2 Struktur und Ablauf wissenschaftlicher Untersuchungen

Was Sie in diesem Kapitel lernen

- Was ist eine wissenschaftliche Hypothese, und wie testet man sie?
- Hat Sigmund Freud jemals einen Abwehrmechanismus gesehen?
- Wie lässt sich wissenschaftlich ermitteln, ob eine Person tabuisierte sexuelle Wünsche verdrängt hat?
- Wie könnte man wissenschaftlich herausfinden, welche Bedeutung das Aufschlitzen von Sitzen in öffentlichen Verkehrsmitteln hat?

Um die Bedeutung von Forschungsmethoden für die Psychologie besser zu verstehen, ist es hilfreich, sie in den Gesamtprozess von Wissenschaft einzuordnen. Wir wollen diese Einordnung hier in zweierlei Hinsicht vornehmen: Zunächst werden wir uns die verschiedenen Ebenen, auf denen wissenschaftliche Aussagen getroffen werden, und die Beziehungen, die zwischen diesen Ebenen bestehen, etwas genauer ansehen. In einem zweiten Schritt werden wir den Ablauf einer typischen wissenschaftlichen Untersuchung nachzeichnen und dabei herausarbeiten, an welchen Stellen welche Methoden in welcher Funktion angewendet werden.

2.1 Hypothesen, Ebenen wissenschaftlicher Aussagen und die Überbrückungsproblematik

Jede Erfahrungswissenschaft trifft, bearbeitet und testet Aussagen auf unterschiedlichen Ebenen. Auf der theoretisch-konzeptionellen Ebene werden Aussagen über Objekte und Ereignisse (in theoretischen Begriffen) sowie die Beziehungen zwischen diesen Objekten und Ereignissen (in Form theoretischer Zusammenhänge) getroffen. Eine theoretische Aussage könnte z. B. lauten: »Wer intelligent ist, der ist auch kreativ.« Intelligenz und Kreativität sind theoretische Begriffe. Die Aussage stellt eine Behauptung dar, deren Wahrheitsgehalt empirisch geprüft werden kann.

2.1.1 Prüfbare und nicht-prüfbare Aussagen

Der Begriff »Empirie« stammt vom griechischen Substantiv »empeiria« und bedeutet Erfahrung im Sinne von sinnlicher Wahrnehmung. Nicht alle Aussagen sind empirisch prüfbar. Manche Aussagen sind praktisch nicht prüfbar. Die Aussage »In der Hölle ist es heißer als 220 Grad« ist nicht prüfbar, da es sich bei der Hölle um einen nicht ohne Weiteres zugänglichen Ort (viele Menschen würden sogar sagen: gar keinen real existierenden Ort) handelt. Mit einem Thermometer in der Hölle Forschung zu betreiben ist also nicht möglich. Etwas technischer gesprochen: Die thermischen Zustände in der Hölle sind nicht erfahrbar. Erfahrbarkeit ist jedoch in den empirischen Wissenschaften eine notwendige Grundvoraussetzung für die Überprüfung theoretischer Aussagen.

Ebenfalls nicht empirisch prüfbar sind Aussagen, die per Definition richtig sein *müssen*. Unverheiratete Männer bezeichnet man als Junggesellen. Empirisch zu prüfen, ob Junggesellen verheiratet sind, ist nicht möglich: Wenn die Definition richtig angewendet wird, kann es keine verheirateten Junggesellen geben. Empirisch nicht prüfbar sind weiterhin Aussagen, die nicht falsch sein können. Dazu gehören Möglichkeitssätze wie »Frauen können schwanger werden«.

Falsifizierbarkeit. Die Feststellung, dass manche Aussagen nicht prüfbar sind, weil sie immer richtig sind, lässt schon erahnen, worauf es bei der Prüfung einer theoretischen Aussage ankommt: auf ihre Falsifizierbarkeit. Eine Aussage muss prinzipiell falsch sein *können,* damit sie empirisch prüfbar ist. Ergibt die Prüfung, dass sie tatsächlich falsch ist, dann hat man ein eindeutiges Ergebnis. Solange jedoch eine Aussage nicht falsifiziert werden kann, muss sie als vorläufig gültig betrachtet werden. Auf die Einschränkung »vorläufig« darf nur dann verzichtet werden, wenn alle Möglichkeiten zur Falsifizierung der Aussage ausgeschöpft sind. Manche Aussagen sind praktisch nicht falsifizierbar. Nehmen wir die Aussage: »Es gibt Menschen, die dreimal am Tag warm essen und trotzdem kein Gramm zunehmen.« Man könnte diese Aussage nur dann falsifizieren, wenn

man nachweisen könnte, dass kein Mensch auf der Welt bei dreimaligem Essen am Tag auch nur ein Gramm zunimmt. Eine solche Form der Falsifizierung ist praktisch unmöglich, weil man dazu alle Menschen, die gegenwärtig auf unserer Welt leben, untersuchen müsste.

2.1.2 Wissenschaftliche Hypothesen

Ein zentrales wissenschaftstheoretisches Konzept ist die Hypothese. Der Begriff »Hypothese« stammt ebenfalls aus dem Griechischen (hypóthesis) und bedeutet Unterstellung, Voraussetzung oder Grundlage. Hypothesen sind Aussagen, die empirisch testbar und somit falsifizierbar sein müssen. Darüber hinaus kommt einer testbaren Aussage nur dann der Rang einer wissenschaftlichen Hypothese zu, wenn sie eine gewisse Allgemeingültigkeit beansprucht, also über den Einzelfall hinausweist. Die Aussage »Peter wird morgen Kopfschmerzen haben, wenn er weiter so viel Alkohol trinkt« ist also noch keine wissenschaftliche Hypothese. Allerdings verbirgt sich hinter dieser Aussage eine (vermutete) allgemeine Gesetzmäßigkeit, nämlich die, dass übermäßiger Alkoholgenuss am Abend im Allgemeinen zu Kopfweh am Morgen führt. Bei dieser Aussage handelt es sich dann um eine wissenschaftliche Hypothese.

Darüber hinaus muss eine wissenschaftliche Hypothese begründet sein. Die Aussage »Wer sich die Haare rot färbt, isst mehr Fisch« wäre zwar testbar und hat eine ausreichende Allgemeingültigkeit, aber man würde dieser Aussage nicht den Rang einer wissenschaftlichen Hypothese zusprechen, solange sie nicht begründet ist.

Definition

Bei einer Aussage handelt es sich um eine wissenschaftliche Hypothese, wenn sie prinzipiell der sinnlichen Erfahrung zugänglich ist, prinzipiell widerlegbar ist, eine gewisse Allgemeingültigkeit beansprucht und theoretisch begründet ist.

Empirische Ebene

Um eine theoretische Aussage bzw. eine wissenschaftliche Hypothese empirisch zu testen, muss man die theoretischen Konzepte und die theoretischen Relationen zwischen diesen Konzepten in konkrete Aussagen überführen, deren Konzepte und Relationen empirisch erfahrbar sind. Dies erfordert einen nicht immer einfachen Übersetzungsprozess, den wir an einem prominenten Beispiel veranschaulichen wollen: dem Konzept des Abwehrmechanismus, einem zentralen Bestandteil der psychoanalytischen Theorie des österreichischen Arztes und Tiefenpsychologen Sigmund Freud (1856–1939).

Theoretische Hypothesen

Abwehrmechanismus ist ein theoretischer Begriff. Niemand hat jemals einen Abwehrmechanismus gesehen, auch Freud nicht. Das Konzept des Abwehrmechanismus ist ein gedankliches Gebilde, eine Schöpfung des Freud'schen Geistes, ein hypothetisches Konstrukt. Freud hat dieses Konstrukt geschaffen, um Beobachtungen, die er in seiner Praxis an Klienten gemacht hat, zu deuten, ihnen einen psychologischen Sinn zu geben.

Freud stellte in seiner praktischen Arbeit häufig fest, dass seine Klienten bestimmte Episoden aus ihrem Leben vergessen hatten und sich erst im Prozess der Psychoanalyse wieder daran erinnerten. In diesen Episoden kamen meistens Handlungen, Vorstellungen, Wünsche oder Fantasien mit sexuellen Inhalten vor. Freud nahm an, dass sexuelle Themen, insbesondere bestimmte sexuelle Wunschvorstellungen (z. B. die eines Knaben, mit der eigenen Mutter erotischen Kontakt zu haben), der Person Angst machen, weil sie tabuisiert sind und sanktioniert werden. So habe etwa der Knabe, der die eigene Mutter begehrt, Angst, vom Vater kastriert zu werden. Abwehrmechanismen dienen nun nach Freuds theoretischen Vorstellungen dazu, solche Ängste zu bewältigen. Eine Möglichkeit besteht darin, die sexuellen Wunschvorstellungen zu verdrängen. Verdrängung ist einer von mehreren Abwehrmechanismen und selbst wiederum ein theoretischer Begriff. Die Verdrängung leistet eine Verschiebung der tabuisierten Wunschvorstellung ins Unbewusste. Auch beim Unbewussten handelt es sich um einen theoretischen Begriff. War die Verdrängung erfolgreich, wird die tabuisierte Wunschvorstellung vergessen und kann erst wieder mit den Mitteln der psychoanalytischen Behandlung (z. B. der freien Assoziation oder der Traumdeutung) der bewussten Erinnerung zugänglich gemacht werden.

Aus diesem Teil der Theorie Freuds lassen sich zahlreiche theoretische Hypothesen ableiten. So sollten tabuisierte Wunschvorstellungen umso wahrscheinlicher verdrängt werden, je stärker sie sind. Daraus folgt, dass der Aufwand an psychoanalytischer Arbeit, der erfor-

derlich ist, um eine tabuisierte Wunschvorstellung in Erinnerung zu rufen, mit der Stärke dieser Wunschvorstellung steigt. Vermittelt werden diese Zusammenhänge theoretisch durch die Angst vor Sanktionen, die bei starken Tabuwünschen größer sein sollte als bei schwachen.

2.1.3 Überbrückungsprobleme

Theorie und Empirie

Um diese theoretischen Hypothesen empirisch prüfen zu können, müssen die Bestandteile der Hypothesen konkretisiert werden. Die Hypothese kann also nur bei konkreten Personen, mit konkreten Tabuwünschen, konkreten Formen der psychoanalytischen Behandlung und einer konkreten Form des Aufwands psychoanalytischer Arbeit angewendet werden. Diese Konkretisierungen sind aber kein Bestandteil der Theorie und lassen sich auch nicht unmittelbar aus den theoretischen Sätzen ableiten. Vielmehr handelt es sich um einen Zuordnungsprozess, in den viele Überlegungen und Entscheidungen einfließen, die außerhalb der Theorie stehen. So sagt die Theorie nichts darüber aus, bei welchen Personen welche Wünsche tabu sind, welche Personen unter welchen Umständen von welchem Mechanismus zur Abwehr ihrer Angst Gebrauch machen und welches Element der psychoanalytischen Theorie bei welchem Tabuthema, bei welchem Klienten, bei welchem Abwehrmechanismus wie effektiv ist.

Eine Übertragung der theoretischen Hypothese in konkrete empirische Hypothesen ist also mit Unsicherheiten behaftet. Die Theorie lässt ihrer Anwendung auf die Wirklichkeit Spielraum und zwingt den Anwender ebenso wie den Forscher, der die Theorie überprüfen möchte, zu Festlegungen, die bis zu einem gewissen Grade beliebig sind. Dies erkennt man u. a. daran, dass unterschiedliche Forscher die gleiche Theorie unterschiedlich anwenden und unterschiedliche empirische Beobachtungen für geeignet halten, die Theorie zu prüfen.

Nehmen wir einmal an, ein Forscher habe sich zur empirischen Prüfung der fraglichen Hypothese entschlossen und die unvermeidlichen Festlegungen getroffen. Beispielsweise habe er sich entschlossen, die Theorie am Beispiel des tabuisierten Wunsches von Knaben nach einer erotischen Beziehung zur eigenen Mutter zu prüfen und die Stärke der Verdrängung über den psychotherapeutischen Aufwand zu bestimmen, der erforderlich ist, um die Erinnerung an das erotische Begehren wieder herzustellen. Nach dieser Festlegung muss sich der Forscher auf die Suche nach Therapeuten machen, die bereit sind, ihm Klienten zu vermitteln, bei denen sich die Hypothese untersuchen lässt. Auch für diese Entscheidung macht die Theorie keine konkreten Vorgaben.

! Das **erste Überbrückungsproblem** besteht darin, die Bestandteile einer theoretischen, wissenschaftlichen Aussage in konkrete, empirisch erfahrbare Aussagen zu überführen. Darüber, wie solche Konkretisierungen zu treffen sind, machen Theorien im Allgemeinen keine Aussage.

