Grundbegriffe der Datenerhebung: Vom Mensch zur Zahl

2

Statistik bezeichnet die, meist hypothesengeleitete, Auswertung von numerischen (quantitativen) Daten, die Rückschlüsse auf gestellte Forschungsfragen zulassen. Doch die Daten und Zahlen, mit denen man bei der Auswertung arbeitet, kommen nicht aus dem luftleeren Raum, sondern müssen zunächst gewonnen werden. In der Datenerhebung – gewissermaßen der „Umwandlung" des Menschen, seines Verhaltens und Erlebens in Zahlen – liegt deshalb eine große Herausforderung. Als Statistiker sollte man den Prozess der Datenerhebung nie aus den Augen verlieren – denn allzu leicht verfällt man sonst dem Trugschluss, dass die Zahlen, mit denen man arbeitet, objektive und zweifelsfreie Aussagen über den Menschen erlauben. Tatsächlich aber wird der Transformationsprozess vom Mensch zur Zahl an vielen Stellen durch die Entscheidungen des Forschers beeinflusst, ob nun bei der Operationalisierung (siehe Abschn. 2.1) oder bei der Wahl der Stichprobe (siehe Abschn. 2.5).

Die Datenerhebung muss übrigens nicht zwangsläufig mit einem Ergebnis in Zahlen enden. Ist das aber der Fall und schließt sich eine statistische Auswertung an, spricht man von *quantitativen Methoden*. Da es in diesem Buch um Statistik geht, ist das quantitative Denken das Feld, in dem wir uns hier bewegen. Neben den quantitativen Methoden existieren auch noch die sogenannten *qualitativen Methoden*, bei deren Anwendung weitgehend auf Zahlen verzichtet wird und alternative Zugänge zum menschlichen Verhalten und Erleben gesucht werden, z. B. in Form von Fallstudien oder Interviews. Bei einigen Fragestellungen hat sich gezeigt, dass diese nur durch qualitative Fragestellungen überhaupt zugänglich gemacht werden können. Der Großteil der psychologischen Forschung fokussiert heute auf den quantitativen Methoden, wenn auch zu beobachten ist, dass die Verwendung qualitativer Methoden in der Psychologie wieder zunimmt.

T. Schäfer, *Methodenlehre und Statistik*,
DOI 10.1007/978-3-658-11936-2_2

Literaturempfehlung

Flick, U., von Kardorff, E., & Steinke, I. (Hrsg.). (2004). *Qualitative Forschung: Ein Handbuch*, (3. Aufl.). Reinbek: Rowohlt.

Kapitel 28 aus Sedlmeier, P., & Renkewitz, F. (2013). *Forschungsmethoden und Statistik*. München: Pearson.

2.1 Ohne Maßband oder Waage: Wie misst man die Psyche?

Da es das Ziel der Psychologie ist, menschliches Erleben und Verhalten zu erklären und zu verstehen, muss sie einen geeigneten Zugang zum Erleben und Verhalten finden, der das Durchführen wissenschaftlicher Untersuchungen erlaubt. In diesem Zugang liegt eine sehr zentrale Herausforderung. Denn vieles, über das wir reden, wenn es um Menschen und ihr Erleben und Verhalten geht, können wir nicht einfach mit einem Mikroskop beobachten oder mit einem Lineal messen. Es gibt natürlich einige Dinge, die man einfach bestimmen oder messen kann, wie beispielsweise das Alter oder das Geschlecht einer Person, ihr Einkommen oder das Geld, das sie pro Tag für Lebensmittel ausgibt. Für andere interessierende Größen ist das nicht so leicht, stattdessen müssen geeignete Instrumente entwickelt werden, mit denen ein solcher Zugang möglich gemacht werden kann. Mit anderen Worten: man benötigt geeignete Messinstrumente für das Erfassen von Emotionen, Verhaltensweisen, Einstellungen, Persönlichkeitsmerkmalen usw. Das Problem dabei besteht – wie man sich leicht vorstellen kann – in der Übersetzung solcher psychologischer Phänomene in Zahlen und Daten. Beispielsweise könnten wir uns für das Thema „Intelligenz“ interessieren. Wie soll man die Intelligenz eines Menschen bestimmen? Was ist Intelligenz überhaupt? Lässt sie sich messen? Und wenn ja, was sagen uns dann die konkreten Zahlen, die nach der Messung übrig bleiben?

Bleiben wir beim Beispiel Intelligenz. Zur Frage, was Intelligenz ist, müssen zuerst theoretische Überlegungen angestellt werden. Und es wird in erster Linie eine Definitionsfrage sein, was eine Gemeinschaft von Forschern unter Intelligenz verstehen möchte und was nicht. Die zweite Frage – ob Intelligenz messbar ist – wird von der Psychologie prinzipiell mit Ja beantwortet. Denn da sie eine Wissenschaft ist, versucht sie ja genau das zu bewerkstelligen: sie versucht, Erleben und Verhalten in wissenschaftlich untersuchbare Teile oder Einzelheiten zu zerlegen.

Im ersten Kapitel haben wir diesen Prozess als *Operationalisierung* kennengelernt: das Einigen auf geeignete Messinstrumente. Der Sinn des Messens ist es, mit Hilfe von Zahlen möglichst genau das abzubilden, was ein Mensch denkt, fühlt oder welche Verhaltensweisen er zeigt. Am Ende soll also eine objektive Zahl für ein meist subjektives oder individuelles Phänomen stehen; die Zahl soll das Phänomen *repräsentieren*.

▶ Messen besteht im Zuordnen von Zahlen zu Objekten, Phänomenen oder Ereignissen, und zwar so, dass die Beziehungen zwischen den Zahlen die analogen Beziehungen der Objekte, Phänomene oder Ereignisse repräsentieren.

Wenn in dieser Definition von Objekten gesprochen wird, so können damit beispielsweise Einstellungen gemeint sein. Eine Einstellung ist die (meist wertende) Überzeugung, die eine Person gegenüber einem gewissen Gegenstand oder Sachverhalt hat. So kann jemand den Umweltschutz befürworten oder kritisieren, und auch die Stärke einer Befürwortung oder einer Kritik kann bei verschiedenen Personen verschieden stark ausgeprägt sein (sie kann also variieren). Will ein Forscher nun die Einstellung verschiedener Personen zum Umweltschutz messen, muss er dafür ein geeignetes Instrument finden oder entwickeln. In diesem Fall könnte er beispielsweise einen Fragebogen entwerfen, auf dem die befragten Personen ihre Meinung auf einer Skala ankreuzen können. Wie solche Skalen aussehen können und welche weiteren Möglichkeiten es gibt, solche Messungen durchzuführen, werden wir im Folgenden sehen. In jedem Fall aber wird der Forscher davon ausgehen wollen, dass das, was er mit seinem Fragebogen erfasst hat, auch dem entspricht, was die befragten Personen wirklich „gemeint“ haben.

Die Übersetzung von Objekten, Phänomenen oder Ereignissen in Zahlen wird in Abb. 2.1 verdeutlicht. Beim Messen werden häufig die Begriffe empirisches und numerisches Relativ verwendet. Das *empirische Relativ* bezieht sich dabei auf die tatsächlichen (empirischen) Verhältnisse oder Tatsachen in der Welt. Beispielsweise könnte ein Forscher die Aggressivität von Personen messen wollen. Die durch eine geeignete Operationalisierung zugänglich und beobachtbar gemachte Aggressivität dieser Personen würde dabei das empirische Relativ bilden. Und es wäre auch möglich, dass zehn verschiedene Personen zehn verschiedene Ausprägungen in der Stärke ihrer Aggressivität haben. Die Idee beim Messen ist es nun, jeder Person einen Zahlenwert für die Stärke ihrer Aggressivität zuzuordnen. Diese Zahlen sollen möglichst gut die tatsächliche Stärke der Aggressivität wiedergeben oder abbilden. Sie bilden dann das *numerische Relativ*. Mit Hilfe der Zahlen ist es nun möglich, Unterschiede oder Verhältnisse zu beschreiben, die die

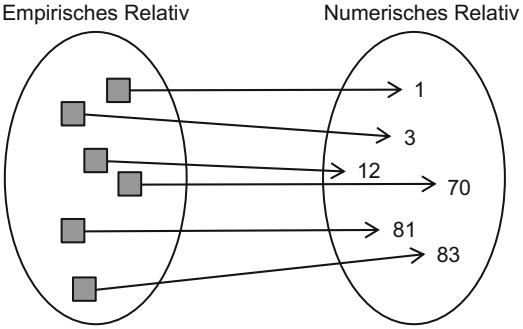

Abb. 2.1 Empirisches und numerisches Relativ beim Messen

Unterschiede und Verhältnisse der tatsächlichen Aggressivität der Personen widerspiegeln.

Die Abbildung eines empirischen in ein numerisches Relativ kann mehr oder weniger gut gelingen. In der Psychologie hat dieses Problem sogar einen Namen: das *Repräsentationsproblem*. Wie dieser Name bereits andeutet, geht es hierbei um die Frage, wie repräsentativ eine Messung für das ist, was gemessen werden soll. Für physikalische Eigenschaften stellt sich dieses Problem nicht: das Körpergewicht eines Menschen lässt sich z. B. zweifelsfrei mit einer Waage feststellen. Außerdem wird sofort klar, was es bedeutet, wenn eine Person 2 Kilogramm schwerer ist als eine andere Person, oder auch, wenn sie „doppelt so schwer" ist. Auch die Eigenschaften Alter und Geschlecht haben wir eben schon genannt; sie sind einfach feststellbar. In der Psychologie sind jedoch die meisten Eigenschaften nicht so eindeutig in Zahlen überführbar. Man kann beispielsweise nicht mehr so einfach behaupten, dass eine Person doppelt so aggressiv sei wie eine andere Person. Was soll mit „doppelt so viel" gemeint sein?

Wie kann die Psychologie das Repräsentationsproblem zumindest annähernd lösen? In der Regel wird dies versucht, indem man den Prozess des Messens so gut und genau wie möglich gestaltet. Was wiederum eine „gute" Messung ist, ist in der Psychologie genauestens definiert. Die Erfüllung sogenannter Gütekriterien (Objektivität, Reliabilität, Validität), auf die hier nicht näher eingegangen werden kann (siehe z. B. Kap. 3 aus Sedlmeier und Renkewitz 2013), spielt dabei eine wichtige Rolle.

Ein wesentliches Ziel der quantitativen Vorgehensweise ist es daher, geeignete Messinstrumente zu entwickeln, die das, was gemessen werden soll, auf einer (numerischen) Skala so genau wie möglich abbilden. Im folgenden Abschnitt werden wir verschiedene Arten von Skalen kennenlernen und sehen, was man mit ihnen machen kann. Vorher jedoch ist es notwendig, einige Begriffe zu klären,

die im Zusammenhang mit Messen und Testen immer wieder auftauchen und für das weitere Verständnis unerlässlich sind.

