Hochschule für Gesundheit

Studienbrief
»Forschungsmethoden in der Logopädie«

-Quantitative Methoden-

Aufbau des weiterbildenden Masterstudiengangs "Evidenzbasierte Logopädie" im Rahmen des BMBF-Verbundprojektes PuG

PuG
Pflege- und Gesundheitswissenschaften
VERBUNDPROJEKTPARTNER
UNIVERSITÄT OLDENBURG, JADE HOCHSCHULE,
HOCHSCHULE FÜR GESUNDHEIT, OSTFALIA HOCHSCHULE

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

AUFSTIEG DURCH BILDUNG
OFFENE HOCHSCHULEN

Prof. Dr. Thomas Hering und Jana Zimmermann

»Forschungsmethoden in der Logopädie«

Quantitative Methoden

IMPRESSUM

Autor*innen:	Prof. Dr. Thomas Hering, Jana Zimmermann
Redaktion:	Sarah Görlich, Juliane Mühlhaus
Herausgeber:	Hochschule für Gesundheit, Bochum
Copyright:	Vervielfachung oder Nachdruck auch auszugsweise zum Zwecke einer Veröffentlichung durch Dritte nur mit Zustimmung des Herausgebers
ISBN:	978-3-946122-08-1

Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 16OH21036 gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei den Autor*innen.

Bochum, September 2016

INHALTSVERZEICHNIS

A PROFIL DER AUTOR*INNEN

PROF. DR. THOMAS HERING

QUALIFIKATIONEN

PROMOTION FREIE UNIVERSITÄT BERLIN
THEMA: „ORGANISATIONSPROFILE UND GESUNDHEIT IM RETTUNGSDIENST“ (2006-2008), VERÖFFENTLICHT MIT DEM TITEL „GESUNDE ORGANISATIONEN IM RETTUNGSDIENST“ IM TECTUM-VERLAG 2009

STUDIUM HOCHSCHULE MAGDEBURG-STENDAL
GESUNDHEITSFÖRDERUNG UND -MANAGEMENT, HEILPÄDAGOGIK UND REHABILITATION (1998-2003)

BERUFSAUSBILDUNGEN
RETTUNGSASSISTENT (1997-1998)
KRANKENPFLEGER (1994-1997)
LOKFÜHRER (1989-1991)

TÄTIGKEITEN

- PROFESSOR FÜR QUANTITATIVE METHODEN, HSG BOCHUM (SEIT 2014)
- REFERENT, MINISTERIUM FÜR ARBEIT UND SOZIALES SACHSEN-ANHALT (2010-2014)
- STUDIENGANGSKOORDINATOR, SPI BERLIN (2009-2010)
- WISSENSCHAFTLICHER MITARBEITER, HOCHSCHULE MAGDEBURG-STENDAL (2003-2009)
- RETTUNGSASSISTENT, LANDKREIS HELMSTEDT, DRK OHREKREIS, JUH MAGDEBURG (1998 UND 2000-2003)
- KRANKENPFLEGER, NEUROLOGISCHES REHAZENTRUM MAGDEBURG (1999)
- LOKFÜHRER, DEUTSCHE BAHN MAGDEBURG (1991-1994)

E-MAIL: THOMAS.HERING@HS-GESUNDHEIT.DE

JANA ZIMMERMANN (M.SC.)

Arbeitsgebiete

Neurogene Sprach-, Sprech-, Stimm- und Schluckstörungen
Quantitative Forschungsmethoden

Kurzvita

- Studium an der Universität Bielefeld „Klinische Linguistik“ mit den Abschlüssen Bachelor of Science und Master of Science

- Sprachtherapeutin im Akutkrankenhaus auf der Stroke Unit, der Frührehabilitationsstation und der HNO-Ambulanz

- Wissenschaftliche Mitarbeiterin an der Hochschule für Gesundheit im Studiengang Logopädie

- Wissenschaftliche Mitarbeiterin an der Hochschule für Gesundheit im Studiengang Evidence-based Health Care

E-Mail: jana.zimmermann@hs-gesundheit.de

B Tabellenverzeichnis

C Abbildungsverzeichnis

D Einleitung

Ein Studienbrief ist kein Lehrbuch im eigentlichen Sinn. Er orientiert sich an der Schwerpunktsetzung eines Moduls. Die Inhalte lehnen sich an vorliegende Veröffentlichungen und Lehrbücher an, geben allerdings lediglich einen Überblick über die Thematik. Für die Vertiefung von Themen ist die Lektüre weiterführender Lehrbücher und Literatur notwendig (s. Anhang IV & V). Dieser Studienbrief erarbeitet das Thema Forschungsmethoden in zehn Kapiteln. Jedes Kapitel beginnt mit der Nennung der Lernergebnisse und einer knappen Kapitelübersicht. Wo es möglich ist, werden die Inhalte grafisch (46 Abbildungen) und tabellarisch (25 Tabellen) veranschaulicht. Ein Kapitel endet mit einer Zusammenfassung der Inhalte. Zur Orientierung im jeweiligen Themenbereich reicht es, die Kapitelzusammenfassung und Kapitelübersicht zu lesen. Notieren Sie sich dabei bereits Fragen, die sich stellen. Innerhalb der Kapitel sind Textfragmente hervorgehoben in denen Zwischenzusammenfassungen erfolgen oder Formulierungsvorschläge enthalten sind. Auch Formeln und Erläuterungen dazu werden im Text hervorgehoben. Jedes Kapitel schließt mit Übungs- und Selbstlernaufgaben ab. Mit diesen Aufgaben festigen Sie Kenntnisse und erwerben Fertigkeiten, Forschung angeleitet zu planen, durchzuführen und Ergebnisse kritisch zu gewichten.

Dieser Studienbrief orientiert sich am Forschungsprozess. Er setzt sich mit Grundlagen der quantitativen Forschung auseinander, die Angehörigen von Gesundheitsberufen im klinischen Alltag häufig begegnet. Qualitative Ansätze werden eingesetzt, wenn beispielsweise unklar ist, welche Ereignisse, unter welchen Bedingungen und warum eintreten – also wenn keine oder nur unzureichende theoretische Erklärung zu Phänomenen gegeben werden kann. Ebenso sind Fragestellungen, die sehr individuelle Entwicklungen und Erklärungen erfordern, wie z.B. über die Bewältigung seltener Erkrankungen, dem Umgang mit Verlust, Trauer und Schmerzen häufig Gegenstand qualitativer Forschungsansätze.

Im ersten Teil werden neben den Zielen empirischer Forschung grundlegende Begriffe eingeführt, die Basis der Kommunikation in der *scientific community* sind. Empirische Forschung muss zahlreichen Kriterien genügen, insbesondere muss sie nachvollziehbar und transparent sein. Bei jedem Schritt müssen Quellen genannt und die Datengrundlage beschrieben werden. Forschung ist frei und darf vieles, muss aber bei Studien mit Menschen die Unversehrtheit der Studienteilnehmer sicherstellen. Ethische Prinzipien setzen den Rahmen, in dem Forschung eingreifen darf.

Fragen danach, wie Wissen entsteht und welche verschiedenen Denkschulen Entstehung von Wissen beleuchten, sind Gegenstände des zweiten Kapitels. Die relevanten Wissensbereiche Alltags- und wissenschaftliches Wissen und ihre Bedeutung werden erörtert. Zudem liegt der Fokus auf wissenschaftstheoretischen Vorstellungen über die Realität oder ihre Konstruktion sowie auf grundsätzlichen Methoden, Wissen zu generieren.

Der Studienbrief orientiert sich am Forschungsprozess in der quantitativen und qualitativen Forschung, der im dritten Kapitel skizziert wird.

Jede Forschungsfrage und jedes Projekt in der quantitativen Forschung gründet sich auf Annahmen darüber, was, warum und in welcher Reihenfolge im Forschungsprozess passieren wird. Im theoretischen Rahmen werden diese Antworten entwickelt. Fragestellungen sind der Fahrplan für den Forschungsprozess, Forschungshypothesen nehmen Ergebnisse einer Studie vorweg.

Das vierte Kapitel definiert Begriffe der empirischen Forschung und erörtert, wie Studienteilnehmer und interessierende Merkmale in der Forschung berücksichtigt werden.

Mit welchem Studiendesign Forschungsfragen genau und zuverlässig beantwortet werden können, wird im fünften Kapitel beleuchtet. Es werden grundsätzlich Studien zu einem Messzeitpunkt und zu mehreren Messzeitpunkten unterschieden. Darüber hinaus variieren Studiendesigns danach, ob beim Studienteilnehmer Maßnahmen durchgeführt werden oder er lediglich beobachtet wird.

Fragen können theoretisch beantwortet werden oder anhand von Daten aus Stichproben. Datenerhebungsmethoden sind Gegenstand des sechsten Kapitels. Ferner wird beleuchtet, wie Merkmale in gesundheits- und sozialwissenschaftlichen Studien in konkrete Fragen und in Zahlen übersetzt werden können (operationalisieren und messen). Messinstrumente liefern aussagekräftige Ergebnisse, wenn sie objektiv, reliabel und valide sind.

Daten werden in Stichproben erhoben. Die Art der Stichprobenrekrutierung entscheidet über die Aussagekraft und die Übertragbarkeit von Forschungsergebnissen. Das siebte Kapitel stellt Verfahren der zufalls- und nicht zufallsbasierten Stichprobengewinnung vor. Es befasst sich mit der Genauigkeit von Stichprobenergebnissen, die von der Stichprobengröße und die Art der Stichprobenziehung beeinfluss wird.

Das achte Kapitel führt in Verfahren der uni- und bivariaten deskriptiven Statistik ein. Sie dienen der Beschreibung der Stichprobe und liefern grundlegende Antworten auf Forschungsfragen. Zu univariaten Verfahren zählen die Bestimmung absoluter und relativer Häufigkeiten, sowie von Lage- und Streuungsmaße. Bivariate Verfahren setzen zwei Merkmale in Beziehung zueinander. Vom Messniveau und der Forschungsfrage hängt ab, welches Verfahren geeignet ist.

Die Wahrscheinlichkeit von Ergebnissen in der Stichprobe unter definierten Annahmen zur Verteilung in der Grundgesamtheit, wird mit Verfahren der Inferenz- (d.h. Prüf-) Statistik bestimmt. Inferenzstatistische Verfahren, die im neunten Kapitel betrachtet werden, lassen keine Schlüsse auf die wahre Konstellation von Kennwerten in der Grundgesamtheit zu. Im neunten Kapitel werden Grundbegriffe erörtert

und Verfahren zum Testen von Unterschieds- und Zusammenhangshypothesen eingeführt.

Abschließend in Kapitel zehn werden Veröffentlichungsvarianten von Studienergebnissen erörtert. Dabei wird die Gliederung wissenschaftlicher Arbeiten ebenso erörtert, wie eine angemessene, knappe wissenschaftliche Sprache und ungeeignete Formulierungen.

Der diesen Studienbrief ergänzende kommentierte Reader vertieft ausgewählte Aspekte qualitativer Forschungsmethodik. Anders als der vorliegende Studienbrief wird der Reader nicht durch die Arbeitsschritte des Forschungsprozesses strukturiert, sondern orientiert sich an verschiedenen Einsatzmöglichkeiten qualitativer Forschungsergebnisse und -methoden in der sprachtherapeutischen Praxis. Zu jedem Bereich wird grundlegende und weiterführende Lektüre erläutert und empfohlen sowie Arbeitsaufgaben zur Vertiefung formuliert.

1 Grundlagen empirischer Forschung

Lernergebnisse:

- Ziele von Forschung können differenziert sowie anhand von Beispielen und konkreten Themen erläutert werden (1.1)
- Grundgesamtheit, Stichprobe, Merkmal und Merkmalsausprägung können definiert und ihre Bedeutung für die Forschungsmethodik beschrieben werden (1.2).
- Die Notwendigkeit guter wissenschaftlicher Praxis kann begründet und wesentliche Gütekriterien von Forschung können definiert werden (1.3).
- Gründe für ethisches Handeln in der Wissenschaft und Instrumente zur Sicherung einer ethischen Wissenschaft können beschrieben werden (1.4).

Die Empirische (auf Erfahrung beruhende) Forschung erhebt und analysiert Daten. Ihre Ergebnisse beschreiben Zustände, vergleichen Gruppen und sind Basis für prognostische Aussagen künftiger Entwicklungen. Ausgehend von offenen Fragen, theoretischen oder praktischen Problemen oder im Rahmen der Entscheidungsfindung werden auf der Basis nachvollziehbarer (Vor-)Annahmen Daten erhoben, systematisiert, Ergebnisse dargestellt und Schlussfolgerungen gezogen. Denkbare Ergebnisse von Forschung sind die (weitgehende) Lösung des Problems, die Antwort auf offene Fragen, der Hinweis auf die passende Entscheidung oder Entwicklungsaussagen. Im Forschungsprozess erhobene und analysierte Daten sind Basis für verschiedene Entscheidungen: Im Gesundheitswesen bilden Daten idealer Weise *eine* Grundlage für Therapieentscheidungen. Der korrekte Umgang mit Annahmen, Fragen, Daten und Ergebnissen, also Transparenz, Ehrlichkeit, Nachvollziehbarkeit, Objektivität, Zuverlässigkeit und Gültigkeit entscheiden mit über die Aussagekraft von Forschungsergebnisse und darüber auch über den Therapieerfolg, die Lebensqualität und vielleicht auch das Überleben von Patienten.

Sozial- und gesundheitswissenschaftliche empirische Forschung „ist die systematische Erfassung und Deutung sozialer (Anm. oder gesundheitlicher) Tatbestände." (Atteslander & Cromm, 2010, S. 3).

Neue Probleme, alternative oder veränderte Lösungswege, offene Fragen jeder Art können in Forschungsvorhaben münden. Im klinischen und gesundheitlichen Kontext interessiert neben der Wirksamkeit neuer Therapieverfahren auch, ob ähnliche Ergebnisse auch bei anderen, nicht untersuchten Patienten wahrscheinlich sind (1.1). Im quantitativen Forschungskontext werden zudem bestimmte grundlegende Begriffe verwendet, wie Grundgesamtheit, Stichprobe, Merkmal und Merkmalsausprägung, die für die Planung, Umsetzung und das Verständnis des Forschungsprozesses wesentlich sind (1.2). Damit Forschungsergebnisse auch außerhalb der untersuchten Stichprobe übertragbar sind und gefundene Veränderungen auch auf die Intervention zurückgeführt werden können, werden Anforderungen an den Forschungsprozess gestellt, die in Gütekriterien zusammengefasst werden (1.3). Auch wenn Forschung hinsichtlich der Wahl von Fragestellung und Methode prinzipiell frei ist, steht das Recht von Lebewesen auf Unversehrtheit und Schmerzfreiheit über den Interessen von Forschung. Die Prüfung ethischer Voraussetzungen sind für alle Forschungsprojekte zwingend, in denen bei Menschen und Tieren Maßnahmen durchgeführt werden oder in denen durch die Datenerhebung Folgen für Menschen und Tiere anzunehmen sind (1.4).

1.1 Ziele von Forschung

Hussy, Schreier und Echterhoff, (2013) definieren mit Beschreiben, Erklären, Vorhersagen und Verändern vier Ziele von Forschung. Synonyme für das *Beschreiben* sind Benennen, Darlegen, Definieren. Beschreiben umfasst die sprachliche Darlegung eines Zustands, die Einordnung in bestehende Begriffe (z.B. Klassifikation von Krankheiten nach ICD sowie10) und die Aufzählung möglicher Erscheinungsformen

eines Gegenstands (Nolting & Paulus, 2008, zit. in Hussy et al., 2013 S. 12). Beschreiben umfasst ferner die Klärung, wie ein Gegenstand empirisch erfasst, d.h. operationalisiert werden kann. Dabei scheint es vergleichsweise einfach zu sein, Messinstrumente für physikalische Merkmale (Temperatur, Druck usw.) zu benennen. Bei sozialwissenschaftlichen Merkmalen ist eine direkte Beobachtung und Messung häufig nicht möglich. Das Sicht- und Messbarmachen eines Merkmals wird Hussy und Kollegen (2013) zufolge als Operationalisierung bezeichnet. Auf gesundheitliche Fragestellungen bezogen wäre beispielsweise interessant, was ein Herzinfarkt ist, wie er sich zeigt, wann er auftreten kann, zu welchen Erkrankungen er gehört und wie man ihn erkennt, also für die Forschung und die Diagnostik sichtbar machen kann. Dabei kann auch ein Herzinfarkt nicht direkt beobachtet werden, sondern es sind neben dem EKG, der Bestimmung bestimmter Blutwerte das Schmerzbild und Aussehen des Patienten von Bedeutung.

