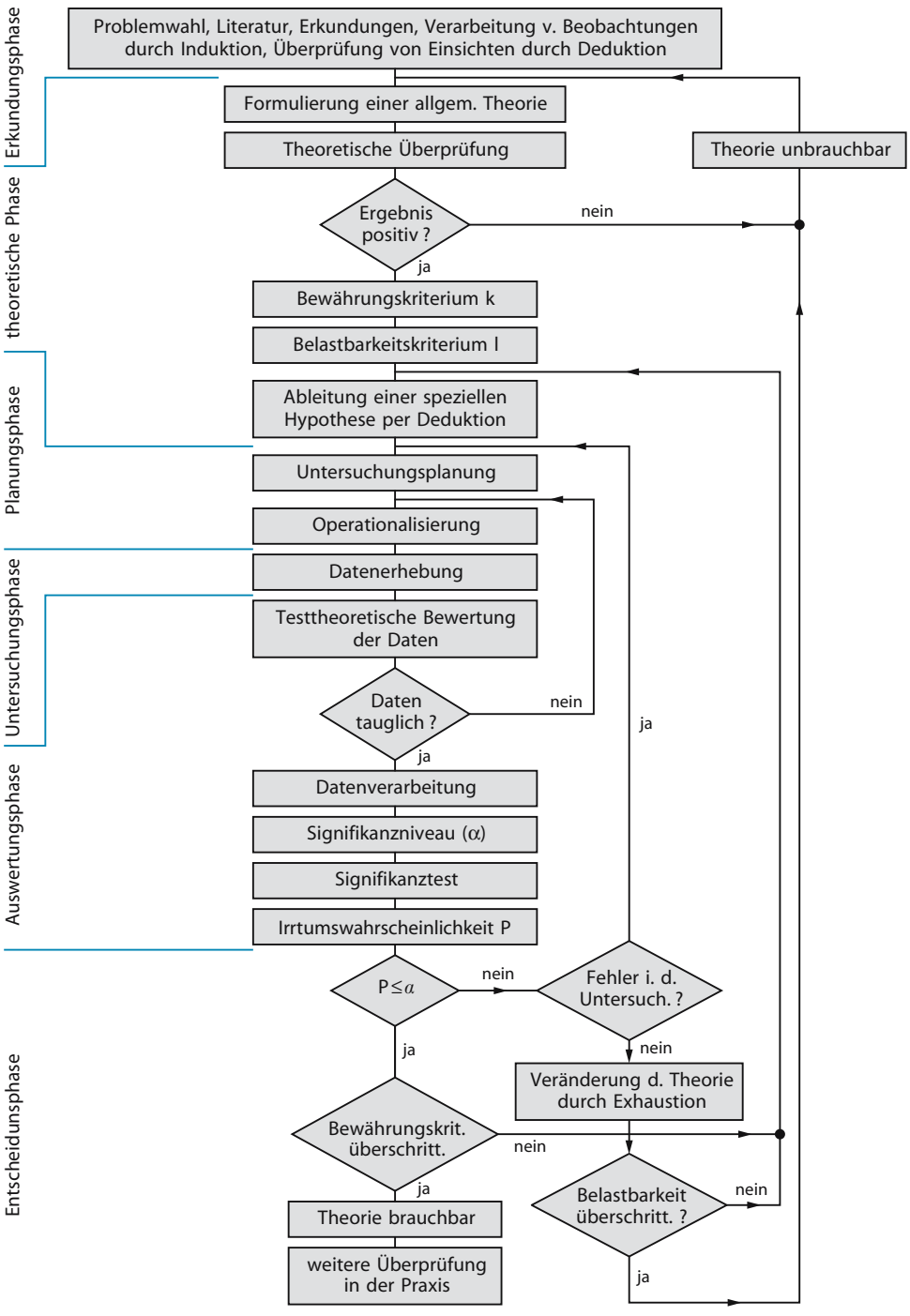


**Abb. 1.** Phasen der empirischen Forschung

bzw. die zu prüfende Theorie einigen formalen bzw. logischen Kriterien genügt. Diese Überprüfung setzt streng genommen voraus, dass die Theorie hinreichend entwickelt und formalisiert ist, um sie exakt nach logischen Kriterien analysieren zu können. Dies trifft auf die wenigsten human- und sozialwissenschaftlichen Theorien zu. Deshalb ist zu erwarten (und dies zeigt auch die derzeitige Forschungspraxis), dass gerade diese Phase in empirischen Untersuchungen eine vergleichsweise geringe Priorität besitzt. Die Prüfkriterien sind jedoch auch für weniger formalisierte Theorien von Bedeutung, denn sie tragen dazu bei, Schwächen des theoretischen Ansatzes bereits vor der empirischen Arbeit aufzudecken, die der empirischen Prüfbarkeit der Hypothesen entgegenstehen könnten.

In Anlehnung an Opp (1999) sollten in der theoretischen Phase folgende Fragen beantwortet werden:

- Ist die Theorie präzise formuliert?
- Welchen Informationsgehalt besitzt die Theorie?
- Ist die Theorie in sich logisch konsistent?
- Ist die Theorie mit anderen Theorien logisch vereinbar?
- Ist die Theorie empirisch überprüfbar?

**Präzision.** Eine Theorie ist wenig tauglich, wenn sie Begriffe enthält, die nicht eindeutig definiert sind. Die Definition der Begriffe sollte sicherstellen, dass diejenigen, die die (Fach-)Sprache beherrschen, mit dem Begriff zweifelsfrei kommunizieren können.

**Informationsgehalt.** Um den Informationsgehalt einer Theorie zu erkunden, werden die Aussagen der Theorie auf die logische Struktur eines „Wenn-dann"- bzw. eines „Je-desto"-Satzes *(Konditionalsätze)* zurückgeführt. (Wenn eine Theorie behauptet, frustrierte Menschen reagieren aggressiv, würde der entsprechende Konditionalsatz lauten: „Wenn Menschen frustriert sind, dann reagieren sie aggressiv.")