2.2 Variablen und Daten

Wir haben bisher oft davon gesprochen, dass man bestimmte Dinge oder Größen messen will. Wenn man etwas misst, dann haben diese „Dinge" oder „Größen" einen Namen; sie heißen *Variablen*.

► Messen bezieht sich immer auf Variablen. „Variable" ist die Bezeichnungen für eine Menge von Merkmalsausprägungen.

Die Variable ist der zentrale Begriff in Methodenlehre und Statistik. Denn letztendlich geht es ja immer um die Erklärung von Phänomenen, die verschiedene Ausprägungen annehmen können, die also *variabel* sind. Etwas, das bei verschiedenen Menschen oder über die Zeit hinweg immer in derselben Ausprägung vorliegt, stellt also keine Variable dar und kann auch nicht gemessen werden. Das hört sich erst mal etwas seltsam an, doch egal womit sich die Psychologie beschäftigt – alles lässt sich als Variable ausdrücken: Bei der Untersuchung von Intelligenz geht es darum zu erklären, warum eine Person intelligenter ist als eine andere. Bei Persönlichkeitsmerkmalen (wie z. B. Großzügigkeit) soll erklärt werden, warum sie bei verschiedenen Personen verschieden stark ausgeprägt sind. Bei psychischen Störungen möchte man wissen, warum der eine sie bekommt, der andere nicht. Und natürlich sucht man bei all diesen Fragen nach den Ursachen, die wiederum auch als Variablen gemessen werden. Variablen, die oft als Ursachen für die Ausprägungen von anderen Variablen in Frage kommen, sind beispielsweise das Alter von Personen, ihr Geschlecht, ihr Bildungsstand, ihre Sozialisationsbedingungen usw. – alles wiederum Größen, die bei verschiedenen Menschen verschieden (variabel) sein können.

Das Besondere an einer Variable ist also, dass sie verschiedene *Ausprägungen* annehmen kann. Je nachdem, welche Ausprägungen eine Variable hat, lassen sich dichotome, kategoriale, diskrete und kontinuierliche Variablen unterscheiden.

Dichotome, kategoriale, diskrete und kontinuierliche Variablen

Jede Variable muss mindestens zwei Ausprägungen haben. Wenn sie genau zwei Ausprägungen hat, dann wird sie auch *dichotome Variable* genannt. Dichotom bedeutet so viel wie Entweder/Oder. Eine typische dichotome Variable ist z. B. das

Geschlecht: es kann nur die Ausprägungen männlich oder weiblich annehmen. Eine Vielzahl von Variablen lässt sich als dichotome Variablen behandeln oder darstellen. Beispielsweise könnte man Menschen ganz grob danach einteilen, ob sie jung sind (z. B. höchstens 40 Jahre alt) oder alt (alle, die älter sind als 40 Jahre). Dann hätte man wieder eine Variable mit zwei Ausprägungen. Eine solche Festlegung von Variablenausprägungen ist natürlich sehr willkürlich, aber sie kann je nach Forschungsfrage ausreichend oder angemessen sein. Ähnlich könnte man demnach auch jeweils zwei Gruppen von intelligenten/nicht intelligenten, aggressiven/friedfertigen oder introvertierten/aufgeschlossenen Personen bilden. In vielen Fällen ist die interessante Frage auch einfach die, ob ein bestimmtes Merkmal vorliegt oder nicht vorliegt, also z. B., ob jemand Raucher ist oder nicht, ob jemand eine bestimmte Krankheit hat oder nicht, ob jemand aus einer Scheidungsfamilie stammt oder nicht, usw.

Wenn nun eine Variable mehr als zwei Ausprägungen hat, dann stellt sich die Frage, wie diese Ausprägungen abgestuft sind. Es gibt dabei zwei prinzipielle Möglichkeiten. Eine Möglichkeit ist, dass die verschiedenen Ausprägungen der Variablen einzelne Kategorien beschreiben. Nehmen wir das Beispiel Haarfarbe, dann könnten wir hier eine Variable definieren, die die Ausprägungen schwarz, blond, braun und rot hat. Diese vier Antwortalternativen entsprechen einfach vier verschiedenen Kategorien. Daher werden solche Arten von Variablen auch *kategoriale Variablen* genannt. Manchmal spricht man auch von *qualitativen Variablen*, weil den verschiedenen Ausprägungen lediglich eine je eigene Qualität zukommt.

Eine andere prinzipielle Möglichkeit ist, dass die Ausprägungen einer Variable nicht bloß Kategorien bilden, sondern quantitativ messbar sind. Dabei kann es sich um diskret oder kontinuierlich messbare Variablen handeln. Einige Variablen haben Ausprägungen, die nur in ganz bestimmten – diskreten – Schritten vorliegen können und daher *diskrete Variablen* genannt werden. Beispielsweise ist die Anzahl von Geschwistern ein diskretes Merkmal, da offensichtlich nur ganzzahlige Ausprägungen sinnvoll sind. Anders ist das bei Variablen, die stufenlos (kontinuierlich) gemessen werden können. In diese Rubrik der *kontinuierlichen Variablen* fallen die meisten Variablen. Einfache Beispiele sind Zeit, Länge oder Gewicht. Diese Variablen kann man kontinuierlich, also in beliebig kleinen Schritten oder Unterteilungen messen. Typisch für diese Variablen ist natürlich, dass man sie in Zahlen ausdrückt, die außerdem beliebig genau sein können (je nachdem, wie viele Stellen nach dem Komma man für diese Zahlen benutzen möchte). So kann die Größe einer Person z. B. 175 cm betragen. Man kann die Größe aber auch genauer angeben, z. B. 175,45 cm. Eine solche Bezeichnung mit Zahlenwerten ist für kontinuierliche Variablen also unumgänglich, während man

kategoriale Variablen zunächst nicht in Form von Zahlenwerten erfasst. Wie wir später noch sehen werden, versucht man in der Psychologie häufig, das Erleben und Verhalten mit Hilfe von kontinuierlichen Variablen zu messen.

Manifeste und latente Variablen

Variablen lassen sich nach einem weiteren Gesichtspunkt unterscheiden, der besonders für die Psychologie sehr wichtig ist. Es geht um die Frage, ob man eine Variable direkt messen kann oder ob sie sozusagen im Verborgenen liegt. Nehmen wir einmal an, wir untersuchen das Kaufverhalten einer Person und wollen wissen, wie der Betrag, den sie an der Supermarktkasse für Lebensmittel ausgibt, von ihrer Einstellung gegenüber gesunder Ernährung abhängt. Den Geldbetrag, den die Person an der Kasse bezahlt, können wir einfach registrieren. Diese Variable manifestiert sich also direkt und wird daher *manifeste Variable* genannt. Die Einstellung der Person gegenüber gesunder Ernährung können wir hingegen nicht so einfach bestimmen; sie ist nach außen nicht sichtbar, sondern liegt in einem subjektiven Werturteil der Person. Wie sollen wir diese Einstellung also messen? Eine Möglichkeit wäre auch hier wieder, einen Fragebogen zu entwerfen, mit dem der Forscher mit Hilfe von ausgewählten Fragen zum Thema Ernährung auf die Einstellung der Person schließen kann. Wir sehen aber, dass diese Einstellung für den Forscher prinzipiell im Verborgenen liegt, also latent ist. Solche Variablen – die man nicht direkt messen kann, sondern durch andere Variablen (z. B. durch die Angaben auf einem Fragebogen) erst erschließen muss – heißen *latente Variablen*.

► Variablen, die man direkt messen kann, heißen manifeste Variablen. Solche, die man nicht direkt messen kann, sondern erst mit Hilfe anderer Variablen erschließen muss, heißen latente Variablen.

In der Psychologie ist die Mehrzahl aller interessanten Variablen latent und muss durch geeignete Instrumente zugänglich gemacht werden. Diesen Schritt haben wir oben als Operationalisierung bezeichnet. Latente Variablen haben auch noch einen anderen Namen, der in der Psychologie sehr gebräuchlich ist: sie heißen auch *Konstrukte*. Konstrukte sind Begriffe, die theoretisch sinnvoll erscheinen, um etwas Interessantes zu beschreiben, was nicht direkt beobachtbar oder messbar (also latent) ist und erst durch andere Variablen erschlossen werden muss. Mit einigen Beispielen für latente Variablen haben wir schon hantiert, beispielsweise Intelligenz, Aggressivität oder Persönlichkeit. Aber auch basale Begriffe wie Wahrnehmung, Lernen, Gedächtnis, Motivation usw. sind Konstrukte: sie beschreiben etwas, was psychologisch interessant ist, was aber erst einmal

lediglich ein Begriff ist und nicht etwas, was man direkt sehen oder messen kann. Wenn Sie schon einmal an einem Intelligenztest teilgenommen haben, dann wissen Sie, dass man dort viele Fragen beantworten und viele Aufgaben lösen muss. All diese Fragen und Aufgaben sind Variablen, die auf das Konstrukt Intelligenz *hindeuten* sollen.

Unabhängige und abhängige Variablen

Eine weitere Unterscheidung, die uns im Rahmen der psychologischen Forschung begleiten wird, ist die zwischen unabhängigen und abhängigen Variablen. Die *abhängige Variable* ist im Forschungsprozess immer diejenige Variable, an deren Erklärung oder Beschreibung man interessiert ist. Wir könnten beispielsweise den Altersdurchschnitt von zwei verschiedenen Städten bestimmt und daraufhin festgestellt haben, dass sich die beiden Durchschnittswerte unterscheiden. Und nun werden wir sehr wahrscheinlich der Frage nachgehen wollen, woran das liegt. Warum ist das Durchschnittsalter in den beiden Städten verschieden? Dafür können mehrere Variablen als Ursache in Betracht kommen – Variablen, die in beiden Städten verschiedene Ausprägungen haben. Beispielsweise könnte die eine Stadt eine Großstadt sein, in der viele junge Leute leben, während die andere Stadt auf dem Land liegt und aufgrund hoher Arbeitslosigkeit weniger attraktiv ist. Diese Variable – nennen wir sie „Urbanisierungsgrad" – würde also als mögliche Erklärung für den Altersunterschied in Frage kommen. Sie wäre dann eine *unabhängige Variable*, denn ihre Ausprägung (z. B. hoher vs. niedriger Urbanisierungsgrad) ist von vornherein durch unsere Fragestellung und die konkrete Untersuchung gegeben, sie ist sozusagen unabhängig von anderen Variablen. Das Entscheidende ist, dass die Ausprägung der abhängigen Variable von der Ausprägung der unabhängigen Variable abhängt. In unserem Beispiel ließe sich das so verallgemeinern: wenn sich der Urbanisierungsgrad einer Stadt verändert, dann verändert sich auch der Altersdurchschnitt ihrer Einwohner.