Erklären erweitert die Beschreibung des Gegenstands um seine Einflussfaktoren und Effekte. Es werden Gründe genannt, wann, warum und wodurch etwas in einem bestimmten Zustand vorliegt. Häufig interessiert in empirischen Studien ein bestimmter Ausschnitt an Einflussfaktoren und Effekten. Nicht alles, was möglicherweise zu einem Ereignis führen kann und was vom betrachteten Ereignis ausgeht wird näher betrachtet, sondern ein inhaltlich und theoretisch begründeter Ausschnitt. Ferner werden Erklärungen für Zusammenhänge zwischen dem interessierenden Merkmal und anderen Gegenständen formuliert (Hussy et al., 2013). Im Kontext Herzinfarkt bei älteren Erwachsenen würde es beispielsweise Sinn machen nach Zigarettenrauchen, Übergewicht oder Diabetes zu fragen. Dies sind theoretisch und physiologisch begründbare Einflussfaktoren. Voraussichtlich sind diese Erklärungen bei einem 18-jährigen Herzinfarktpatienten nicht sinnvoll. Hier wäre, ebenfalls physiologisch begründbar, die Betrachtung von Fettstoffwechselstörungen, Gefäßfehlbildungen, familiärer Häufung oder genetischer Prädisposition sinnvoller. In den Bereich der Erklärung fällt außerdem der Effekt des Gegenstands. Beim Herzinfarkt lassen sich theoretisch und physiologisch begründbar eine Reihe potenzieller Effekte erklärend heranziehen, Erwerbsunfähigkeit, körperliche Beeinträchtigung oder Tod beispielsweise. Im Rahmen einer qualitativen Studie kann sich das Erklärungsinteresse z.B. auf die Frage richten, wie es dazu kommt, dass Patienten nach einer erfolgreichen Bypass-Operation und einem medizinisch positiv bewerteten Gesundheitszustand nicht in ihr Berufsleben zurückkehren (Borgetto, 1999).

Bei der *Vorhersage* werden Aussagen zum zeitlichen Ablauf von Ereignissen getroffen (Hussy et al., 2013). Aus einer inhaltlichen Aussage können über statistische Analysen Wahrscheinlichkeiten für das Eintreten eines Ereignisses bei Vorliegen einer oder mehrerer Voraussetzungen berechnet werden. Am Beispiel Herzinfarkt wäre beispielsweise von Interesse, um wievielmal häufiger Raucher verglichen mit Nichtrauchern betroffen sind.

In der klinischen Interventionsforschung interessieren ferner *Veränderungen* von Merkmalsausprägungen. Die bereits beschriebenen Zusammenhänge im Zielbereich Erklären sowie die definierte, theoretisch und empirisch begründete zeitliche Reihenfolge von Ereignissen im Zielfeld Vorhersage, lassen Rückschlüsse auf mögliche Maßnahmen zu, mit denen Ereignisse verändert werden können (Hussy et al., 2013). In klinischen Studien werden daher zunächst Voraussetzungen definiert, also um welche Patienten es sich handelt und was der Endpunkt ist, beispielsweise kein Herzinfarkt bei einem konkreten Patienten oder ein insgesamt geringeres Herzinfarktrisiko für Risikogruppen. Im Rahmen des Zielbereichs Veränderungen werden ferner Intervention und Therapieverfahren durchgeführt und definiert, was mit der Therapie erreicht werden soll. Bei der Prävention des Herzinfarkts spielt z.B. eine Rolle, ob das Herzinfarktrisiko zurückgeht, wenn ein bestimmtes Ernährungsregime eingehalten wird oder ob ein bestimmtes Blutdruckpräparat den Blutdruck nachhaltig senkt und damit einen Risikofaktor des Herzinfarkts abmildert.

Forschung beschreibt und erklärt Veränderungen von Merkmalen und sie gibt einen Ausblick auf künftige Ereignisse unter bestimmten Vorbedingungen. Im Rahmen der Beschreibung erfolgt eine nachvollziehbare Begriffsbestimmung, die Einordnung in bestehende Kategorien, die Nennung möglicher Erscheinungsformen und Klärung möglicher Verfahren zur empirischen Erfassung eines Merkmals (Operationalisierung). Mit dem Erklären wird die Beziehung eines Merkmals zu anderen Merkmalen hervorgehoben. Theoretisch begründet werden Einflussfaktoren und Effekte des interessierenden Merkmals bestimmt. Im Zielbereich Vorhersage interessiert, wie die zeitliche Abfolge von Ereignissen ist. Über statistische Analysen soll zudem die Bedeutung eines Einflussfaktors auf einen Effekt quantifiziert werden. Über derartige Analysen lassen sich künftige Ereignisse, z.B. ein Herzinfarkt, mit einiger Wahrscheinlichkeit bestimmen. Der Zielbereich Verändern betrachtet die Bedeutung konkreter Einflussnahme durch den Forscher. Die klinische Interventionsforschung setzt – wiederum theoretisch begründet – bestimmte Maßnahmen, Therapieverfahren und Interventionen ein, um ihre Wirksamkeit in der Prävention oder Therapie bestimmter Krankheiten zu bestimmen.

1.2 Gegenstand empirischer Forschung – Grundgesamtheit, Stichprobe, Merkmal und Merkmalsausprägung

Ergebnisse quantitativ empirischer Forschung sollten für (nahezu) alle betroffenen Menschen gelten. Wenn im klinischen Kontext Therapieverfahren bei essentieller Hypertonie untersucht werden, sollten Therapieempfehlungen für alle von essentieller Hypertonie betroffenen Menschen zutreffen. Dabei stellt sich die Frage, auf welcher Grundlage entsprechende Aussagen getroffen werden dürfen, wenn doch immer nur ein kleiner Ausschnitt aller essentiellen Hypertoniker für Forschung greifbar ist? Im besten Fall werden alle Hypertoniker untersucht. Allerdings werden Forscher so rasch auf ein strukturelles Problem stoßen: ca. 44% der Frauen und 51% der Männer sind von Hypertonie betroffen, ältere in höherem Maß als junge Men-

schen. Auf Deutschland bezogen müssten ca. 45% der Bevölkerung untersucht werden, also 36.900.000 Menschen. Kein noch so engagiertes Forscherteam, wird diese Menge an Probanden untersuchen können. Müssen sie auch nicht. Denn ausgehend von den Daten eines Teils der großen Menge von Hypertonie betroffener Menschen können zumindest Rückschlüsse für die Eignung oder Nichteignung von Therapieverfahren gezogen werden.

Dieses kurze Beispiel weist auf die Gegenstände und grundlegende Begriffe in der empirischen Forschung hin. Forschung betrachtet einen Ausschnitt der Realität, in unserem Fall Menschen mit essentieller Hypertonie. Menschen mit essentieller Hypertonie bilden die Population oder Grundgesamtheit, aus der eine (repräsentative) Stichprobe gezogen wird (s. 7). Empirische Forschung ist selten in der Lage eine vollständige Population bzw. die Grundgesamtheit zu untersuchen. In repräsentativen Stichproben kann mit Verfahren der schließenden (Inferenz-) Statistik (s. 9) berechnet werden, wie wahrscheinlich bestimmte Ergebnisse in Stichproben sind, wenn in der Grundgesamtheit bestimmte Zustände vorliegen (i.d.R. werden keine Unterschiede, Zusammenhänge oder Veränderungen erwartet und sog. Nullhypothesen getestet, s. 4.4, 9.1). Im Forschungsprozess interessiert ferner ein theoretisch begründeter Ausschnitt aller untersuchbaren Merkmale von Menschen und nicht alle Eigenschaften und Merkmale. Anhand des Beispiels würden also theoretisch begründbare Einflussfaktoren auf die essentielle Hypertonie und die essentielle Hypertonie selber als Merkmale betrachtet. Während Körpergewicht, Größe, Ernährungsverhalten, Raucherstatus usw. hochrelevant wären für die skizzierte Fragestellung, interessieren dagegen der bevorzugte Kleidungsstil, die Religion, das Vermögen oder der Bildungsstand weniger (es sei denn sie lassen sich theoretisch begründen). Im Forschungsprozess können zudem nur Merkmale untersucht werden, die in unterschiedlichen Ausprägungen vorliegen. Es ist unmöglich Männer und Frauen hinsichtlich ihres Rauchverhaltens zu vergleichen, wenn eine untersuchte Stichprobe nur ein Geschlecht umfasst. Merkmalsausprägungen werden von Messinstrumenten erfasst (s. 6.2). Messergebnisse sind in der quantitativen Forschung Zahlen.

Grundlegende Begriffe in der Statistik und quantitativen Forschung sind Grundgesamtheit, Stichprobe, Merkmal und Merkmalsausprägung.

Als *Grundgesamtheit* wird eine Population (Menschen) mit bestimmten Eigenschaften verstanden. Grundgesamtheit kann alle Menschen umfassen, in der Forschungspraxis werden allerdings Teilpopulationen definiert, die von bestimmten Merkmalen geprägt sind (Abbildung 1). Idealerweise ist die Grundgesamtheit namentlich bekannt, wie z. B. in Einwohnermelderegistern. In der klinischen Forschung kann die Grundgesamtheit zwar theoretisch beschrieben werden, die zur Grundgesamtheit gehörenden Personen häufig jedoch nicht explizit benannt werden. Dies ist im Sinne einer repräsentativen Stichprobenziehung nicht unproblematisch (s. 7.1).

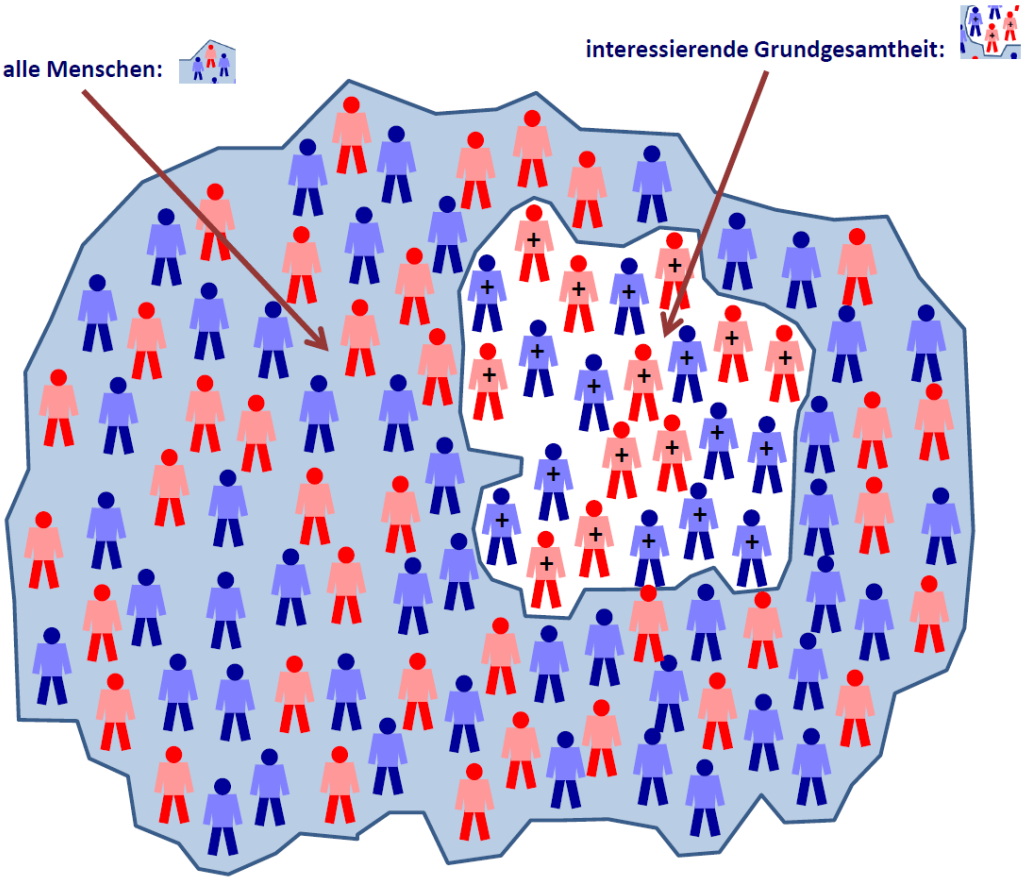


Abbildung 1: Die interessierende Grundgesamtheit sind nicht alle Menschen! (eigene Darstellung)

Aus der Grundgesamtheit wird für konkrete Forschungsvorhaben zufallsbasiert oder willkürlich eine Stichprobe gezogen. Die Stichprobe sollte die Grundgesamtheit repräsentieren, also in wesentlichen untersuchungsrelevanten Kriterien nicht von der Struktur der Grundgesamtheit abweichen (z.B. Geschlechterverteilung, Bildungsstand, Einkommen, Schweregrad der Erkrankung, Therapieregime). Bei der Auswahl und Befragung darf ein bestimmter Teil der Population nicht benachteiligt oder übervorteilt werden. Stichproben sind repräsentativ, wenn die Auswahl aus der interessierenden Grundgesamtheit auf Zufall beruht, jedes Mitglied aus der Grundgesamtheit also dieselbe Chance hatte, an der Studie beteiligt zu werden (s. 7.1).

Quantitativ empirische Forschung in den Gesundheitsberufen ist an bestimmten Eigenschaften, d.h. *Merkmalen*, von Menschen interessiert. Welche Merkmale konkret untersucht werden hängt von der Fragestellung der Untersuchung ab. Merkmale können Körpergröße, Gewicht, Ernährungsverhalten, Lebensqualität oder Blutdruck sein. Mit verschiedenen Messverfahren werden *Merkmalsausprägungen* erhoben. Bei der Erhebung von Merkmalsausprägungen werden qualitative Größen, wie das Gewicht oder die Lebensqualität in Zahlen, also quantitative Größen, übersetzt. Dazu muss das Messinstrument geeignet sein, die Ausprägung des Merkmals

zu erfassen und genau abzubilden. Für viele Merkmale stehen Messinstrumente zur Verfügung: Waagen für das Gewicht, ein Thermometer für die Temperatur oder ein Blutdruckmesser für den Blutdruck. Wie aber misst man Lebensqualität, Stress oder Glück? Die Entwicklung von Messinstrumenten für nicht sofort offensichtliche Merkmale erfolgt in einem mehrstufigen Prozess, der in 6.2 beschrieben wird. Ausgangspunkt dafür ist die *Definition* eines Merkmals. Definitionen enthalten neben der konkreten Beschreibung des Merkmals auch die Erklärung, unter welchen Bedingungen bestimmte Merkmalsausprägungen erkennbar werden. Daneben umfasst die Definition bedeutende Einflussfaktoren eines Merkmals, stellt Gemeinsamkeiten zu ähnlichen und grenzt es verschiedenen Merkmalen ab.