Operationalisierung und Messung

Kommen wir zum zweiten Überbrückungsproblem, zu der Übersetzung empirischer Aussagen in die Sprache der Zahlen. Wie kann und wie sollte der Forscher, wenn er genügend Klienten gefunden hat, die zur Teilnahme an der Untersuchung bereit sind, die relevanten Größen, die in der Hypothese vorkommen, registrieren, die erhobenen Informationen in die Symbolsprache der Zahlen übersetzen und diese Zahlen auf eine Weise auswerten, die eine bestmögliche Prüfung der Hypothese gewährleistet?

Operationalisierung. Auf diese Frage gibt es viele Antworten, und es ist schwer zu entscheiden, welche davon die beste ist. Beispielsweise könnte der Forscher so vorgehen, dass er die Mütter der Klienten dazu befragt, wie oft ihr Sohn in einem bestimmten Moment Kontakt zu ihr gesucht hat, den man als erotische Annäherung bezeichnen könnte. Blenden wir die offensichtlichen Schwierigkeiten, die mit dieser Methode verbunden sind, einmal aus und nehmen an, die Mütter der Klienten würden die Frage beantworten. Dann könnte der Forscher die Häufigkeit von Annäherungsversuchen als Maß für die Stärke des tabuisierten erotischen Begehrens nehmen. Eine solche Art der Übersetzung wird in der Methodensprache häufig als Operationalisierung (Messbarmachung) bezeichnet. Später werden wir den Begriff des Messens präzise definieren und ausführlich erläutern (Abschn. 5.2).

Im gegenwärtigen Beispiel stellt die Menge der natürlichen Zahlen die Symbole bereit, in denen wir den ersten Teil der zu prüfenden Behauptung formulieren

könnten. In ähnlicher Weise könnte der Forscher sich entscheiden, den psychotherapeutischen Aufwand durch die Zahl der therapeutischen Sitzungen anzugeben, die laut den Aufzeichnungen der Psychoanalytiker erforderlich waren, um den verdrängten Tabubruch in Erinnerung zu rufen. Auch die mit dieser Entscheidung verbundenen Probleme lassen wir im Moment beiseite und halten fest, dass sich auch der zweite Teil der theoretisch erwarteten empirischen Verhältnisse in der Sprache der natürlichen Zahlen ausdrücken ließe.

! Das **zweite Überbrückungsproblem** besteht darin, diejenigen Bestandteile, über deren Relation die Hypothese eine Aussage macht, zu quantifizieren, d. h. in messbare Größen (die wiederum mithilfe von Zahlen darstellbar sind) zu übertragen.
Darüber, wie solche Operationalisierungen vorzunehmen sind, machen Theorien im Allgemeinen keine Aussage.

Datenerhebung und Datenauswertung

Nehmen wir schließlich an, der Forscher habe die beiden fraglichen Größen auf die beschriebene Weise bei einer Gruppe von Klienten ermittelt. Er verfügt dann über zwei Zahlenreihen, die mit einer der Hypothese angemessenen Methode ausgewertet werden müssen.

Was bedeuten in diesem Zusammenhang »auswerten« und »angemessen«? Auswerten heißt, die Zahlen in einer Weise zu ordnen, dass sie eine Aussage ergeben. Angemessen heißt, dass die Form dieser Aussage jener der empirischen Aussage möglichst genau entspricht, also der Aussage, dass die Häufigkeit von Annäherungsversuchen als Maß für die Stärke des tabuisierten erotischen Begehrens mit der Zahl der therapeutischen Sitzungen zusammenhängt, die laut den Aufzeichnungen der Psychoanalytiker erforderlich waren, um den verdrängten Tabubruch in Erinnerung zu rufen. Wie wir später erfahren werden, könnte die Korrelationsanalyse diesen Zweck erfüllen. Auf der Basis des Ergebnisses dieser Analyse würde der Forscher beurteilen, ob bzw. wie genau die theoretischen Vorhersagen eingetroffen sind.

Die Qualität der Schlussfolgerung, die der Forscher aus seinen Daten zur Beurteilung der Theorie zieht, hängt davon ab, wie gut er die Theorie in empirische Hypothesen übersetzt hat, wie gut seine empirischen Beobachtungen waren, wie gut er seine empirischen Beobachtungen in die Zahlensprache übersetzt hat und wie gut sich die gewählte Auswertungsmethode für die verfügbaren Daten und die Fragestellung eignet. Der Forscher – und dies gilt für den Praktiker in vergleichbarer Weise – muss also im Forschungsprozess zahlreiche Entscheidungen fällen, deren Güte sich nicht in jedem Fall mit letzter Sicherheit beurteilen lässt. Dies ist einer der vielen Gründe, weshalb Theorien durch empirische Untersuchungen nie definitiv und abschließend als richtig oder falsch beurteilt werden können. Wenn z. B. die Daten mit einer unpassenden oder mangelhaften Methode erhoben oder ausgewertet wurden, sagen die Ergebnisse nichts über die Theorie aus, die geprüft werden sollte. Deshalb ist es wichtig, die besten der verfügbaren Methoden für empirische Untersuchungen zu finden und den Untersuchungsprozess so transparent wie möglich zu gestalten, damit andere Wissenschaftler oder Praktiker die Entstehung der Ergebnisse nachvollziehen, sich ein eigenes Urteil über die Qualität der Untersuchung bilden und die Fragestellung gegebenenfalls mit besseren Methoden erneut untersuchen können.

2.2 Schritte im Forschungsprozess

Das Beispiel, mit dem wir im letzten Abschnitt die Ebenen wissenschaftlicher Aussagen und die Schwierigkeiten, diese Ebenen ineinander zu überführen, illustriert haben, ließ bereits in groben Zügen den Ablauf einer wissenschaftlichen Untersuchung erkennen. Betrachten wir nun diesen Ablauf etwas genauer.

2.2.1 Entstehung eines Erkenntnisinteresses

Der Forschungsprozess beginnt in der Regel mit einem Erkenntnisinteresse. Dessen Quellen können vielfältig sein:

- Man stößt auf Ungereimtheiten in den Befunden verschiedener Untersuchungen und möchte sie klären.
- Man glaubt einen Befund nicht, den man liest, und möchte ihn selbst nachprüfen.
- Man erhält die Anfrage eines Praktikers, der mit einem Problem konfrontiert ist, für das er keine Lösung kennt.
- Man erhält von einem Auftraggeber den gezielten Auftrag, eine Fragestellung zu klären.

- Man macht im Alltag eine Beobachtung, die man sich nicht erklären kann und für die man auch in der Literatur keine Erklärung findet.
- Man stößt in der Literatur auf viele unterschiedliche Erklärungen für ein Phänomen, das man gerne verstehen möchte.

Die eigene Forschung könnte dann das Ziel verfolgen herauszufinden, welche Erklärung sich am besten bewährt oder unter welchen Randbedingungen welche Erklärung zutrifft.

Nehmen wir einmal an, eine Forscherin, die normalerweise keine öffentlichen Verkehrsmittel benutzt, würde ausnahmsweise an mehreren Tagen mit der Straßenbahn fahren und dabei würde ihr auffallen, dass viele Sitzbezüge aufgeschlitzt sind. Nachdem sich ihre erste Empörung gelegt hat, beginnt die Forscherin neugierig zu werden und sich nach den Ursachen zu fragen. Sie beschließt, der Frage wissenschaftlich nachzugehen.

2.2.2 Sammlung verfügbaren Wissens

Die meisten Wissenschaftlerinnen und Wissenschaftler werden in einer solchen Situation damit beginnen nachzudenken, ob sie über das erklärungsbedürftige Phänomen vielleicht schon etwas wissen. Falls ihnen keine Theorie oder Untersuchung zum Thema einfällt, werden sie vielleicht das Gespräch mit Kolleginnen oder Kollegen suchen, denen sie zutrauen, etwas von der Sache zu verstehen.

Gleichzeitig wird man sich als Wissenschaftler in der Literatur auf die Suche nach Arbeiten machen, die sich mit dem Phänomen oder verwandten Phänomenen befasst haben. Geeignet für diesen Zweck sind Literaturdatenbanken, Lehr- und Handbücher sowie Zeitschriften, die Überblicksarbeiten (sog. Review-Artikel) publizieren (z. B. die Zeitschrift »Annual Review of Psychology«). Mithilfe von passenden Suchbegriffen – im gegenwärtigen Beispiel etwa »Vandalismus«, »mutwillige Zerstörung« oder »Zerstörung öffentlichen Eigentums« – wird man versuchen, möglichst aktuelle Arbeiten zum Thema zu finden. Wird man fündig, beginnt man diese Arbeiten zu lesen und prüft dabei, ob sie das eigene Erkenntnisinteresse hinreichend befriedigen können oder nicht. Meistens wird man in den gelesenen Artikeln Hinweise auf weitere Arbeiten finden, die man mit der ersten Suchstrategie nicht entdeckt hat. Auch diese Arbeiten wird man lesen und nach brauchbaren Informationen für die eigene Fragestellung durchforsten. Wenn das Erkenntnisinteresse im Zuge dieser Lektüre befriedigt werden kann, hat sich eine eigene Untersuchung erübrigt. Häufig jedoch bleiben Fragen offen oder es entstehen neue. In diesem Prozess kann es also bereits zu einer Veränderung oder Präzisierung des eigenen Erkenntnisinteresses kommen.

Eine weitere Strategie in dieser frühen Erkundungsphase besteht darin, zu überlegen und systematisch anhand der verfügbaren Literatur zu prüfen, ob sich Erkenntnisse aus anderen Gebieten der Psychologie, die vordergründig mit dem interessierenden Phänomen nichts zu tun zu haben scheinen, übertragen lassen. Möglicherweise gehört das Aufschlitzen von Sitzen in öffentlichen Verkehrsmitteln zu einer Kategorie, die viele psychologisch gleichwertige Verhaltensweisen umfasst. Und möglicherweise gibt es zu anderen Verhaltensweisen aus dieser Kategorie bereits brauchbare Erkenntnisse, die sich auf die eigene Fragestellung anwenden lassen. Probehalber könnte man das Aufschlitzen von Sitzen als eine von vielen Formen des Auslebens von Aggressionen oder als eine von vielen normabweichenden Verhaltensweisen interpretieren. Dann würde man den Erkundungsprozess auf diese allgemeineren Phänomene (Verhaltensklassen) ausdehnen und erneut in die Literatursuche eintreten, diesmal jedoch mit allgemeineren Suchbegriffen wie »Aggression« oder »normabweichendes Verhalten«.

2.2.3 Entwicklung einer Fragestellung oder Hypothese

In vielen Fällen wird man bei einer solchen erweiterten Suche nach bereits verfügbaren Erkenntnissen fündig. In manchen Fällen wird man sie für so gut übertragbar halten, dass sich eine eigene Untersuchung erübrigt. Andernfalls lässt man sich durch die gefundenen Arbeiten dazu anregen, vorhandene Theorien auf die eigene Fragestellung zu übertragen. Diese Übertragung ist jedoch mit einem Risiko behaftet: Es bleibt die Ungewissheit, ob sich die Theorie, die sich in anderen Anwendungen bewährt hat, auch für die eigene Fragestellung eignet. Eine eigene Untersuchung könnte dann das Ziel haben, diese Ungewissheit zu beseitigen.

Für den Fall, dass sowohl die spezifische als auch die erweiterte Literatursuche ergebnislos bleibt, müssen

neue Erklärungen entwickelt werden. Meistens wird man in einem solchen Fall jedoch nicht gleich eine ausgearbeitete Theorie anstreben, sondern zunächst die Beobachtung des erklärungsbedürftigen Phänomens ausdehnen und systematischer vornehmen. Dies geht zwar nicht ganz ohne theoretische Vorannahmen; die Untersuchung verfolgt dennoch eher die Klärung einer Fragestellung als die Prüfung einer ausgearbeiteten Theorie.

In unserem Beispiel könnten sinnvolle Fragestellungen etwa lauten:

- Zeigen Personen, die Sitze aufschlitzen, noch andere Verhaltensweisen, die man in die gleiche psychologische Kategorie einordnen kann? Beschädigen sie z. B. auch Straßenlampen und öffentliche Toiletten?
- Richten sich die Verhaltensweisen nur gegen öffentlichen oder auch gegen privaten Besitz? Tendieren die betreffenden Personen z. B. auch dazu, Autoantennen und Scheibenwischer an privaten Pkws abzubrechen?
- Werden von den Personen auch andere normwidrige Verhaltensweisen gezeigt, die nichts mit Vandalismus zu tun haben? Begehen sie z. B. auch Ladendiebstähle, erpressen sie ihre Mitschüler oder Arbeitskollegen, hinterziehen sie Steuern, begehen sie Versicherungsbetrug?