Prinzipiell lässt sich fast jede Erkenntnis, die die wissenschaftliche Psychologie aufgrund empirischer Daten erlangt, in der Form *unabhängige Variable → abhängige Variable* beschreiben. Wie wir noch sehen werden, ist es oft das Ziel psychologischer Forschung, unabhängige Variablen ausfindig zu machen oder sogar selbst zu manipulieren und den Effekt auf die abhängige Variable zu untersuchen. Die Aufgabe der Forschungsmethoden und vor allem der Statistik ist es dabei, den Zusammenhang zwischen unabhängiger und abhängiger Variable mathematisch zu beschreiben und zu verallgemeinern. Wann immer wir nach Erklärungen für ein psychologisches Phänomen suchen, wird diese Erklärung in Form einer unabhängigen Variable formuliert sein.

Tab. 2.1 Verschiedene Arten von Variablen

Variablen lassen sich einteilen...	Beschreibung	Beispiele
nach der Art ihrer Ausprägungen		
dichotom	nur 2 mögliche Ausprägungen	Geschlecht, Raucher/ Nichtraucher, Atomgegner/ Atombefürworter
kategorial	mehrere Ausprägungen, die verschiedenen Kategorien entsprechen	Schulabschluss, Wohngegend, Musikgeschmack
diskret	Ausprägungen, die sich der Größe nach ordnen lassen	Anzahl von Geschwistern, Schulnoten
kontinuierlich	Stufenlose Ausprägungen, die sich der Größe nach ordnen lassen	Alter, Intelligenz
nach ihrer Beobachtbarkeit bzw. Messbarkeit		
manifest	direkt messbar oder beobachtbar	Alter, Geschlecht, präferiertes Fernsehprogramm
latent	nicht direkt messbar oder beobachtbar, muss erschlossen werden	Intelligenz, Einstellung gegenüber Ausländern, Glücklichkeit
nach ihrer Rolle im Forschungsprozess		
unabhängig	wird beobachtet oder systematisch variiert	Hintergrundmusik in Kaufhaus A, aber nicht in Kaufhaus B
abhängig	wird als Effekt der UV gemessen	Umsatz in Kaufhaus A und Kaufhaus B

► Die unabhängige Variable *(UV)* ist die Variable, die während einer Untersuchung fokussiert oder während eines Experimentes systematisch variiert oder manipuliert wird. Die abhängige Variable *(AV)* ist die Variable, mit der der Effekt festgestellt wird, der auf die UV zurückführbar ist.

Die verschiedenen Unterteilungen von Variablen sind in Tab. 2.1 noch einmal zusammengefasst.

Wir wissen jetzt, was Variablen sind und dass sich Messen immer auf Variablen bezieht. Wenn wir Variablen gemessen und bestimmte Ergebnisse erhalten haben, dann werden diese Ergebnisse *Daten* genannt. Daten sind damit Ausschnitte der

Wirklichkeit, die als Grundlage für empirisch-wissenschaftliche Erkenntnisse benötigt werden. Die Daten bilden letztendlich die Basis für jede Art von Aussage, die ein Forscher über einen bestimmten Gegenstand machen kann.

2.3 Daten auf unterschiedlichem Niveau: das Skalenproblem

Skalen und Skaleneigenschaften

Wie wir gesehen haben, können wir die Ausprägung einer Variable messen (den empirischen Relationen numerische zuordnen). Dabei kann diese Messung ganz unterschiedlich aussehen: sie kann darin bestehen, dass man danach fragt, ob eine bestimmte Variablenausprägung vorliegt oder nicht, ob sie in eine bestimmte Kategorie fällt, oder man sucht einen konkreten Zahlenwert, wenn die Variablenausprägung diskret oder kontinuierlich gemessen werden kann. Offenbar haben wir es hier also mit ganz unterschiedlichen Arten von Messung zu tun, und die Daten (also das Ergebnis der Messung) liegen in ganz verschiedenen Formaten vor. Diese Unterschiede kommen daher, dass wir Messungen auf verschiedenen *Skalen* machen können. Der Begriff „Skala" beschreibt die Beschaffenheit des empirischen und des numerischen Relativs sowie eine Abbildungsfunktion, die die beiden verbindet. Dabei geht es um die Frage, wie das, was durch das empirische Relativ erfasst wird, durch ein numerisches Relativ (also durch Zahlen) sinnvoll repräsentiert werden kann. Je nach Beschaffenheit des empirischen Relativs sind verschiedene Abbildungsfunktionen in Zahlenwerte möglich bzw. sinnvoll. Insgesamt kann man vier Arten von Skalen unterscheiden; man spricht auch von *Skalenniveaus*: Nominal-, Ordinal-, Intervall- und Verhältnisskala. Von Skalen „niveaus" spricht man deshalb, weil der Informationsgehalt und die mathematische Güte über die vier Skalen hinweg steigen. Doch schauen wir uns zunächst an, was es mit diesen Skalen auf sich hat.

Die *Nominalskala* ist die einfachste Skala. Auf ihr werden dichotome und kategoriale Variablen gemessen, und sie ist lediglich dazu geeignet, die Gleichheit oder Ungleichheit von Variablenausprägungen zu beschreiben. Betrachten wir noch einmal das Beispiel Haarfarbe mit den Ausprägungen schwarz, blond, braun und rot. Wenn wir mehrere Personen hinsichtlich ihrer Haarfarbe untersuchen, dann können wir sagen, dass eine bestimmte Anzahl von Leuten z. B. schwarze Haare hat und dass diese Leute in der Haarfarbe schwarz übereinstimmen. Ein anderes Beispiel könnte das Genre von Musikstücken sein: z. B. Klassik, Pop, Electro. Jedes Musikstück lässt sich für diese Variable in eine Kategorie einordnen.

Wenn zwei Musikstücke in der gleichen Kategorie landen, dann wissen wir, dass sie hinsichtlich ihres Genres übereinstimmen. Das ist alles. Wir können mit Variablen, die auf einer Nominalskala gemessen wurden, keinerlei weitere mathematische Berechnungen anstellen. Wir könnten zwar den verschiedenen Variablenausprägungen Zahlen zuordnen (z. B. eine 1 für schwarze Haare, eine 2 für blonde Haare, eine 3 für braune Haare und eine 4 für rote Haare), aber diese Zahlen drücken keine quantitativen Beziehungen aus. Wir können nicht etwa sagen, dass blonde Haare „doppelt so viel" sind wie schwarze Haare, weil 2 doppelt so viel ist wie 1. Und wir können auch nicht sagen, dass rote Haare irgendwie „mehr" oder „besser" sind als schwarze. Diese Aussagen machen offenbar keinen Sinn. Daten auf Nominalskalenniveaus lassen also nur qualitative Aussagen zu.

Eine zweite Art von Variablen lässt sich so messen, dass man auch quantitative (also mengenmäßige) Aussagen über ihre Ausprägungen machen kann, weil sie bestimmte Relationen erkennen lassen. Ein gutes Beispiel sind die Ränge bei einem sportlichen Wettkampf. Wenn die drei Sieger die Ränge 1, 2 und 3 bekommen, dann wissen wir, wer der Beste war, wer der Zweitbeste und wer der Drittbeste. Mit den Rängen 1, 2, 3 können wir also eine Relation deutlich machen, die einen quantitativen Unterschied beschreibt. Man kann auch von einer größer-kleiner Relation sprechen. Daten, die solche Aussagen über Relationen zulassen, befinden sich auf *Ordinalskalenniveau*. Obwohl wir hier schon mathematisch von größer-kleiner Beziehungen sprechen können, sind wir aber immer noch nicht in der Lage, mit solchen Daten die genauen numerischen Distanzen zwischen Variablenausprägungen zu beschreiben. Wenn wir beim Beispiel der Ränge 1, 2, 3 bleiben, wissen wir also hier nicht, „um wie viel besser" der Sportler mit Rang 1 als der Sportler mit Rang 2 war. Er könnte z. B. doppelt so schnell oder dreimal so schnell gewesen sein, oder aber auch nur wenige Millisekunden schneller. Und wir wissen auch nicht, ob der Abstand zwischen den Sportlern mit den Rängen 1 und 2 genauso groß war wie der zwischen den Sportlern mit den Rängen 2 und 3. Über diese *absoluten* Unterschiede und über die Größe der Differenzen erfahren wir also nichts, sondern müssen uns damit begnügen, nur etwas über die *relativen* Unterschiede zwischen den Variablenausprägungen zu erfahren.

Um tatsächlich etwas über absolute Unterschiede herausfinden zu können, müssen wir unsere Daten mindestens auf einer *Intervallskala* messen. Die Bezeichnung „Intervall" drückt aus, dass auf dieser Skala die genauen Intervalle (also Abstände) zwischen den einzelnen Variablenausprägungen gemessen werden können. Ein Beispiel ist die Messung von Intelligenz mit Hilfe des Intelligenzquotienten (IQ). Der IQ wird auf einer Skala gemessen, die mehr oder weniger willkürlich festgelegt wurde. Sie ist so angelegt, dass die meisten Menschen auf dieser Skala einen Wert von ca. 100 erreichen. IQ-Werte, die kleiner oder größer sind als

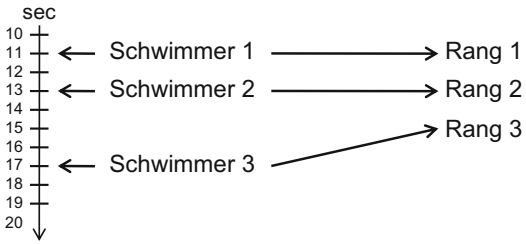

Abb. 2.2 Rangvergabe nach den Zeiten für drei Schwimmer auf 25 Meter

100, sind nicht mehr so häufig und solche, die sehr stark von 100 abweichen (z. B. 180 oder 65) sind schon sehr selten. Das Entscheidende ist aber, dass man mit Hilfe der IQ-Skala die absoluten Unterschiede zwischen Personen bestimmen kann und dass man außerdem etwas über die Gleichheit oder Ungleichheit von Differenzen sagen kann. Wenn eine Person einen IQ von 110 und eine andere Person einen IQ von 120 hat, dann weiß man nicht nur, dass Person 2 intelligenter ist als Person 1, sondern man hat auch eine Vorstellung darüber, was dieser Unterschied inhaltlich bedeutet (sofern man weiß, was genau in dem Test gemacht wurde). Außerdem weiß man, dass sich diese beiden Personen in ihrer Intelligenz genauso stark unterscheiden wie zwei andere Personen, die einen IQ von 90 und einen IQ von 100 haben: in beiden Fällen beträgt die Differenz 10, und auf Intervallskalenniveau bedeutet das, dass beide Differenzen inhaltlich identisch sind. Mit Daten, die auf Intervallskalenniveau gemessen wurden, kann man deshalb auch mathematische Berechnungen anstellen, die über einfache größer-kleiner Beziehungen hinausgehen. Man kann hier addieren und subtrahieren: wenn man den IQ von Person 1 vom IQ der Person 2 abzieht, dann erhält man die Differenz von 10, die Auskunft über den absoluten Intelligenzunterschied gibt. Eine solche Berechnung lässt sich mit Daten auf Ordinalskalenniveau nicht anstellen. Wenn Ränge addiert oder subtrahiert werden, dann erhält man kein inhaltlich interpretierbares Ergebnis, weil man nicht weiß, welche konkreten Zahlenwerte sich hinter den Rängen verbergen. Abbildung 2.2 verdeutlicht dieses Problem noch einmal.