Eine Je-desto-Formulierung resultiert, wenn zwei *kontinuierliche Merkmale* miteinander in Beziehung gesetzt werden, wie z. B. in der Aussage: „Mit zunehmendem Alter sinkt die Sehtüchtigkeit des erwachsenen Menschen." Der Konditionalsatz hierzu lautet: „Je älter ein Erwachsener, desto schlechter ist seine Sehtüchtigkeit."

Der Informationsgehalt eines Wenn-dann-Satzes (entsprechendes gilt für Je-desto-Sätze) nimmt zu, je mehr Ereignisse denkbar sind, die mit der Aussage des Dann-Teiles im Widerspruch stehen. Ereignisse, die mit dem Dann-Teil der Aussage nicht vereinbar sind, werden als potenzielle ***Falsifikatoren der Theorie*** bezeichnet. Der Satz „Wenn der Alkoholgehalt des Blutes 0,5‰ übersteigt, dann hat dies positive oder negative Auswirkungen auf die Reaktionsfähigkeit", hat demnach einen relativ geringen Informationsgehalt, da sowohl verbesserte Reaktionsfähigkeit als auch verschlechterte Reaktionsfähigkeit mit dem Dann-Teil übereinstimmen. Die Aussage hat nur einen potenziellen Falsifikator, nämlich „gleichbleibende Reaktionsfähigkeit". Der Informationsgehalt dieses Satzes könnte gesteigert werden, wenn der Dann-Teil weniger Ereignisse zulässt, sodass die Anzahl der potenziellen Falsifikatoren steigt. Dies wäre der Fall, wenn beispielsweise eine verbesserte Reaktionsfähigkeit durch den Dann-Teil ausgeschlossen wird.

Der Informationsgehalt eines Satzes hängt auch von der Präzision der verwendeten Begriffe ab. Betrachten wir hierzu den Satz: „Wenn sich eine Person autoritär verhält, dann wählt sie eine konservative Partei". Der Informationsgehalt dieses Satzes hängt davon ab, wie die Begriffe „autoritär" und „konservativ" definiert sind. Für jemanden, der den Begriff „konservativ" sehr weit fasst und eine Vielzahl von Parteien konservativ nennt, hat der Satz wenig potenzielle Falsifikatoren und damit weniger Informationsgehalt als für jemanden, der den Begriff „konservativ" sehr eng fasst und nur eine begrenzte Zahl von Parteien darunter zählt.

**Logische Konsistenz.** Führt die logische Überprüfung einer theoretischen Aussage zu dem Ergebnis, dass diese immer wahr ist, so ist die entsprechende Aussage logisch inkonsistent. Wir bezeichnen derartige Aussagen als analytisch wahr bzw. als ***tautologisch***. Ein tautologischer Satz besitzt keine potenziellen Falsifikatoren. Beispielsweise wäre der Satz: „Wenn ein Mensch einen Intelligenzquotienten über 140 hat, dann ist er ein Genie", tautologisch, falls der Begriff „Genie" durch eben diese Intelligenzhöhe definiert ist. Dieser Satz ist bei jeder Beschaffenheit der Realität immer wahr, *er hat keine potenziellen Falsifikatoren*.

Nicht immer ist der tautologische Charakter einer Aussage offensichtlich. Die Wahrscheinlichkeit einer „verkappten“ Tautologie nimmt zu, wenn in einem Satz unpräzise Begriffe enthalten sind.

Ebenfalls nicht offensichtlich ist die Tautologie von so genannten „Kann“-Sätzen. Betrachten wir beispielsweise die folgende Aussage: „Wenn jemand ständig erhöhtem Stress ausgesetzt ist, dann kann es zu einem Herzinfarkt kommen.“ Bezogen auf eine einzelne Person ist dieser Satz nicht falsifizierbar, da sowohl das Auftreten als auch das Nichtauftreten eines Herzinfarktes mit dem Dann-Teil der Aussage vereinbar ist. Beziehen wir den Satz auf alle Menschen, so wäre er nur falsifizierbar, wenn unter allen Menschen, die jemals an irgendeinem Ort zu irgendeiner Zeit gelebt haben, leben oder leben werden, kein einziger durch erhöhten Stress einen Herzinfarkt erleidet. Da eine solche Überprüfung niemals durchgeführt werden kann, sind Kann-Sätze für praktische Zwecke tautologisch.

Überprüfbar und damit wissenschaftlich brauchbar wird ein Kann-Satz erst durch die Spezifizierung bestimmter Wahrscheinlichkeitsangaben im Dann-Teil, wenn also die Höhe des Risikos eines Herzinfarktes bei ständigem Stress genauer spezifiziert wird. Lautet der Satz beispielsweise: „Wenn jemand ständig erhöhtem Stress ausgesetzt ist, dann kommt es mit einer Wahrscheinlichkeit von mindestens 20% zu einem Herzinfarkt“, dann ist diese Aussage zwar ebenfalls, auf eine einzelne Person bezogen, nicht falsifizierbar. Betrachten wir hingegen eine Gruppe von hundert unter ständigem Stress stehenden Menschen, von denen weniger als 20 einen Herzinfarkt erleiden, dann gilt dieser Satz als falsifiziert. (Genauer werden wir dieses Problem im Kap. 3 behandeln, in dem es u.a. um die Verallgemeinerung und Bewertung von Stichprobenergebnissen geht.)

Im Gegensatz zu einer tautologischen Aussage ist eine *kontradiktorische Aussage* immer falsch. Sie kann empirisch niemals bestätigt werden, d. h. sie hat keine potenziellen Konfirmatoren. Kontradiktorisch ist beispielsweise der Satz: „Wenn eine Person keinen Wein trinkt, dann trinkt sie Chardonay.“ Aus der Tatsache, dass Chardonnay ein spezieller Wein ist, folgt, dass dieser Satz analytisch falsch ist. Auch kontradiktorische Sätze sind natürlich wissenschaftlich unbrauchbar.

Neben tautologischen und kontradiktorischen Aussagen gibt es Sätze, die deshalb unwissenschaftlich sind, weil sie aus anderen Sätzen ***logisch falsch abgeleitet*** sind. So wird man beispielsweise leicht erkennen, dass die Aussage „Alle Christen sind Polizisten“ logisch falsch aus den Sätzen „Christen sind hilfsbereite Menschen“ und „Polizisten sind hilfsbereite Menschen“ erschlossen wurde.