Im *qualitativen Forschungsstil* unterscheidet sich die Stichprobenziehung – oftmals als Fallauswahl bezeichnet – deutlich vom quantitativen Vorgehen. Qualitativ Forschende arbeiten in der Regel mit einer kleinen Anzahl von Fällen, die zielgerichtet ausgewählt und kontextnah untersucht werden. Qualitative Forschung zielt nicht darauf ab, repräsentative und allgemeingültige Ergebnisse vorzulegen, sondern komplexe Phänomene (wie z.B. das oben beispielhaft genannte Phänomen des Ausstiegs aus dem Berufsleben trotz erfolgreicher Bypass-Operation) zu verstehen. Der Geltungsbereich qualitativer Forschungsergebnisse ist kontextgebunden. Im kommentierten Reader werden das Vorgehen in der qualitativen Fallauswahl und die Diskussion um die Verallgemeinerbarkeit qualitativer Ergebnisse aufgenommen und vertieft.

1.3 Gütekriterien von Forschung

Mit der Erhebung und Analyse von Daten ist eine hohe Verantwortung verbunden. An Datenerhebung und Datenanalyse werden somit hohe Anforderungen gestellt. Fehlschlüsse aus ungeeigneten Studien können, wie auch korrekte Schlüsse aus geeigneten Studien, eine hohe Tragweite für einzelne Menschen und Gruppen haben. Für die sozial-, gesundheits- und therapiewissenschaftliche Forschung ist die Subjektivität von Haltungen, Bewertungen und Prognosen von Menschen eine besondere Schwierigkeit bei der Datenerhebung, -auswertung und -interpretation. Der Versuch eine Aussage, z.B. die eines Politikers zu deuten, wird ganz verschiedene Interpretationen ergeben.

Empirische Forschung sollte daher *nachvollziehbar*, *transparent*, *objektiv*, *zuverlässig* und *gültig* sein. Sie darf Informationen nicht überhöhen und muss auch zunächst nicht passende Ergebnisse berichten. Kurz gefasst: empirische (erfahrungsbasierte) Forschung unterliegt in der Planung, Durchführung und Auswertung Regeln, die eine wesentliche Basis für aussagekräftige Forschungsergebnisse bilden. Wissenschaftler, Praktiker, jeder, der auf der Basis von Datenerhebungen und -analysen Aussagen treffen und untermauern möchte, muss diese Regeln kennen und sie einhalten. Methodische Fehler reduzieren nicht nur die Aussagekraft von Studien, sie können zu falschen Schlüssen verleiten, zur Anwendung ungeeigneter Therapien oder zu politischen Fehlschlüssen führen (Balzert, Schröder, Schäfer & Motte, 2011).

Nachvollziehbarkeit bedeutet, jeder Untersuchungsschritt ist begründbar, ergibt sich aus der Fragestellung und ist wiederholbar.

Unter *Transparenz* wird die Offenlegung aller Quellen, Daten, Analysen und (statistischer) Methoden verstanden. Transparenz ist notwendige Voraussetzung für Nachvollziehbarkeit.

Studien sind *objektiv*, wenn die Ergebnisse unabhängig vom Untersucher zustande kommen. Bei gleichen Zuständen des Untersuchungsgegenstands müssen alle Untersucher zum selben Schluss kommen. Auf das Gesundheitswesen übertragen bedeutet dies, alle Untersucher müssten bei ein und demselben Patienten unter Anwendung derselben Untersuchungsmethoden *dieselbe Diagnose stellen*. Mit einer *standardisierten* Datenerhebung, -systematisierung, -analyse und Ergebnisdarstellung werden objektive Ergebnisse erzielt (Hussy, Schreier & Echterhoff, 2013). Bei der Datenerhebung bedeutet Standardisierung, dass Fragen vorformuliert sind und Antwortkategorien vorgegeben sind (standardisierte Messinstrumente) (s. 6.2).

Zuverlässigkeit bzw. *Reliabilität* bezieht sich auf vergleichbare Messergebnisse bei unveränderten Zuständen des gemessenen Merkmals. Ein Fieberthermometer sollte bei zehn Patienten mit derselben Körpertemperatur auch immer dasselbe Ergebnis anzeigen. Objektivität ist die wesentliche Voraussetzung für Reliabilität (Hussy, Schreier & Echterhoff, 2013).

Die Gültigkeit oder *Validität* bezieht sich auf das Messinstrument und das Studiendesign. Soll ein Merkmal gemessen werden, muss das Messinstrument geeignet sein, dieses Merkmal abzubilden. Valide Messergebnisse lassen eindeutige Rückschlüsse auf die Ausprägung eines zu messenden Merkmals zu. Beispielsweise ist zur Messung der Körpertemperatur eine Personenwaage ungeeignet, ein Quecksilberthermometer dagegen passt. Wenn Messinstrumente objektiv und reliabel sind, können sie auch valide sein (Hussy, Schreier & Echterhoff, 2013). Auch Studien können gültige oder ungültige Ergebnisse liefern. Es wird in interne und externe Validität von Studien unterschieden (s. 5.6). Interne Validität fragt, ob ein Effekt zweifelsfrei auf einen untersuchten Einflussfaktor zurückgeführt werden kann. Interessant ist beispielsweise, ob ein Antibiotikum die Heilung von einem grippalen Infekt erzielte, oder ob Patienten nicht auch ohne medikamentöse Behandlung gesund würden. Interne Validität kann durch die Berücksichtigung von Kontrollgruppen und durch die Untersuchung zu mehreren Messzeitpunkten erreicht werden. Bei der externen Validität interessiert ob Studienergebnisse übertragbar sind auf Menschen in einer Grundgesamtheit mit vergleichbaren Eigenschaften. Dazu muss die Stichprobenziehung bestimmte Anforderungen erfüllen und auf Zufall beruhen (s. 7.1).

Forschung geht Fragen allgemeinen Interesses nach. Sie beschreibt Phänomene, liefert Erklärungen und trägt im besten Fall zum Erkenntnisgewinn bei. Die angesprochenen Regeln sind notwendig, weil sich Daten in der sozialwissenschaftlichen und klinischen Forschung auf subjektiven Wahrnehmungen gründen – sie können

also verzerrt sein. Über nachvollziehbare, transparente, objektive, zuverlässige, gültige Methoden und Forschungsdesigns soll – bei aller verbleibenden Fehleranfälligkeit der Ergebnisse – Aussagekraft, Übertragbarkeit und Vergleichbarkeit der Resultate von Forschung hergestellt werden.

Ob und inwieweit die beschriebenen Gütekriterien auch für *qualitative Forschung* Geltung haben, ist Gegenstand einer langanhaltenden Fachdiskussion. Die Debatte ist dabei durch drei Grundpositionen gekennzeichnet (Steinke, 2000): Die erste Position geht davon aus, dass die zentralen Kriterien des quantitativen Forschungsstils auch für die qualitative Forschung gelten sollten. Die zweite Position argumentiert, dass wissenschaftstheoretische und methodische Besonderheiten qualitativer Forschung Ausgangspunkte für die Formulierung spezifischer Kriterien sein sollten. Wie unter Kapitel 1.2 bereits angesprochen, arbeitet qualitative Forschung beispielsweise mit kleinen, gezielt ausgewählten Stichproben. Externe Validität kann also nicht über die Größe und Zufallsauswahl der Stichprobe erreicht und bewertet werden. Die dritte Position lehnt allgemeine Qualitätskriterien grundsätzlich ab. Sie geht davon aus, dass in einer „sozial-konstruierten Welt“ jegliche Standards für höherwertiges Wissen oder Erkennen kritisch hinterfragt werden müssen. In der qualitativen Gesundheitsforschung kann die zweite Position aktuell als dominierend bezeichnet werden. Die Diskussion um angemessene Gütekriterien wird aber weiterhin intensiv geführt und im nachfolgenden kommentierten Reader erneut aufgegriffen.

1.4 Ethische Grundsätze von Forschung am Menschen

Forschung in den Gesundheitsberufen befasst sich mit der Wirksamkeit von Therapieverfahren sowie mit der Bedeutung von Risiko- und Schutzfaktoren bei Menschen. Bei der Beantwortung der Frage, welches Therapieverfahren bei einer bestimmten Krankheit hilft, genügt es nicht, biochemische und physiologische Überlegungen zu berücksichtigen. Vielmehr müssen und werden die Wirksamkeit von Therapieverfahren im Rahmen von Studien an Lebewesen überprüft werden. Grundsätzlich sind derartige Experimente und Forschungen am Lebewesen und Menschen nicht unproblematisch, sie können im besten Fall Nutzen haben, im schlimmsten Fall schaden sie. Die Beachtung ethischer Grundsätze dient dem Schutz der Unversehrtheit von Untersuchungsteilnehmern und soll Missbrauch zu Forschungszwecken vermeiden. In der Geschichte gibt es eklatante Beispiele unethischer Forschung, die bewusst oder unbeabsichtigt Schaden herbeiführten. Die Menschenversuche im Dritten Reich (1933 bis 1945) sind Beispiele unethischer und unmenschlicher Forschung. Um vergleichbare Entgleisungen künftig zu verhindern, orientieren sich die Wissenschaftsdisziplinen an Prinzipien, die in Form von Deklarationen (Deklaration von Nürnberg, 1947, Deklaration von Helsinki, 1996 und die Deklaration von Edinburgh, Herschel, 2013) oder Grundsätzen guter wissenschaftlicher Praxis (Fiedler, 2016) vorliegen. Ethisches Handeln in der Forschung bedeutet also die Achtung und Wahrung der Integrität und des Selbstbestimmungsrechts von Menschen. Dies setzt die Einwilligung von Studienteilnehmern und ihre gewissenhafte Aufklä-

rung über alle bekannten Chancen und Risiken, die mit einer Studie verbunden sind voraus. Hussy und Kollegen (2013, S. 44ff) stellen sechs ethische Prinzipien für die Forschung vor:

1. Gewährleistung der psychischen und physischen Unversehrtheit der Studienteilnehmer.

2. Vollständige Transparenz der Untersuchung für die Studienteilnehmer, sie sind über alle geplanten Maßnahmen, Risiken und Chancen informiert. Über Änderungen im Forschungsprozess wird informiert, alle Kontakte zwischen Studienleitung und Studienteilnehmern werden dokumentiert.

3. Täuschung von Studienteilnehmern soll vermieden werden. Dies lässt sich je nach Art der Studie nicht vermeiden. Beispielsweise ist die Verblindung der Studienteilnehmer streng genommen eine Täuschung, denn sie werden im Unklaren gelassen, ob sie der Interventions- oder der Kontrollgruppe angehören (s. 5.3). Bei diesem Prinzip schlagen Hussy und Kollegen (2013) ein Abwägen der wertrationalen (es darf nicht getäuscht werden) und der zweckrationalen Position (ohne Verblindung keine validen Ergebnisse) vor, bei dem grundsätzlich die Schädigungsfreiheit des Studienteilnehmers gewährleistet sein muss.

4. Die Teilnahme an Studien ist freiwillig. Studienteilnehmer dürfen die Untersuchung jederzeit ohne Angaben von Gründen beenden.

5. Untersuchungsergebnisse werden vertraulich behandelt. Veröffentlichungen geben keinen Rückschluss auf einzelne Studienteilnehmer.

6. Nach Abschluss der Studie werden Studienteilnehmer über alle Details, die Ergebnisse und bei Verblindung über die Gruppenzugehörigkeit (Kontroll- vs. Interventionsgruppe) informiert.

Die Bewertung und das Abwägen zwischen Nutzen durch ein Forschungsprojekt und möglichen Risiken für die Studienteilnehmer erfolgt bei Studien an Lebewesen, Menschen und Patienten durch Ethikkommissionen, die in Forschungs- und Behandlungsinstitutionen gegründet werden. Sie orientieren sich an den vorliegenden Deklarationen und Leitlinien der jeweiligen Disziplin (z.B. Deklaration von Helsinki, 1996, Herschel, 2013). Erst mit einem positiven Votum der Ethikkommission dürfen Forschungsprojekte mit Interventionen an Lebewesen starten.

Zusammenfassend dienen ethische Prinzipien in der Forschung dem Schutz von Menschen, die an Studien teilnehmen. Sie fordern die Gewährleistung von Unversehrtheit, Freiwilligkeit, Transparenz, Vertraulichkeit und die Vermeidung von Täuschung.

Zusammenfassung und Aufgaben

Ziele, Gegenstände, Gütekriterien und ethische Prinzipien von Forschung wurden beleuchtet. Ziel von Forschung ist die Beschreibung, Erklärung, Vorhersage und die Veränderung von Merkmalen. Mit der *Beschreibung* werden alle relevanten Besonderheiten eines Gegenstands dargestellt. Daneben erfolgt seine Einordnung in bestehende begriffliche Kategorien. Ähnlichkeiten und Unterschiede mit vergleichbaren Merkmalen werden genannt. Mit der Identifikation beschreibender Merkmale eines Gegenstands wird die Basis für die Entwicklung von Messinstrumenten gelegt. *Erklärung* hebt die Beziehung eines Gegenstands zu anderen Gegenständen hervor. Zusammengefasst werden die Einflussfaktoren auf den interessierenden Gegenstand und seine Bedeutung für andere Merkmale beschrieben. Dies erfolgt theoriegeleitet mit verschiedenen qualitativen und quantitativen Analysemethoden. Im Kontext *Verändern* wird die Bedeutung der Einflussnahme durch Forscher hervorgehoben. Insbesondere in der klinischen Interventionsforschung werden theoretisch begründet Therapieverfahren eingesetzt. Dabei interessiert insbesondere die Eignung und Wirksamkeit von Therapieverfahren.

Im quantitativen Forschungsprozess interessieren ausgehend von der Fragestellung bestimmte Konstellationen von Eigenschaften beim untersuchten Gegenstand. Alle interessierenden Merkmalsträger (Bluthochdruckpatienten, Männer mit Stirnglatze, bestimmte Autotypen, …) formen die *interessierende Grundgesamtheit*. Zur Grundgesamtheit können insbesondere bei epidemiologischen Studien auch alle Menschen eines umschriebenen Gebiets gezählt werden. Forschung kann dabei kaum alle Merkmalsträger untersuchen, sondern rekrutiert Stichproben, die einer interessierenden Grundgesamtheit entstammen. Über die zufällige Ziehung von Studienteilnehmern aus der bekannten Grundgesamtheit können *repräsentative Stichproben* gebildet werden. Von Studienteilnehmern werden diejenigen Merkmale erhoben, die theoriegeleitet notwendig sind. Soll beispielsweise die Ursache für regionale Durchfallerkrankungen untersucht werden, interessieren gefahrene PKW's, die Schulbildung oder Religion nicht, sondern Ernährungsverhalten, Mobilität oder Vorerkrankungen.

Forschung geht Fragen von allgemeinem Interesse nach. Sie beschreibt Phänomene, liefert Erklärungen und trägt im besten Fall zum Erkenntnisgewinn bei. Insbesondere in der sozialwissenschaftlichen, teilweise auch in der klinischen Forschung stützen sich Ergebnisse auf subjektive Wahrnehmungen und Bewertungen von Menschen – sie sind nicht zwangsläufig fehlerfrei. Über nachvollziehbare, transparente, objektive, zuverlässige und gültige Erhebungs- und Analysemethoden sowie Forschungsdesigns soll – bei aller verbleibenden Fehleranfälligkeit der Ergebnisse – Aussagekraft, Übertragbarkeit und Vergleichbarkeit der Resultate von Forschung gewährleistet werden.

Da in der sozialwissenschaftlichen und klinischen Forschung Lebewesen und insbesondere Menschen untersucht werden, werden ethische Prinzipien formuliert, die Studienteilnehmer vor übereifriger oder nachlässiger Forschung und vor Schaden

schützen sollen. Forschung muss in jedem Schritt die Unversehrtheit von Studienteilnehmern sichern, sie beruht auf Freiwilligkeit, Transparenz, Vertraulichkeit und die Vermeidung von Täuschung. Bei sehr strikter Einhaltung dieser Prinzipien wird Forschung erschwert oder unmöglich gemacht. Daher werden heikle Projekte und Methoden sowie alle Studien, die Interventionen an Menschen untersuchen, durch Ethikkommissionen bewertet und genehmigt. Das sogenannte Ethikvotum ist eine Voraussetzung, damit Forschungsprojekte starten dürfen.