Wenn wir unsere Daten auf Intervallskalenniveau gemessen haben, können wir also schon interessante Berechnungen mit ihnen anstellen, wie beispielsweise die Berechnung von Mittelwerten (siehe Abschn. 3.3). Mittelwerte sind nur auf Intervallskalenniveau sinnvoll interpretierbar. Und wir wissen jetzt auch, dass wir mit solchen Daten etwas über die Gleichheit oder Ungleichheit von Differenzen sagen können. Was wir jedoch noch nicht können, ist eine Aussage darüber treffen, in welchem *Verhältnis* zwei Messwerte stehen. Ein Verhältnis geht über die bloße Differenz zweier Messwerte hinaus, es beschreibt vielmehr die relative Lage dieser

Messwerte in Bezug auf den Nullpunkt der Skala. Gehen wir noch einmal zu unserem Beispiel mit dem Intelligenztest zurück. Wenn zwei Personen einen IQ von 80 und 160 haben, dann wissen wir zwar, dass sie sich mit einer Differenz von 80 IQ-Punkten unterscheiden, wir können aber nicht sagen, dass die zweite Person „doppelt so intelligent" ist wie die erste. Eine solche Aussage ist deshalb nicht möglich, weil die Intelligenzskala keinen natürlichen Nullpunkt hat. Genauer gesagt, kann niemand einen IQ von Null haben. Wie schon erwähnt, wurde die Intelligenzskala relativ willkürlich festgelegt, ihr Mittelwert liegt bei 100 und die im Test geringsten möglichen IQ-Werte liegen bei etwa 30 bis 40 Punkten. Wenn ein solcher Nullpunkt fehlt oder er mehr oder weniger willkürlich auf einen bestimmten Wert festgelegt wurde, sind also keine sinnvollen Aussagen über Verhältnisse zwischen Messwerten möglich. Bei Skalen, die einen solchen natürlichen Nullpunkt besitzen, kann man die Verhältnisse von Messwerten angeben. Beispiele für solche *Verhältnisskalen* sind Temperatur (auf der Kelvin-Skala), Körpergröße, Alter, Anzahl usw. Hier kann man also Aussagen über die Gleichheit oder Ungleichheit von Verhältnissen machen. Beispielsweise ist eine dreißigjährige Person natürlich doppelt so alt wie eine fünfzehnjährige Person. Gleichermaßen würde eine Person mit 3 Stunden Fernsehkonsum pro Tag dreimal so lang fernsehen wie eine Person mit einer Stunde Fernsehkonsum. Wir können hier also Verhältnisse wie 1:2 oder 1:3 angeben.

Da man mit den verschiedenen Skalen, die wir kennengelernt haben, Messungen auf unterschiedlichen Niveaus machen kann, spricht man auch oft vom *Messniveau* einer Skala oder vom Messniveau der Daten. Man unterscheidet hier entsprechend nominales Messniveau (für Daten von Nominalskalen), ordinales Messniveau (für Daten von Ordinalskalen) und metrisches Messniveau (für Daten von Intervall- und Verhältnisskalen). Der Begriff „metrisch" deutet dabei an, dass Daten mindestens auf Intervallskalenniveau gemessen wurden und daher schon die gebräuchlichsten Berechnungen mit ihnen durchgeführt werden können. Manchmal spricht man auch einfach von *Intervalldaten* oder benutzt synonym den Begriff *metrische Daten*, sobald Intervallskalenniveau erreicht ist. In Tab. 2.2 sind die Skalenarten und Skaleneigenschaften noch einmal zusammengefasst.

In der Forschung ist man nun häufig bestrebt, Daten auf einem möglichst hohen Messniveau zu erheben. Dabei wird in den meisten Fällen mindestens Intervallskalenniveau angestrebt. Den Grund dafür haben wir nun schon mehrfach angedeutet: erst auf Intervallskalenniveau werden viele statistische Kennwerte (wie z. B. Mittelwerte) überhaupt berechenbar oder interpretierbar. Damit sind auch erst Daten auf diesem Messniveau für die statistischen Auswertungen geeignet, die wir noch kennenlernen werden. Außerdem können Daten im Nachhinein von

Tab. 2.2 Skalenarten und ihre Eigenschaften

Skalenart	Mess-niveau	Mögliche Aussagen	Rechen-operationen	Beispiele
Nominalskala	nominal	Gleichheit oder Ungleichheit	$=/\neq$	Familienstand, Wohnort
Ordinalskala	ordinal	größer-kleiner Relationen	$</>$	Ranking von Hochschulen, Tabellenplatz im Sport
Intervallskala	metrisch	Gleichheit oder Ungleichheit von Differenzen	$+/-$	Intelligenzquotient, Feindseligkeit gegenüber Ausländern
Verhältnisskala		Gleichheit oder Ungleichheit von Verhältnissen	$:/\cdot$	Länge, Gewicht, Alter

einem höheren auf ein niedrigeres Messniveau transformiert werden, was umgekehrt jedoch nicht funktioniert.

Ratingskalen

In der psychologischen Forschung versucht man meist, Intervallskalenniveau durch die Konstruktion geeigneter Fragebögen zu erreichen. Diese Fragebögen enthalten Fragen, deren Antwortmöglichkeiten auf Intervallskalen erfasst werden können. Solche Skalen, auf denen ein Befragter eine Antwort (ein sogenanntes Rating) abgeben muss, werden *Ratingskalen* genannt.

► Ratingskalen verwendet man, um Urteile über einen bestimmten Gegenstand zu erfragen. Es wird ein Merkmalskontinuum vorgegeben, auf dem der Befragte die Merkmalsausprägung markiert, die seine subjektive Empfindung am besten wiedergibt.

„Gegenstand" eines solchen Urteils kann die eigene Person sein (z. B. wenn man seinen eigenen Charakter einschätzen soll), eine oder mehrere andere Personen (z. B. Ausländer) oder ein abstraktes Einstellungsobjekt (z. B. die Einstellung gegenüber Umweltschutz). Ratingskalen können ganz verschieden gestaltet sein, und jede dieser Gestaltungsmöglichkeiten kann Vorteile und Nachteile haben. Typische Ratingskalen sehen meist so aus wie in Abb. 2.3. Diese Skala hat zehn Stufen, also zehn Antwortmöglichkeiten, zwischen denen der Befragte wählen kann. Um mit Hilfe von Ratingskalen tatsächlich intervallskalierte Daten zu

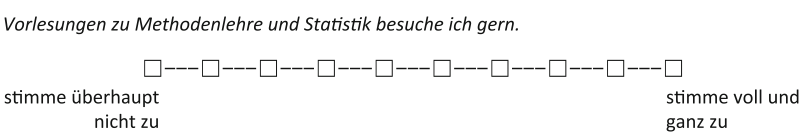

Abb. 2.3 Eine typische Ratingskala

erhalten, empfiehlt es sich die Unterteilung der Skala nicht zu grob zu gestalten. Hat die Skala nur vier Stufen, ist die inhaltliche Differenzierung des erfragten Sachverhaltes eingeschränkt. Mit anderen Worten: Personen mit unterschiedlichen aber doch ähnlichen Einstellungen müssen alle denselben Skalenwert ankreuzen, während sie bei einer feineren Skalierung eventuell verschiedene Skalenwerte angekreuzt hätten. Es macht daher mehr Sinn, eine Skala mit beispielsweise zehn Skalenwerten zu konstruieren. Voraussetzung für das Erlangen intervallskalierter Daten ist aber stets, dass das Phänomen, welches man messen möchte, eine solche Quantifizierung zulässt.

2.4 Fragebögen und Tests

In den vorangegangenen Abschnitten haben wir das Prinzip des Messens in der Psychologie ausführlich beleuchtet. Vor allem haben wir ein häufig verwendetes Messinstrument, die Ratingskala, kennengelernt. Nun ist es aber selten der Fall, dass man einer Person nur eine einzige Frage stellt oder ihr nur eine einzige Ratingskala vorlegt. In der Regel hat man eine ganze Sammlung von Fragen, auf die eine Person antworten soll – die *Fragebögen*. Fragebögen messen in aller Regel Eindrücke, Einstellungen, Meinungen, Gefühle, Gedankeninhalte oder auch persönliche Daten wie Alter und Geschlecht. Beim Ausfüllen von Fragebögen gibt es keine Zeitvorgabe und keine richtigen oder falschen Antworten. Neben den Ratingskalen kommen in Fragebögen auch Fragen mit Mehrfachantworten, ja/nein-Fragen oder Fragen mit offenen Antwortfeldern zum Einsatz. Die Konstruktion von Fragebögen folgt keinem festgelegten Schema; Wissenschaftler können Fragen selbst entwerfen und ein geeignetes Layout für die Antwortmöglichkeiten entwickeln.

Während Fragenbögen in der Regel nur Meinungen oder Einstellungen abfragen, sind Forscher oft an mehr interessiert und wollen einzelne Individuen so genau wie möglich charakterisieren. Zur Messung individueller Eigenschaften, Fähigkeiten oder Leistungen eignen sich Fragebögen manchmal nicht so gut, ganz einfach weil die befragte Person nur eingeschränkten Zugang dazu hat. Wenn man

etwa die Fähigkeit sich über einen längeren Zeitraum zu konzentrieren (Konzentrationsfähigkeit) einer Person messen möchte, dann ist es wenig sinnvoll, sie danach zu fragen. Sie könnte zwar auf einer Ratingskala beurteilen, für wie konzentriert sie sich hält, aber es wäre wesentlich sinnvoller, die Konzentrationsfähigkeit durch bestimmte Aufgaben genau zu erfassen. Die Messung von Eigenschaften, Fähigkeiten oder Leistungen von Individuen erfolgt durch *Tests*. Es lassen sich Persönlichkeits- und Leistungstests unterscheiden. *Persönlichkeitstests* laufen auch ohne Zeitdruck ab, und es gibt keine richtigen oder falschen Antworten. Sie sind aber nach einem festgelegten Schema konstruiert und normiert. Normiert bedeutet, dass man von einer recht großen Zahl von Menschen aus der Bevölkerung (etwa 2000) die Werte kennt, die sie in diesem Test erreichen. So kann man den Wert, den eine bestimmte Person erreicht hat, genau einordnen und mit den Werten anderer vergleichen. Bei *Leistungstest* gibt es in der Regel eine Zeitbegrenzung und natürlich richtige und falsche Antworten. Solche Tests beinhalten also neben Fragen auch verbale, mathematische, grafische oder praktische Aufgaben, die gelöst werden müssen. Intelligenztests sind also z. B. typische Leistungstests. Die Fragen und Aufgaben in einem Test werden auch *Items* genannt. Manchmal werden aber auch die Fragen aus einem Fragebogen als Item bezeichnet.