Die Ermittlung des Wahrheitswertes derartiger abgeleiteter Sätze ist Gegenstand eines Teilbereiches der Wissenschaftstheorie, der formalen Logik, mit dem wir uns nicht weiter auseinandersetzen wollen (Literatur zur Logik: Carnap, 1960; Cohen u. Nagel, 1963; Kyburg, 1968; Stegmüller, 1969, Kap. 0; Tarski, 1965).

**Logische Vereinbarkeit.** Der Volksmund rät angehenden Paaren: „Gleich und Gleich gesellt sich gern“. Er sagt aber auch: „Gegensätze ziehen sich an.“ Wir haben es hier offenbar mit zwei widersprüchlichen theoretischen Aussagen zu tun. Theorien, die sich logisch widersprechen, müssen bzgl. ihrer internen Logik, ihres Informationsgehalts und ihrer Präzision verglichen werden. Sind die Theorien hinsichtlich dieser Kriterien gleichwertig, ist diejenige Theorie vorzuziehen, die empirisch am besten abgesichert erscheint oder sich in einem kritischen Vergleichsexperiment als die bessere erweist. Außerdem solle man – wie im o. g. Beispiel – überprüfen, ob *beide* Theorien, unter jeweils spezifischen Randbedingungen, Gültigkeit beanspruchen können.

Widerspruchsfreiheit der verglichenen Theorien bedeutet keineswegs, dass die Theorien wahr sind. Es lassen sich Theorien konstruieren, die zwar in keinem logischen Widerspruch zueinander stehen, die aber dennoch falsch sind. *Der Wahrheitsgehalt einer Theorie kann nur durch empirische Überprüfungen ermittelt werden.* Dies setzt allerdings voraus, dass die Theorie unbeschadet ihrer logisch fehlerfreien Konstruktion überhaupt empirisch überprüfbar ist.

**Empirische Überprüfbarkeit.** Die Forderung nach empirischer Überprüfbarkeit einer Theorie ist eng an die Forderung nach ihrer Falsifizierbarkeit geknüpft. Es sind aber Aussagen denkbar, die zwar im Prinzip falsifizierbar, aber (noch) nicht empirisch überprüfbar sind. Zur Verdeutlichung nehmen wir folgende Aussage: „Alle Menschen

sind von Natur aus aggressiv. Wenn sich die Aggressivität im Verhalten nicht zeigt, dann ist sie verdrängt.“ Unabhängig von der mangelnden Präzision der verwendeten Begriffe kann diese Aussage nur dadurch falsifiziert werden, dass ein Mensch gefunden wird, der weder aggressives Verhalten zeigt noch seine Aggressionen verdrängt hat. Die Falsifizierbarkeit hängt somit ausschließlich von der Möglichkeit ab, nachweisen zu können, dass jemand weder manifeste noch verdrängte Aggressionen hat.

Eine solche Theorie kann unbeschadet ihrer potenziellen Falsifizierbarkeit und unbeschadet ihres möglichen Wahrheitsgehaltes nur dann empirisch überprüft werden, wenn ein wissenschaftlich anerkanntes Instrument zum Erkennen verdrängter und manifester Aggressionen existiert. So gesehen ist es durchaus denkbar, dass wissenschaftliche Theorien zwar falsifizierbar, aber beim derzeitigen Stand der Forschung noch nicht empirisch überprüfbar sind. Die Überprüfung der Theorie muss in diesem Falle die Entwicklung geeigneter Messinstrumente abwarten.

Erweist sich die Theorie hinsichtlich der genannten Kriterien (Präzision, Informationsgehalt, logische Konsistenz, logische Vereinbarkeit, empirische Überprüfbarkeit) als unbrauchbar, sollte auf dem fortgeschrittenen Informationsstand eine neue Erkundungsphase eröffnet werden. Ein positiver Ausgang der theoretischen Überprüfung ermöglicht die endgültige Festlegung des Untersuchungsgegenstandes.

Ein Beispiel soll diese Zusammenhänge erläutern. Einer Untersuchung sei der folgende theoretische Satz vorangestellt: „Autoritärer Unterricht hat negative Auswirkungen auf das Sozialverhalten der Schüler.“ Wenn diese Behauptung richtig ist, dann müssten sich Schüler aus 8. Schulklassen, in denen Lehrer autoritär unterrichten, weniger kooperationsbereit zeigen als Schüler 8. Schulklassen mit nicht autoritär unterrichtenden Lehrern (zum Hypothesenbegriff vgl. z. B. Groeben u. Westmeyer, 1975 oder Hussy u. Möller, 1996).

Diese Hypothese ist durch drei Deduktionsschlüsse mit der Theorie verbunden: Erstens wurde aus allen möglichen autoritären Unterrichtsformen der Unterrichtsstil von Lehrern 8. Klassen herausgegriffen, zweitens wurde auf einen bestimmten Personenkreis, nämlich Schüler der 8. Klasse, geschlossen und drittens wurde als eine Besonderheit des Sozialverhaltens die Kooperationsbereitschaft ausgewählt.

Neben dieser einen Hypothese lassen sich natürlich weitere Hypothesen aus der Theorie ableiten, womit sich das Problem stellt, wie viele aus einer Theorie abgeleitete Hypothesen überprüft werden müssen, damit die Theorie als bestätigt gelten kann. Auf diese Frage gibt es keine verbindliche Antwort. Der Allgemeinheitsanspruch einer Theorie lässt es nicht zu, dass eine Theorie auf Grund empirischer Überprüfungen endgültig und eindeutig als „wahr“ bezeichnet werden kann (vgl. S. 12).

Aus heuristischen Gründen wurden im Flussdiagramm (vgl. Abb. 1) ein theoretisches *Bewährungskriterium* k und ein theoretisches *Belastbarkeitskriterium* l aufgenommen. Diese Kriterien sollen angeben, nach wie vielen Hypothesen bestätigenden Untersuchungen der Konsens über die Brauchbarkeit (Bewährungskriterium) bzw. über die Unbrauchbarkeit (Belastbarkeitskriterium) der Theorie hergestellt sein sollte. Auf diese Kriterien wird in der Entscheidungsphase (s. unten) ausführlicher eingegangen.