Hinter solchen Fragestellungen stehen in aller Regel schon Vermutungen, die mehr oder weniger konkret und präzise sein können. Beispielsweise könnte hinter der dritten Fragestellung die Annahme stehen, dass es eine interindividuell unterschiedliche und über viele Verhaltensbereiche generalisierte Bereitschaft gibt, soziale Normen zu übertreten.

In diesem Falle hätte die Fragestellung bereits den Charakter einer konkreten Hypothese: Das Aufschlitzen von Sitzen ist eines von vielen Anzeichen einer Disposition (Neigung) zu abweichendem Verhalten. Dies ist jedoch nur eine von vielen denkbaren Erklärungen für das beobachtete Phänomen. Weitere Hypothesen könnten lauten:

- Die Täter wissen vielleicht nicht, dass das Aufschlitzen von Sitzen unerwünscht und verboten ist.
- Das Aufschlitzen von Sitzen könnte eine Mutprobe sein und dazu dienen, Freunden zu imponieren.
- Das Aufschlitzen von Sitzen könnte ein Zeichen von Neugier sein. Vielleicht haben die Täter ein brennendes Interesse herauszufinden, wie die Sitze aufgebaut sind, wie robust sie sind oder wie leistungsfähig das neu erworbene Taschenmesser ist.
- Das Aufschlitzen von Sitzen könnte eine symbolische Botschaft an andere Fahrgäste sein, dass man sich der Gesellschaft nicht zugehörig fühlt und ihre Spielregeln nicht akzeptiert.

Viele weitere Erklärungen sind denkbar. Wichtig ist es, bei der Formulierung von Hypothesen im Auge zu behalten, dass sie empirisch prüfbar bzw. widerlegbar sind, eine gewisse Allgemeingültigkeit aufweisen und theoretisch begründet sind. Die theoretische Begründung ergibt sich meist aus der Literaturrecherche. Um Hypothesen einer empirischen Prüfung unterziehen zu können, ist es ferner von Vorteil, alle theoretischen Konzepte, deren man sich bedient, so präzise wie möglich zu definieren. Im Falle der Hypothese, das Aufschlitzen von Sitzen diene dazu, Freunden zu imponieren, müsste definiert werden, was unter »imponieren« zu verstehen ist. Im Falle der Hypothese, das Aufschlitzen von Sitzen befriedige Neugier, müsste definiert werden, was »Neugier« bedeutet, etc. Je präziser die theoretischen Konzepte, die Bestandteil einer Hypothese sind, definiert sind, desto leichter fällt die Hypothesenprüfung.

2.2.4 Planung einer Untersuchung

Um solche Fragestellungen bzw. Hypothesen zu klären, bedarf es nach dem Selbstverständnis unserer wissenschaftlichen Disziplin einer empirischen Untersuchung. Zur weiteren Illustration des Forschungsprozesses wählen wir die Dispositionshypothese aus, die besagt, dass das Aufschlitzen von Sitzen eines von vielen Anzeichen einer generalisierten Bereitschaft zur Übertretung von sozialen Normen ist. Die Konzentration auf diese Hypothese macht es zunächst erforderlich, andere Formen normabweichenden Verhaltens zu bestimmen. Einige Beispiele wurden oben gegeben (Versicherungsbetrug etc.).

Auswahl einer Erhebungsmethode

Anschließend muss überlegt und entschieden werden, wie man über diese Verhaltensweisen Informationen gewinnen möchte. Beispielsweise könnte man sich für eine Verhaltensbeobachtung oder für eine Befragung entscheiden. Weitere Möglichkeiten werden wir später kennenlernen (Kap. 3). Würde man sich für eine

Befragung entscheiden, müsste man einen Fragebogen entwickeln, der eine Vielzahl normabweichender Verhaltensweisen enthält. Man könnte zu jeder Verhaltensweise fragen, ob sie schon einmal gezeigt wurde, wann das letzte Mal und wie oft in einem definierten Zeitraum. Da normabweichendes Verhalten von den meisten Menschen aus Angst vor Strafe oder Ablehnung nur ungern zugegeben wird, müsste man mit geeigneten Maßnahmen sicherstellen, dass die befragten Personen ehrlich antworten. Dies könnte man etwa durch anonyme Befragungen erreichen, wie sie in der Kriminologie zur Ermittlung von Dunkelziffern durchgeführt werden. Alternativ dazu könnte man die Aussagen der Personen überprüfen, indem man Bezugspersonen wie Freunde, Verwandte und Lehrer über das gleiche Verhalten der Person befragt und die Angaben der Zielperson mit jenen dieser Fremdbeurteiler vergleicht.

Festlegung der Population und Auswahl einer Stichprobe

Als Nächstes muss für die Erhebung eine Stichprobe ausgewählt und deren Größe festgelegt werden. Wie wir später lernen werden, ist es dabei wichtig, dass man zunächst die Grundgesamtheit (Population) derjenigen Personen definiert, für die die zu prüfende Hypothese gelten soll. In unserem Beispiel könnte die Grundgesamtheit aus Jugendlichen und jungen Erwachsenen bestehen. Aus dieser Grundgesamtheit muss dann eine repräsentative Stichprobe gezogen werden. Die Stichprobengröße ergibt sich aus dem Genauigkeitsanspruch der Untersuchung. Wenn die Untersuchung eine erste Erkundung sein soll und v. a. den Zweck verfolgt, weiterführende Ideen zu generieren, genügt eine kleine Zahl von Personen. Wenn hingegen eine Theorie getestet werden soll oder aus den Befunden weitreichende Schlussfolgerungen gezogen werden sollen, von denen die Grundgesamtheit betroffen ist, wird die Stichprobe größer sein müssen. Warum das so ist und wie man bei der Planung der Stichprobengröße vorgeht, werden wir später erfahren (Kap. 8).

Probleme bei der Versuchsdurchführung

Nachdem man sich für eine bestimmte Form der empirischen Prüfung der Hypothese entschieden hat, sollte man sich als Nächstes die Frage stellen, mit welchen Problemen bei der Versuchsdurchführung potenziell zu rechnen ist.

Mangelnde Validität. Eine Schwierigkeit haben wir bereits kennengelernt: Fragt man Menschen nach ihrer Neigung zu normabweichendem Verhalten, besteht die Gefahr, dass man keine ehrlichen Antworten erhält; es könnte also sein, dass zwei Personen, die auf die Frage, wie oft sie im letzten Jahr Ladendiebstahl begangen haben, »noch nie« bzw. »drei Mal« antworten, sich nicht wirklich in ihrer Normbruchneigung unterscheiden, sondern lediglich in ihrer Ehrlichkeit. Das wäre misslich, denn man hätte nicht das gemessen, was man messen wollte: Statt Unterschieden in der Normbruchneigung hätte man Unterschiede in der Ehrlichkeit erfasst. Wir werden später sehen, dass das Ausmaß, in dem eine empirische Messung tatsächlich das erfasst, was sie erfassen soll, eine Eigenschaft der Messung ist und als »Validität« bezeichnet wird (Abschn. 3.4). Wir werden auch sehen, wie man die Validität einer Messung quantifizieren kann und mit welchen Strategien man die Validität der Messung erhöhen kann.

Systematisch fehlende Werte. Eine weitere Schwierigkeit könnte darin bestehen, dass Personen, die man in die Stichprobe gezogen hat, sich weigern, die gestellten Fragen zu beantworten. Dann gäbe es fehlende Werte in den Daten. Solange solche fehlenden Werte unsystematisch über die Personen hinweg verteilt sind und es mehr oder weniger Zufall ist, wer die Auskunft verweigert und wer nicht, gibt es keine großen Probleme. Schwierig wird es hingegen, wenn nur bestimmte Personen die Auskunft verweigern, z. B. diejenigen, die in sehr starkem Maße zu normabweichendem Verhalten neigen. Wenn also Werte aus einem bestimmten Spektrum systematisch fehlen, kann das die Aussagekraft der Daten erheblich mindern.

Ethische Unbedenklichkeit

Eine zentrale Frage, die man sich bei der Planung der eigenen Untersuchung stellen sollte, ist die der ethischen Unbedenklichkeit. Wissenschaftlerinnen und Wissenschaftler haben ein Erkenntnisinteresse, das im Idealfall der Gesellschaft dient, Fragen beantwortet und Probleme lösen hilft. Aber manchmal steht dieses Erkenntnisinteresse im Konflikt mit Prinzipien der ethischen Verantwortlichkeit gegenüber den Versuchspersonen. Ethisch bedenklich sind Untersuchungen, wenn sie die Menschenwürde verletzen oder mit potenziellen Gefahren für Leib, Leben und Wohlergehen verbunden sind. Die Frage, wie Menschen Traumata und schweren

psychischen Stress bewältigen, mag wissenschaftlich und gesellschaftlich hoch relevant sein, aber sie berührt einen sensiblen Bereich. So verbietet es sich, Traumata gezielt experimentell auszulösen, um zu untersuchen, wie deren kognitive Verarbeitung funktioniert. Auch die bloße Befragung zu zurückliegenden traumatischen Erfahrungen kann ethisch bedenklich sein, wenn die Gefahr besteht, dass die Befragungssituation von der befragten Person als belastend erlebt wird. Ethisch bedenklich wäre in diesem Fall andererseits der Verzicht auf eine Untersuchung, deren Ergebnisse den betroffenen Personen helfen könnten, ihre traumatischen Erfahrungen besser zu verarbeiten.

Die Deutsche Gesellschaft für Psychologie (DGPs) und der Berufsverband Deutscher Psychologinnen und Psychologen (BDP) haben gemeinsam ethische Richtlinien herausgegeben, die nicht nur die Unbedenklichkeit wissenschaftlicher Fragestellungen und empirischer Herangehensweisen betreffen, sondern auch allgemeine Fragen der beruflichen Praxis von Psychologinnen und Psychologen. Der aktuelle Text ist im Internet abrufbar. (Einen Link darauf finden Sie in unseren Online-Materialien.)

2.2.5 Durchführung der Untersuchung

Zunächst muss das Untersuchungsmaterial erstellt werden, z. B. der Fragebogen, mit dem die Häufigkeit normabweichenden Verhaltens ermittelt werden soll. Die Konstruktion von Fragebögen ist eine Kunst für sich, die wir hier nicht im Detail behandeln können. Auf einige Aspekte, die es dabei zu beachten und entscheiden gilt, werden wir später genauer eingehen (Abschn. 3.3.3). Außerdem gibt es Bücher, die sich ausschließlich mit der Konstruktion von Fragebögen befassen (z. B. Moosbrugger & Kelava, 2012; Mummendey & Grau, 2014).

Als Nächstes wird die Stichprobe rekrutiert. In unserem Beispiel könnte man etwa versuchen, über Schulen an die jüngeren Untersuchungsteilnehmer zu gelangen. Man könnte aber auch Personen auf der Straße oder in öffentlichen Verkehrsmitteln ansprechen und um ihre Teilnahme bitten. Weiterhin könnte man sich von der Einwohnerbehörde eine Stichprobe von Personen aus dem Melderegister ziehen lassen. Schließlich könnte mithilfe von Anzeigen in der Tagespresse zur Teilnahme an der Untersuchung aufgerufen werden. Welche dieser Strategien man verfolgt und wie man sie konkret anwendet, hängt von der Fragestellung, der Definition der Grundgesamtheit sowie von Annahmen über Gründe der Teilnahmebereitschaft und befürchteten Störfaktoren ab. Rechnet man z. B. mit regionalen Unterschieden im fraglichen Verhalten, muss man die Region bei der Stichprobenziehung systematisch berücksichtigen.

Sobald die Stichprobe gezogen ist, kann mit der Datenerhebung begonnen werden. In unserem Beispiel bedeutet dies, dass man den Fragebogen in Schulklassen austeilt, jenen Personen, die man auf der Straße oder in öffentlichen Verkehrsmitteln angesprochen hat, den Fragebogen nebst einem frankierten Rücksendeumschlag aushändigt oder jenen Personen, deren Anschrift man von der Einwohnerbehörde bekommen hat bzw. die sich auf die Anzeige in der Zeitung hin gemeldet haben, den Fragebogen zusammen mit einem Rücksendeumschlag, einem Begleitschreiben und einer Instruktion zuschickt.