Schlüsselwörter: *Beschreiben, Erklären, ethische Prinzipien, externe Validität, Definition, Grundgesamtheit, Interne Validität, Merkmale, Merkmalsausprägungen, Nachvollziehbarkeit, Objektivität, qualitativer Forschungsstil, qualitative Forschung, repräsentative Stichprobe, Reliabilität, Standardisierung, Transparenz, Validität, Verändern, Vorhersagen*

Aufgaben zur Lernkontrolle

1. Welche Gütekriterien gibt es in der Forschung und wie werden sie definiert?
2. Welche 6 ethischen Prinzipien stellen Hussy und Kollegen (2013) für die Forschung vor?
3. Was wird unter Grundgesamtheit verstanden?
4. Was bedeutet der Begriff „Merkmalsausprägung“?

Aufgabe zur Berufspraxis

5. Recherchieren Sie eine Studie aus Ihrem sprachtherapeutischen Interessensgebiet und markieren Sie in dieser Studie, was beschrieben, erklärt und verändert wird.

Literatur

Atteslander, P., & Cromm, J. (2010). *Methoden der empirischen Sozialforschung* (13., neu bearbeitete und erweiterte Auflage ed.). Berlin: Erich Schmidt Verlag.

Balzert, H., Schröder, M., Schäfer, C., & Motte, P. (2011). *Wissenschaftliches Arbeiten Ethik, Inhalt & Form wiss. Arbeiten, Handwerkszeug, Quellen, Projektmanagement, Präsentation* (2., um 50 Prozent erw. und aktualisierte Aufl. ed.). Herdecke [u.a.]: W3L-Verl.

Borgetto, B. (1999). *Berufsbiographie und chronische Krankheit Handlungsrationalität am Beispiel von Patienten nach koronarer Bypassoperation*. Opladen [u.a.]: Westdt. Verl.

Fiedler, K. (2016). Empfehlungen der DGPs-Kommission „Qualität der psychologischen Forschung“. *Psychologische Rundschau, 67*(1), 59-74. doi: doi:10.1026/0033-3042/a000316

Herschel, M. (2013). *Das KliFo-Buch Praxisbuch klinische Forschung* (2., vollst. überarb. und erw. Aufl. ed.). Stuttgart: Schattauer.

Hussy, W., Schreier, M., & Echterhoff, G. (2013). *Forschungsmethoden in Psychologie und Sozialwissenschaften für Bachelor: mit 23 Tabellen* (2., überarb. Aufl. ed.). Berlin u.a.: Springer.

Steinke, I. (2000). Gütekriterien qualitativer Forschung. In U. Flick, E. v. Kardorff, & I. Steinke (Eds.), *Qualitative Forschung. Ein Handbuch* (pp. 319-331). Reinbek bei Hamburg: Rowohlt.

2 Wissen und Erkenntnisprozess in der wissenschaftlichen Forschung

Lernergebnisse:

- Wissensbereiche können in Alltags- und wissenschaftliches Wissen differenziert und ihre Bedeutung anhand von Beispielen erläutert werden (2.1, 2.2).
- Induktion und Deduktion können definiert sowie qualitativen und quantitativen Forschungstraditionen zugeordnet werden (2.2).
- Vorstellungen von der Realität, ihrer Konstruktion sowie der Wissensgewinnung können definiert und wissenschaftstheoretischen eingeordnet werden (2.3).

Als Kind saß ich wie gebannt vor einem eigenartigen Gerät. Es machte Geräusche, zeigte Bilder, Menschen sprachen mal aus Kairo und im nächsten Moment aus Buenos Aires. Sie machten Witze, schlugen und liebten sich, starben oder waren unverwundbar. Wie kann das funktionieren? Mir wurde nach und nach klar, all das kann sich unmöglich auf einer Bühne mit Minimenschen innerhalb des Geräts in unserem Wohnzimmer abspielen. Wie kam ich darauf? In vielen Haushalten standen diese Geräte. Es müsste also eine hinreichend große Anzahl an Minimenschen geben, die sich alle in Kammern innerhalb dieser Geräte aufhielten. Das hielt ich nicht für logisch. Immerhin müssten sie, wenn sie sind wie ich, wie wir „großen Menschen“, essen, trinken, sich waschen und hin und wieder auf die Toilette. Auch müssten sie einkaufen gehen und es müssten viele von ihnen geben. Leider sah ich außerhalb des Fernsehers keine Kerlchen, die sind wie wir, nur eben kleiner. Sie müssten zudem (wie wir auch) eine Unmenge Müll produzieren, der irgendwo gelagert und entsorgt werden muss. So viele Menschen müssten sich kleiden, irgendwo schlafen und wohnen. Ich schloss daraus, ohne es freilich zu wissen, diese kleinen Kerle gibt es nicht. Zumindest können sie nicht in unserem Fernseher wohnen. Mit einiger kindlicher Neugier und etwas Phantasie schaffte ich einen induktiven und deduktiven Schluss, der in etwa diese Form hatte:

	Ich bin groß und mache Müll Ich bin ein Mensch		Alle Menschen sind groß und machen Müll Ich bin groß und mache Müll
Induktion	Alle Menschen sind groß und machen Müll	**Deduktion**	Ich bin ein Mensch

Abbildung 2: Induktion und Deduktion am Beispiel Fernseher (Struktur aus Wikipedia, 2016)

Wenn auf mich zutrifft: groß zu sein, Müll zu machen und ich demnach ein Mensch bin, können mit den gegebenen Erklärungen im Fernseher keine Menschen wohnen. Es muss also ein anderes Geheimnis hinter diesem eigenartigen Gerät stecken, denn mit einfacher Logik kam ich nicht weiter.

Die nächste Eskalationsstufe kindlichen Forschergeists beginnt bei einem Schraubenzieher und endet mit einer gehörigen Standpauke von den Eltern, die zusammenfassend in der Frage mündete „Was hast du dir bloß dabei gedacht?“. Grund war meine relativ unsystematische Untersuchung des Inneren dieses seltsamen Geräts. Ich wollte durch Beobachten, Experimentieren, kurz durch das Zerlegen des Fernsehers dem Geheimnis auf die Spur kommen, wie er Geräusche und Bilder machen kann. Das Geheimnis lüftete sich irgendwann im Physikunterricht. Hier wurde die Erklärung für das ungelöste Rätsel erarbeitet (Induktion). Ich denke, ich habe es

verstanden und würde heute allenfalls anhand der Anschlüsse versuchen herauszufinden, ob die Übertragung per Internet, digital, terrestrisch oder per Satellit erfolgt (Deduktion) (s. 2.2).

Wissenschaft ist, ebenso wie der Autor in jungen Jahren, daran interessiert, beständiges Wissen zu generieren und Antworten auf interessierende Fragen zu geben – jede Disziplin für sich (Hussy et al., 2013). Jedoch erfolgt die Formung von Wissen und die Formulierung von Antworten auf offene Fragen nicht ausschließlich durch Wissenschaftler – sie wählen lediglich spezielle Methoden und knüpfen an den Erkenntnisprozess bestimmte Anforderungen (s. 5.3) Wissen generieren *alle* Menschen täglich, in dem sie Entscheidungen treffen (müssen). Dies läuft häufig sehr spontan, oft unbewusst und überwiegend unsystematisch „aus einer Laune“ heraus. Im Alltag ist die Fähigkeit, schnell auf Grundlage einer sehr überschaubaren Datenbasis Entscheidungen zu treffen, sehr vorteilhaft und nicht selten überlebenswichtig.

Neben dem *wissenschaftlichen Wissen* existiert ein weiter Bereich von *Alltagswissen*. Beide Wissensbereiche beruhen auf Erkenntnis- und Problemlöseprozessen. Sie unterscheiden sich sehr deutlich in der Systematik, im Tempo, im Methodenspektrum und nicht zuletzt im Allgemeingültigkeitsanspruch.

2.1 Wissensbereiche I: Alltagswissen

Das Fernsehbeispiel weist auf unterschiedliche Wege der Erkenntnisgewinnung hin. Wenig überraschend gehen nur sehr wenige Menschen bei der Entscheidungsfindung im Alltag systematisch vor oder nutzen Methoden, die nachgewiesener Weise geeignet sind, um bestimmte Rückschlüsse zu ziehen. Im Alltag werden Entscheidungen nicht selten ad hoc und auf der Basis einer sehr überschaubaren Informationsgrundlage getroffen: Spontanes Einkaufen der Abendmahlzeit, Auswahl der Bekleidung für den Tag, Nutzung des eigenen PKW, des Fahrrads oder öffentlicher Verkehrsmittel und vieles andere mehr. Kaum einer dieser „profanen“ Alltagsentscheidung liegt ein systematischer Prozess zugrunde, vielmehr spielen Erfahrung, Überzeugungen, Gewohnheiten oder soziale Ansteckung eine große Rolle. Selbst ob, wann und in wen man sich verliebt, unterliegt einer zufälligen Konstellation von Affekten, Gedanken, Ereignissen, Handeln und eigentlich (hoffentlich) nie einem systematischen Entscheidungsprozess, auch wenn Portale wie Parship® etwas anderes suggerieren und meinen, Partnervorschlägen wissenschaftlich überzeugende Algorithmen zugrunde zu legen (wie z.B. Clusteranalysen).

Das sogenannte Alltagswissen ist Basis für das Handeln im Alltag, für die Wahl von Nahrungsmitteln, Bekleidungsentscheidungen, Kaufentscheidungen, die Entscheidung zu trinken zu rauchen oder dies eben nicht zu tun. Menschen sind ständig mit Situationen konfrontiert, die ihre Reaktion erfordern. Ständig werden (teils unbewusst, teils bewusst) Entscheidungen getroffen, weil Entscheidungen notwendig sind um im Alltag handeln zu können. Holzkamp (1968) spricht vom „Zwang zur

handelnden Daseinsbewältigung“ (zit. in Rott, 1995, S. 12). Rasches Treffen von Entscheidungen gelingt nicht ohne Irrtum. Alltagswissen deckt sich oftmals auch nicht mit wissenschaftlichem Wissen. Sie kennen sicher den Verweis von Rauchern auf Helmut Schmidt. Oder sorgfältig getroffene Kaufentscheidungen für ein Auto, bei dem Käufer vielleicht auch sehr systematisch vorgingen, Magazine und Testberichte lasen, in Internetforen aktiv waren usw., werden verworfen, wenn ein guter Bekannter oder ein geachtetes Familienmitglied energisch auf Basis einer Einzelerfahrung oder auf Basis von „Hören-Sagen“ vom Kauf abrät. Alltagswissen und Alltagsentscheidungen sind emotional gefärbt und stark von Gefühlen beeinflusst (Rott, 1995).

Alltagswissen beruht auf subjektiven (Lebens-) Erfahrungen, auf Meinungen, Vermutungen, Behauptungen, Überzeugungen und Traditionen. Denn wer weiß eigentlich, dass Kakerlaken nicht schmecken? Niemand, dieses Nahrungsmittel entspricht lediglich nicht unseren Traditionen und Ernährungsgewohnheiten, die die meisten von Kindesbeinen an prägten. Der Erkenntnisprozess beruht auf einem Abgleich mit bereits gespeichertem Wissen, auf Bewertung und die Reaktion auf Handeln. Dabei wird von der Allgemeingültigkeit von Beobachtungen und Erkenntnissen ausgegangen, die alles andere als zutreffend sein muss (z.B. kennt sicher jeder Menschen, deren erster Eindruck ihnen nicht gerecht wurde – positiv und negativ).

Alltagswissen formt sich auch durch „überzeugt werden“ und mit der Orientierung an Autoritäten (Hussy et al., 2013). In der Überzeugungsstrategie wird die Antwort auf eine Frage energisch und überzeugt vertreten, ohne dass die genaue Antwort bekannt ist. „Die Renten sind sicher“ war eine Überzeugung des in den 1980er Jahren zuständigen Arbeitsministers Norbert Blüm. Man muss fast sagen: wider besseren Wissens. Auch Eltern überzeugen ihre Kinder vom Nutzen der Zahnbürste, warmer Kleidung im Winter, von regelmäßiger Körperpflege und der gesundheitsförderlichen Wirkung von Spinat. Die wenigsten dieser alltagspraktischen Fertigkeiten werden vermittelt, weil Eltern systematisch recherchiert haben und den Nutzen ihrer Empfehlungen theoretisch begründen können (Hussy et al., 2013).

Daneben stellt die Orientierung an und die Berufung auf Autoritäten eine Strategie zur Entwicklung von Alltagswissen dar. Hussy et al. (2013) nennen dabei renommierte Wissenschaftler und Experten als Quelle. Doch auch diese Strategie ist fehleranfällig. Denn die Orientierung an einem Experten oder einen Wissenschaftler kann nicht nur zur Verbreitung falscher Schlüsse, sondern auch zum Scheitern in verschiedenen Situationen führen. Zudem irren sich Wissenschaftler, wie alle Menschen. Nicht zuletzt können sich Rahmenbedingungen ändern, die es notwendig machen, bislang (auch wissenschaftlich untermauerte) Annahmen zu überdenken. Ein Beispiel dafür kann in der erklärten Notwendigkeit von Wirtschaftswachstum gesehen werden, die vor dem Hintergrund endlicher Ressourcen, einer (zumindest in der ersten Welt) alternden und schrumpfenden Bevölkerung nicht logisch und tragfähig erscheint.

Bauernregeln oder in Sprichworte gefasste Verhaltensnormen sind überlieferbare Ergebnisse alltagswissenschaftlicher Erkenntnisprozesse und unsystematischer Beobachtungen, die teilweise bis heute handlungsleitend sind.

Zusammengefasst beruht Alltagswissen auf einem lebensnotwendigen aber hochgradig irrtumsanfälligen Erkenntnisprozess, der auf Annahmen basiert, die sich aus wenigen Informationen, vielen subjektiven Erfahrungen und Gefühlen speisen. Alltagswissen entsteht unsystematisch, ist dabei aber fundamentale Voraussetzung um Anforderungen des Alltags zu bewältigen. Überzeugend vertretene Aussagen und die Orientierung an Autoritäten formt Alltagswissen.

2.2 Wissensbereiche II: Wissenschaftliches Wissen

Wissenschaft sucht Antworten auf alltagsrelevante Fragen, für das Zusammenleben von Menschen, für die Vermeidung von Gefahren, für Prävention und die Behandlung von Krankheiten. Dazu muss sie erklären, was, warum, unter welchen Bedingungen und in welcher Reihenfolge passiert. Dabei ist zu Beginn des Forschungsprozesses oft noch vieles unklar. Durch systematische Beobachtung, Befragung oder einen gezielten Eingriff (Intervention) klären sich die eingangs gestellten Fragen nach dem „Was“, „Warum“, „unter welchen Bedingungen“ und „in welcher Reihenfolge“ Phänomene auftreten. Diese Art der Forschung wird unter Grundlagenforschung zusammengefasst, ihr Ziel ist die Entwicklung von Theorien, also von Erklärungen, warum Dinge unter bestimmten Bedingungen genau so und nicht anders passieren. Wissenschaftstheoretisch handelt es sich um *induktiv-explorative* Forschungsansätze.

Wissenschaft muss andererseits klarstellen, ob die beschriebenen Phänomene und Theorien allgemeingültig sind, also für alle Menschen oder für bestimmte Gruppen mit vergleichbaren Merkmalen zutreffend sind. Bei dieser Anwendungsforschung steht prinzipiell schon fest, was passieren muss. Die offene Frage ist hier lediglich ob ein Phänomen in der beschriebenen Weise auch für bestimmte Gruppen gilt. Therapie- und Gesundheitsforschung ist häufig genau diese Anwendungsforschung. Therapeuten wenden Maßnahmen an, von denen sie den physiologischen (theoretischen) Funktionsmechanismus kennen. Sie nutzen fundierte Theorien und Erklärungen für Fälle und Krankheitsbilder. Sie gehen der Frage nach, ob eine Intervention, die theoretisch funktionieren müsste, auch in einem konkreten Fall oder einer Gruppe von Fällen erfolgreich ist. Wissenschaftstheoretisch wird hier von *deduktiv-explanativen* Ansätzen gesprochen.