### Planungsphase

Nachdem das Thema festliegt, müssen *vor Beginn der Datenerhebung* Aufbau und Ablauf der Untersuchung vorstrukturiert werden. Durch eine sorgfältige Planung soll verhindert werden, dass während der Untersuchung Pannen auftreten, die in der bereits laufenden Untersuchung nicht mehr korrigiert werden können.

**Auswahl der Variablen.** Die Planung beginnt mit einer *Aufstellung von Variablen*, die für die Untersuchung relevant sind. Wir verstehen unter einer ***Variablen*** ein Merkmal, das – im Unterschied zu einer ***Konstanten*** – in mindestens zwei Abstufungen vorkommen kann. Eine zweistufige Variable wäre beispielsweise das Geschlecht (männlich, weiblich), eine dreistufige Variable die Schichtzugehörigkeit (Unter-, Mittel-, Oberschicht) und eine Variable mit beliebig vielen Abstufungen das Alter. (Das Problem der Variablenklassifikation wird in Kap. 1, S. 18 ff. ausführlich behandelt.)

Als nächstes erfolgt eine Klassifikation der Variablen. Wir unterscheiden

- unabhängige Variablen,

- abhängige Variablen und
- Kontrollvariablen.

(Ausführlicher hierzu vgl. Bortz u. Dörung, 2002, Kap. 1.1.1.)

Unter den unabhängigen Variablen werden diejenigen Merkmale verstanden, deren Auswirkungen auf andere Merkmale – die abhängigen Variablen – überprüft werden sollen. Im Allgemeinen ist bereits auf Grund der Fragestellung festgelegt, welche der relevanten Variablen als abhängige und welche als unabhängige Variablen in die Untersuchung eingehen sollen. Darüber hinaus wird die Liste der relevanten Variablen jedoch häufig weitere Variablen enthalten, die weder zu den abhängigen noch zu den unabhängigen Variablen zu zählen sind. Es muss dann entschieden werden, ob diese Variablen als Kontrollvariablen mit erhoben werden sollen, ob nur eine Ausprägung der Variablen (z. B. nur weibliche Personen) erfasst (was als Konstanthalten einer Variablen bezeichnet wird) oder ob die Variable überhaupt nicht berücksichtigt werden soll.

Für das o. g. Beispiel wäre folgende Variablengruppierung denkbar:

*Unabhängige Variable:* Art des Unterrichtsstils („autoritär" vs. „demokratisch").

Bei der Festlegung der unabhängigen Variablen ist darauf zu achten, dass nicht nur die eigentlich interessierende Merkmalsausprägung – hier also autoritärer Unterrichtsstil – untersucht wird. Um den Begriff „Variable" rechtfertigen zu können, sind (mindestens) zwei Ausprägungen (also mindestens zwei Unterrichtsformen) als Stufen der unabhängigen Variablen in die Untersuchung einzubeziehen, denn nur so kann das Besondere des autoritären Unterrichtsstils im Vergleich zu anderen Unterrichtsformen herausgearbeitet werden.

Für eine Hypothesen prüfende Untersuchung ist es zudem erforderlich, für jede Stufe der unabhängigen Variablen mehrere Untersuchungseinheiten vorzusehen, d.h., für unser Beispiel benötigen wir eine Auswahl autoritär unterrichteter und eine Auswahl demokratisch unterrichteter Schulklassen.

*Abhängige Variable:* Kooperationsbereitschaft.
Die Frage, wie die abhängige Variable genau erfasst bzw. „operationalisiert" wird, behandeln wir später (s. S. 9).

*Kontrollvariablen:* Erziehungsstil der Eltern, Anzahl der Geschwister, soziale Schicht der Kinder, Geschlecht der Kinder.
Diese Variablen werden miterhoben, um später prüfen zu können, ob sie den Zusammenhang zwischen Unterrichtsstil und Kooperationsbereitschaft beeinflussen bzw. „moderieren". Die Kontrollvariablen werden deshalb gelegentlich auch *Moderatorvariablen* genannt.

*Konstant gehaltene Variablen:* Alter der Kinder (14 Jahre oder 8. Schulklasse), Größe der Schulklasse (16–20 Kinder), Geschlecht des Lehrers (männlich), Unterrichtszeit (8 bis 9 Uhr bzw. 1. Unterrichtsstunde), Art des Unterrichtsstoffes (Mathematik).
Es ist zu beachten, dass ein Untersuchungsergebnis um so weniger generalisierbar ist, je mehr Variablen konstant gehalten wurden. Es gilt in unserem Beispiel nur für 8. Schulklassen mit 16–20 Jungen, die in der 1. Stunde Mathematikunterricht haben. Wir werden dieses Thema unter dem Stichwort „Labor- oder Felduntersuchung" erneut aufgreifen.

*Nicht berücksichtigte Variablen:* Alter des Lehrers, Intelligenz der Kinder, Motivation der Kinder, Lärmbelästigung etc.
Auch dies sind Variablen, die die Kooperationsbereitschaft der Kinder zumindest potenziell beeinflussen können. In diesem Falle würden sie den eigentlich interessierenden Zusammenhang zwischen Unterrichtsstil und Kooperationsverhalten „stören" bzw. dessen Interpretation erschweren. Die potenziell bedeutsamen, aber in der Untersuchung nicht berücksichtigten Variablen werden deshalb häufig *Störvariablen* genannt.

**Labor- oder Felduntersuchung.** Diese Untersuchungsvarianten markieren die Extreme eines Kontinuums, das durch eine unterschiedlich starke Kontrolle untersuchungsbedingter Störvariablen gekennzeichnet ist. Wenn in einer Untersuchung äußere Einflüsse, die den Untersuchungsablauf stören könnten, weitgehend kontrolliert oder ausgeschaltet sind, sprechen wir von einer Laboruntersuchung. Findet umgekehrt die Untersuchung in einem natürlichen („biotischen") Umfeld statt, das durch äußere Eingriffe des Untersuchenden nicht verändert wird, handelt es sich um eine Felduntersuchung.