► Items sind Fragen oder Aufgaben, die beantwortet bzw. gelöst werden müssen. Tests bestehen aus einer Zusammenstellung von Items.

2.5 Stichproben und Population

Die Psychologie strebt in der Regel nach Erkenntnissen, die auf größere Personengruppen anwendbar sind. Zum Beispiel sucht man nach Möglichkeiten zur optimalen Förderung von Kindern im Vorschulalter oder nach einer Erklärung, warum Menschen depressiv werden. In beiden Fällen bezieht sich die Fragestellung auf sehr große Personengruppen, z. B. alle in Deutschland lebenden Kinder im Alter von 4–6 Jahren. Diese große Gruppe, nach der in einer Untersuchung gefragt wird, wird *Population* genannt.

Von praktischer Seite betrachtet wird jedem schnell einleuchten, dass man in einer kleinen psychologischen Untersuchung nicht alle Vorschulkinder der Bundesrepublik untersuchen kann, sondern sich auf einen Auszug beschränken muss. Diesen Auszug bezeichnet man als *Stichprobe*. Obwohl man in der Psychologie immer nur mit (teilweise sehr kleinen) Auszügen aus einer Population arbeitet,

hegt man doch den Wunsch, die Ergebnisse aus der Stichprobe auf die gesamte Population zu verallgemeinern (man sagt auch: zu *generalisieren*).

Das ist ein großer Anspruch. Offensichtlich kann eine solche Generalisierung von Ergebnissen von einer Stichprobe auf eine Population nur dann sinnvoll gelingen, wenn die Personen in der Stichprobe in all ihren Eigenschaften den Personen entsprechen, die die Population ausmachen. Das heißt, die Personen in der Stichprobe sollten möglichst *repräsentativ* für die Population sein. Überspitzt formuliert würde es kaum Sinn machen, eine Fragestellung nur an Frauen zu untersuchen und anschließend das gefundene Ergebnis auf Männer zu verallgemeinern. Schließlich hätte die Studie bei Männern zu völlig anderen Ergebnissen führen können. Sind Stichproben kein repräsentatives Abbild der Population, so können wir unsere Ergebnisse nicht sinnvoll verallgemeinern. Stattdessen würden unsere Ergebnisse immer nur auf die „Art" von Personen zutreffen, die auch in der Stichprobe waren. Führen wir etwa eine Befragung per Post durch, bekommen wir meist nicht von allen angeschriebenen Personen eine Antwort. Es besteht also das Risiko, dass nur ganz bestimmte Personen auf die Umfrage antworten. Wenn z. B. nur extravertierte Personen antworten (weil sich introvertierte nicht trauen), dann hätten wir keine repräsentative, sondern eine sogenannte *selektive* Stichprobe vorliegen und könnten ein gefundenes Ergebnis streng genommen nur auf die Population von extravertierten Personen verallgemeinern. Die Gefahr, selektive Stichproben zu ziehen, besteht immer. Machen Sie sich deutlich, dass die Mehrzahl der Forschungsergebnisse in der Psychologie an Psychologiestudierenden gewonnen wurde und damit eigentlich gar nicht auf die Gesamtbevölkerung verallgemeinerbar ist! Wenn die Repräsentativität von Stichproben so wichtig ist, was können wir dann tun, um solche Stichproben zu bekommen? Die Antwort ist verblüffend einfach: wir ziehen die Leute für die Stichprobe *zufällig* aus der Population. Bei einer zufälligen Ziehung von Personen aus einer Population kommt uns der Zufall – siehe auch Abschn. 2.7 – dadurch zu Hilfe, dass er alle möglichen Merkmale und Besonderheiten, die Personen aufweisen können, zu gleichen Anteilen auch in unsere Stichprobe einbringt. Betrachten wir das Prinzip der Zufallsstichproben an Abb. 2.4.

Das Auswahlverfahren besteht im Ziehen einer Zufallsstichprobe. Ein einfaches Beispiel ist das Geschlecht. In der Population gibt es etwa gleich viele Männer wie Frauen. Der Zufall sollte dafür sorgen, dass in der Stichprobe der Anteil von Frauen und Männern ebenfalls 50:50 ist. Genauso verhält es sich mit allen anderen Merkmalen. So werden z. B. unterschiedlich intelligente Menschen, Menschen unterschiedlichen Alters, ledige und verheiratete Menschen, Gesunde und Kranke, Extravertierte und Introvertierte usw. in demselben Verhältnis in unserer Stichprobe auftauchen, wie sie auch in der Population vorliegen.

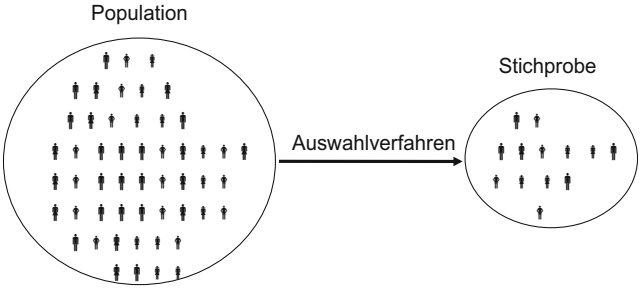

Abb. 2.4 Ziehen einer Stichprobe aus einer Population

Wenn wir also sichergehen wollten, dass in einer Studie mit Schulkindern diese tatsächlich repräsentativ sind für die Population aller Schulkinder, könnten wir nicht einfach in eine Schulklasse gehen, sondern müssten von allen deutschen Schülern eine zufällige Stichprobe ziehen. Sie sehen, dass das Ziehen von Zufallsstichproben mit ziemlich viel Aufwand verbunden sein kann. Daher wird vor allem in der Grundlagenforschung oft auf Zufallsstichproben verzichtet. Bei sehr anwendungsorientierten Studien sind Zufallsstichproben aber in der Regel unerlässlich, um verallgemeinerbare Ergebnisse zu erzielen. Ein häufig zitiertes Beispiel sind Wahlumfragen, bei denen man durch die Befragung einer kleinen Stichprobe eine Hochrechnung des Anteiles von Wählern verschiedener Parteien erhalten möchte. Hierbei ist das Verwenden einer Zufallsstichprobe so einfach wie effektiv. Die Population besteht hier aus den Stimmberechtigten einer ganzen Nation. Repräsentative Stichproben werden dabei durch eine Zufallsauswahl aus allen deutschen Haushalten gezogen. Oder aber, das Umfrageunternehmen stellt sich selbst einen repräsentativen Pool von Personen zusammen, deren in einer Datenbank registrierte Merkmale in der Stichprobe so verteilt werden, dass sie auch der Verteilung in der Population entsprechen. Bei einer so sorgfältig gezogenen repräsentativen Stichprobe ist es möglich, durch eine Umfrage an nur 2000 Personen eine ziemlich exakte Hochrechnung des Wahlergebnisses für über 60 Millionen Wahlberechtigte zu erhalten!

In der Psychologie ist es die Regel, dass man mit eher kleinen Stichproben arbeitet, teilweise mit 20–100 Versuchsteilnehmern. Damit läuft man Gefahr, dass ein Effekt, den wir in unserer Stichprobe gefunden haben, eventuell nur durch Zufall zustande kam. Das heißt, der Effekt könnte für unsere Stichprobe gelten, nicht aber für die Population. Um zu prüfen, wie gut wir aufgrund von Stichproben in der Lage sind, einen Effekt in der Population zu schätzen, brauchen wir statistische Methoden, die unter dem Begriff *Inferenzstatistik* zusammengefasst

werden (siehe Kap. 5). Sie können also schon im Hinterkopf behalten, dass die Inferenzstatistik die Verallgemeinerbarkeit von Ergebnissen aus Studien auf die Population prüft. Die deskriptive und die explorative Datenanalyse hingegen beziehen sich vor allem auf die Beschreibung und Analyse von Stichprobendaten, in die noch keine Überlegungen zur Generalisierbarkeit eingeflossen sind.

2.6 Methoden der Datenerhebung I: Befragungen und Beobachtungen

Die Kenntnisse zum Messen und Testen aus den vorangegangenen Abschnitten sind die Grundlage für die konkreten Methoden, mit denen man Daten erheben kann. Diesen Methoden – Befragen, Beobachten und Experiment – wollen wir uns jetzt zuwenden. Allen drei Methoden liegt die Idee des Messens zugrunde, und meist werden Fragebögen oder Tests verwendet. Während sich also Messen und Testen eher auf den theoretischen Aspekt der Datenerhebung beziehen, geht es beim Befragen, Beobachten und Experimentieren um die praktische Durchführung und um den Kontext, in dem die Datenerhebung stattfindet. Dem Experiment werden wir uns etwas ausführlicher zuwenden, da die Prinzipien beim Experimentieren einen unmittelbaren Einfluss auf die spätere statistische Auswertung der Daten haben.

Befragungen

Wenn es um die Untersuchung von Sachverhalten geht, die man einfach erfragen kann – wie die Erfassung von Einstellungen, Gewohnheiten, Persönlichkeitsmerkmalen usw. – dann ist die Befragung die entsprechende Methode der Datenerhebung. Befragungen kann man auf vielfältige Art und Weise gestalten und durchführen. Das Spektrum reicht vom Einholen einfacher Informationen (z. B. eine Befragung, wie gern jemand ein bestimmtes Produkt mag oder wie viel Geld er dafür bezahlen würde) bis hin zu formalen Befragungssituationen, in denen man konkrete Tests einsetzt, von denen wir oben gesprochen hatten.

Befragungen können mündlich oder schriftlich durchgeführt werden. Die mündliche Befragung hat in aller Regel die Form eines Interviews, bei der ein Interviewer entweder eine Person (Einzelinterview) oder gleich mehrere Personen (Gruppeninterview) befragt. Eine typische praktische Anwendung von Interviews sind Bewerbungssituationen. In der Forschung dagegen werden Interviews nur dort angewendet, wo man über ein bestimmtes Themengebiet noch wenig oder gar nichts weiß. In diesem Fall werden Interviews genutzt, um von den Befragten

interessante Ideen zu bekommen oder auf Aspekte zu stoßen, auf die man selbst nicht gekommen wäre. Sie können damit ein Hilfsmittel zur Generierung von Hypothesen oder Theorien sein.