Insbesondere wenn man Experimente durchführt, ist es wichtig, den Versuchsablauf penibel zu dokumentieren. Eine gute Protokollierung hilft dabei, im Nachhinein unerwartete Schwierigkeiten oder Probleme bei der Versuchsplanung bzw. der Versuchsdurchführung zu erkennen oder diejenigen Versuchspersonen zu identifizieren, die aus bestimmten Gründen (z. B. mangelnde Motivation, Hypothese korrekt erraten, Probleme bei der Datenspeicherung etc.) von der Datenanalyse ausgeschlossen werden müssen.

Arbeitet man mit standardisiertem Versuchsmaterial (wie z. B. einem Fragebogen, einem Test oder einer computergesteuerten Apparatur), ist es sinnvoll, eine Zwischenanalyse der Daten vorzunehmen und zu überprüfen, ob es Schwierigkeiten gibt (z. B. zu schwere oder missverständliche Fragen), die es im Extremfall nötig machen, den Versuch abzubrechen.

Auch bei der Versuchsdurchführung ist die Einhaltung ethischer Standards essenziell (vgl. die ethischen Richtlinien der DGPs und des BDP). Hierzu gehören v. a.:

- die Aufklärung der Versuchspersonen über potenzielle Risiken vor Beginn des Experiments,
- der Hinweis darauf, dass die Teilnahme an der Untersuchung freiwillig ist und jederzeit ohne Angabe von Gründen abgebrochen werden kann,
- die Zusicherung von Vertraulichkeit bei der Aufbereitung und Auswertung der Daten,

- das Einholen einer Einwilligungserklärung seitens der Versuchspersonen,
- eine ausführliche und lückenlose Aufklärung der Versuchspersonen über den Zweck der Untersuchung spätestens nach Abschluss der Untersuchung.

2.2.6 Auswertung der Daten

Sobald die Erhebung abgeschlossen ist, kann die Auswertung beginnen. Als Erstes müssen die erhobenen Informationen in einer Weise aufbereitet werden, die sich zur Auswertung eignet. In den meisten psychologischen Untersuchungen werden die erhobenen Informationen durch numerische Codierung in Zahlen übersetzt. In unserem Beispiel könnte die Bejahung der Frage nach normabweichenden Verhaltensweisen mit der Zahl 1, die Verneinung mit der Zahl 0 codiert werden. Antworten auf die Frage nach der Häufigkeit normabweichender Verhaltensweisen in einem definierten Zeitraum könnte man zahlenmäßig übernehmen. Den Zeitraum bis zur letzten Übertretung einer Norm könnte man in Abschnitte einteilen (z. B. 1 Monat, 3 Monate, 6 Monate, 1 Jahr, 3 Jahre usw.) und für jeden Abschnitt eine Zahl festlegen (1, 2, 3, 4, 5 usw.) oder aber den Zeitraum in der Anzahl der vergangenen Monate codieren (1, 3, 6, 12, 36 usw.). Die Art der Codierung hängt davon ab, welche Auswertungsmethode angewendet werden soll. Um Fehler bei der Übertragung zu vermeiden, empfiehlt es sich, einen Codierplan zu erstellen.

Datenmatrix: Darstellung der Daten

Die Speicherung der codierten Informationen erfolgt heute in elektronischen Datenbanken. Meistens werden die Daten in Matrixform geordnet und dargestellt. Diese Form ist übersichtlich und entspricht der Struktur, die viele Auswertungsmethoden voraussetzen. Eine Matrix ist die Anordnung von Zahlen in Tabellenform. Die Tabelle ist durch die Anzahl der Zeilen und Spalten definiert. Einer bewährten Konvention entsprechend schreibt man in der Psychologie die codierten Informationen über eine Person nebeneinander. Personen stellen in einer Datenmatrix also die Zeilen dar. Die Spalten entsprechen der Art der Information, die erhoben wurde, in unserem Beispiel also den einzelnen Fragen des Fragebogens. In einer Zelle der Matrix steht dann die numerisch codierte Antwort, die eine Person auf eine bestimmte Frage des Fragebogens gegeben hat.

Deskriptivstatistik: Beschreibung der Daten

Die eigentliche Auswertung besteht darin, dass die Zahlen, die in der Datenmatrix stehen, nach bestimmten Regeln kombiniert und zusammengefasst werden. In der Regel beginnt man mit einer einfachen deskriptiven (beschreibenden) Analyse der Daten. In unserem Beispiel wird man sich vielleicht zunächst dafür interessieren, wie häufig bestimmte Verhaltensweisen vorkommen. Um dies zu ermitteln, könnte man die Summe der Spalten ermitteln, in denen die Antworten der Probanden auf die Frage stehen, wie oft sie eine bestimmte Norm in einem definierten Zeitraum verletzt haben, wie oft sie z. B. in den vergangenen fünf Jahren Steuern hinterzogen und Versicherungen betrogen haben. Die Summe der jeweiligen Spalte sagt aus, wie häufig das entsprechende Verhalten in der Gruppe insgesamt vorkam.

Zentrale Tendenz. Dividiert man diese Summe durch die Zahl der Probanden, ergibt sich die durchschnittliche Häufigkeit der jeweiligen Verhaltensweise. Dieser Durchschnittswert beschreibt die Gruppe insgesamt, nicht mehr eine einzelne Person. Man spricht von einer zentralen Tendenz.

Streuung. Weiterhin könnte man in einer ersten Analysephase auch die Frage klären, wie stark sich Personen in der Häufigkeit, mit der sie ein bestimmtes normabweichendes Verhalten in einem definierten Zeitraum zeigen, voneinander unterscheiden. Möglicherweise gibt es eine größere Gruppe, die das Verhalten nie zeigt, eine zweite größere Gruppe, die das Verhalten selten zeigt, und eine kleine Gruppe, die das Verhalten sehr häufig zeigt. Analysen dieser Art intendieren die Beschreibung der Verteilung oder Streuung von Verhaltensweisen in der untersuchten Gruppe.

Kovariation. In einem weiteren Analyseschritt könnte man sich der Frage zuwenden, ob zwei Verhaltensweisen (z. B. Steuerhinterziehung und Versicherungsbetrug) im Sinne der Dispositionshypothese etwas miteinander zu tun haben. Um dies zu ermitteln, könnte man auszählen, wie viele Personen beide Verhaltensweisen (Fall 1), nur eine der beiden Verhaltensweisen (Fall 2) oder keine der beiden Verhaltensweisen

14 Unterschiede zwischen mehreren abhängigen Stichproben: Varianzanalyse mit Messwiederholung und verwandte Verfahren

Was Sie in diesem Kapitel lernen

- Wie testet man, ob Bedingungsmittelwerte bei Experimenten mit intraindividueller Bedingungsvariation signifikant voneinander abweichen?
- In welche Bestandteile wird bei der Varianzanalyse mit Messwiederholung die Gesamtvariation zerlegt?
- Wie können bei der Varianzanalyse mit Messwiederholung spezifische Hypothesen über Mittelwertsunterschiede getestet werden?
- Wie werden Interaktionseffekte zwischen zwei messwiederholten Faktoren getestet?
- Wie werden Interaktionseffekte zwischen einem messwiederholten und einem nicht-messwiederholten Faktor getestet?
- Wie testet man auf der Basis eines nonparametrischen Tests, ob die Mediane mehrerer messwiederholter Bedingungen signifikant voneinander abweichen?

In Kapitel 13 haben wir die Varianzanalyse als Möglichkeit kennengelernt, Mittelwertsunterschiede aus mehreren unabhängigen Stichproben auf ihre statistische Bedeutsamkeit hin zu testen. Die Variation in den Messwerten, die ein Faktor (bzw. mehrere Faktoren und ihre Wechselwirkung) verursacht, haben wir als Effekte (Haupteffekte, Interaktionseffekte) bezeichnet und gesehen, dass solche Effekte auf der Basis eines *F*-Tests inferenzstatistisch abgesichert werden können. Wir haben festgestellt, dass eine notwendige Voraussetzung für die Anwendung des in Kapitel 13 beschriebenen *F*-Tests darin besteht, dass es zwischen Messwerten, die aus unterschiedlichen Stichproben (d. h. aus unterschiedlichen Stufen des Faktors) stammen, keinerlei gegenseitigen Abhängigkeiten gibt. Es muss sich also bei den Stufen eines Faktors jeweils um *unabhängige* Stichproben handeln: In den unterschiedlichen Faktorstufen müssen sich unterschiedliche Personen befinden.

In diesem Kapitel werden wir nun sehen, wie die Varianzanalyse funktioniert, wenn es sich bei den Faktorstufen um *abhängige* Stichproben handelt. Was mit abhängigen Stichproben gemeint ist (und was nicht mit ihnen gemeint ist), haben wir bereits in Kapitel 12 ausführlich behandelt. Abhängige Stichproben liegen z. B. vor, wenn es sich um Experimente mit intraindividueller Bedingungsvariation handelt (sog. messwiederholte Faktoren) oder wenn unterschiedliche Versuchspersonen in den unterschiedlichen Faktorstufen einander zugeordnet werden können (aufgrund einer »natürlichen Beziehung« zwischen ihnen oder aufgrund gleicher Ausprägungen auf einer Kontrollvariablen; vgl. die Technik des Parallelisierens in Abschn. 4.3.3). Da die wiederholte Messung an den gleichen Personen den typischen Anwendungsfall abhängiger Stichproben darstellt, wird die entsprechende Auswertungsprozedur als »Varianzanalyse mit Messwiederholung« (engl. *repeated-measures analysis of variance*, kurz: RM-ANOVA oder auch *within-subjects* ANOVA) bezeichnet. Den Fall einer einfaktoriellen Varianzanalyse mit Messwiederholung werden wir in Abschnitt 14.1 detailliert behandeln.

Im Falle von mehrfaktoriellen Designs können Messwiederholungen bei keinem, einigen oder allen Faktoren vorliegen. Sind alle Faktoren messwiederholt, spricht man von einem komplett messwiederholten Design; sind nur einige Faktoren messwiederholt, andere hingegen nicht, spricht man von einem partiell messwiederholten Design. Wir werden uns in Abschnitt 14.2 auf zwei Fälle beschränken: ein zweifaktorielles Design mit vollständiger Messwiederholung und ein zweifaktorielles Design mit Messwiederholung auf einem Faktor. Abschließend werden wir in diesem Kapitel noch einen nonparametrischen Test auf Unterschiede zwischen mehreren Medianen im Falle eines messwiederholten Faktors vorstellen (Abschn. 14.3) und auf Verfahren für kategoriale Variablen verweisen (Abschn. 14.4).

14.1 Einfaktorielle Varianzanalyse mit Messwiederholung

Wir haben in Abschnitt 13.1 darauf hingewiesen, dass die einfaktorielle Varianzanalyse *ohne* Messwiederholung eine Erweiterung (oder Verallgemeinerung) des *t*-Tests für unabhängige Stichproben ist (vgl. Abschn. 11.1.2). Dementsprechend ist die einfaktorielle Varianzanalyse *mit* Messwiederholung eine Erweiterung (oder Verallgemeinerung) des *t*-Tests für abhängige Stichproben (vgl. Abschn. 12.1.1). Sie wird verwendet, wenn die Nullhypothese getestet werden soll, dass sich die Mittelwerte mehrerer abhängiger Stichproben (z. B. experimenteller Bedingungen) nicht voneinander unterscheiden.

Der klassische Fall von abhängigen Stichproben ist die intraindividuelle Bedingungsvariation: Von allen Personen in der Stichprobe werden wiederholt Messwerte unter unterschiedlichen experimentellen Bedingungen erhoben. Beispielsweise könnte man die kognitive Leistungsfähigkeit von Personen dreimal erheben und miteinander vergleichen: (1) ohne Stimmungsinduktion, (2) nach positiver Stimmungsinduktion und (3) nach negativer Stimmungsinduktion. Mit einem solchen Design könnte die Hypothese überprüft werden, dass sich positive Stimmung förderlich und negative Stimmung hemmend auf die kognitive Leistungsfähigkeit auswirken. Da in allen drei experimentellen Bedingungen die gleichen Personen getestet werden, handelt es sich um abhängige Stichproben: Die Messwerte werden über die drei Bedingungen hinweg miteinander kovariieren, da die kognitive Leistungsfähigkeit eben nicht nur von der Stimmung, sondern auch von personengebundenen Variablen abhängt, die über die drei Stimmungsbedingungen hinweg stabil bleiben (Intelligenz, Teilnahmemotivation etc.).