2.2.1 Induktion

Im induktiven Ansatz wird auf der Basis von Einzelbeobachtungen auf allgemeine Gesetzmäßigkeiten bzw. sich wiederholende Muster und Strukturen geschlossen (induktiv/logischer Schluss). Unter diesem Ansatz wird geforscht, wenn Beobach-

tungen nicht zu bislang beschriebenen Regeln, Theorien und Gesetzmäßigkeiten passen oder ihnen eklatant widersprechen. Ziel induktiver Forschungsansätze ist es, Theorien zu entwickeln – also Erklärungen was, unter welchen Bedingungen und in welcher Reihenfolge passieren wird.

Beispiel für ein einflussreiches Ergebnis induktiver Forschung sind die Arbeiten von Aaron Antonovsky in den 1960er und 1970er Jahren zum Thema Salutogenese und damit zur Frage: Wie entsteht Gesundheit (Antonovsky, 1991, Antonovsky & Franke, 1997, Bengel, Strittmatter & Willmann, 2006)? In dieser Zeit beherrschten Vorstellungen von Stress und seinen Folgen die wissenschaftliche Forschung, die zusammengefasst davon ausgingen, dass Menschen, die großen (traumatischen) Stress erlebten, mit einer sehr großen Wahrscheinlichkeit krank sind und langfristig Probleme haben werden. Antonovsky war jüdischer Sozialwissenschaftler. Er erlebte in seiner Praxis insbesondere Frauen, die den Holocaust überlebten, in Konzentrationslagern gefangen gehalten wurden, die unvorstellbares Leid gesehen haben und ertragen mussten. Im Gegensatz zur herrschenden Theorie waren diese Frauen nicht krank, erfreuten sich bester Gesundheit, führten ein erfülltes Leben und waren glücklich. Diesen Widerspruch konnten die bis dahin veröffentlichten Stresstheorien und Vorstellungen von Gesundheit und Krankheit nicht auflösen. Es gab keine schlüssige wissenschaftliche Erklärung für dieses Phänomen außer Zufall. Da es sich jedoch um eine hinreichend große Zahl von Frauen handelte, war Zufall keine gute Erklärung. Als Wissenschaftler suchte er systematisch und mit transparenten Methoden nach dem Grund für diese unerwarteten Beobachtungen. Er befragte die Frauen persönlich im Rahmen narrativer (erzählender) Interviews. Es handelt sich dabei um ein qualitatives Vorgehen, bei dem das Ziel verfolgt wird, die Befragten zu intensiven und detaillierten Erzählungen ihrer Lebensgeschichte anzuregen und (in diesem Fall die wider Erwarten gesunde Frauen) dazu zu bringen, Gründe für ihren Gemüts- und Gesundheitszustand anzuführen. Dies wurde nicht nur bei einer Betroffenen, sondern bei vielen Frauen durchgeführt, auf die diese unerwarteten Beobachtungen zutrafen. Antonovsky versuchte also, das bis dato unzusammenhängende Chaos an Beobachtungen zu ordnen und zu systematisieren. Bei der qualitativen Auswertung der Interviews stellte Antonovsky ein Muster an Merkmalen bei diesen Frauen fest, das möglicherweise Grund für ihren unerwartet guten Gesundheitszustand sein konnte: der Kohärenzsinn. Die Mehrzahl der untersuchten Frauen verstand, was ihnen zustieß (Verstehbarkeit), konnte auch den quälenden Erfahrungen eine Bedeutung beimessen (Sinnhaftigkeit) und fühlte sich auch in der sehr ausweglosen Situation handlungsfähig (Handhabbarkeit) (Antonovsky & Franke, 1997). Die Komponenten subsummierte er unter dem Begriff „Kohärenzgefühl“. Menschen mit starkem Kohärenzgefühl scheinen demnach in der Lage zu sein, selbst stark herausfordernde und belastende Situationen gelingend zu bewältigen und langfristig gesund zu bleiben.

Die auf der Basis von Induktion erarbeiteten Erklärungen sollen über die beobachteten Fälle hinaus gültig sein, das heißt auf vergleichbare Situationen zutreffen (Abbildung 2). Diese in der Logik fundierte Annahme ist in den Sozialwissenschaften

nicht 1:1 umsetzbar und haltbar. Denn es wird immer Einzelfälle geben, auf die eine sozial- oder gesundheitswissenschaftliche Theorie nicht zutrifft. Sie werden beispielsweise immer sehr gesunde und uralte Kettenraucher finden bzw. sehr kranke und jung gestorbene vormals gesund lebende Menschen. Induktive Schlüsse in der Sozialwissenschaft haben also keinen (naturwissenschaftlichen) Gesetzescharakter. Vielmehr werden sie als gültig betrachtet, wenn sie für einen hinreichenden Anteil von Beobachtungen zutreffen. Was hinreichend ist, lässt sich ebenfalls nicht beziffern. Sozialwissenschaftliche Erklärungen werden als probabilistische Theorien verstanden, ihr zutreffen muss hinreichend wahrscheinlich (*probabilistisch*) sein. Dagegen haben induktive Schlüsse in der Naturwissenschaft Gesetzescharakter. Im Unterschied zu den probabilistischen sozialwissenschaftlichen Theorien wäre eine konkurrierende Beobachtung in einer naturwissenschaftlichen Theorie Grund dafür, die naturwissenschaftliche Theorie zurückzuweisen. Würde beispielsweise ein Metall gefunden, das sich bei Wärme nicht ausdehnt, sondern kleiner wird, müssten Bereiche der Physik und die Annahme von der Bewegung der Atome überdacht werden. Naturwissenschaftliche Erklärungen werden als *deterministische* Theorien (Theorien mit Gesetzescharakter) bezeichnet. Dabei ist das Vorgehen bei der Entwicklung naturwissenschaftlicher Theorien ähnlich mit dem sozialwissenschaftlichen Vorgehen. Bis dato allgemeingültige Annahmen werden durch konkurrierende Beobachtungen in Zweifel gestellt, neue systematische Beobachtungen sind Basis für die Beschreibung eines Alternativmodells. Nach Galileis Beobachtungen beispielsweise zeigte der Planet Venus Phasen. Demnach kann sie sich nicht um die Erde drehen, wie das im Mittelalter prominente geozentrische Weltbild annahm. Galileis und Keplers Beobachtungen waren Basis der modernen Astronomie und führten mit einiger Zeit Verzögerung zur Aufgabe des geozentrischen Weltbildes zugunsten eines heliozentrischen.

In dem beschriebenen induktiven Vorgehen werden zusammengefasst folgende Merkmale deutlich:

1. Bestehende Regeln und wissenschaftliche Theorien liefern keine Erklärung für eine Beobachtung oder systematische, wiederholte Beobachtungen widersprechen bislang als gültig betrachtete Theorien.

2. Ausgewählte Fälle, die bestimmte, nicht theoriekonforme, bisher nicht erklärbare Besonderheiten aufweisen werden intensiv untersucht. Es wird nach Gründen für diese Besonderheiten gesucht.

3. Auf der Grundlage einer größeren Anzahl von Interviews und Befragungen von Menschen, mit den beschriebenen Eigenschaften wird (sofern erkennbar) ein Muster an Erklärungen für das beobachtete Phänomen entwickelt.

4. Am Ende des induktiven Prozesses steht ein alternatives Muster an Erklärungen, also eine neue bzw. eine modifizierte Theorie.

Induktive Schlüsse in der Sozialwissenschaft sind Grundlage für probabilistische Theorien, die nicht aufgrund konkurrierender Einzelbeobachtungen zurückgewiesen werden können. Die Entwicklung von Theorien erfolgt auf der Grundlage qualitativer Forschungsansätze. Diese Ansätze erheben und analysieren verbale Daten (z.B. im Rahmen narrativer Interviews) und visuelle Daten (z.B. im Rahmen teilnehmender Beobachtungen). Sie sind die Verfahren der Wahl, wenn Beobachtungen bestehenden Theorien widersprechen oder mit existierenden Annahmen nicht erklärt werden können. Qualitative Forschungsansätze unter einem induktiven Design weisen nicht nur darauf hin, dass etwas passiert, sie liefern Erklärungen warum ein Phänomen auftritt. Naturwissenschaftliche Induktion liefert deterministische Theorien, die bereits bei einer widersprechenden (systematischen) Einzelbeobachtung vollständig revidiert werden müssen. Der Prozess in der Naturwissenschaft beruht ebenfalls auf der Beobachtung und Beschreibung des Phänomens – also auf einem quasi-qualitativen Ansatz.

2.2.2 Deduktion

Im induktiven Prozess wird ein wahrnehmbares, bislang unerklärtes „Chaos“ geordnet. Die bislang nicht erklärte Realität wird auf der Basis von Beobachtungen beschrieben. Gründe für genau diese beobachtete Realität (z.B. der Zustand eines Patienten) werden ergründet und erklärt. Der Anspruch an induktiv-explorative Forschung ist es, über die Einzelbeobachtung hinausgehende Erklärungen zu finden. Wenn, wie am Beispiel Antonovskys, aber ausschließlich den Holocaust überlebende Frauen befragt wurden und ein darauf basierendes Modell beschrieben wurde, kann auf dieser Basis auch auf andere Personengruppen geschlossen werden, die ebenfalls traumatische Erfahrungen gemacht haben? Treffen die Aussagen zur Salutogenese und zum Kohärenzgefühl also auch auf Eltern zu, deren Kinder gestorben sind, auf Überlebende von Kriegen, Unruhen und Verkehrsunfällen? Diese Frage kann durch die Fortsetzung des induktiven Prozesses beantwortet werden. Allerdings wäre die Fortsetzung rein qualitativer Forschungsansätze bei größeren Personengruppen sehr zeitaufwändig und nicht zuletzt unnötig, um Antworten zu liefern. Denn im Unterschied zum Ausgangspunkt induktiv-explorativer Forschung liegt nun ein Modell, eine Theorie vor, die Hinweise liefert, warum Menschen auch nach großen Belastungen gesund bleiben. Im Rahmen deduktiv-explanativer Forschung erfolgen nun die Erkundung und die Überprüfung der Theorie auf Allgemeingültigkeit. Deduktive Forschungsansätze fragen also nicht nach dem *Warum*, sondern *ob* eine Erklärung auch für andere Situationen, als die im induktiv-explorativen Vorgehen betrachtete, gültig ist. Bezogen auf das Beispiel Galileis könnte gefragt werden, ob nicht auch der Mars oder der Merkur Phasen zeigt, um das heliozentrische Weltbild zu untermauern. Deduktive Forschung erfolgt in quantitativen Forschungsdesigns. Auf der Basis einer Theorie werden Fragebögen/Messinstrumente entwickelt, mit denen die interessierenden Merkmale (z.B. des Kohärenzgefühls) abgebildet werden. Die qualitativen Merkmale werden also operationalisiert (s. 6.2). Mit den erarbeiteten Fragebögen wird eine große Zahl zufällig ausgewählter Menschen befragt.

Die Daten werden in Statistikprogrammen eingelesen und mathematisch ausgewertet, wobei das Ziel quantitativer Forschung neben der Beschreibung der untersuchten Personen und Merkmale auch die Analyse der zugrundeliegenden Theorie ist. Im Prinzip steht das Ergebnis quantitativ-deduktiver Forschung bereits fest, bevor eine Analyse erfolgt ist, denn die Theorie liefert bereits die Erklärung für das was erkennbar werden soll und für den Grund der Ergebnisse. Ziel deduktiver Ansätze ist die Überprüfung der Theorie und der Gültigkeit von annahmen für bestimmte Subgruppen, also von Menschen, die sich hinsichtlich ihrer Merkmale von denen im induktiven Ansatz untersuchten unterscheiden. Liefern Ergebnisse deduktiver Forschung Hinweise auf unzureichende Theorien erfolgt die Revision, Erweiterung, Ergänzung oder Veränderung einzelner Aussagen in den Theorien. Einzelne der theoretischen Annahmen widersprechende Forschungsergebnisse sind in den Sozialwissenschaften kein Grund dafür, eine Theorie zurückzuweisen. Stattdessen werden Erklärungen gesucht, warum sich aussagekräftige Theorien nicht oder nur zum Teil untermauern lassen. Erklärungen dafür finden sich in Veröffentlichungen, die zu ähnlichen Schlüssen kamen, sie können begründet sein in der untersuchten Stichprobe oder dem verwendeten Fragebogen. Theorien, die sich in vielen Forschungsprojekten nicht widerlegen ließen, wird eine große Aussagekraft beigemessen. Sie werden jedoch auch bei einer großen Anzahl von Studien nicht als „bestätigt“, „belegt“ oder „bewiesen“ bewertet, sondern ihnen wird eine große Aussagekraft beigemessen (Hussy et al., 2013).

Zusammengefasst geht Deduktion von bereits vorliegenden Erklärungen und Theorien aus, die in induktiven Forschungsprozessen erarbeitet wurden. Es wird überprüft, ob die Theorie allgemeingültig ist, also für ganz verschiedene Situationen zutrifft. Deduktive Forschung ohne Theorie ist unzweckmäßig, sie beantwortet keine Fragen und ist überflüssig. Deduktiv-quantitative Designs sind nur sinnvoll, wenn die Erklärung für eine Annahme bereits im Vorfeld vorliegt, also eine Theorie vorhanden ist. Theoriefreie quantitative Forschung kann Stichproben beschreiben, nicht jedoch Gründe für Phänomene liefern oder Hinweise auf die Wirksamkeit von Therapien liefern.

2.3 Wissenschaftstheorie oder wie erklärt sich die Entstehung von Wissen?

Die bisherigen Ausführungen betrachteten verschiedene Wissensbereiche. Das Alltagswissen ist eine, zwar höchst fehleranfällige, zugleich aber notwendige und nicht zuletzt bewährte Grundlage für das Funktionieren von Menschen im Alltag. Im Unterschied zum unsystematischen Alltagswissen geht Wissenschaft mit bestimmten Methoden vor, um Wissen zu entwickeln. Wissenschaftliches Wissen hat den Ruf verlässlich, nachvollziehbar und wahr zu sein. Die unterschiedlichen Wissenschaftsdisziplinen beruhen, was die Entstehung und den Erklärungsgehalt von Wissen angeht, auf unterschiedlichen philosophischen Denkschulen. Wissenschaftstheorie beschäftigt sich als ein Teilbereich der Philosophie mit den Menschenbildern in der Wissenschaft. Schäfer (2016) differenziert zwei Extrempositionen dazu, *was* mit

Wissenschaft eigentlich herausgefunden werden kann (Realismus vs. Konstruktivismus) und auf welchem Weg, also *wie* Erkenntnis eigentlich entsteht (Rationalismus vs. Empirismus). Bei der Frage danach, was eine Wissenschaft herausfinden kann, existiert nach der Position des *Realismus* eine vom Menschen „…unabhängige, unveränderliche Wirklichkeit…“ (Schäfer, 2016, S. 6). Wissenschaftler sind in der Lage diese Wirklichkeit unter Nutzung wissenschaftlicher Methoden vollständig erfassen. Bis die Realität vollständig beschrieben ist vergeht Zeit. Der Realismus ist wissenschaftstheoretisches Paradigma der meisten Naturwissenschaften. Im Unterschied dazu hebt die Position des *Konstruktivismus* das Bewusstsein, das Erleben und Wahrnehmen von Menschen hervor. Wirklichkeit *ist nicht* sondern entsteht durch bewusste (subjektive) Wahrnehmung im Individuum. Menschen ist eine allgemeingültige und objektive Beschreibung einer Realität, die nach dieser Position nicht als interindividueller Konsens existiert, schlicht nicht möglich, weil Wahrnehmungen subjektiv sind und von Lebenserfahrung, Aufmerksamkeit, Gefühlen (…) beeinflusst werden (Schäfer, 2016).