Wenn allerdings die Fragen bzw. Aufgaben, die man untersuchen möchte, bereits feststehen – und das ist wie gesagt in der Forschung der häufigere Fall – so kann man auf die zeitintensive Durchführung von Interviews verzichten und statt dessen eine schriftliche Befragung einsetzen. Der Vorteil bei schriftlichen Befragungen ist, dass kein Interviewer anwesend sein muss und die Befragung daher an vielen Personen gleichzeitig und beispielsweise auch per Post oder im Internet durchgeführt werden kann. Ein Nachteil bei Befragungen per Post ist allerdings die sogenannte Rücklaufquote, also der Anteil von ausgefüllten Fragebögen, die der Forscher tatsächlich zurückerhält. Die Rücklaufquote ist meist eher gering (manchmal nur 30 %), und man weiß dann nicht, ob diejenigen Personen, die geantwortet haben, dies aus einem bestimmten Grund getan haben. Das heißt, man kann sich dann nicht mehr sicher sein, dass man mit den zurückerhaltenen Fragebögen eine repräsentative Stichprobe vorliegen hat.

Befragungen können mehr oder weniger *standardisiert* sein. Das bedeutet, dass die Durchführung entweder konkret festgelegt ist und beispielsweise die gestellten Fragen schon feststehen oder völlig offen ist und der Befragte im Prinzip frei assoziieren und berichten kann, was ihm zu einem bestimmten Thema einfällt. Wenig standardisierte Befragungen führen meist zu größeren Datenmengen (also längeren Texten) und einer Vielzahl unterschiedlichster Aussagen. Sie sind daher schwerer auszuwerten als stärker standardisierte Befragungen, bei denen sich die meisten Aussagen auf die konkreten, vorher festgelegten Fragen des Forschers beziehen.

Beobachtungen

Nicht immer ist es sinnvoll, zur Erhebung von Daten die entsprechenden Personen zu fragen, z. B. wenn es um Verhaltensweisen geht, die in einer konkreten Situation auftreten. Beispielsweise könnte ein Therapeut das Verhalten eines Patienten in sozialen Situationen unter die Lupe nehmen wollen. In einem solchen Fall wäre eine Befragung eher unzweckmäßig. Eine bessere Möglichkeit ist die Beobachtung von konkreten Situationen (also z. B. eine Situation, in der der Patient einen Fremden nach der Uhrzeit fragen soll). Der Beobachter kann das Verhalten der beobachteten Person bzw. Personen nach relevanten Verhaltensweisen, Äußerungen, nonverbalen Gesten usw. untersuchen, um Antworten auf bestimmte Fragen zu erhalten (z. B. ob sich der Patient freundlich gegenüber dem Fremden verhält).

Wenn es um eine komplexe Beobachtungssituation (mit vielen Fragestellungen oder mit vielen zu beobachtenden Personen) geht, ist es immer sinnvoll die

Beobachtung auf Video aufzuzeichnen. Die Auswertung von Beobachtungen, egal ob live oder per Videomaterial, gestaltet sich dabei ähnlich schwierig wie die Auswertung unstandardisierter Interviews. Der Beobachter muss das relevante Verhalten identifizieren, kategorisieren und versuchen, die für ihn entscheidenden Informationen zu extrahieren. Und oft ist gar nicht so klar, was genau eigentlich der Gegenstand der Beobachtung ist. Soll untersucht werden, was jemand sagt, wie viel und wie er es sagt, wie er dabei Blickkontakt mit seinem Gegenüber hält, welche Gesten er macht, welche Körperhaltung er einnimmt, oder gar alles zusammen? Es empfiehlt sich daher immer, das Ziel der Beobachtung vorher genau festzulegen und die Beobachtung genauestens zu protokollieren. Eine Videoaufzeichnung bietet sich auch dann an, wenn ein einzelner Beobachter mit einer live-Situation leicht überfordert sein könnte.

Beobachtungen können wiederum ganz unterschiedlich gestaltet sein. Der Beobachter kann Teil des beobachteten Geschehens sein (*teilnehmende* Beobachtung) oder außerhalb des Geschehens stehen (*nicht-teilnehmende* Beobachtung). Die Beobachteten können von der Befragung wissen (*offene* Beobachtung) oder sie werden nicht darüber informiert, dass es eine Beobachtung gibt (*verdeckte* Beobachtung). Und nicht zuletzt ist neben *Fremdbeobachtungen*, bei denen eine außenstehende Person andere Menschen beobachtet, die *Selbstbeobachtung* der eigenen Person möglich.

In den vergangenen Jahren haben mehr und mehr physiologische Messungen in die psychologische Forschung Einzug gehalten, darunter vor allem die Messung von Blickbewegungen, der Herzaktivität (EKG, Blutdruck) oder der Funktion und Struktur des Gehirns mit Hilfe bildgebender Verfahren. All diese Verfahren liefern ebenfalls Beobachtungsdaten, auch wenn es hier weniger der Forscher selbst ist, der beobachtet, als vielmehr sein Messgerät.

2.7 Methoden der Datenerhebung II: Experimente

Bei Beobachtungen und Befragungen ist ein wesentlicher Punkt im Verborgenen geblieben, der aber für psychologische Untersuchungen von zentraler Bedeutung ist: die Kausalität. Psychologen fragen oft nach den Ursachen für menschliches Verhalten und Erleben. Diese sind aber oft viel schwerer zu ermitteln, als man auf den ersten Blick meinen könnte. Der einzige Weg, um kausale Aussagen über Ursachen und Wirkungen treffen zu können, ist die Durchführung eines Experiments. Sehen wir uns an, worin genau das Problem mit der Kausalität besteht, und wenden uns dann dem Grundgedanken des Experiments zu.

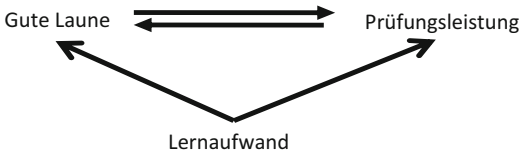

Abb. 2.5 Beispiel für Zusammenhänge von Variablen

Kausalität

Nehmen wir an, wir hätten beobachtet, dass Schüler mit guter Laune bessere Klausuren schreiben als schlechtgelaunte Schüler. Diese Beobachtung mag uns interessant erscheinen, aber was verbirgt sich eigentlich hinter ihr? Auf den ersten Blick würden wir wahrscheinlich sagen: Ist doch klar, gute Laune verbessert die Prüfungsleistungen, z. B. weil man sich bei besserer Laune mehr zutraut oder weil man konzentrierter ist. Das Problem bei dieser Interpretation ist aber, dass wir schlichtweg nicht wissen, ob sie stimmt. Es gibt nämlich auch andere Interpretationsmöglichkeiten, die auf Basis der vorliegenden Beobachtung möglich sind. Um genau zu sein, gibt es in jedem Fall drei mögliche Interpretationen, wenn zwei Variablen – so wie in unserem Beispiel – einen Zusammenhang aufweisen (siehe Abb. 2.5).

Die erste Möglichkeit hatten wir bereits formuliert: gute Laune könnte die Ursache für bessere Prüfungsleistungen sein. Die zweite Möglichkeit geht in die entgegengesetzte Richtung: Schüler, die generell bessere Noten haben, könnten deswegen generell auch eher gute Laune haben. Und schließlich gibt es noch eine dritte Möglichkeit: es könnte eine dritte Variable geben, die den Zusammenhang von guter Laune und Prüfungsleistung hervorgerufen hat. In unserem Beispiel könnte dies die Variable Lernaufwand sein. Schüler, die einen größeren Lernaufwand betreiben, könnten sich durch diese Anstrengung besser fühlen, und gleichzeitig würde der höhere Lernaufwand zu besseren Prüfungsleistungen führen. Gute Laune und Prüfungsleistungen hätten dann überhaupt keine direkte Verbindung – sie wären *kausal unabhängig* voneinander.

► Kausalität beschreibt die Ursache-Wirkungs-Beziehung zweier Ereignisse oder Variablen. Dafür sind ein zeitliches Nacheinander von Ursache und Wirkung und der Ausschluss alternativer Erklärungen unverzichtbare Voraussetzungen.

Es kann natürlich Beobachtungen geben, bei denen die Richtung der Kausalität klar ist. So ist die Straße nass (Wirkung), weil es vorher geregnet hat (Ursache) und nicht umgekehrt. Höheres Alter ist die Ursache für mehr Erfahrungswissen und nicht umgekehrt. Aus diesen Beispielen können wir die allgemeinen Kriterien

ableiten, die für Kausalität erfüllt sein müssen: A verursacht B kausal, wenn (1) A zeitlich *vor* B auftritt, (2) A und B „kovariieren" (eine Veränderung von A mit einer Veränderung von B einhergeht) und (3) der Einfluss von Drittvariablen (Alternativerklärungen) ausgeschlossen werden kann.

Diese Kriterien klingen vielleicht ziemlich theoretisch, sie sind aber praktisch sehr einleuchtend. Nehmen wir an, in unserem Beispiel ist Möglichkeit 1 die zutreffende (gute Laune verursacht bessere Prüfungsleistungen). Diese Aussage können wir nur mit Sicherheit machen, wenn (1) die gute Laune vor der Prüfung da war, (2) gute Laune zu guten und schlechte Laune zu schlechteren Prüfungsleistungen führt und (3) und es keine Drittvariablen gibt, die den Zusammenhang erklären könnten.

In den meisten Fällen wissen wir all diese Dinge nicht und können daher durch die bloße Beobachtung von Variablen noch nichts über ihre Kausalität sagen. Wie in jeder Wissenschaft ist es aber auch in der Psychologie das höchste Ziel, Kausalaussagen über den Zusammenhang von Variablen zu treffen. Noch genauer: meist sind wir an den Ursachen von bestimmten Variablen interessiert. Wie aber können wir es methodisch anstellen, etwas über die Kausalitätsrichtung zu erfahren? Hier kommt eine einfache wie geniale Methode ins Spiel: das Experiment.