In der Untersuchungsplanung muss nun entschieden werden, ob die Untersuchung eher La-

bor- oder eher Feldcharakter haben soll. Beide Varianten sind mit Vor- und Nachteilen verbunden. Die Kontrolle von untersuchungsbedingten Störvariablen in der Laboruntersuchung gewährleistet, dass die Untersuchungsergebnisse weitgehend frei von störenden Einflüssen und damit eindeutiger interpretierbar sind. In diesem Sinne haben Laboruntersuchungen eine hohe interne Validität bzw. Gültigkeit.

**Eine Untersuchung ist intern valide, wenn ihr Ergebnis eindeutig interpretierbar ist. Die interne Validität sinkt mit wachsender Anzahl plausibler Alternativerklärungen für das Ergebnis auf Grund nicht kontrollierter Störvariablen.**

Der Nachteil einer Laboruntersuchung liegt in ihrer eingeschränkten Generalisierbarkeit, denn Untersuchungsergebnisse, die für ein „steril" gehaltenes Untersuchungsumfeld gültig sind, können nur bedingt auf natürliche Lebenssituationen übertragen werden. Laboruntersuchungen verfügen in der Regel über eine geringere externe Validität.

**Eine Untersuchung ist extern valide, wenn ihr Ergebnis über die besonderen Bedingungen der Untersuchungssituation und über die untersuchten Personen hinausgehend generalisierbar ist. Die externe Validität sinkt mit wachsender Unnatürlichkeit der Untersuchungsbedingungen bzw. mit abnehmender Repräsentativität der untersuchten Stichproben.**

Angesichts dieser Gültigkeitskriterien ist es häufig schwierig, für die zu prüfende Fragestellung eine geeignete Untersuchungskonzeption zu entwickeln. Oft wird man sich – wie in unserem Beispiel – mit einem Planungskompromiss begnügen müssen, der Feld- und Laborelemente in einer der Fragestellung angemessenen Weise kombiniert. Man beachte allerdings, dass ein Mindestmaß an interner Validität für jede wissenschaftliche Untersuchung erforderlich ist.

**Experimentelle oder quasiexperimentelle Untersuchung.** Während das Kontinuum Labor vs. Feld das Ausmaß der Kontrolle untersuchungsbedingter Störvariablen beschreibt, kennzeichnet die Unterscheidung von experimenteller und quasiexperimenteller Untersuchung das Ausmaß der Kontrolle von Personen bedingten Störvariablen. In unserem Beispiel wären dies Variablen wie Intelligenz oder Motivation der Schüler, die Anzahl der Geschwister, der Erziehungsstil der Eltern etc.

In einer experimentellen Untersuchung ist dafür Sorge zu tragen, dass die Personen bezogenen Störvariablen unter allen Untersuchungsbedingungen (d.h. unter allen Stufen der unabhängigen Variablen) annähernd gleich ausgeprägt sind. Dies ist dadurch zu erreichen, dass die Personen den Untersuchungsbedingungen nach Zufall zugeordnet werden. Diese Vorgehensweise wird Randomisierung genannt.

**Unter Randomisierung versteht man die zufällige Zuordnung der Untersuchungsteilnehmer zu den Untersuchungsbedingungen.**

Da es durch die Randomisierung der Personen zu einem „statistischen Fehlerausgleich" kommt, hat dieser Untersuchungstyp natürlich eine höhere interne Validität als Untersuchungen ohne Randomisierung. Die Personen-bezogene externe Validität wäre durch eine repräsentativ auszuwählende Stichprobe sicherzustellen (vgl. hierzu 3.1).

Bei einer quasiexperimentellen Untersuchung muss auf eine Randomisierung verzichtet werden, da hier „natürliche" bzw. bereits existierende Gruppierungen untersucht werden. Beispiele hierfür sind Vergleiche von weiblichen und männlichen Personen, von Abiturienten und Realschülern, von Autofahrern und Nichtautofahrern etc. In diesen Fällen ist die Zugehörigkeit der Untersuchungsteilnehmer zu den Stufen der unabhängigen Variablen vorgegeben, d.h. eine Randomisierung ist ausgeschlossen.

Unser Schülerbeispiel ließe sich vermutlich auch nur quasiexperimentell realisieren, es sei denn, die ausgewählten Schulklassen erhalten durch Zufall einen autoritären oder demokratischen Lehrer. Da dies der üblichen Schulpraxis widerspricht, wird man bereits bei der Auswahl der Schulklassen darauf achten, welche Klassen eher von einem als autoritär bzw. demokratisch zu bezeichnenden Lehrer unterrichtet werden.

Gegenüber einem experimentellen Ansatz birgt diese Vorgehensweise jedoch die Gefahr, dass die vom Untersuchungsleiter nicht hergestellte Schulklassengruppierung von Störvariablen überlagert ist, die die spätere Interpretation der Ergebnisse erschweren. Beispielsweise könnten die sog. autoritären Lehrer älter sein als die sog. demokratischen

Kollegen und deshalb ein anderes didaktisches Unterrichtskonzept vertreten; hier wäre also das Alter die eigentlich relevante Variable.

Diese Hinweise mögen genügen, um zu verdeutlichen, dass quasiexperimentelle Untersuchungen intern weniger valide sind als experimentelle Untersuchungen.

**Experimentelle Untersuchungen haben eine höhere interne Validität als quasiexperimentelle Untersuchungen.**

Die interne Validität einer quasiexperimentellen Untersuchung lässt sich jedoch erhöhen, wenn es gelingt, die zu vergleichenden Gruppen nach relevanten Störvariablen zu *parallelisieren*. Um im Beispiel zu bleiben, könnten die Schulklassengruppen paarweise so zusammengestellt werden, dass der autoritäre und der demokratische Lehrer in jedem Schulklassenpaar ungefähr gleichaltrig sind. Auf diese Weise aufgestellte Stichproben bezeichnet man als „*matched samples*".

**Operationalisierung.** Von entscheidender Bedeutung für den Ausgang der Untersuchung ist die Frage, wie die unabhängigen Variablen, die abhängigen Variablen und die Kontrollvariablen operationalisiert werden. Durch die Operationalisierung wird festgelegt, welche Operationen (Handlungen, Reaktionen, Zustände usw.) wir als indikativ für die zu messende Variable ansehen wollen und wie diese Operationen quantitativ erfasst werden. Anders formuliert: Nachdem festgelegt wurde, *welche* Variablen erfasst werden sollen, muss durch die Operationalisierung bestimmt werden, *wie* die Variablen erfasst werden sollen. Bezogen auf unser Beispiel stellt sich z. B. die Frage, wie wir die Kooperationsbereitschaft der untersuchten Schüler messen bzw. den Unterrichtsstil der Lehrer erfassen können.