Auf einem zweiten Kontinuum stellt Schäfer (2016) die Methode der Erkenntnisgewinnung in den Mittelpunkt und wirft die fragt nach dem „Wie“ der Entstehung von Erkenntnis auf. Auch hier stehen sich zwei wissenschaftstheoretische Positionen gegenüber. Auf der einen Seite entsteht nach dem *Rationalismus* neues Wissen und Erkenntnis durch Nachdenken und die intensive Nutzung des eigenen Verstandes. Nicht viele Disziplinen sehen diese Position als ihre wissenschaftstheoretische Basis ihres Handelns. Für die Mehrzahl sozial-, gesundheitswissenschaftlicher und klinischer Forscher beruht Erkenntnis auf Erfahrung (Empirie). Verbunden mit dieser als *Empirismus* bezeichneten Position ist die Annahme, über eine hinreichend große Menge an Erfahrung zu einem Gegenstand entstehe ein mehr oder weniger objektives Bild der Realität. Im Unterschied zum reinen Nachdenken im Rationalismus sind Erfahrungen zugänglich für jedermann und damit auch nachprüfbar (Schäfer, 2016). Wobei Nachdenken auch nicht schadet, wie in der Position des kritischen Rationalismus erkennbar wird.

Allerdings wird die auf dem Empirismus fundierte Annahme, Realität könne durch Beobachtung hinreichend genau abgebildet werden, kritisiert. Die Kritik gründet sich auf die Subjektivität von Wahrnehmung und damit auch und in viel stärkerem Maß auf die Schlüsse aus subjektiven Wahrnehmungen und Beobachtungen. Beobachtungen sind nicht universell, können also nicht alle Bedingungen und Einflussfaktoren einschließen. In der Position des *kritischen Rationalismus* (Popper, 1989, zit. in Döring, Bortz & Pöschl, 2016) wird diese Kritik aufgegriffen. Wissen im Empirismus entsteht durch einen induktiven Schluss (s. 2.2.1). Den Annahmen des kritischen Rationalismus zufolge werden durch systematische Überlegungen und die Nutzung des Verstands (ratio) Vermutungen über die Realität angestellt und Theorien formuliert. Die sich darauf gründenden Annahmen werden mit gewonnenen Daten überprüft, wobei eine grundsätzliche Zurückweisung der Theorie einkalkuliert wird (Falsifikation). Eine kritische Position gegenüber eigenen Überlegungen wird eingenommen, Zweifel ist forschungsleitend. Der kritische Standpunkt bezieht die

eingesetzten Forschungsmethoden ein. Der kritische Rationalismus ist handlungsleitende Wissenschaftstheorie in der quantitativen Forschung. Theorien beschreiben die Realität, wie sie logischer Weise sein sollte, nicht aber muss.

Die *qualitative Forschungslandschaft* ist durch eine intensive und anhaltende Diskussion erkenntnistheoretischer und methodologischer Fragen gekennzeichnet. Es gibt eine Vielzahl qualitativer „Schulen", die sich auf unterschiedliche philosophische Forschungstraditionen beziehen. Gemeinsam ist diesen Ansätzen die Kritik an der – oftmals als positivistisch oder reduktionistisch bezeichneten – Vorstellung, dass eine Wirklichkeit jenseits der Ideen, der Sprache und der Wahrnehmungen der Menschen existiert und diese Wirklichkeit durch vorstrukturierte Messungen und Beobachtungen objektiv erfasst werden kann. Qualitativ Forschende gehen von der Existenz vielfältiger Realitäten aus, deren Komplexität und Bedeutung nicht in standardisierten Laborsituationen, sondern in alltäglichen Kontexten menschlicher Erfahrung und menschlichen Verhaltens erschlossen werden kann. Wie Forschende menschliche Realität wahrnehmen, beschreiben und erklären stellt unvermeidbar eine weitere Form der sozialen Konstruktion dar.

Zwischen den Extrempolen des „Was" (Realismus vs. Konstruktivismus) und des „Wie" (Rationalismus vs. Empirismus) existieren zahlreiche weitere wissenschaftstheoretische Positionen. Chalmers, Bergemann und Altstötter-Gleich (2007) bereiten die Gegenstände der verschiedenen wissenschaftstheoretischen Schulen in einer sehr gut lesbaren und unterhaltsamen Form auf.

2.4 Zusammenfassung und Aufgaben

Ziel von Forschung ist die Beantwortung relevanter Fragen. Wissenschaft beschreibt und erklärt Phänomene, sie untersucht Gründe für das Eintreten von Ereignissen und sie stellt heraus, wie z.B. Krankheiten, Störungen oder Schäden begegnet werden kann. Das weite Wissensfeld des Alltagswissens bewegt sich außerhalb jeder wissenschaftlichen Systematik. Es beruht auf Überzeugungen und beruft sich auf Autoritäten (Hussy et al., 2013). Auch wenn es unsystematisch, ad hoc, ungeprüft und intransparent entsteht, ist Alltagswissen die Grundlage des Handelns und des Funktionierens von Menschen. Die Mehrzahl der Entscheidungen wird einigermaßen spontan und willkürlich getroffen. Dennoch ist Alltagswissen notwendige Basis handelnder Daseinsbewältigung (Holzkamp, 1968, zit. in Rott, 1995).

Wissenschaftliches Wissen entsteht auf Grundlage systematischer, überprüf- und nachvollziehbarer Prozesse. Offene Fragen, die nicht auf der Basis bestehender Regeln, Annahmen und Theorien beantwortet werden können, werden in induktiv-explorativen Studien erforscht. Erklärungen werden auf Grundlage eines qualitativen Forschungsprozesses geliefert. Ergebnisse beantworten nicht nur die Frage danach, was passiert, sondern warum, unter welchen Bedingungen und in welcher Reihenfolge Phänomene auftreten. Induktive Forschung schafft Ordnung in einem Chaos an Annahmen, Vermutungen und Behauptungen. In der sozial- und gesund-

heitswissenschaftlichen Forschung entstehen auf diese Weise *probabilistische Theorien*, also Annahmen, die mit einiger Wahrscheinlichkeit für eine Vielzahl von Menschen mit vergleichbaren Eigenschaften zutreffen.

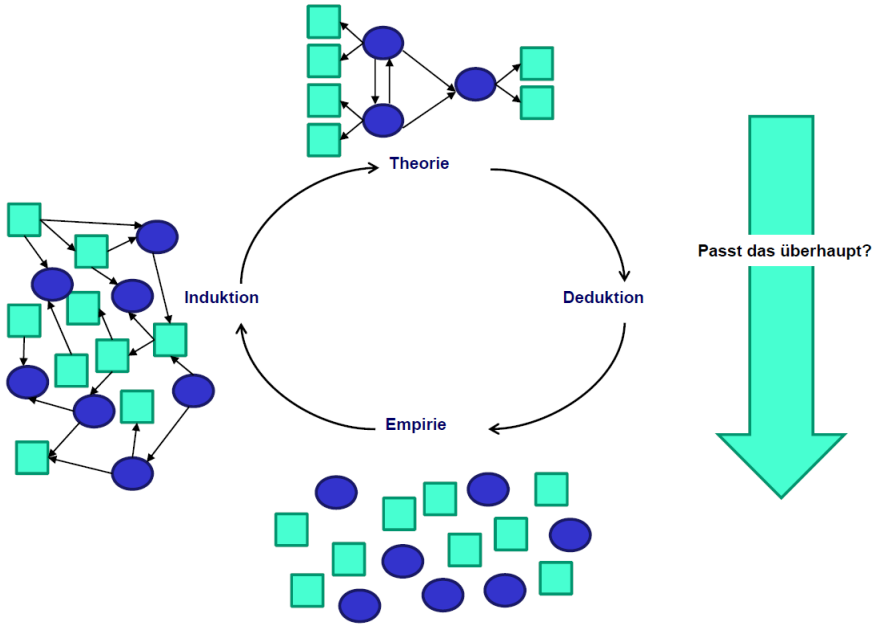


Abbildung 3: Wissenschaftliches Wissen: Induktion und Deduktion (eigene Darstellung).

Deduktiv-explanative Forschung versucht die Tragweite bereits existierender Erklärungen in großen Stichproben zu untersuchen. Dabei werden quantitative Forschungsdesigns angewendet, die erhobenen Daten (Zahlen) werden statistisch analysiert, die Stichprobe wird beschrieben und die Aussagekraft der Theorie kritisch beleuchtet. Deduktive Forschung stützt Theorien, untermauert Wissen oder liefert Hinweise, wo der Erklärungsbeitrag für Theorien unzureichend ist. Ergebnisse deduktiver Forschung können Ausgangspunkt für weitere induktive oder deduktive Forschungsprozesse sein. Daher ist deduktive Forschung auch bedeutend als Initiator für die Revision oder Neuinterpretation bestehender wissenschaftlicher Annahmen. Sozialwissenschaftliche, probabilistische Theorien lassen sich aufgrund einer oder weniger (deduktiver) Studien mit konkurrierenden Ergebnissen nicht zurückweisen. Diese Theorien können aber auf Basis widersprüchlicher Ergebnisse revidiert, erweitert oder ergänzt werden (Abbildung 3).

Die Wissenschaftstheorie befasst sich schließlich damit, was durch Wissenschaft beschrieben werden kann und wie Erkenntnis entsteht. Der Position des Realismus folgend existiert eine objektive Realität, die vom Forscher beschrieben werden kann. Aus einer konstruktivistischen Position wird der objektiven Realität widersprochen. Vielmehr wird Realität durch intensive Wahrnehmung und Erleben konstruiert. Demnach ist Realität subjektiv, weshalb eine interindividuelle Realität nicht existiert, sondern eine primär subjektive, individuelle Wirklichkeit, die durch Erfah-

rung auch revidiert werden kann. Zwei weitere Positionen bilden Grundlage des Erkenntnisprozesses. Der Rationalismus hebt bei der Entstehung von Wissen den Verstand des Menschen hervor, Wissen entsteht durch Nachdenken. Dem gegenüber entsteht Wissen nach dem Empirismus durch die wiederholte Beobachtung und Erfahrung. Erfahrungen sind nachprüfbar und wiederholbar, Gedanken nicht. Zwischen den beschriebenen Extrempositionen existieren weitere wissenschaftstheoretische Zwischenpositionen (Chalmers, Bergemann & Altstötter-Gleich, 2007).

Schlüsselwörter: *Alltagswissen, Deduktion, deduktiv-explanativen Ansätzen, deterministische Theorien, Empirismus, Induktion, induktiv-explorative Forschungsansätze, Konstruktivismus, kritischen Rationalismus, probabilistische Theorien, qualitative Forschungslandschaft, Rationalismus, Realismus, Wissenschaftliches Wissen*

Aufgaben zur Lernkontrolle

1. Was wird unter Alltagswissen verstanden?

2. Grenzen Sie Rationalismus vom Empirismus ab.

Rationalismus	Empirismus

3. Grenzen Sie Realismus vom Konstruktivismus ab.

Realismus	Konstruktivismus

Aufgabe zur Berufspraxis

4. Was verstehen Sie unter Induktion und was unter Deduktion?
   Versuchen Sie, die Begriffe anhand eines sprachtherapeutischen Beispiels zu erklären.

Literatur

Antonovsky, A. (1991). *Health, stress, and coping* (1st ed.). San Francisco [u.a.]: Jossey-Bass Publ.

Antonovsky, A., & Franke, A. (1997). *Salutogenese zur Entmystifizierung der Gesundheit*. Tübingen: DGVT-Verlag.

Bengel, J., Strittmatter, R., & Willmann, H. (2006). *Was erhält Menschen gesund? Antonovskys Modell der Salutogenese - Diskussionsstand und Stellenwert ; eine Expertise* (9., erw. Neuaufl ed.). Köln: BZgA.

Chalmers, A. F., Bergemann, N., & Altstötter-Gleich, C. (2007). *Wege der Wissenschaft Einführung in die Wissenschaftstheorie* (6., verb. Aufl. ed.). Berlin [u.a.]: Springer.

Döring, N., Bortz, J., & Pöschl, S. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften* (5. vollständig überarbeitete, aktualisierte und erweiterte Auflage ed.). Berlin Heidelberg: Springer.

Hussy, W., Schreier, M., & Echterhoff, G. (2013). *Forschungsmethoden in Psychologie und Sozialwissenschaften für Bachelor: mit 23 Tabellen* (2., überarb. Aufl. ed.). Berlin u.a.: Springer.

Rott, C. (1995). Basiskarte: Alltagswissen. In K.-E. Rogge (Ed.), *Methodenatlas* (pp. 12-18). Heidelberg: Springer.

Schäfer, T. (2016). *Methodenlehre und Statistik Einführung in Datenerhebung, deskriptive Statistik und Inferenzstatistik*. Wiesbaden: Springer.

Wikipedia. (2016). Induktion (Philosophie). Retrieved 18.5.2016, from https://de.wikipedia.org/wiki/Induktion_(Philosophie)

3 Der Forschungsprozess

Lernergebnisse:

- Der Forschungsprozess kann beschrieben werden. Unterschiede im Forschungsprozess zwischen qualitativen und quantitativen Forschungsansätzen können erläutert werden.

- Bestandteile des Forschungsprozesses können geordnet und in einen Zeitverlauf gebracht werden.

Forschung ist grundsätzlich kreativ und frei in der Wahl ihres Schwerpunkts und ihrer Methoden (innerhalb der unter 1.3 und 1.4 beschriebenen Grenzen). Das Beschreiten neuer Wege bei der Untersuchung von Problemen und offenen Fragen ist nicht nur erlaubt, sondern erwünscht und notwendig, um neue Erkenntnisse zu generieren. Mit den Zielen von Forschung: der Beschreibung, dem Erklären, Vorhersage u.a. der Veränderung von Phänomenen (s. 1.1), sind Anforderungen verbunden. Gütekriterien von Forschung (1.3) setzen Vorgaben, die eine aussagekräftige, transparente und valide Forschung sicherstellen sollen. Auch der Forschungsprozess folgt einer Logik. Dabei bauen einzelne Schritte im Forschungsprozess bauen aufeinander auf. In der Themensuche wird ausgehend von eigenen, wissenschaftlichen und praktischen Interessen ein Untersuchungsrahmen definiert. An das Thema werden ebenfalls bestimmte Ansprüche gestellt (s. 4.1).

Im quantitativen Forschungsprozess (Abbildung 4) erfolgt daran anschließend auf der Basis einer systematischen Literaturrecherche die Formulierung eines theoretischen Rahmens (s. 4.2). Eine Theorie beschreibt zunächst wie Merkmale miteinander verknüpft sind und in welchem zeitlichen Zusammenhang sie zu sehen sind. Basis dafür sind die Grundannahmen und Gesetze einer (oder mehrerer) Disziplinen. Sie erklären, warum eine Verknüpfung so und nicht anders auftreten muss. Forschungsfragen skizzieren den Ausschnitt einer Theorie, der in einem Forschungsvorhaben untersucht werden sollen (s. 4.3). Aus Theorie und Fragestellung ergeben sich Hypothesen. Hypothesen sind einzelne Aussagen, die in einer Theorie systematisch zusammengefasst sind. Sie skizzieren die Ergebnisse eines Forschungsvorhabens auf den beschriebenen theoretischen Grundlagen (s. 4.4). In Hypothesen werden die interessierenden Merkmale genannt. Die Wahl des Studiendesigns orientiert sich an der gewünschten Aussagekraft der Ergebnisse und nicht zuletzt an wirtschaftlichen Aspekten (1.). Kausale Zusammenhänge, also Ursache-Wirkungsbeziehungen, können beispielsweise nur auf der Grundlage von Längsschnittstudien (5.2) ermittelt werden.