Kovariation zwischen den Messungen. Ein Teil der Variation in den Messwerten ist also dadurch zu erklären, dass es sich um die gleichen Personen (und damit teilweise um die gleichen personengebundenen Einflüsse auf die Messwerte) handelt. Wir haben in Kapitel 12.1 die Kovarianz als einen statistischen Kennwert kennengelernt, der angibt, wie groß der Einfluss solcher personengebundenen Merkmale ist (s. hierzu ausführlich Abschn. 16.3.1). Genauer gesagt: Die Kovarianz gibt an, wie stabil Unterschiede zwischen Personen über zwei Messungen hinweg bleiben. Liegen mehr als zwei Messungen vor, muss diese Stabilität anders quantifiziert werden. In Abschnitt 14.1.2 werden wir sehen, wie diese »Varianz zwischen Personen« bestimmt werden kann und welche Rolle sie für die Zerlegung der Gesamtvariation spielt. Durch die Kovariation wiederholter Messungen können stabile Störvariablen, die an die Person gebunden sind (z. B. Persönlichkeitsmerkmale), statistisch kontrolliert werden. Hierdurch reduziert sich der Anteil unerklärter Varianz. Folglich gewinnt man an Teststärke (Power) und benötigt eine geringere optimale Stichprobengröße als bei einer Varianzanalyse ohne Messwiederholung. Dies ist ein großer Vorteil von Messwiederholungsdesigns.

Sequenzeffekte. Ein Nachteil experimenteller Designs mit intraindividueller Bedingungsvariation besteht darin, dass die Messwerte nicht vor systematischen Verfälschungen gefeit sind. So könnte es sein, dass Versuchspersonen im Laufe der Untersuchung die Hypothese erraten, dass ihre Motivation, am Experiment teilzunehmen, zunehmend nachlässt oder dass sie Antwort- bzw. Lösungsstrategien von einer Messung auf die nächste übertragen bzw. sich an ihre früheren Antworten erinnern. All diese systematischen Verfälschungen, die typisch für Messwiederholungsdesigns sind, werden unter dem Begriff *Sequenzeffekte* zusammengefasst. Für eine Systematik solcher Sequenzeffekte und den Umgang mit ihnen (etwa die Berücksichtigung bestimmter Kontrollstrategien wie die intra- oder interindividuelle Ausbalancierung) verweisen wir auf einschlägige Lehrbücher zur Versuchsplanung (z. B. Huber, 2013; Hussy & Jain, 2002; Shadish et al., 2002).

Veränderungsmessung. In vielen psychologischen Fragestellungen geht es nicht um Unterschiede, die durch die Manipulation einer experimentellen Variablen hervorgerufen werden, sondern lediglich um Unterschiede im Laufe der Zeit. Anders gesagt: Oft ist die Zeit selbst die unabhängige Variable, und man interessiert sich für die Veränderung in den (durchschnittlichen) Messwerten über die Zeit hinweg. Beispielsweise interessiert man sich in der allgemeinen Entwicklungspsychologie dafür, ob und wie sich psychologische Merkmale mit zunehmendem Alter verändern. Dies erfordert die wiederholte Messung der fraglichen Merkmale. Werden über die Zeit hinweg immer wieder Messwerte von den gleichen Personen erhoben, handelt es sich um eine sog. Längsschnittuntersuchung. Auch hier kann man sich mithilfe

einer Varianzanalyse mit Messwiederholung der Frage widmen, ob die Unterschiede in den Mittelwerten zwischen den Messzeitpunkten signifikant variieren.

Evaluationsforschung. Interessiert man sich im Rahmen einer Untersuchung zur Wirksamkeit einer psychologischen Intervention dafür, ob und inwieweit sich durch die Intervention die Merkmalsausprägung über die Zeit hinweg verändert hat (z. B. inwieweit durch eine Psychotherapie das Ausmaß subjektiver Belastungen reduziert wurde oder inwieweit durch ein Lerntraining die Schulleistung von Kindern verbessert wurde), so sind wiederum Messwiederholungsdesigns angezeigt. Messwiederholungsdesigns sind typisch für die Evaluationsforschung, v. a. für die Prozess- und die Wirksamkeitsevaluation (vgl. Gollwitzer & Jäger, 2014).

Datenbeispiel

Um die Grundidee der einfaktoriellen Varianzanalyse mit Messwiederholung zu veranschaulichen, beginnen wir auch hier mit einem einfachen Datenbeispiel. Stellen wir uns vor, das in Abschnitt 13.1 geschilderte Experiment zum Modelllernen sei nicht auf der Basis einer interindividuellen Bedingungsvariation, sondern vielmehr auf der Basis einer *intra*individuellen Bedingungsvariation durchgeführt worden. Insgesamt werden fünf Personen untersucht, und zwar jeweils unter drei Bedingungen. Zunächst sehen die Probanden einen Film, in dem eine Person ihres Alters für ein bestimmtes Verhalten (z. B. Aggression) belohnt wird. Im Anschluss daran wird mit einem Selbstberichtsmaß (z. B. einer visuellen Analogskala mit Ausprägungen zwischen 0 und 100) gemessen, wie stark die Personen dazu tendieren, das vom Modell gezeigte Verhalten nachzuahmen. Anschließend sehen die gleichen Personen einen zweiten Film, in dem die Modellperson für ihr Verhalten bestraft wird; wiederum wird die Nachahmungstendenz per Selbstauskunft gemessen. Schließlich sehen die Probanden einen dritten Film, in dem das Verhalten der Modellperson ohne Konsequenzen bleibt. Die Nachahmungstendenz wird hier ein drittes Mal gemessen.

Die unabhängige Variable (UV) ist also die Verhaltenskonsequenz (Belohnung, Bestrafung, keine Konsequenz). Die abhängige Variable (AV) ist die Nachahmungstendenz. Sollte die Hypothese, die aus der Theorie des Modelllernens von Bandura (1976, 1977) abgeleitet wurde, korrekt sein, müsste die Nachahmungstendenz nach dem ersten Film (Belohnungsbedingung) am stärksten und nach dem zweiten Film (Bestrafungsbedingung) am schwächsten sein. Dass es sich hier zugegebenermaßen um ein ziemlich problematisches Versuchsdesign handelt, da mit erheblichen Verzerrungen durch Sequenzeffekte zu rechnen ist, dürfte klar sein. Trotzdem werden wir mit diesem Beispiel weiterarbeiten.

Die experimentellen Bedingungen erhalten einen Laufindex von $j = 1, \ldots, J$ (mit $J = 3$ in unserem Beispiel). Die Personen erhalten einen Laufindex von $m = 1, \ldots, n$ (mit $n = 5$ in unserem Beispiel). Den Index j können wir bei n hier weglassen, da die Anzahl der beobachteten Werte pro Bedingung aufgrund der Messwiederholung der Gesamtanzahl aller Versuchspersonen entspricht. Die (fiktiven) Rohdaten des Versuchs stehen in Tabelle 14.1. Beachten Sie, dass es sich hier um die gleichen Rohdaten handelt wie in Tabelle 13.1 (in Abschn. 13.1). Wir haben bewusst die gleichen Daten gewählt, um den Unterschied zwischen der Varianzanalyse ohne Messwiederholung und der Varianzanalyse mit Messwiederholung (und die statistischen Implikationen dieses Unterschiedes) deutlich zu machen.

Personen als Stufen eines »zufälligen Faktors«

Der Unterschied zwischen Tabelle 13.1 und Tabelle 14.1 besteht darin, dass wir hier eine zusätzliche Spalte eingefügt haben, in der die *durchschnittliche* Nachahmungstendenz einer Person über die drei experimentellen Bedingungen hinweg angegeben ist. In Tabelle 13.1 wäre es aufgrund des interindividuellen Designs nicht sinnvoll gewesen, einen solchen Personmittelwert zu berechnen, da den unterschiedlichen experimentellen Bedingungen unterschiedliche Personen zugeordnet waren. Beim intraindividuellen Design hingegen stammen alle Messwerte innerhalb der gleichen Zeile des Versuchsplans von ein und derselben Person. Daher ist hier nicht nur der Bedingungsmittelwert $\bar{x}_{\bullet j}$ (Spaltenmittelwert; s. die untere Zeile in Tab. 14.1), sondern auch der Personmittelwert $\bar{x}_{m\bullet}$ (Zeilenmittelwert; s. die rechte Spalte in Tab. 14.1) sinnvoll interpretierbar.

In Abschnitt 13.1.11 hatten wir die Unterscheidung zwischen »festen« und »zufälligen« Effekten (bzw. Faktoren) kennengelernt. Bei den drei experimentellen Bedingungen in unserem Beispiel handelt es sich zweifelsohne um Stufen eines festen Faktors, denn das variierte Merkmal kann nur eine bestimmte Anzahl möglicher

Ausprägungen haben, und die realisierten Faktorstufen entsprechen genau diesen Ausprägungen. Bei Messwiederholungsdesigns kann man nun auch die unterschiedlichen Personen als »Stufen« eines »Faktors« auffassen. Der Personfaktor umfasst dabei alle möglichen Unterschiede zwischen den Personen. Da die Personen eine Zufallsstichprobe darstellen und sich daher auch zufällig hinsichtlich der Personmerkmale unterscheiden, handelt es sich formal gesehen um einen zufälligen Faktor (s. Abschn. 13.1.11). Anders als der Bedingungsfaktor ist der Personfaktor von untergeordnetem Interesse für die Hypothesenprüfung, aber die Vorstellung, dass es sich bei den Personen um Stufen eines zufälligen Faktors handelt, hilft uns, die Logik der Quadratsummenzerlegung besser zu verstehen. Wir können also die einfaktorielle Varianzanalyse mit Messwiederholung als quasi-zweifaktorielles Design mit einem (festen) Faktor »Bedingung« und einem zweiten (zufälligen) Faktor »Person« auffassen.

14.1.1 Messwertzerlegung

In Tabelle 14.1 gibt es drei Quellen der Variation: (1) Unterschiede zwischen den experimentellen Bedingungen, (2) Unterschiede zwischen den Personen und (3) den Anteil der Variation, der weder auf Haupteffekte der Bedingungen noch auf Personeffekte zurückgeführt werden kann. Dieser Restanteil (Residualanteil) geht auf Interaktionen zwischen den Personen und den Bedingungen und andere unsystematische Störeinflüsse wie z. B. den Messfehler zurück.

Der Messwert x_{mj} einer Person m in einer Bedingung a_j lässt sich daher wie folgt zerlegen:

$$x_{mj} = \bar{x} + t_j + p_m + e_{mj} \qquad \text{(F 14.1)}$$

In dieser Gleichung bezeichnet $\bar{x}$ den Gesamtmittelwert aller Werte, $t_j = \bar{x}_{\bullet j} - \bar{x}$ ist der Haupteffekt der j-ten Stufe des Faktors A, d. h. die Abweichung des Bedingungsmittelwerts $\bar{x}_{\bullet j}$ vom Gesamtmittelwert $\bar{x}$. Mit $p_m = \bar{x}_{m\bullet} - \bar{x}$ wird der Haupteffekt der m-ten Person bezeichnet, d. h. die Abweichung des Mittelwerts $\bar{x}_{m\bullet}$ einer Person m (über alle J Bedingungen hinweg) vom Gesamtmittelwert $\bar{x}$. Der Residualwert $e_{mj} = x_{mj} - \bar{x} - t_j - p_m$ einer Person m in Bedingung a_j ist der Anteil am Wert der Person, der übrig bleibt, wenn man den Gesamtmittelwert und die Bedingungs- und Personeffekte abzieht.

14.1.2 Quadratsummenzerlegung

Die Gesamtvariation (totale Quadratsumme) der 15 Messwerte berechnet sich analog zu dem in Abschnitt 13.1.4 beschriebenen Vorgehen. Sie ist definiert als die Summe der quadrierten Abweichungen der einzelnen Messwerte vom Gesamtmittelwert:

$$QS_{\text{tot}} = \sum_{j=1}^{J} \sum_{m=1}^{n} \left(x_{mj} - \bar{x}\right)^2 \qquad \text{(F 14.2)}$$

Sie beträgt in unserem Beispiel $QS_{\text{tot}} = 4.532$.