Die Operationalisierung wird um so schwieriger, je komplexer die einbezogenen Variablen sind. Während einfache Variablen wie z. B. „Anzahl der Geschwister" problemlos zu ermitteln sind, kann es oftmals notwendig sein, komplexe Variablen wie z. B. „kooperatives Verhalten" durch mehrere operationale Indikatoren zu bestimmen. Fundierte Kenntnisse über bereits vorhandene Messinstrumente (Tests, Fragebögen, Versuchsanordnungen usw.) können die Operationalisierung erheblich erleichtern, wenngleich es häufig unumgänglich ist, unter Zuhilfenahme der einschlägigen Literatur über Test- und Fragebogenkonstruktion eigene Messinstrumente zu entwickeln. Hinweise hierzu und weiterführende Literatur findet man bei Bortz u. Döring (2002, Kap. 4).

Hinsichtlich der unabhängigen Variablen muss zweifelsfrei entschieden werden können, welchen Unterrichtsstil ein Lehrer praktiziert. Dies kann z. B. durch Verhaltensbeobachtung, Interviews oder Fragebögen (vgl. z. B. Mummendey, 1995) geschehen. Auch diese Datenerhebungstechniken werden bei Bortz u. Döring (2002, Kap. 4) ausführlich beschrieben.

Ist entschieden, wie die einzelnen Variablen zu operationalisieren sind, können die entsprechenden Untersuchungsmaterialien bereitgestellt werden. Wenn neue Messinstrumente entwickelt werden müssen, sollten diese unbedingt zuvor an einer eigenen Stichprobe hinsichtlich des Verständnisses der Instruktion, der Durchführbarkeit, der Eindeutigkeit in der Auswertung, des Zeitaufwandes usw. getestet werden.

**Stichprobengröße.** Eine dem Statistiker häufig gestellte Frage lautet: Wie viele Untersuchungsteilnehmer oder „Versuchspersonen" (abgekürzt: „Vpn") werden für die Untersuchung benötigt? Allgemein bezieht sich diese Frage auf die Anzahl der Untersuchungseinheiten bzw. – in unserem Beispiel – auf die Anzahl der Schulklassen, die erforderlich ist, um eine Hypothese verlässlich überprüfen zu können. Die einfachste Antwort auf diese Frage wäre: So viele wie möglich.

Präziser kann die Antwort des Statistikers nicht sein, es sei denn, er erhält genauere Informationen über den Kontext der Untersuchung. Dazu zählen:

- eine Mindestangabe über die Größe des Effektes, den der Untersuchende für praktisch bedeutsam halten würde (im Beispiel: Wäre es von praktischer Bedeutung, wenn demokratisch unterrichtete Schüler nur um 3% kooperativer sind als autoritär unterrichtete Schüler?);
- eine Einschätzung der Folgen, die sich ergeben, wenn aus der Untersuchung falsche Schlüsse gezogen werden (im Beispiel: Welche Konsequenzen hätte es, wenn auf Grund der Untersuchung fälschlicherweise behauptet wird, autoritär unterrichtete Schüler seien weniger ko-

operativ als demokratisch unterrichtete Schüler?).

Wie mit diesen Informationen umgegangen wird, um eine begründete Entscheidung über den zu wählenden Stichprobenumfang treffen zu können, behandeln wir im Kap. 4.

**Planung der statistischen Auswertung.** Die Planungsphase endet mit Überlegungen zur statistischen Auswertung des Untersuchungsmaterials. Es müssen diejenigen statistischen Auswertungstechniken festgelegt werden, mit denen über die Brauchbarkeit der Hypothesen entschieden werden soll. Manchmal wird auf eine Planung der statistischen Auswertung verzichtet, in der Hoffnung, dass sich nach der Datenerhebung schon die geeigneten Auswertungsverfahren finden werden. Diese Nachlässigkeit kann dazu führen, dass sich die erhobenen Daten nur undifferenziert auswerten lassen, wobei eine geringfügige Änderung in der Datenerhebung (z. B. verbessertes *Skalenniveau*, vgl. Kap. 1.1) den Einsatz differenzierterer Auswertungstechniken ermöglicht hätte.

### Untersuchungsphase

Wurde die Untersuchung in der Planungsphase gründlich vorstrukturiert, dürfte die eigentliche Durchführung der Untersuchung keine prinzipiellen Schwierigkeiten bereiten. Wir wollen deshalb auf eine Erörterung dieser Phase verzichten unter Verweis auf die eingangs (S. 2) erwähnte Literatur zur Planung und Durchführung empirischer Untersuchungen.

Ein besonderes Problem psychologischer Untersuchungen sind sog. Versuchsleiter-(Vl-)Artefakte, also mögliche Beeinträchtigungen des Untersuchungsergebnisses durch das Verhalten des Versuchsleiters. Hierzu findet man ausführliche Informationen bei Rosenthal (1966) bzw. Rosenthal u. Rosnow (1969) oder zusammenfassend bei Bortz u. Döring (2002, Kap. 2.5).

### Auswertungsphase

In der Auswertungsphase werden die erhobenen Daten statistisch verarbeitet. Zuvor sollte man sich jedoch – zumindest bei denjenigen Fragebögen, Tests oder sonstigen Messinstrumenten, die noch nicht in anderen Untersuchungen erprobt wurden – einen Eindruck von der *testtheoretischen Brauchbarkeit der Daten* verschaffen.