Ausgehend von der Definition der interessierenden Merkmale erfolgen die Entwicklung von Messinstrumenten und die Formulierung konkreter Items (Fragen) für Fragebögen (6.2). Mit einem gültigen und zuverlässigen Messinstrument werden Daten zur Beantwortung der Fragestellungen erhoben. Diese Untersuchung erfolgt normalerweise nicht bei allen Menschen einer Population, sondern in der interessierenden Grundgesamtheit (1.2), aus der (möglichst auf Basis von Zufall) eine Stichprobe gezogen wird (s. 7.1). Von der Stichprobe werden Fragebögen ausgefüllt oder bei ihren Mitgliedern Messungen vorgenommen. In der anschließenden Datenanalyse werden Daten zusammengefasst und systematisiert (s. 8). Die Eignung theoretischer Annahmen wird im Rahmen quantitativer Forschungsdesigns mit prüf-, bzw. inferenzstatistischen Verfahren überprüft (s. 9). Der Test auf statistische Bedeutsamkeit und Signifikanz im Rahmen inferenzstatistischer Prüfverfahren ergibt eine Wahrscheinlichkeitsaussage darauf, ob ein Ergebnis (d.h. Unterschied, Zusammenhang oder Veränderung) in der Stichprobe auftreten kann, wenn in der Grundgesamtheit kein Ergebnis (d.h. kein Unterschied, kein Zusammenhang oder keine Veränderung)

vorliegt (Krauss & Wassner, 2001). Diese Irrtumswahrscheinlichkeit umschreibt, mit welcher Wahrscheinlichkeit die Entscheidung für das Zurückweisen der Nullhypothese in der Stichprobe falsch ist (s. 9.1).

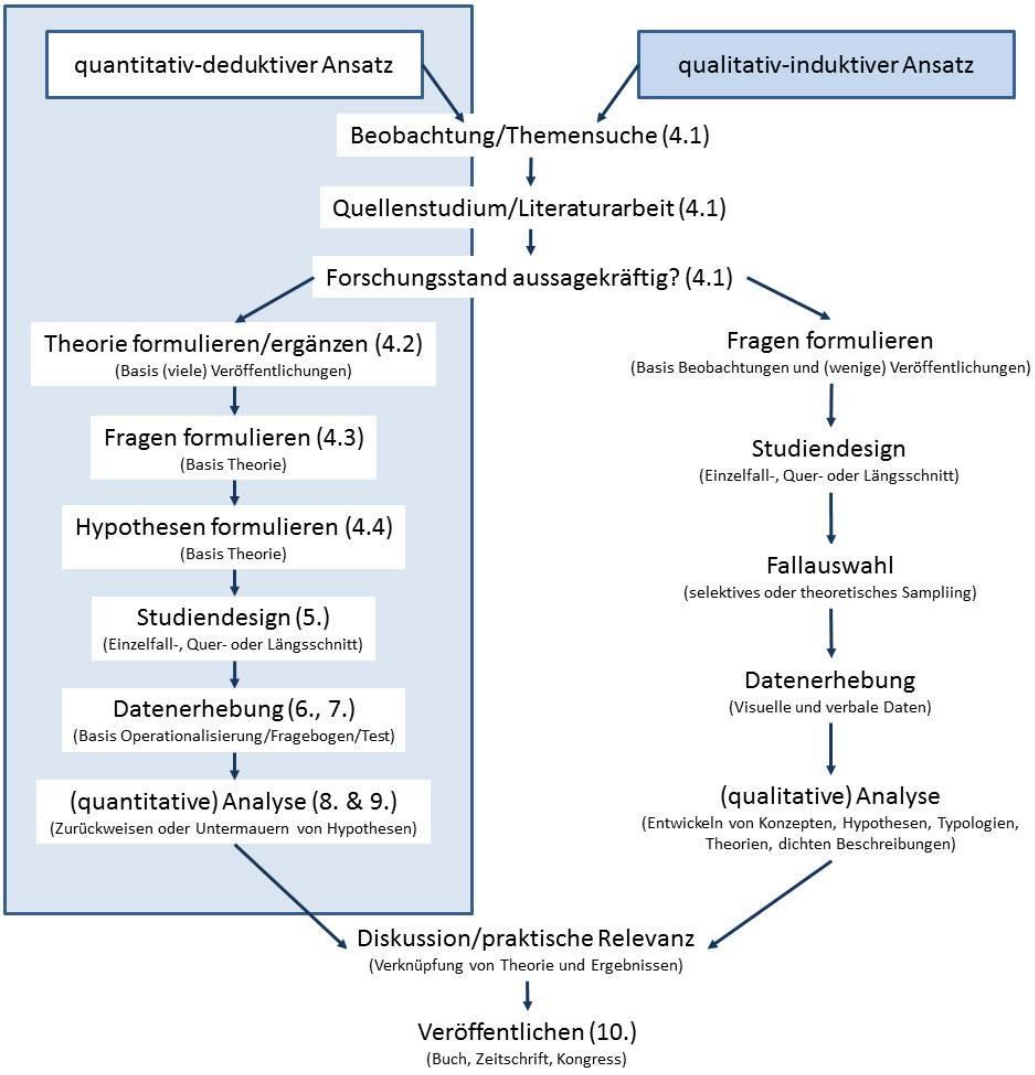


Abbildung 4: Skizze des Forschungsprozesses mit den Gliederungspunkten in diesem Studienbrief (eigene Darstellung)

Während quantitative Forschungsprozesse teils linear verlaufen, sind qualitative Forschungsprozesse durch eine zirkuläre und flexiblere Vorgehensweise gekennzeichnet. Vor der Datenerhebung wird ein theoretisches und methodologisches Vorverständnis erarbeitet und eine vergleichsweise offene Forschungsfrage formuliert. Wie durch die Pfeile in Abbildung 4 angedeutet, ist es in qualitativen Studien möglich und vielfach auch notwendig, die Forschungsfrage in der Auseinandersetzung mit den erhobenen Daten kontinuierlich zu diskutieren und ggf. anzupassen. Die Formulierung der leitenden Forschungsfragen orientiert sich dabei in der Regel an den methodologischen Grundannahmen und methodischen Zugängen ausge-

wählter qualitativer Forschungstraditionen (wie z.B. der Grounded Theory oder der Ethnografie). Ein weiteres Beispiel für die Zirkularität des qualitativen Forschungsprozesses ist das sogenannte theoretische Sampling. Bei dieser Vorgehensweise werden die untersuchten Fälle nicht vorab, sondern basierend auf den Ergebnissen der Datenauswertung fortlaufend ausgewählt. Kennzeichnend für die qualitative Datenerhebung ist ein hoher Grad an Offenheit, der es ermöglichen soll, Unerwartetes und bisher nicht Bekanntes sichtbar werden zu lassen. Beispielsweise verfolgen alle qualitativen Interviewformen das grundlegende Ziel, den Befragten ausreichend Raum für die Darstellung ihrer individuellen Relevanzsetzungen, Sichtweisen und Erfahrungen zu eröffnen. Auch teilnehmende Beobachtungen setzen i.d.R. zunächst keine Beobachtungskriterien oder -schemata ein. Die erhobenen verbalen oder visuellen Daten werden mit Hilfe vielfältiger interpretativer Verfahren ausgewertet. Interpretative Analysen und Rekonstruktionen bilden die Basis für die Entwicklung von Konzepten, Hypothesen, Typologien, Theorien oder dichten Beschreibungen. Weiterführende Informationen zu ausgewählten Arbeitsschritten des qualitativen Forschungsprozesses beinhaltet der nachfolgende kommentierte Reader.

Der letzte Schritt des Forschungsprozesses umfasst die Veröffentlichung der Forschungsergebnisse in Zeitschriften, in Buchform oder die Präsentation auf Kongressen in Form von Vorträgen oder als Poster (s. 10).

Schlüsselwörter: *qualitativer Forschungsprozess, quantitativer Forschungsprozess*

Aufgaben zur Lernkontrolle

1. Worin unterscheidet sich der quantitativ-deduktive vom qualitativ-induktiven Ansatz in einem Forschungsprozess?

2. Was haben die beiden Ansätze im Forschungsprozess gemein?

Aufgabe zur Berufspraxis

3. Recherchieren Sie sprachtherapeutische Studien, die auf dem quantitativ- deduktivem und auf dem qualitativ-induktivem Ansatz basieren.

Literatur

Krauss, S., & Wassner, C. (2001). Wie man das Testen von Hypothesen einführen sollte. *Stochastik in der Schule, 21*(1), 29-34.

4 Begriffe empirischer Forschung

Lernergebnisse:

- Die Bedeutung und der Prozess der Themensuche können beschrieben werden. Ein klinisches Thema kann ausgehend vom PICO-Schema erarbeitet werden (4.1).

- Der Theoriebegriff kann definiert und Bestandteile von Theorien differenziert werden. Die Bedeutung von Theorien für die qualitative und quantitative Forschung kann diskutiert werden(4.2).

- Die Bedeutung der Fragestellung in der empirischen Forschung kann beschrieben werden. Theorie und Fragestellung können inhaltlich verknüpft werden (4.3).

- Theorie, Fragestellung und Hypothese können in Verbindung gebracht werden. Arten von Hypothesen können differenziert und ihr Einsatzfeld beschrieben werden (4.4)

- Der Variablenbegriff kann definiert werden. Es kann inhaltlicher Bezug zu Grundbegriffen empirischer Forschung (1.2) genommen werden (4.5).

- Variablen können hinsichtlich ihres Stellenwerts in Untersuchungen, der Art ihrer Merkmalsausprägung und ihrer empirischen Zugänglichkeit differenziert werden (4.5).

Im Entstehungsprozess einer wissenschaftlichen Arbeit steht die Suche nach einem geeigneten Thema am Anfang. Häufig wird bereits in der Festlegung eines Themas ein komplizierter Prozess gesehen, der Zeit kostet und stets Grundsatzfragen aufwirft: Ist mein Thema angemessen? Kann das so untersucht werden? Ist das nicht zu wenig/zu viel? Ein Thema sollte in der Tat untersuchbar sein und sich zumindest grundsätzlich theoretisch fundieren lassen (4.1). Die Formulierung eines untersuchbaren theoretischen Modells ist eine zentrale Basis im quantitativen Forschungsprozess. Theorien erklären welche Merkmale, wie miteinander verbunden sind und nehmen vorweg, warum sich bestimmte Effekte zeigen. Dabei beleuchten Theorien lediglich Ausschnitte einer komplexen Realität. Theorien entstehen nicht ausschließlich durch gründliches Nachdenken, sondern sie beruhen auf Erfahrungen – sind also auch empirisch begründet und bereits (zumindest ansatzweise) untersucht worden. Aus der Theorie entwickelt sich die Forschungsfrage. Forschungsfragen orientieren sich an einzelnen Elementen von Theorien. Sie fokussieren auf bestimmte Beziehungen zwischen den in der Theorie definierten Merkmalen. Die Entwicklung von Fragestellungen ist entscheidend, weil wissenschaftliche Probleme in Antworten münden sollen. Die Formulierung von Antworten setzt die Formulierung von Fragen voraus. Auf der Basis einer Theorie lassen sich diese Fragen bereits im Vorfeld beantworten (4.3). Theoretisch müsste gar nicht empirisch geforscht werden, denn alle Antworten liegen in der Theorie (4.2). Allerdings lassen sozial- und gesundheitswissenschaftliche Theorien nur bedingt allgemeingültige, also für alle potenziellen Besonderheiten von Menschen zutreffende Aussagen zu. Mit der Formulierung von Hypothesen werden demnach theoretisch begründete Antworten auf die Fragestellung gegeben, die sich in der eigenen Forschung mit einiger Wahrscheinlichkeit zeigen sollten – aber nicht müssen. Hypothesen stellen heraus, was theoretisch begründet Einflussfaktor und was Effekt ist (4.4). Forschungsmethodisch handelt es sich bei Einflussfaktoren und Effekten um Merkmale von Personen oder Organisationen, die als Variablen in Untersuchungen berücksichtigt werden. Variablen werden ausgehend von ihrem Stellenwert in Untersuchungen, der Art der theoretisch möglichen Merkmalsausprägung und ihrer empirischen Zugänglichkeit differenziert (4.5).

Die folgenden Gliederungspunkte skizzieren den Prozess von der Entwicklung eines Forschungsthemas, zur Beschreibung der theoretischen Basis, bis zur Formulierung von Fragestellungen und von Hypothesen. Im Forschungsprozess berücksichtigte Gegenstände und Merkmale werden als Variablen in Untersuchungen berücksichtigt, die theoretisch fundiert in einer Beziehung zueinander stehen und in einer bestimmten Reihenfolge stattfinden müssten.

Wie bereits im vorangehenden Kapitel beschrieben, unterscheiden sich die Arbeitsschritte quantitativer und qualitativer Forschung. Wichtig ist an dieser Stelle der Hinweis, dass theoretische Grundlagen und Modelle in qualitativen Forschungsprozessen i.d.R. nicht genutzt werden, um Hypothesen abzuleiten, sondern vielfach als sensibilisierende Perspektiven für die Datenerhebung und -auswertung dienen.

4.1 Themensuche und Thema in der empirischen Forschung

Die Formulierung eines Themas steht am Beginn einer wissenschaftlichen Arbeit und ist häufig eine nicht ganz einfache Aufgabe. Studierende haben anfangs Schwierigkeiten, die Relevanz und Eignung von eigenen Themen abzuschätzen. Bei der ersten Formulierung eines Forschungsthemas wird häufig auf eine umfassende Erklärung von Phänomenen abgezielt, die teils sehr abstrakt ist und kaum hinreichend untersuchbar. Das Problem ist häufig nicht eine zu geringe, sondern eine zu große Komplexität oder Tiefe des Themas. Im Thema sollten folgende Elemente erkennbar werden (Döring et al., 2016):

- Was soll untersucht werden? Hier interessieren also ein Endpunkt oder Besonderheiten von bestimmten Gruppen. Beispielsweise wäre aus gesundheitswissenschaftlicher Perspektive interessant zu untersuchen, wie Menschen auch unter Belastungen gesund bleiben können. Dieser Schwerpunkt – also Gesundheit (die noch näher beschrieben werden müsste) – könnte auch in klar definierten Gruppen untersucht werden, wie beispielsweise bei alleinerziehenden Müttern, beim Pflegepersonal in der Intensivpflege oder Menschen nach einer Chemotherapie.

- Welche Einflussfaktoren beeinflussen den Endpunkt (in der definierten Gruppe)? Theoretisch fundiert werden Einflussfaktoren beschrieben, die in unserem Beispiel Gesundheit verändern können. Dies wären beispielsweise genetische Voraussetzungen, erlebte Belastungen, bestimmte Verhaltensweise usw.

- Gibt es potenzielle Vergleichs- und Kontrollgruppen in denen abweichende Forschungsergebnisse denkbar sind?

Döring et al. (2016) zufolge werden im Forschungsthema zumindest der Untersuchungsgegenstand genannt. Beispiele aus der gesundheitswissenschaftlichen und klinischen Forschung sind: „Herz-Kreislauf-Erkrankungen und Rauchverbote“, „Arbeitsbedingungen und Arbeitszufriedenheit in der ambulanten Logopädie“, „Sprachentwicklungsverzögerung in der Kindertagesstätte“ (…). Die Themen sollten zudem den disziplinären Bezug eines Themas erkennbar werden lassen. Im Forschungsproblem (Döring et al., 2016) bzw. der Fragestellung (s. 4.3) soll erkennbar werden, „…welche Erkenntnisse zu welchen Aspekten des Untersuchungsgegenstandes auf welcher theoretischen, empirischen und methodischen Basis gewonnen werden sollen.“ (Döring et al., S. 144).