Die Idee des Experiments

Machen wir zunächst ein Gedankenexperiment (im wahrsten Sinne des Wortes). Stellen Sie sich vor, Sie sind ein Forscher, der den Zusammenhang der Variablen in unserem Beispiel untersuchen möchte. Sie haben die Hypothese, dass gute Laune die Ursache für bessere Prüfungsleistungen ist. Wie könnten Sie vorgehen? Sagen wir, Sie haben 20 Schüler einer Schulklasse zur Verfügung, mit denen Sie einen Test schreiben können. Laut unserer Definition von Kausalität müssen Sie zuerst sicherstellen, dass die gute Laune *vor* der Prüfungssituation auftritt. Das könnten Sie tun, indem Sie über einen Fragebogen bei jedem Schüler seine aktuelle Laune ermitteln, bevor Sie den Test schreiben. Zweitens sollten Schüler mit besserer Laune bessere Testergebnisse haben und Schüler mit schlechterer Laune schlechtere Ergebnisse (*Kovariation*). Hier kommt eine zentrale Idee des Experimentes ins Spiel: Sie müssen die Laune in irgendeiner Art und Weise *variieren,* um dieses Kriterium zu prüfen. Wenn Sie Glück haben, gibt es in der Klasse bereits Schüler mit guter und Schüler mit schlechter Laune. Wenn Sie Pech haben, sind alle Schüler schlecht gelaunt. Sie müssen daher bei einem Teil der Schüler dafür sorgen, dass sie bessere Laune haben. Das könnten Sie tun, indem Sie diesen Schülern einen kurzen lustigen Film zeigen. Danach müssten Sie mithilfe des Fragebogens prüfen, ob diese Manipulation geklappt hat und ein Teil der Schüler jetzt wirklich besser gelaunt ist. Sie können nun prüfen, ob die gutgelaunten

Schüler tatsächlich bessere Noten im Test erreichen. Ist das der Fall, besteht Ihre letzte Aufgabe im *Ausschließen von Alternativerklärungen*. Sie müssen zeigen, dass der Zusammenhang zwischen guter Laune und Testergebnis nicht durch eine andere Variable hervorgerufen wurde. Dafür müssen Sie sich überlegen, welche Variablen hier in Frage kommen. Oben hatten wir gesagt, dass beispielsweise der Lernaufwand vor dem Test sowohl gute Laune als auch bessere Prüfungsleistungen bewirken könnte. Wie könnten Sie das prüfen? Anders ausgedrückt: wie könnten Sie den Einfluss des Lernaufwandes „ausschalten"? Zunächst müssen Sie den Lernaufwand jedes Schülers erfassen. Das könnten Sie wieder mit einem Fragebogen tun. Was aber, wenn alle Schüler, die von guter Laune berichten, auch mehr gelernt haben? Dann stehen Sie vor einem Problem und kommen nicht weiter. Sie müssten stattdessen dafür sorgen, dass Schüler mit verschieden großem Lernaufwand sowohl in der Gruppe von gutgelaunten als auch in der Gruppe von schlechtgelaunten Schülern vorkommen. Wenn sich die Gruppen dann immer noch in ihrem Testergebnis unterscheiden, dann wissen Sie, dass das nicht mehr am Lernaufwand liegen kann, da der jetzt in beiden Gruppen gleich ist – man sagt, er ist *konstant gehalten*. Um das zu bewerkstelligen, könnten Sie nun eine Art Trick anwenden und sich der Methode von oben bedienen: Sie teilen die Klasse zuerst in zwei Hälften, in denen sich jeweils Schüler mit durchschnittlich gleich hohem Lernaufwand befinden. Dann hätten Sie in diesen beiden Gruppen den Lernaufwand konstant gehalten. Und nun der „Trick": da Sie in der einen Gruppe ja Schüler mit guter und in der anderen Gruppe Schüler mit schlechter Laune haben wollten, müssen Sie mit Hilfe des lustigen Filmes gute Laune in der einen Hälfte hervorrufen. Da sich in der anderen (der schlechtgelaunten) Gruppe eventuell auch ein paar Leute mit guter Laune befinden werden, können Sie die gleiche Methode anwenden und mit Hilfe eines unangenehmen oder langweiligen Filmes alle Schüler dieser Gruppe in schlechte Laune versetzen. Nun schreiben Sie den Test. Wenn die gutgelaunten Schüler bessere Leistungen erzielen als die schlechtgelaunten, können Sie nun mit großer Sicherheit sagen, dass die gute Laune tatsächlich die Ursache für den Prüfungserfolg war. Sie haben ein echtes Experiment durchgeführt.

An diesem einfachen Beispiel haben wir gesehen, welche Grundidee dem Experiment zugrunde liegt.

► Experimente sind künstliche Eingriffe in die natürliche Welt mit dem Ziel systematische Veränderungen in einer unabhängigen Variable (UV) hervorzurufen, die ursächlich zu einer Veränderung in einer abhängigen Variable (AV) führen. Alternativerklärungen werden dabei ausgeschlossen.

An dieser Definition wird der Unterschied zwischen Beobachtungen und Befragungen auf der einen Seite und Experimenten auf der anderen Seite deutlich: Experimente begnügen sich nicht mit dem Gegebenen, sondern sie stellen sozusagen eine bestimmte „Wirklichkeit" gezielt und künstlich her. In unserem Gedankenexperiment haben Sie z. B. gute und schlechte Laune durch einen Eingriff (den Film) einfach hergestellt oder induziert. Das Entscheidende dabei ist, dass die Variable, die uns als potenzielle Ursache einer anderen Variable interessiert, *systematisch variiert* wird. Wenn sie wirklich die Ursache der anderen Variable ist, muss diese systematische Variation zu einer Veränderung in dieser Variable führen. Diese Art von Kausalitätsprüfung ist beim Beobachten und Befragen nicht möglich. Das Experiment wird daher oft als „Königsweg" der Datenerhebung bezeichnet. Wenn es um das Aufdecken von Ursache-Wirkungs-Beziehungen geht, ist das Experiment meist die einzige Möglichkeit.

Das Experiment hat aber noch einen anderen großen Vorteil. Beim Experimentieren können wir sämtliche Bedingungen, die das Experiment stören könnten, selbst ausschalten oder kontrollieren. Man spricht dabei auch vom Ausschalten oder Kontrollieren von *Störvariablen*, denen wir uns jetzt zuwenden wollen.

Störvariablen

In unserem Gedankenexperiment hatten wir versucht, die Alternativerklärung – dass der Lernaufwand ebenfalls eine Ursache für unterschiedliche Prüfungsleistungen sein kann – auszuschließen. Das mussten wir deswegen tun, weil wir sonst nicht zweifelsfrei hätten behaupten können, dass gute Laune die kausale Ursache für bessere Prüfungsleistung ist. Wir mussten also sicherstellen, dass die Beziehung zwischen den beiden Variablen nicht durch eine dritte Variable (den Lernaufwand) *gestört* wird.

► Störvariablen sind Merkmale der Person oder der Situation, die eventuell ebenfalls die abhängige Variable (AV) beeinflussen. Ihr Effekt soll im Experiment ausgeschaltet werden, weil sie den Effekt der unabhängigen Variable (UV) stören könnten. Man spricht dabei auch von *experimenteller Kontrolle* von Störvariablen.

Konstanthalten und Parallelisieren

Wir hatten versucht, diesen störenden Effekt dadurch auszuschalten, dass wir verschieden hohen Lernaufwand gleichmäßig auf die beiden Gruppen aufgeteilt haben, in denen wir später gute bzw. schlechte Laune induziert hatten. Dieses *Konstanthalten*, wie wir es genannt hatten, sorgt dafür, dass sich die Gruppen hinsichtlich des Merkmals Lernaufwand nicht mehr unterscheiden. Folglich kann unterschiedlich hoher Lernaufwand nicht mehr die Ursache für unterschiedliche

Prüfungsleistungen zwischen unseren beiden Gruppen sein. Da man die unterschiedlichen Ausprägungen der Störvariable sozusagen parallel auf die beiden Gruppen aufgeteilt hat, spricht man anstelle vom Konstanthalten der Störvariablen auch oft vom *Parallelisieren* der Gruppen hinsichtlich der Störvariablen.

Das Konstanthalten von potenziellen Störvariablen ist schon eine gute und einfache Lösung von experimenteller Kontrolle. Leider kann es aber zwei Probleme geben, die das Konstanthalten von Störvariablen unmöglich machen.

Das erste Problem tritt auf, wenn es zu viele potenzielle Störvariablen gibt. Es könnte z. B. sein, dass in unserer Schulklasse die Mädchen generell bessere Prüfungsleistungen erbringen als die Jungen. Nun könnte es passieren, dass wir fast alle Mädchen in die gute-Laune-Gruppe getan haben und die meisten Jungen in die schlechte-Laune-Gruppe, oder umgekehrt. Das würde offensichtlich dazu führen, dass unterschiedliche Prüfungsleistungen in beiden Gruppen jetzt genauso gut auf das Merkmal Geschlecht zurückgeführt werden könnten und nicht unbedingt auf unsere Manipulation (gute versus schlechte Laune). Wir müssten nun also – zusätzlich zum Lernaufwand – auch noch das Geschlecht konstanthalten, indem wir den Anteil von Jungen zu Mädchen in beiden Gruppen gleich verteilen. Eine weitere Störvariable könnte aber auch noch die Intelligenz sein. Es ist sogar sehr wahrscheinlich, dass intelligentere Schüler bessere Prüfungsleistungen erzielen. Wir müssten also das Merkmal Intelligenz ebenfalls konstanthalten. An dieser Stelle wird deutlich, dass der Aufwand der experimentellen Kontrolle schnell anwächst, wenn die Anzahl potenzieller Störvariablen steigt. Es kann sogar sein, dass es technisch unmöglich wird, all diese Störvariablen gleich auf die beiden Gruppen zu verteilen – vor allem, wenn man nur 20 Personen zur Verfügung hat (was in Experimenten häufig der Fall ist). In den meisten Fällen wird es so sein, dass es nicht nur eine potenzielle Störvariable gibt. Es gibt Merkmale, die so gut wie immer als Störvariablen betrachtet werden, da von ihnen bekannt ist, dass sie auf fast alle abhängigen Variablen einen Effekt ausüben: darunter Alter, Geschlecht und Intelligenz.

Bevor wir zu einer Lösung dieses Problems kommen, sehen wir uns noch das zweite Problem beim Konstanthalten an, das noch verzwickter ist als das erste. Bisher hatten wir überlegt, wie wir die potenziellen Störvariablen gleichmäßig auf unsere Gruppen aufteilen. Das setzt allerdings voraus, dass wir diese Störvariablen auch kennen! Bei einer Vielzahl von Fragestellungen wissen wir schlichtweg nicht, welche möglichen Störvariablen es geben könnte. Folglich sind wir auch nicht in der Lage, die Gruppen im Experiment hinsichtlich der Störvariablen zu parallelisieren. Wie könnten wir es dennoch schaffen, dass alle potenziellen Störvariablen gleich auf die beiden Gruppen verteilt werden?

Randomisierung

Hier kommt uns eine der wichtigsten Techniken zu Hilfe, die es bei der Durchführung von Studien gibt: die *Randomisierung*. Das englische Wort random bedeutet zufällig.

► Bei der Randomisierung werden die Versuchspersonen zufällig den verschiedenen Versuchsbedingungen (den Gruppen des Experimentes) zugeteilt.