Im einfachsten Fall wird man sich damit begnügen zu überprüfen, ob das Untersuchungsmaterial eindeutig quantifizierbar ist bzw. ob verschiedene Auswerter den Vpn auf Grund der Untersuchungsergebnisse die gleichen Zahlenwerte zuordnen. Dieses als ***Objektivität*** des Untersuchungsinstrumentes bezeichnete Kriterium ist bei den meisten im Handel erhältlichen Verfahren gewährleistet. Problematisch hinsichtlich ihrer Objektivität sind Untersuchungsmethoden, die zur Erfassung komplexer Variablen nicht hinreichend standardisiert sind. So wäre es in unserem Beispiel möglich, dass verschiedene Auswerter – bedingt durch ungenaue Operationalisierungen – zu unterschiedlichen Einstufungen der Kooperationsbereitschaft der Schüler gelangen oder dass Lehrer nicht übereinstimmend als demokratisch oder autoritär bezeichnet werden. Ein Untersuchungsmaterial, das eine nur geringe Objektivität aufweist, ist für die Überprüfung der Hypothesen wenig oder gar nicht geeignet. Sobald sich solche Mängel herausstellen, sollte die Untersuchung abgebrochen werden, um in einem neuen Versuch zu Operationalisierungen zu gelangen, die eine objektivere Datengewinnung gestatten.

In größer angelegten Untersuchungen ist zusätzlich zur Objektivität auch die ***Reliabilität*** der Untersuchungsdaten zu überprüfen. Über dieses Kriterium, das die Genauigkeit bzw. Zuverlässigkeit der erhobenen Daten kennzeichnet, sowie über weitere Gütekriterien wird in der testtheoretischen Literatur berichtet. Auch eine zu geringe Reliabilität des Untersuchungsmaterials sollte eine bessere Operationalisierung der Variablen veranlassen.

Genügen die Daten den testtheoretischen Anforderungen, werden sie in übersichtlicher Form *tabellarisch* zusammengestellt bzw., falls die Auswertung mit einem statistischen Programmpaket geplant ist, in geeigneter Weise aufbereitet (vgl. Anhang E, S. 733 als Beispiel für die Aufbereitung einer SPSS-Datei). Die sich anschließende statistische Analyse ist davon abhängig, ob eine Hypothesen erkundende oder Hypothesen prüfende Untersuchung durchgeführt wurde. Für Hypothesen erkundende Untersuchungen nimmt man üblicherweise Datenaggregierungen vor, die in Kap. 1 zusammengestellt sind. Hypothesen prüfende Unter-

suchungen werden mit den vielfältigen, in diesem Buch dargestellten Methoden der schließenden Statistik oder Inferenzstatistik ausgewertet.

Mit der Anwendung eines inferenzstatistischen Verfahrens bzw. eines „Signifikanztests“ wird eine Entscheidung über die zu prüfende Hypothese herbeigeführt. Hierzu errechnet man eine sog. Irrtumswahrscheinlichkeit P, die angibt, mit welcher Wahrscheinlichkeit man sich irren würde, wenn man die fragliche Hypothese akzeptiert. Um die Hypothese annehmen zu können, sollte diese Irrtumswahrscheinlichkeit natürlich möglichst klein sein.

Die Größe der maximal tolerierbaren Irrtumswahrscheinlichkeit liegt allerdings nicht im Ermessen des Untersuchenden, sondern ist durch eine allgemein gültige Konvention festgelegt. Man bezeichnet diese Grenze, die von der Irrtumswahrscheinlichkeit P nicht überschritten werden darf, als „Signifikanzniveau“ und verwendet hierfür das Symbol $\alpha$. Die üblichen Werte für das Signifikanzniveau sind $\alpha = 5\%$ oder sogar $\alpha = 1\%$. Der Untersuchende muss vor Durchführung des Signifikanztests festlegen, welches $\alpha$-Niveau für die Untersuchung angemessen ist.

## Entscheidungsphase

Ein Vergleich der ermittelten Irrtumswahrscheinlichkeit P mit dem zuvor fest gelegten Signifikanzniveau $\alpha$ zeigt, ob das Ergebnis der Untersuchung signifikant ($P \leq \alpha$) oder nicht signifikant ($P > \alpha$) ist. Zunächst wollen wir uns einem *nicht signifikanten Ergebnis* zuwenden.

Bei einem nicht signifikanten Ergebnis gilt die geprüfte Hypothese – wir werden sie unter 4.1 als Alternativhypothese bzw. als $H_1$ bezeichnen – als nicht bestätigt. Diese Aussage basiert auf einer sehr vorsichtigen Entscheidungsregel, nach der eine Hypothese bereits dann als nicht bestätigt gelten soll, wenn man im Falle ihrer Annahme mit einer Wahrscheinlichkeit von nur 5% oder mehr (bzw. gar 1% oder mehr) eine Fehlentscheidung riskiert.

Diese Konvention gewährleistet, dass die Hypothese erst dann als bestätigt angesehen wird, wenn das empirische Ergebnis in sehr überzeugender Weise für die Richtigkeit dieser Hypothese spricht. „Nicht signifikant“ bedeutet also nicht, dass die Hypothese ($H_1$) falsch ist; „nicht signifikant“ heißt lediglich, dass die Untersuchung nicht geeignet war, die Gültigkeit der Hypothese zu belegen.

Vor einer endgültigen Ablehnung der eigenen Hypothese ist zunächst zu überprüfen, ob in der Untersuchung Fehler begangen wurden, auf die das nicht signifikante Ergebnis zurückgeführt werden kann. Wird im Nachhinein erkannt, dass beispielsweise bestimmte relevante Variablen nicht hinreichend berücksichtigt wurden, dass Instruktionen falsch verstanden wurden, dass sich die Vpn nicht instruktionsgemäß verhalten haben oder dass die untersuchten Stichproben zu klein waren, kann die gleiche Hypothese in einer Wiederholungsuntersuchung, in der die erkannten Fehler korrigiert sind, erneut überprüft werden.

Problematischer ist ein nicht signifikantes Ergebnis, wenn Untersuchungsfehler praktisch auszuschließen sind. Ist der deduktive Schluss von der Theorie auf die überprüfte Hypothese korrekt, muss an der allgemeinen Gültigkeit der Theorie gezweifelt werden. Wenn in unserem Beispiel die allgemeine Theorie richtig ist, dass sich ein autoritärer Unterrichtsstil negativ auf das Sozialverhalten von Schülern auswirkt, und wenn Kooperationsbereitschaft eine Form des Sozialverhaltens ist, dann muss die Kooperationsbereitschaft auch bei den untersuchten Kindern durch den autoritären Unterrichtsstil negativ beeinflusst werden. Andernfalls ist davon auszugehen, dass die der Untersuchung zugrunde liegende Theorie fehlerhaft ist.