Wie kann die totale Quadratsumme bei der einfaktoriellen Varianzanalyse mit Messwiederholung zerlegt werden? Um das zu veranschaulichen, machen wir von der Vorstellung Gebrauch, es handele sich bei den unterschiedlichen Personen um Stufen eines zufälligen

Tabelle 14.1 Fiktive Daten eines Experiments zum Modelllernen (mit intraindividueller Bedingungsvariation)

Person m	Stufe a_j des Faktors			Personmittelwert $\bar{x}_{m\bullet}$
	Belohnung (a_1)	Bestrafung (a_2)	Keine Konsequenz (a_3)	
1	57	18	36	37
2	45	15	27	29
3	49	13	43	35
4	69	37	29	45
5	70	37	55	54
Bedingungsmittelwert $\bar{x}_{\bullet j}$	58	24	38	$\bar{x} = 40$

Faktors. Genau wie jeder andere Faktor kann also auch jede Person einen Haupteffekt, d. h. einen von der experimentellen Bedingung unabhängigen (und insofern »unbedingten«) Effekt, haben. Die Haupteffekte der Personen sind darauf zurückzuführen, dass diese sich über die unterschiedlichen Messungen hinweg konsistent (d. h. gleichbleibend) in Personmerkmalen unterscheiden, welche die AV beeinflussen. Zusätzlich zu den Haupteffekten der Personen gibt es für jede Stufe des Faktors *A* einen Haupteffekt der experimentellen Bedingung, der von der jeweiligen Person unabhängig ist. Und schließlich kann es sein, dass der Effekt der experimentellen Manipulation bei unterschiedlichen Personen unterschiedlich ausfällt. Unklar bleibt jedoch, ob es sich hierbei um einen echten Interaktionseffekt zwischen Personmerkmalen und der experimentellen Bedingung handelt oder lediglich um unsystematische Effekte wie Messfehler.

Bei der einfaktoriellen Varianzanalyse mit Messwiederholung wird die totale Quadratsumme QS_{tot} also in drei Teile zerlegt:

- eine Quadratsumme, welche die Variation zwischen Personen angibt (QS_{zwP}),
- eine Quadratsumme, welche die Variation zwischen den Stufen des Faktors *A* angibt (QS_{zwA}), und
- eine Quadratsumme, welche diejenige Variation angibt, die weder durch Haupteffekte der Bedingungen noch durch Haupteffekte der Personen erklärt werden kann. Diese Quadratsumme werden wir im Folgenden als *Residualquadratsumme* (QS_{Res}) bezeichnen.

Variation zwischen Personen

Die Variation zwischen Personen ist derjenige Teil der Gesamtvariation, der auf Unterschiede zwischen den Personen – unabhängig von den Stufen des experimentellen Faktors – zurückgeht. Es handelt sich also um diejenigen Unterschiede zwischen Personen, die sich über alle Faktorstufen hinweg konsistent zeigen. Die Variation zwischen Personen ist ein unbedingter (d. h. von Faktor *A* unabhängiger) Effekt des »Faktors« Person.

Die Variation zwischen Personen entspricht im Falle von zwei Bedingungen der Kovarianz beim *t*-Test für abhängige Stichproben (s. Abschn. 12.1.1). Sie ist ein Maß für den Einfluss jener Merkmale, die über die experimentellen Bedingungen hinweg konsistente Unterschiede zwischen den Personen produzieren. Solche konsistenten Unterschiede zwischen Personen manifestieren sich in der Variation der Personmittelwerte (s. rechte Spalte in Tab. 14.1).

Die Zwischen-Personen-Quadratsumme QS_{zwP} basiert also auf den quadrierten Abweichungen der Personmittelwerte vom Gesamtmittelwert:

$$QS_{zwP} = J \cdot \sum_{m=1}^{n} \left(\bar{x}_{m\bullet} - \bar{x}\right)^2 \qquad \text{(F 14.3)}$$

Sie beträgt in unserem Beispiel $QS_{zwP} = 1.128$.

Variation zwischen Faktorstufen

Die Variation zwischen den Faktorstufen ist derjenige Teil der Gesamtvariation, der auf systematische Unterschiede zwischen den experimentellen Bedingungen zurückgeführt werden kann. Diese Unterschiede manifestieren sich – genau wie bei der einfaktoriellen Varianzanalyse ohne Messwiederholung – in der Variation der Bedingungsmittelwerte (s. untere Zeile in Tab. 14.1). Die Zwischen-Quadratsumme des Faktors *A* (QS_{zwA}) basiert also auf den quadrierten Abweichungen der Bedingungsmittelwerte vom Gesamtmittelwert:

$$QS_{zwA} = n \cdot \sum_{j=1}^{J} \left(\bar{x}_{\bullet j} - \bar{x}\right)^2 \qquad \text{(F 14.4)}$$

Sie beträgt in unserem Beispiel $QS_{zwA} = 2.920$.

Variation zwischen Personen in Bezug auf den Effekt des Faktors

Wenn wir die Personen und die Bedingungsvariation als Faktoren auffassen, dann besteht auch die Möglichkeit, dass diese miteinander interagieren. Inhaltlich bedeutet diese Interaktion, dass sich Personen darin unterscheiden, wie groß die Unterschiede in ihren Messwerten zwischen den drei Bedingungen sind. Grafisch kann man die Idee einer solchen Interaktion zwischen Person und Bedingung veranschaulichen, wenn man die Messwerte in einem Liniendiagramm abträgt, wobei die Personen als Ausprägungen auf der Abszisse und die experimentellen Bedingungen als drei unterschiedliche Linien dargestellt werden können (s. Abb. 14.1). Man sieht deutlich, dass die Linien nicht parallel verlaufen; es gibt also (Person-)Unterschiede in den (Bedingungs-)Unterschieden.

Machen wir uns die Idee einer Interaktion zwischen Person und Bedingung an Abbildung 14.1 klar. Wir sehen z. B., dass die Messwerte von Person 2 näher beieinanderliegen als die Messwerte von Person 1. Der Effekt der experimentellen Manipulation ist bei Per-

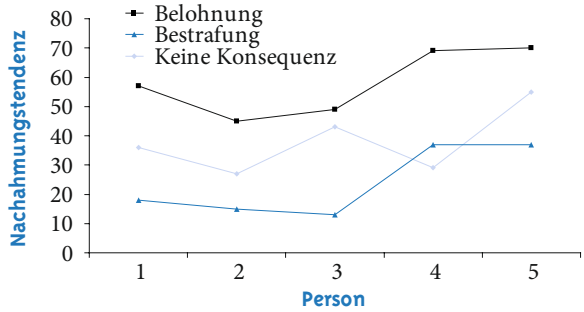

Abbildung 14.1 Grafische Darstellung der Interaktion zwischen Person und Bedingung (Datenbeispiel aus Tab. 14.1)

son 1 größer als bei Person 2. Möglicherweise ist dieser Unterschied zwischen den beiden Personen darauf zurückzuführen, dass Person 2 ihr Verhalten generell weniger stark an stellvertretenden Verhaltensbekräftigungen ausrichtet. Und möglicherweise ist der Effekt der experimentellen Manipulation bei Person 1 genau deshalb stärker, weil diese Person ihr Verhalten sehr stark daran ausrichtet, ob andere Menschen für das gleiche Verhalten belohnt oder bestraft werden. Diese Überlegung würde nahelegen, dass es eine echte Interaktion zwischen der experimentellen Manipulation (stellvertretende Verhaltenskonsequenzen) und Merkmalen der beobachtenden Personen (hier: Sensibilität für stellvertretende Verhaltenskonsequenzen) gibt. Das Problem ist: Wir können nicht testen, ob es sich um eine echte Interaktion handelt oder lediglich um zufällige Schwankungen bzw. Messfehler, die überhaupt nicht auf systematische Personunterschiede zurückzuführen sind. Der Grund dafür, dass wir systematische Person-Bedingungs-Interaktionen nicht testen können, ist der, dass wir in jeder Kombination von Person und Bedingung jeweils nur einen einzigen Messwert haben. Wie stark dieser Messwert von Messfehlern vs. von systematischen Person-Bedingungs-Interaktionseffekten beeinflusst ist, bleibt unbekannt. Insofern werden wir der Einfachheit halber die entsprechende Quadratsumme mit QS_{Res} bezeichnen.

Wie quantifiziert man die Residualquadratsumme QS_{Res}? Grundsätzlich erfolgt die Berechnung nicht anders, als in Abschnitt 13.2.3 beschrieben. Es handelt sich um die Variation zwischen den Messwerten, die weder auf einen unbedingten Effekt der Person noch auf einen unbedingten Effekt der Bedingung zurückgeführt werden kann. Diese Variation manifestiert sich also in den Abweichungen der Messwerte vom Gesamtmittelwert, nachdem der Haupteffekt einer Person und der Haupteffekt der Bedingung a_j von diesen Abweichungen abgezogen wurden (vgl. auch Formel F 13.97a):

$$QS_{Res} = \sum_{j=1}^{J} \sum_{m=1}^{n} \left((x_{mj} - \bar{x}) - (\bar{x}_{\bullet j} - \bar{x}) - (\bar{x}_{m\bullet} - \bar{x}) \right)^2$$

$$= \sum_{j=1}^{J} \sum_{m=1}^{n} \left(x_{mj} - \bar{x}_{\bullet j} - \bar{x}_{m\bullet} + \bar{x} \right)^2 \qquad \text{(F 14.5)}$$

Sie beträgt in unserem Beispiel $QS_{Res} = 484$.

Die Summe aus QS_{zwP}, QS_{zwA} und QS_{Res} entspricht der totalen Quadratsumme.

! Additivität der Quadratsummen

Bei der einfaktoriellen Varianzanalyse mit Messwiederholung lässt sich die totale Quadratsumme QS_{tot} in drei Teile zerlegen:

- einen Teil, der die Variation zwischen Personen ausdrückt (»Haupteffekte« der Person; QS_{zwP}),
- einen Teil, der die Variation zwischen Bedingungen ausdrückt (Haupteffekte des Faktors A; QS_{zwA}), und
- einen unerklärten Teil (QS_{Res}):

$$QS_{tot} = QS_{zwP} + QS_{zwA} + QS_{Res} \qquad \text{(F 14.6)}$$

Variation zwischen und innerhalb von Personen

Versuchen wir nun, die Quadratsummenzerlegung bei der einfaktoriellen Varianzanalyse mit Messwiederholung mit jener ohne Messwiederholung zu vergleichen. Bei der einfaktoriellen Varianzanalyse ohne Messwiederholung haben wir die totale Quadratsumme in zwei Teile zerlegt: einen, der Variation zwischen Bedingungen (QS_{zw}), und einen, der Variation innerhalb von Bedingungen anzeigt (QS_{inn}). Bei der einfaktoriellen Varianzanalyse mit Messwiederholung haben wir die totale Quadratsumme in drei Teile zerlegt: einen, der Variation zwischen Bedingungen (QS_{zwA}), einen, der Variation zwischen Personen (QS_{zwP}), und einen dritten, der unerklärte Variation anzeigt (QS_{Res}).

Variation zwischen Bedingungen. Die Variation zwischen Bedingungen ist in beiden varianzanalytischen Modellen die gleiche, und sie wird auch gleich be-

rechnet (vgl. die Formeln F 13.6b und F 14.4). Da unser Datenbeispiel in Tabelle 14.1 exakt dem in Tabelle 13.1 entspricht, resultiert für die Variation zwischen den Bedingungen in beiden Fällen der gleiche Wert, nämlich $QS_{zwA} = 2.920$. Das bedeutet auch, dass der Anteil der Variation, der auf die experimentelle Manipulation zurückzuführen ist (also der Anteil der QS_{zwA} an der totalen Quadratsumme), für beide Modelle gleich ist. Darauf kommen wir später zurück, wenn es um die Bestimmung der Effektgröße geht. Der einzige formale Unterschied besteht darin, dass es sich bei der QS_{zw} der Varianzanalyse ohne Messwiederholung um eine Variation *zwischen Personen* handelt, während es sich bei der QS_{zwA} der Varianzanalyse mit Messwiederholung aufgrund der intraindividuellen Bedingungsvariation um eine Variation *innerhalb von Personen* handelt.

Variation innerhalb Bedingungen. Bei der einfaktoriellen Varianzanalyse ohne Messwiederholung haben wir die Innerhalb-Quadratsumme wie folgt hergeleitet: Wir hatten einen Fall konstruiert, in dem es nur Unterschiede *innerhalb* der, nicht aber *zwischen* den experimentellen Bedingungen gibt (s. Abschn. 13.1.4). Die Quadratsumme dieser an ihrem jeweiligen Bedingungsmittelwert zentrierten Messwerte betrug gemäß Formel F 13.7 $QS_{inn} = 1.612$. Ein Vergleich mit den Quadratsummen, die wir in diesem Kapitel berechnet haben, zeigt: Dieser Wert entspricht der Summe aus der Zwischen-Personen-Quadratsumme QS_{zwP} und der Residualquadratsumme QS_{Res} (1.128 + 484 = 1.612). Mit anderen Worten: Die QS_{inn} bei der Varianzanalyse ohne Messwiederholung entspricht der Summe aus QS_{zwP} und QS_{Res} bei der Varianzanalyse mit Messwiederholung:

$$QS_{inn} = QS_{zwP} + QS_{Res} \qquad \text{(F 14.7)}$$

Bei der QS_{zwP} handelt es sich um Variation *zwischen Personen,* während es sich bei der QS_{Res} um Variation *innerhalb von Personen* handelt.