Die Versuchspersonen sind in unserem Beispiel die Schüler. Sie sollen nun nach dieser Definition zufällig (z. B. durch Lose) auf die beiden Gruppen aufgeteilt werden, in denen wir später gute bzw. schlechte Laune induzieren wollen. Aber wie löst dieses Vorgehen unsere beiden Probleme? Ganz einfach: Alle potenziellen Störvariablen – und zwar auch solche, die wir gar nicht kennen – werden durch den Zufall gleichmäßig auf beide Gruppen verteilt. Konkret heißt das, dass bei einer zufälligen Zuordnung der 20 Schüler in zwei Gruppen in beiden Gruppen gleich viele Schüler mit hohem und niedrigem Lernaufwand, gleich viele Jungen und Mädchen, sowie gleich viele intelligentere und weniger intelligente Schüler vorkommen. Das Gleiche passiert auch mit allen anderen Merkmalen, die wir gar nicht kennen. Wir müssen uns also gar nicht überlegen, welche Störvariablen es geben könnte, sondern wir überlassen dem Zufall die Arbeit, der für eine mehr oder weniger perfekte Parallelisierung sorgt. Natürlich werden per Zufall nicht immer *genau* gleich viele Jungen und Mädchen oder *genau* gleich viele intelligentere und weniger intelligente Schüler in die beiden Gruppen gelangen. Aber eine ungefähre Gleichverteilung reicht schon aus, um den Effekt der Störvariablen zu kontrollieren. Wichtig dabei ist, dass die Stichprobe ausreichend groß ist, denn sonst können die „ausgleichenden Kräfte des Zufalls“ nicht richtig wirken (siehe Abschn. 3.6).

Sie sollten die Technik der Randomisierung gut im Hinterkopf behalten, da sie das wichtigste Grundprinzip für das Durchführen experimenteller Studien ist und oft auch eine Art Gütesiegel für methodisch korrekt durchgeführte Studien darstellt. In Abb. 2.6 ist der gesamte Ablauf beim Vorgehen unseres Experimentes noch einmal dargestellt.

Quasiexperimente

In unserem Schulklassen-Beispiel ist es kein Problem gewesen, zunächst zwei Gruppen von Schülern zufällig zu ziehen und danach das uns interessierende Merkmal (gute bzw. schlechte Laune) zu induzieren. Nun kann es allerdings auch Fälle geben, in denen es nicht möglich ist, das relevante Merkmal selbst zu beeinflussen. Nehmen wir an, wir wollen untersuchen, ob Menschen, die rauchen,

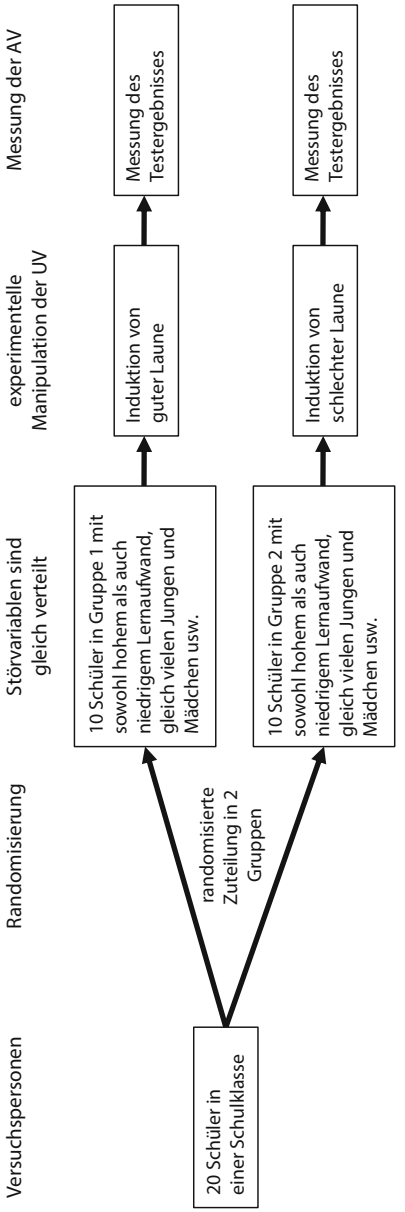

Abb. 2.6 Überblick über das experimentelle Vorgehen für die Beispielstudie

auch mehr Geld für Alkohol ausgeben als Menschen, die nicht rauchen. In diesem Fall hätten wir als unabhängige Variable wieder zwei Gruppen, nämlich Raucher und Nichtraucher (die abhängige Variable wäre der Geldbetrag für gekauften Alkohol). Wie man aber schon sehen kann, sind wir hier nicht in der Lage, das Merkmal Raucher/Nichtraucher einfach zu induzieren. (Streng genommen könnten wir natürlich wieder per Zufall zwei Gruppen von Leuten zusammenstellen und der einen Gruppe sagen, sie soll pro Tag 10 Zigaretten rauchen, während die andere Gruppe nicht rauchen darf. Aber ein solches Vorgehen verstößt offensichtlich gegen jegliche Forschungsethik und ist ausgeschlossen.) Stattdessen müssen wir uns wohl damit begnügen, die Gruppe von Rauchern und die Gruppe von Nichtrauchern so zu nehmen, wie sie sind. Das hat aber wiederum zur Folge, dass wir nicht sicher sein können, dass es keine Störvariablen gibt, in denen sich die beiden Gruppen unterscheiden. Da wir keine Randomisierung vornehmen können, sind wir daher wieder auf das Konstanthalten möglicher Störvariablen angewiesen. Wir müssten also wieder nach potenziellen Störvariablen schauen und versuchen, jeweils Raucher und Nichtraucher zu finden, für die alle Störvariablen gleich ausgeprägt sind. Sie sehen aber schon, dass wir auf diese Weise nicht in der Lage sind, alle Störvariablen mit Sicherheit auszuschalten. Man kann daher bei solchen Untersuchungen streng genommen nicht von Experimenten sprechen, da diese das Ausschalten von Störvariablen verlangen. Deshalb werden solche Arten von Untersuchungen *Quasiexperimente* genannt – im Gegensatz zu den *echten Experimenten*, von denen wir bisher gesprochen haben.

► Echte Experimente setzen das randomisierte Aufteilen von Versuchspersonen auf die Versuchsbedingungen voraus. Ist die Gruppeneinteilung jedoch von Natur aus vorgegeben und daher keine Randomisierung möglich, spricht man von Quasiexperimenten.

In der Grundlagenforschung sind die interessierenden unabhängigen Variablen meist manipulierbar bzw. induzierbar. Je anwendungsbezogener die Fragestellungen werden, desto eher hat man es mit Variablen zu tun, die schon vorgegeben sind und die man daher nur quasiexperimentell untersuchen kann. Ein häufiges Beispiel sind Untersuchungen, bei denen Männer und Frauen verglichen werden. Auch hier ist die Gruppeneinteilung vorgegeben. Entsprechend müssen alle Störvariablen parallelisiert werden. Manchmal kann es vorkommen, dass sich Störvariablen nicht vollständig parallelisieren lassen. Wenn beispielsweise in einer Untersuchung an Männern und Frauen die Aggressivität als Störvariable berücksichtigt werden soll, kann es schwierig sein, das Aggressionslevel in beiden Gruppen gleich zu

verteilen, wenn Männer im Durchschnitt aggressiver sind als Frauen. Diesen Unterschied muss man vorerst in Kauf nehmen. Es ist aber in jedem Fall sinnvoll, die Ausprägung aller möglichen Störvariablen in der Untersuchung mit zu erheben und zu dokumentieren.

Gütekriterien bei Experimenten

Wie wir gelernt haben, sind Experimente eine unverzichtbare Methode, um Kausalitäten auf den Grund zu gehen. Aus den Erläuterungen sollte aber auch hervorgegangen sein, dass beim Experimentieren immer wieder Schwierigkeiten auftreten und man viele Fehler machen kann. Die sogenannten Gütekriterien dienen der Beurteilung der Qualität eines Experiments.

Das erste Gütekriterium wird als *interne Validität* eines Experiments bezeichnet. Wir hatten gefordert, dass durch Randomisieren bzw. Parallelisieren die Effekte potenzieller Störvariablen ausgeschaltet werden sollen. Wenn wir das geschafft haben, können wir sicher sein, dass ein Effekt in der AV auch tatsächlich auf die Veränderung der UV zurückgeht.

► Interne Validität liegt vor, wenn die Veränderung in der AV *eindeutig* auf die Veränderung in der UV zurückgeführt werden kann.

Wenn wir in einem intern validen Experiment einen Effekt gefunden haben, bleibt noch die Frage offen: Können wir dieses Ergebnis verallgemeinern? Das Ziel von Studien ist es immer, eine generelle Aussage über die Wirkung von Manipulationen zu treffen. Mit anderen Worten: die Ergebnisse, die anhand einer Stichprobe von Versuchsteilnehmern gewonnen wurden, sollen nicht nur für die untersuchte Stichprobe gelten, sondern auf die Allgemeinheit übertragen – man sagt auch *generalisiert* – werden. Mit Allgemeinheit ist dabei die jeweilige Gruppe von Personen gemeint, über die man eine Aussage treffen möchte (auch *Population* genannt, siehe Abschn. 2.5). In unserem Schulklassen-Beispiel könnte die relevante Population aus allen Schülerinnen und Schülern bestehen. Wenn wir in einer Studie mit Hilfe einer repräsentativen Stichprobe ein auf die Population verallgemeinerbares Ergebnis gefunden haben, dann sprechen wir von einer *extern validen* Studie.

► Externe Validität liegt vor, wenn das in einer Stichprobe gefundene Ergebnis auf andere Personen bzw. auf die Population verallgemeinerbar ist. Sie wird durch repräsentative Stichproben erreicht, die am einfachsten durch eine zufällige Ziehung der Stichprobenmitglieder zustande kommen.

Literaturempfehlung
Huber, O. (2005). *Das psychologische Experiment: Eine Einführung* (4. Aufl.). Bern: Huber.

Der Zusammenhang der Methoden der Datenerhebung
Bevor wir dieses Kapitel abschließen, soll noch etwas zum Zusammenhang der verschiedenen Methoden der Datenerhebung gesagt werden. Sicher ist Ihnen aufgefallen, dass wir der Beschreibung des Experimentes sehr viel Raum geschenkt haben. Das hat zwei Gründe. Zum einen ist das Experiment – wie wir gesehen haben – der Königsweg der Datenerhebung. Wann immer möglich, sollte man sich für die Durchführung eines Experimentes entscheiden, weil nur mit dieser Methode das Aufdecken von kausalen Zusammenhängen möglich ist. Zum anderen *beinhaltet* das Experiment meist die anderen Methoden – Beobachtung und Befragung. Zur Messung des Effektes in Experimenten werden fast immer Tests oder Fragebögen eingesetzt. Auch kann das Verhalten der Versuchsteilnehmer durch Beobachtung erfasst werden. Und die erwähnten biopsychologischen Messungen wie EKG oder Hirnscan stellen ebenfalls Beobachtungen dar.