Konsequenterweise ist in Abb. 1 auf Grund eines nicht signifikanten Ergebnisses, das nicht auf Untersuchungsfehler zurückzuführen ist, ein Pfeil eingezeichnet, der besagt, dass *die Theorie verändert werden muss*. Die veränderte Theorie sollte jedoch nicht nur an die alte Theorie anknüpfen, sondern auch die Erfahrungen berücksichtigen, die durch die Untersuchung gewonnen wurden. So könnte beispielsweise die hier skizzierte Untersuchung, von der wir einmal annehmen wollen, dass sich der Zusammenhang zwischen autoritärem Unterrichtsstil und unkooperativem Verhalten als nicht signifikant herausgestellt habe, zur Vermutung Anlass geben, dass das Kooperationsverhalten nur bei Schülern aus der Oberschicht durch den Unterrichtsstil beeinflusst wird, während die beiden Merkmale bei anderen Schülern keinen Zusammenhang aufweisen. Anlässlich eines solchen Befundes würden wir durch *Indukti-*

*onsschluss* den Geltungsbereich der ursprünglichen Theorie auf Oberschichtschüler begrenzen. Formal stellt sich diese Veränderung der Theorie so dar, dass *der Wenn-Teil der theoretischen Aussage konjunktiv um eine Komponente erweitert wird:* „Wenn autoritär unterrichtet wird *und* die Schüler der Oberschicht entstammen, dann wird das Sozialverhalten negativ beeinflusst." Derartige Modifikationen einer Theorie auf Grund einer falsifizierten Hypothese bezeichnen wir in Anlehnung an Holzkamp (1968, 1971) bzw. Dingler (1923) als ***Exhaustion.***

Es ist nun denkbar, dass auch die Überprüfung weiterer, aus der exhaurierten Theorie abgeleiteten Hypothesen zu nicht signifikanten Ergebnissen führen, sodass sich die Frage aufdrängt, durch wie viele Exhaustionen eine Theorie „*belastet*" (Holzkamp, 1968) werden kann bzw. wie viele exhaurierende Veränderungen eine Theorie „erträgt". Theoretisch findet ein sich zyklisch wiederholender Exhaustionsprozess dann ein Ende, wenn durch ständig zunehmende Einschränkung der im Wenn-Teil genannten Bedingungen eine „Theorie" resultiert, deren Informationsgehalt praktisch gegen Null geht. So könnten weitere Exhaustionen an unserem Modellbeispiel zu einer Theorie führen, nach der sich eine ganz spezifische Form des autoritären Unterrichts nur bei bestimmten Schülern zu einer bestimmten Zeit unter einer Reihe von besonderen Bedingungen auf einen Teilaspekt des Sozialverhaltens negativ auswirkt. Eine solche Theorie über die Bedingungen von Sozialverhalten ist natürlich wenig brauchbar. (Koeck, 1977, diskutiert die Grenzen des Exhaustionsprinzips am Beispiel der Frustrations-Aggressions-Theorie.)

Die Wissenschaft wäre allerdings nicht gut beraten, wenn sie jede schlechte Theorie bis zu ihrem, durch viele Exhaustionen bedingten, natürlichen Ende führen würde. Das Interesse an der Theorie wird auf Grund wiederholter Falsifikationen allmählich nachlassen, bis sie in Vergessenheit gerät. *Das Belastbarkeitskriterium der Theorie ist überschritten.*

Als nächstes wollen wir überprüfen, welche Konsequenzen sich mit einem *signifikanten Ergebnis* verbinden. Bei einem signifikanten Ergebnis riskieren wir mit der Annahme der untersuchten Hypothese ($H_1$) eine Fehlentscheidung, deren Wahrscheinlichkeit nicht größer als 5% (1%) ist. Man ist sich also ziemlich sicher, mit einer Entscheidung zugunsten der geprüften Hypothese keinen Fehler zu begehen, aber auch nur „ziemlich" sicher und nicht „völlig" sicher, denn es verbleibt eine Restwahrscheinlichkeit von 5% (1%) für eine Fehlentscheidung. Dennoch ist es Konvention, die geprüfte Hypothese in diesem Falle als bestätigt anzusehen.

Hinsichtlich der Theorie besagt eine durch ein signifikantes Ergebnis bestätigte Hypothese, dass wir keinen Grund haben, an der Richtigkeit der Theorie zu zweifeln, sondern dass wir vielmehr der Theorie nach der Untersuchung eher trauen können als vor der Untersuchung. Die absolute Richtigkeit der Theorie ist jedoch damit nicht erwiesen; dafür müssten letztlich unendlich viele aus der Theorie abgeleitete Einzelhypothesen durch Untersuchungen verifiziert werden – eine Forderung, die in der empirischen Forschung nicht realisierbar ist. *Somit kann durch empirische Forschung auch die absolute Richtigkeit einer Theorie nicht nachgewiesen werden.*

Dennoch regulieren neue, durch empirische Forschung gewonnene Erkenntnisse mehr oder weniger nachhaltig unseren Alltag. Genauso, wie eine schlechte Theorie allmählich in Vergessenheit gerät, kann sich eine gute Theorie durch wiederholte Bestätigung zunehmend mehr bewähren, bis sie schließlich Eingang in die *Praxis* findet. *Das Bewährungskriterium ist überschritten.*

„So ist die empirische Basis der objektiven Wissenschaft nichts ‚Absolutes'; die Wissenschaft baut nicht auf Felsengrund. Es ist eher ein Sumpfland, über dem sich die kühne Konstruktion ihrer Theorien erhebt; sie ist ein Pfeilerbau, dessen Pfeiler sich von oben her in den Sumpf senken – aber nicht bis zu einem natürlichen ‚gegebenen' Grund. Denn nicht deshalb hört man auf, die Pfeiler tiefer hineinzutreiben, weil man auf eine feste Schicht gestoßen ist: Wenn man hofft, dass sie das Gebäude tragen werden, beschließt man, sich vorläufig mit der Festigkeit der Pfeiler zu begnügen" (Popper, 1966; S. 75f.).