QS_{zwP} und QS_{innP}. Während also bei der Varianzanalyse ohne Messwiederholung eine Unterscheidung in Variationsquellen zwischen den bzw. innerhalb der *Bedingungen* im Vordergrund steht, wird bei der Varianzanalyse mit Messwiederholung zunächst danach unterschieden, ob es sich um Variationsquellen zwischen oder innerhalb der *Personen* handelt. Die Variation zwischen Personen drückt sich in der Zwischen-Personen-Quadratsumme QS_{zwP} aus, die Variation innerhalb von Personen in der Zwischen-Bedingungen-Quadratsumme QS_{zwA} und der Quadratsumme QS_{Res}, deren Summe die Innerhalb-Personen-Quadratsumme QS_{innP} ergibt. Sie berechnet sich in unserem Beispiel zu 2.920 + 484 = 3.404. Die Quadratsummenzerlegung bei der einfaktoriellen Varianzanalyse mit und ohne Messwiederholung ist in Abbildung 14.2 grafisch veranschaulicht.

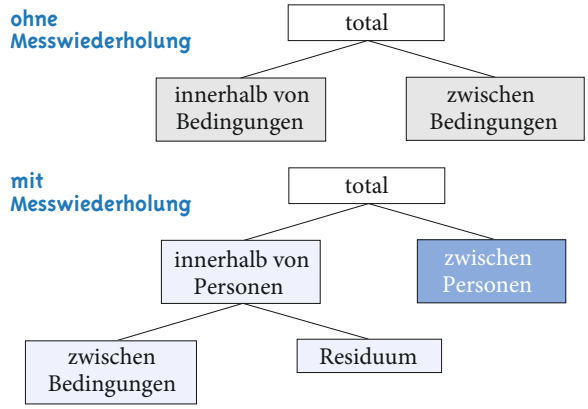

Abbildung 14.2 Quadratsummenzerlegung bei der einfaktoriellen Varianzanalyse mit und ohne Messwiederholung

14.1.3 Effektgrößenmaße

Als Effektgröße bietet sich auch bei der einfaktoriellen Varianzanalyse mit Messwiederholung an, denjenigen Anteil der Varianz der Messwerte zu quantifizieren, der auf den Effekt des Faktors *A* zurückzuführen ist. Insoweit stimmt die Bedeutung des Konzepts »Effektgröße« bei Varianzanalysen mit Messwiederholung genau mit der Bedeutung bei Varianzanalysen ohne Messwiederholung überein. Bei der Schätzung dieser Effektgröße aus den Daten kann das Ausmaß der stabilen Personunterschiede entweder berücksichtigt werden oder nicht. Genau das macht den Unterschied zwischen dem partiellen Effektgrößenschätzer $\hat{\eta}_p^2$ und dem nicht-partiellen Effektgrößenschätzer $\hat{\eta}^2$ aus. Wir werden beide Maße im Folgenden behandeln und im Anschluss daran diskutieren, ob und wann es sinnvoll ist, die Abhängigkeit der Stichproben bei der Effektgrößenschätzung mit zu berücksichtigen oder nicht.

Nicht-partielles Effektgrößenmaß $\hat{\eta}^2$. Beim sog. »nicht-partiellen« Effektgrößenmaß $\hat{\eta}^2$ wird der Quotient aus der Quadratsumme QS_{zwA} und der totalen Quadratsumme gebildet:

$$\hat{\eta}^2 = \frac{QS_{zwA}}{QS_{tot}} = \frac{QS_{zwA}}{QS_{zwA} + QS_{zwP} + QS_{Res}} \qquad \text{(F 14.8)}$$

Partielles Effektgrößenmaß $\hat{\eta}^2_p$. Beim partiellen Effektgrößenmaß $\hat{\eta}^2_p$ wird der Anteil der Gesamtvarianz, der auf stabile Unterschiede zwischen den Personen zurückgeht, nicht mit berücksichtigt. Bei der Schätzung der Effektgröße geht die QS_{zwP} also nicht mehr mit in die Gleichung ein:

$$\hat{\eta}^2_p = \frac{QS_{zwA}}{QS_{zwA} + QS_{Res}} \qquad \text{(F 14.9)}$$

Da beim partiellen Effektgrößenmaß der Nenner niemals einen größeren Wert annehmen kann als beim nicht-partiellen Effektgrößenmaß, kann $\hat{\eta}^2_p$ niemals kleiner werden als $\hat{\eta}^2$. Der Unterschied zwischen beiden wird umso größer, je größer der Anteil der Gesamtvariation ist, der auf stabile Personunterschiede zurückgeführt werden kann.

14

Welches Effektgrößenmaß ist informativer?

Stellen wir uns vor, das zu Beginn dieses Kapitels geschilderte Experiment zum Modelllernen mit drei experimentellen Bedingungen wäre (1) einmal mit einer »echten« Messwiederholung (intraindividuelle Bedingungsvariation: alle Personen durchlaufen alle Bedingungen des Faktors A), (2) einmal mit einer »Quasi-Messwiederholung« aufgrund einer Parallelisierung der Versuchspersonen anhand eines Vortests (interindividuelle Bedingungsvariation, wobei Personen mit gleichen Ausprägungen auf der Vortestvariablen jeweils einer der drei experimentellen Bedingungen zugewiesen werden) und (3) ein drittes Mal ohne Messwiederholung (interindividuelle Bedingungsvariation mit randomisierter Zuweisung der Personen zu einer der drei Bedingungen) durchgeführt worden. Im ersten Fall ist eine hohe Zwischen-Personen-Quadratsumme QS_{zwP} zu erwarten; im zweiten Fall dürfte die QS_{zwP} kleiner sein, da die Abhängigkeit zwischen den Messwerten hier nur noch auf eine einzige Variable (nämlich diejenige, die im Vortest gemessen wurde) zurückzuführen ist; und im dritten Fall ist die QS_{zwP} definitionsgemäß gleich 0. Die QS_{tot} wäre in allen drei Fällen die gleiche, und auch die Haupteffekte der Bedingungen wären in allen drei Fällen exakt identisch. Das nicht-partielle Effektgrößenmaß $\hat{\eta}^2$ würde in allen drei Fällen also den gleichen Wert annehmen. Das partielle Effektgrößenmaß $\hat{\eta}^2_p$ hingegen wäre im ersten Fall größer als im zweiten Fall und dort wiederum größer als im dritten Fall, weil eben die Varianz, die auf konsistente Personenunterschiede zurückgeht (QS_{zwP}), hier nicht mit in den Ausdruck im Nenner eingeht und der Ausdruck im Nenner dementsprechend im ersten Fall kleiner ist als im zweiten Fall und dort wiederum kleiner als im dritten Fall.

Das Beispiel zeigt, dass mit dem partiellen Effektgrößenmaß $\hat{\eta}^2_p$ ein Problem verbunden ist: Untersucht man den gleichen Effekt mit unterschiedlichen Designs (intraindividuelle vs. interindividuelle Bedingungsvariation), dann unterscheidet sich das partielle Effektgrößenmaß $\hat{\eta}^2_p$ zwischen diesen Designs auch dann, wenn die Unterschiede zwischen den Bedingungsmittelwerten in den Designs exakt identisch sind. Unterschiedliche Untersuchungen, die zwar das Gleiche untersuchen, aber unterschiedliche Designs verwenden, sind hinsichtlich ihrer Effektgrößen nicht mehr miteinander vergleichbar. Das ist v. a. dann ein Problem, wenn man im Rahmen einer sog. Metaanalyse versucht, die Effektgrößen, die in vielen unterschiedlichen Primärstudien gefunden wurden, zusammenzufassen (Dunlap et al., 1996). Und dabei sollen standardisierte Effektgrößen ja gerade so definiert sein, dass sie auch über unterschiedliche Studien hinweg miteinander verglichen werden können.

Welches der beiden Effektgrößenmaße informativer ist, hängt auch von der konzeptuellen Bedeutung des Faktors ab, den man untersuchen will. Handelt es sich bei dem Faktor A um eine experimentell manipulierte Variable, die man sowohl mit einem intraindividuellen als auch mit einem interindividuellen Design untersuchen kann, sollte man beide Effektgrößenmaße berichten. Das nicht-partielle Effektgrößenmaß $\hat{\eta}^2$ erlaubt dann einen besseren Vergleich mit Studien, die mit einem interindividuellen Design gearbeitet haben. Handelt es sich bei dem Faktor A hingegen um die Zeit und ist man daran interessiert, wie viel Varianz in den Messwerten durch Veränderungen in der Merkmalsausprägung über die Zeit hinweg aufgeklärt wird, so ist das partielle Effektgrößenmaß $\hat{\eta}^2_p$ informativer, da Unterschiede in der Größe der Personeffekte für die Schätzung des Veränderungseffekts irrelevant sind.

Statistikprogramme wie SPSS geben nur das partielle Effektgrößenmaß an; will man zusätzlich das nicht-partielle Effektgrößenmaß berichten, so muss man es anhand von Formel F 14.8 selbst berechnen.

Beispiel

Effektgrößenmaße für das Beispiel zum Modelllernen

Wie groß ist der Effekt der experimentellen Manipulation (Faktor A) in unserem Datenbeispiel aus Tabelle 14.1? Zunächst berechnen wir das geschätzte nicht-partielle Effektgrößenmaß nach Formel F 14.8:

$$\hat{\eta}^2 = QS_{zwA}/QS_{tot} = 2.920/4.532 = 0{,}64$$

Das partielle Effektgrößenmaß beträgt nach Formel F 14.9:

$$\hat{\eta}^2_p = QS_{zwA}/(QS_{zwA} + QS_{Res}) = 2.920/3.404 = 0{,}86$$

14.1.4 Populationsmodell der einfaktoriellen Varianzanalyse mit Messwiederholung

Das Populationsmodell der einfaktoriellen Varianzanalyse mit Messwiederholung besagt, dass ein Messwert beeinflusst wird durch

- den unbedingten Populationsmittelwert (μ),
- den Populationseffekt derjenigen Bedingung, unter der der Wert erhoben wurde (τ_j),
- Effekte, die auf Eigenschaften der jeweiligen Person zurückgehen (π_m),
- den bedingten Effekt der Bedingung, gegeben eine spezifische Person (Interaktion Person × Bedingung; $(\pi\tau)_{mj}$), und
- alle unsystematischen Einflüsse einschließlich des Messfehlers (ε_{mj}).

Formal:

$$x_{mj} = \mu + \tau_j + \pi_m + (\pi\tau)_{mj} + \varepsilon_{mj} \qquad \text{(F 14.10a)}$$

Da sich die Einflussgrößen $(\pi\tau)_{mj}$ und ε_{mj} mit diesem Design empirisch nicht trennen lassen, gehen wir im Folgenden davon aus, dass alle Interaktionseffekte $(\pi\tau)_{mj}$ gleich 0 sind. Damit verkürzt sich Formel F 14.10a wie folgt:

$$x_{mj} = \mu + \tau_j + \pi_m + \varepsilon_{mj} \qquad \text{(F 14.10b)}$$

Diese Annahme vereinfacht die weitere Ableitung des Modells.

Haupteffekt der Bedingung a_j. Der Koeffizient τ_j wird – genau wie bei der einfaktoriellen Varianzanalyse ohne Messwiederholung – als Haupteffekt einer Bedingung a_j bezeichnet. Er ist definiert als die Abweichung eines Bedingungsmittelwerts ($\mu_{\bullet j}$) vom Gesamtmittelwert:

$$\tau_j = \mu_{\bullet j} - \mu \qquad \text{(F 14.11)}$$

Diese Definition impliziert, dass die Summe der Haupteffekte über alle J Bedingungen hinweg immer 0 ergeben muss:

$$\sum_{j=1}^{J} \tau_j = 0 \qquad \text{(F 14.12)}$$

Die Haupteffekte τ_j variieren nicht über Personen hinweg; sie sind für alle Personen in der Population konstant. Insofern trägt ihre Varianz auch nichts zur Varianz der Messwerte innerhalb einer Bedingung bei.

Haupteffekt der Person m. Der Koeffizient π_m kennzeichnet den unbedingten Effekt einer Person m. Hierunter fallen alle Merkmale dieser Person, die einen Einfluss auf die abhängige Variable haben und von der experimentellen Manipulation unabhängig sind, also über die Messungen hinweg stabil bleiben. Der Haupteffekt π_m ist definiert als die Abweichung eines Personmittelwerts ($\mu_{m\bullet}$) vom Gesamtmittelwert:

$$\pi_m = \mu_{m\bullet} - \mu \qquad \text{(F 14.13)}$$

Auch hier gilt, dass die Summe der Personen-Haupteffekte über alle n Personen hinweg immer 0 ergeben muss:

$$\sum_{m=1}^{n} \pi_m = 0 \qquad \text{(F 14.14)}$$

Da Haupteffekte der Personen zufällige Effekte sind, lässt sich ihre Varianz (σ^2_π) nicht von vornherein kontrollieren; vielmehr handelt es sich um einen Populationsparameter, der aus den Daten geschätzt werden muss. Wichtig ist, dass die Personen-Haupteffekte einen Teil der Unterschiede zwischen den Personen (und damit einen Teil der Varianz innerhalb einer Bedingung) erklären.

Residuum. Wie wir gesehen haben, setzt sich das Residuum ε_{mj} aus allen unsystematischen Einflüssen wie z. B. dem Messfehler zusammen. Es ist derjenige Teil in der Variation der Messwerte, der weder durch Bedingungseffekte noch durch Personeffekte erklärt werden kann. Setzt man die Formeln F 14.11 und F 14.13 in Formel F 14.10b ein und löst nach ε_{mj} auf, ergibt sich:

$$\begin{aligned}\varepsilon_{mj} &= x_{mj} - \mu - \tau_j - \pi_m \\ &= x_{mj} - \mu - \left(\mu_{\bullet j} - \mu\right) - \left(\mu_{m\bullet} - \mu\right) \\ &= x_{mj} - \mu - \mu_{\bullet j} + \mu - \mu_{m\bullet} + \mu \\ &= x_{mj} - \mu_{\bullet j} - \mu_{m\bullet} + \mu \end{aligned} \qquad \text{(F 14.15)}$$

Das Residuum ist also nichts anderes als die Abweichung der einzelnen Messwerte vom Gesamtmittelwert, nachdem sowohl der Bedingungs- als auch der Personeffekt kontrolliert (herausgerechnet) wurden. Die Varianz dieses Residuums (σ_ε^2) ist ein Populationsparameter, der aus den Daten geschätzt werden muss; sie ist ein Teil der Varianz der Messwerte innerhalb der Bedingungen.

Varianz der Messwertvariablen

Unter der Annahme, dass das Modell in Formel F 14.10b gültig ist, lässt sich die Varianz der Messwertvariablen X_{mj} anhand der siebten Rechenregel für Varianzen (vgl. Formel F 7.35) zerlegen. Die Variable X_{mj} repräsentiert die potenziellen Werte einer zufällig gezogenen Person m in einer Bedingung a_j. Der individuelle Wert x_{mj} ist die Realisierung der Variablen X_{mj} in einer konkreten Studie. Wenden wir Formel F 7.35 auf unser Problem an, erhalten wir:

$$\begin{aligned}Var(X_{mj}) &= Var(\mu + \tau_j + \pi_m + \varepsilon_{mj}) \\ &= Var(\mu) + Var(\tau_j) + Var(\pi_m) \\ &\quad + Var(\varepsilon_{mj}) + Cov(\mu, \tau_j) \\ &\quad + Cov(\mu, \pi_m) + Cov(\mu, \varepsilon_{mj}) \\ &\quad + Cov(\tau_j, \pi_m) + Cov(\tau_j, \varepsilon_{mj}) \\ &\quad + Cov(\pi_m, \varepsilon_{mj}) \end{aligned} \qquad \text{(F 14.16a)}$$

Der Gesamtmittelwert μ und die Bedingungs-Haupteffekte τ_j variieren nicht zwischen Personen; der Effekt einer Bedingung a_j ist für alle Personen in der Population gleich. Innerhalb einer Bedingung a_j gibt es also keine Varianz von τ_j. Bei τ_j handelt es sich somit innerhalb von a_j um eine Konstante.

Die Feststellung, dass μ und τ_j Konstanten sind, hat verschiedene Implikationen:

- Die beiden Ausdrücke $Var(\mu)$ und $Var(\tau_j)$ sind beide gleich 0.
- Alle Kovarianzen, an denen μ und τ_j beteiligt sind, sind gleich 0; denn wenn eine Variable eine Konstante ist, kann sie auch nicht mit anderen Variablen kovariieren (s. Abschn. 16.3.1).

Damit verkürzt sich Formel F 14.16a wie folgt:

$$Var(X_{mj}) = Var(\pi_m) + Var(\varepsilon_{mj}) + Cov(\pi_m, \varepsilon_{mj}) \qquad \text{(F 14.16b)}$$

Annahmen

Bei der einfaktoriellen Varianzanalyse mit Messwiederholung werden in Bezug auf die drei Größen auf der rechten Seite von Formel F 14.16b die folgenden drei zusätzlichen Annahmen getroffen (Fahrmeir et al., 1996a):

(1) Die zufälligen Personeffekte π_m sind unabhängig und identisch normalverteilt mit $N(0, \sigma_\pi^2)$.
(2) Die Residuen ε_{mj} sind unabhängig und identisch normalverteilt mit $N(0, \sigma_\varepsilon^2)$.
(3) Die Kovarianz der Personeffekte und der Residuen ist gleich 0: $Cov(\pi_m, \varepsilon_{mj}) = 0$.

Aus diesen Annahmen folgt, dass die Varianz von X_{mj} der Varianz $\sigma_{X_j}^2$ entspricht mit

$$\sigma_{X_j}^2 = \sigma_\pi^2 + \sigma_\varepsilon^2. \qquad \text{(F 14.17)}$$

Bei Gültigkeit der Modellannahmen in der Population muss die Varianz des Merkmals in allen Faktorstufen identisch sein, nämlich $\sigma_\pi^2 + \sigma_\varepsilon^2$. Aufgrund der Annahme, dass die Variablen unabhängig und identisch verteilt sind, lassen wir im Folgenden bei der weiteren Betrachtung der Variablen des Populationsmodells aus Gründen der Vereinfachung den Index m für die Person weg. Mit X_j bezeichnen wir die Variable, deren Werte die individuellen Messwerte x_{mj} in einer Bedingung (bzw. einem Messzeitpunkt) a_j sind, π bezeichnet die Personeffektvariable und ε_j die Residualvariable in der Bedingung a_j.

Kovarianz zwischen den Faktorstufen

Bei Designs mit intraindividueller Bedingungsvariation (Messwiederholung) sind die Messwerte zwischen den Faktorstufen nicht unabhängig voneinander, da sie von den gleichen Personen stammen. Das Ausmaß dieser Abhängigkeit kann über die Kovarianz quantifiziert werden. Das hatten wir bereits in Abschnitt 12.1.1

im Rahmen des *t*-Tests für abhängige Stichproben festgestellt, und in Abschnitt 16.3.1 werden wir die Kovarianz im Detail behandeln. Die Kovarianz ist dann positiv, wenn Individuen, die in einer Bedingung überdurchschnittliche Werte haben, auch in einer anderen Bedingung überdurchschnittliche Werte aufweisen und umgekehrt. In einem Design mit intraindividueller Bedingungsvariation und $J = 3$ Faktorstufen können demnach drei Kovarianzen berechnet werden: $Cov(X_1, X_2)$, $Cov(X_1, X_3)$ und $Cov(X_2, X_3)$. Bei intraindividuellen Designs gibt es also nicht nur Varianzen innerhalb der Faktorstufen, sondern auch Kovarianzen zwischen den Faktorstufen. Varianzen und Kovarianzen werden in Form einer Matrix (Varianz-Kovarianz-Matrix oder einfach *Kovarianzmatrix*) dargestellt. Eine Kovarianzmatrix wird mit dem griechischen Großbuchstaben Σ (Sigma) symbolisiert. Um deutlich zu machen, dass es sich um eine Matrix handelt, wird Σ fett geschrieben. Für den Fall, dass es $J = 3$ Faktorstufen gibt, sieht die Kovarianzmatrix der Variablen X_j wie folgt aus:

$$\boldsymbol{\Sigma}_X = \begin{pmatrix} Var(X_1) & Cov(X_1, X_2) & Cov(X_1, X_3) \\ Cov(X_2, X_1) & Var(X_2) & Cov(X_2, X_3) \\ Cov(X_3, X_1) & Cov(X_3, X_2) & Var(X_3) \end{pmatrix} \quad \text{(F 14.18)}$$

Eine Kovarianzmatrix ist immer quadratisch und hat J Zeilen und J Spalten. In der Hauptdiagonale der Matrix stehen die Varianzen innerhalb einer jeweiligen Faktorstufe. In den restlichen Zellen der Matrix stehen die Kovarianzen zwischen zwei der J Faktorstufen.

In der einfaktoriellen Varianzanalyse mit Messwiederholung wird angenommen, dass die Residuen unterschiedlicher Faktorstufen voneinander unabhängig sind und daher alle eine Kovarianz von 0 aufweisen. Unter den Modellannahmen der einfaktoriellen Varianzanalyse mit Messwiederholung gibt es daher nur einen Grund, aus dem die Messwerte über die Bedingungen hinweg kovariieren: nämlich dass es stabile Personunterschiede gibt. Diese sind auf die Personen-Haupteffekte π_m zurückzuführen; sie manifestieren sich also in der Varianz σ_π^2. Da die Personen-Haupteffekte über alle Messzeitpunkte hinweg konstant sind, müssen alle Kovarianzen gleich sein und der Personvarianz σ_π^2 entsprechen. Mit dieser Feststellung und der Zerlegung der Varianzen in Formel F 14.17 lässt sich die Kovarianzmatrix in F 14.18 wie folgt reformulieren:

$$\boldsymbol{\Sigma}_X = \begin{pmatrix} \sigma_\pi^2 + \sigma_\varepsilon^2 & \sigma_\pi^2 & \sigma_\pi^2 \\ \sigma_\pi^2 & \sigma_\pi^2 + \sigma_\varepsilon^2 & \sigma_\pi^2 \\ \sigma_\pi^2 & \sigma_\pi^2 & \sigma_\pi^2 + \sigma_\varepsilon^2 \end{pmatrix} \quad \text{(F 14.19)}$$

! Kovarianzstruktur des Modells der einfaktoriellen Varianzanalyse mit Messwiederholung

Aus den genannten Annahmen der einfaktoriellen Varianzanalyse mit Messwiederholung folgt, dass alle Faktorstufen in der Population eine konstante Varianz aufweisen, die der Summe aus der Personvarianz und der Residualvarianz entspricht, und dass die Kovarianz der Messwerte zwischen zwei beliebigen Faktorstufen der Personvarianz entspricht.

14.1.5 Schätzung der Populationsparameter

Es gibt vier Populationsparameter, die wir aus den Stichprobendaten schätzen müssen: den unbedingten Populationsmittelwert (μ), den Effekt einer experimentellen Bedingung (τ_j), den Effekt einer Person (π_m) bzw. die Personvarianz (σ_π^2) und das Residuum ε_{mj} bzw. die Populationsresidualvarianz (σ_ε^2).

Unbedingter Populationsmittelwert μ. Den unbedingten Populationsmittelwert μ schätzen wir wieder aus dem Gesamtmittelwert $\bar{x}$ (s. auch Formel F 13.23):

$$\hat{\mu} = \bar{x} = \frac{\sum_{j=1}^{J} \sum_{m=1}^{n} x_{mj}}{J \cdot n} \quad \text{(F 14.20)}$$

Bedingungs-Haupteffekte τ_j. Die Bedingungs-Haupteffekte τ_j schätzen wir aus den Differenzen der Bedingungsmittelwerte $\bar{x}_{\bullet j}$ vom Gesamtmittelwert $\bar{x}$ (vgl. Formel F 13.24):

$$\hat{\tau}_j = \bar{x}_{\bullet j} - \bar{x} \quad \text{(F 14.21)}$$

Personen-Haupteffekte π_m. Die Personen-Haupteffekte π_m schätzen wir aus den Abweichungen der Personenmittelwerte $\bar{x}_{m\bullet}$ vom Gesamtmittelwert $\bar{x}$:

$$\hat{\pi}_m = \bar{x}_{m\bullet} - \bar{x} \quad \text{(F 14.22)}$$

Residuum ε_{mj}. Nachdem die Abweichungen $\bar{x}_{\bullet j} - \bar{x}$ und $\bar{x}_{m\bullet} - \bar{x}$ zur Schätzung der Populationsparameter τ_j und π_m verwendet werden, bleibt nur noch eine Quelle der