Bei der Wahl des Forschungsthemas sollten folgende Kriterien berücksichtigt werden (Döring et al., 2016, S. 149ff):

1. Am Thema besteht ein hinreichendes persönliches Interesse. Diese Empfehlung ist wesentlich, da die Anfertigung wissenschaftlicher Arbeiten zeitaufwändig und phasenweise durchaus anstrengend sein kann. Kaum etwas kann unter diesen Bedingungen mehr frustrieren, als ein langweiliges und persönlich uninteressantes Thema zu bearbeiten (Döring et al., 2016).

2. Für das Thema existiert eine hinreichende theoretische, methodische empirische Fundierung. Insbesondere für Haus- und Bachelorarbeiten ist eine weitgehend offene, schwach fundierte theoretische Basis problematisch, weil die Formulierung theoretische Grundlagen auf einen methodisch anspruchsvollen Prozess aus Literaturanalyse und qualitativen Analysen von Befragungen basiert. Dieser induktive Forschungsprozess bedarf Erfahrung und einiger Übung (s. 2.2) (Döring et al., 2016).

3. Das Thema ist wissenschaftlich relevant, ist also bereits Gegenstand von Diskussionen in der *scientific community*. Ziel sollte also weniger sein, ein bislang völlig unbeforschtes Feld zu bearbeiten, sondern an existierenden wissenschaftlichen Diskussionslinien anzuknüpfen und sie zu erweitern. Dabei kann auch die Wiederholung einer bereits abgeschlossenen Untersuchung von hoher wissenschaftlicher Relevanz sein, denn Studien, die bei gleichen Fragestellungen vergleichbare Ergebnisse und Schlussfolgerungen liefern untermauern die Aussagekraft der zugrundeliegenden Theorien (Döring et al., 2016).

4. Das Thema hat praktische Bedeutung, es liefert also Anhaltspunkte für die Veränderungen und die Optimierung von Therapien, die Gestaltung von Arbeitsbedingungen oder für technische Anwendungen (Döring et al., 2016).

5. Das Thema muss erfahrbar und empirisch untersuchbar sein. Eine Reihe von Gründen kann die tatsächliche wissenschaftliche Umsetzung einer Forschungsfrage erschweren. Döring und Kollegen (2016) beschreiben ethische Grenzen (z.B. Stammzellforschung), die mögliche politische Brisanz von Themen (Integration behinderter Menschen, Mitarbeiterbefragungen, die heikle Rahmenbedingungen in Unternehmen einbeziehen), einen zu hohen Aufwand für die Untersuchungsteilnehmer und die Forschenden (hier u.a. die Notwendigkeit von Vollerhebungen), schwer erreichbare Zielgruppen (Drogenmissbrauch durch das Pflegepersonal), übermäßig große Abhängigkeit von Kooperationspartnern oder die Notwendigkeit der Nutzung schwer oder nicht verfügbarer technischer Hilfsmittel (Döring et al., 2016).

6. Für ein Thema muss sich auch der zukünftige Gutachter der Arbeit begeistern können und das Thema sollte an Arbeiten des Gutachters anknüpfen. Dies vermeidet nicht nur Überraschungen im Betreuungsprozess, sondern kann Grundlage einer längeren, für beide Seiten gewinnbringenden Zusam-

menarbeit sein (wissenschaftliche Weiterqualifizierung, Drittmitteleinwerbung, Publikationstätigkeit) (Döring et al., 2016).

Ein Thema muss zusammenfassend den Untersuchungsgegenstand, Einflussfaktoren und die untersuchte Population offenlegen. Darüber hinaus müssen Untersuchbarkeit, theoretische, empirische und methodische Fundierung, wissenschaftliche und praktische Relevanz sowie die empirische Untersuchbarkeit gewährleistet sein. Schließlich sollte sich auch ein Gutachter für das Thema begeistern können.

4.2 Theorien und Modelle als Basis empirischer Forschung

Die Notwendigkeit der theoretischen Fundierung eines empirischen Forschungsthemas wurde bereits angesprochen. Quantitative empirische Forschung ist in der Lage standardisiert Ergebnisse zu liefern, mit quantitativer Forschung können Populationen beschrieben, Zusammenhänge und Veränderungen aufgedeckt werden. In diesen Ergebnissen findet sich allerdings kein Erkenntnisgewinn, wenn nicht auch deutlich wird, warum bestimmte Beobachtungen gemacht, Zusammenhänge bestehen oder Veränderungen erfolgten. Theoretisch nicht fundiert ist beispielsweise ein rechnerisch nachweisbarer enger Zusammenhang (r= 0,79) zwischen produzierten Kinowerbefilmen und die Mortalität bei alkoholbedingten Störungen (Die Zeit, 2013, Nr. 13). Ein weiteres Beispiel ist die Verbindung zwischen Anzahl der Störche und der Geburtenzahl. Beide Zusammenhänge bestehen empirisch und mathematisch, sie sind aber theoretisch unerklärt. Die untersuchten Ereignisse treten also zufälliger Weise gemeinsam auf. Bislang fand sich keine inhaltlich schlüssige Erklärung für die gemeinsame Zunahme produzierter Kinowerbefilme und einer höheren Sterblichkeit aufgrund von alkoholbedingten psychischen und Verhaltensstörungen. Es ist unnötig und wissenschaftlich falsch, Zusammenhänge zu diskutieren, für die keine schlüssige theoretische Basis, keine wissenschaftlich korrekte Begründung existiert. Allenfalls müsste eine solche Erklärung über induktive Forschungsdesigns entwickelt werden (s. 2.2.1, Abbildung 2).

Die Begriffe Theorie und Modell werden häufig Synonym verwendet. Teilweise wird in der Begriffsbestimmung auch auf unterschiedliche Abstraktionsgrade von Theorien und Modellen eingegangen (Fawcett & Erckenbrecht, 1999). Modelle, genauer konzeptuelle Modelle, bewegen sich in ihren Aussagen auf einem höheren Abstraktionsgrad als Theorien. Modelle bilden danach die Basis für die Ausformulierung konkreterer Theorien. Eine Grundhaltung, in der Menschen als kommunizierende Wesen betrachtet werden (Modell), wird in einer theoretischen Haltung münden, Interventionen mit Patienten zu besprechen. Sie wird in der Kommunikation eine wesentliche Grundlage für therapeutische und pflegerische Interventionen sehen (Theorie). Demzufolge können Theorien als konkret untersetzte, empirisch untersuchbare Aussagesysteme verstanden werden. Konzeptuelle Modelle sind dagegen

vager formuliert und zeichnen ein umfassendes, jedoch wenig konkretes und untersuchbares Bild unter Einbeziehung des Grundverständnisses einer Disziplin und ihres Menschenbildes. Ob eine solche Untergliederung sinnvoll ist oder nicht, wird kontrovers diskutiert. Für die quantitative empirische Forschung sind konkrete, beobachtbare und damit untersuchbare Aussagen notwendig. Ob diese in Modellen oder in Theorien gesehen werden scheint zunächst wenig entscheidend. In der Literatur spielt diese begriffliche Unterscheidung jedenfalls häufig keine Rolle. Theorie und Modell werden als Begriff augenscheinlich synonym verwendet und inhaltlich wenig unterschieden.

In den bisherigen Ausführungen blieb vage, worum es sich bei einer Theorie handelt. Zunächst einmal sei auf das Ziel von Forschung verwiesen, Phänomene zu erklären, zu beschreiben und vorherzusagen. An einem Beispiel zusammengefasst interessiert sich Forschung für die psychischen Folgen von Alkoholmissbrauch. Es soll dabei nicht nur herausgestellt werden, dass Alkohol Depressivität begünstigt, sondern auch warum Alkohol dies vermag – aus der Perspektive biochemischer oder psychologischer Erklärungen. Zudem sollen Forschungsergebnisse auch Anhaltspunkte für andere Menschen und künftige Entwicklungen liefern. Ausgehend von unserem Beispiel interessiert demnach auch, ob und unter welchen Bedingungen Alkoholmissbrauch Depressionen hervorrufen kann und warum das psychologisch oder biochemisch möglich ist.

Raithel (2008, S. 16) zufolge ist eine Theorie „…ein System logisch miteinander verbundener, widerspruchsfreier Aussagen oder Sätze (…), das mehrere Hypothesen und Gesetze umfasst." Theorien beschreiben Wirklichkeit nicht vollumfänglich, sondern konzentrieren sich auf die für ein Thema relevanten Aussagen – sie bilden somit einen „verkleinerten Ausschnitt der Realität". Folgende Komponenten sind in Theorien aufgenommen (Raithel, 2008, S. 16):

1. Grundlegende Annahmen, also die wesentlichen Hypothesen (einzelne Aussagen über Einflussfaktor und Effekt) und die Definition der verwendeten Begriffe. Die beschriebenen Annahmen dürfen existierenden und empirisch untermauerten Aussagen nicht wesentlich widersprechen. Eine widersprechende Annahme muss in jedem Fall theoretisch untermauert werden. Die Definition der Merkmale umfasst die beschreibenden Kriterien und stellt heraus, wie sich ein Merkmal von anderen Merkmalen abgrenzt.

2. Hypothesen und Regeln, wie die in den Hypothesen aufgenommenen Merkmale gemessen werden können (s. 6.2).

An sozial- und gesundheitswissenschaftliche Theorien werden Anforderungen geknüpft. Atteslander und Cromm (2010, S. 33) zufolge müssen Theorien in der empirischen Sozialforschung:

1. Logisch aufgebaut und in sich widerspruchsfrei sein. Theorien sollten zudem bestehenden Gesetzmäßigkeiten nicht widersprechen bzw. Erklärungen für Widersprüche beinhalten und begründen.

2. In allen Aussagen (Hypothesen) der empirischen Forschung und der bewussten Erfahrung zugänglich sein.

3. Bestehende Theorien Kenntnisse und Erklärungen ergänzen oder einen höheren Erklärungsbeitrag für ein Phänomen leisten.

Theorien können bereits beschrieben und veröffentlicht sein (Salutogenemodell/-theorie). Zudem können theoretische Aussagen und ein Forschungsmodell aus veröffentlichten Arbeiten entwickelt werden. Die Aussagekraft bereits veröffentlichter und hinreichend untersuchter Theorien ist größer als die „ad-hoc“ entwickelter Modelle. Aus dem Forschungsstand entwickelte Forschungsmodelle können bestehende Theorien sinnvoll erweitern und um eigene thematische Interessen ergänzen. Von der Entwicklung und der Untersuchung eines Modells ohne jegliche beschriebene theoretische Fundierung ist für das Feld der quantitativen Forschung ist insgesamt abzuraten.

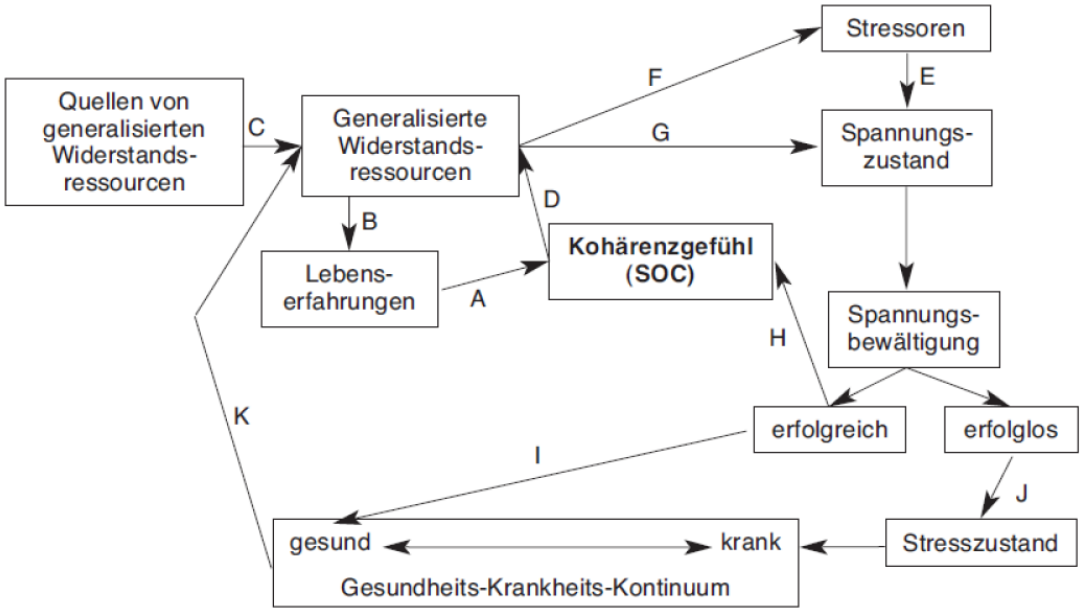


Abbildung 5: Theorie am Beispiel des Salutogenesemodells von Antonovsky (1991) (Quelle Bengel et al., 2006, S. 36).

Abbildung 5 zeigt die schematische und stark vereinfachte Darstellung der Salutogenese-Theorie (Antonovsky & Franke, 1997, Bengel et al., 2006, S. 36). Die Kästen enthalten die aufgenommenen Merkmale, also Aspekte, von denen Einfluss erwartet wird bzw. die von verschiedenen Einflussfaktoren beeinflusst werden. Zwischen den Merkmalen verlaufen Pfeile, die als „steht in Verbindung mit“ bzw. „wird beeinflusst durch“ interpretiert werden können. Das Schema trifft zunächst Aussagen über die Verbindung zwischen einzelnen Merkmalen, aus welchen wissenschaftlich

und empirisch nachvollziehbaren Gründen diese Verbindungen auftreten, zeigt das Schema nicht. Diese Erläuterungen werden in der Beschreibung der Theorie ergänzt. Sie enthält neben den erwarteten Verbindungen auch die (psychologischen, soziologischen, medizinischen, biochemischen (…)) Gründe für einen Zusammenhang.

Zusammengefasst liefern Theorien die Erklärung für (potenzielle) empirische Befunde. Theorien beschreiben den für ein Forschungsvorhaben interessierenden Ausschnitt der Realität. Sie sind so komplex, wie es für die Beantwortung einer Forschungsfrage notwendig ist (die Erklärung des Herzinfarktrisikos erfordert die Berücksichtigung genetischer und von Verhaltensaspekten). Theorien sind ein System einzelner Aussagen (s. 4.4), die logisch und widerspruchsfrei sein, empirisch untersuchbar sein und zusätzliche/neue Erklärungen liefern müssen.

4.3 Fragestellung als programmatischer Rahmen von Forschung

Forschung soll Antworten liefern. Fragestellungen sind der Analyseplan in der empirischen Forschung. Fragen beruhen auf einzelnen Aussagen in der Theorie (4.2). In der quantitativen Forschung liefern Hypothesen (s. 4.4) die vorweggenommenen Antworten auf aufgeworfene Forschungsfragen. Fragestellungen werden zudem aus offenen, theoretisch nicht fundierten Beobachtungen geschlossen. Unter einem qualitativ-induktiven Forschungsansatz bilden Fragestellungen die Basis der Untersuchung. Ein wichtiges Ergebnis qualitativ-induktiver Forschung können Hypothesen bzw. Einzelaussagen sein, die im weiteren Forschungsverlauf in die Formulierung von Theorien und Modellen münden können. In der quantitativen Forschung werden Fragen durch Einzelaussagen, sog. Hypothesen, beantwortet, die auf einer klar umschriebenen theoretischen Basis beruhen. Quantitative Forschung versucht die Aussagen in Hypothesen und damit die der Theorien durch die erhobenen Daten systematisch zu untermauern (oder zu widerlegen).

In der klinischen Forschung existiert mit dem PICO-Schema ein Hilfsmittel zur Entwicklung der Fragestellung für eine Untersuchung. Die Orientierung der Fragestellung erfolgt an (Mangold, 2013, S. 45):

- Patienten, d.h. am gesundheitlichen Problem, einer Diagnose verknüpft mit dem Lebensalter oder dem Geschlecht des Patienten

- Intervention, also welche Maßnahme wird als überlegen betrachtet und soll am Patienten zur Therapie eines konkreten Gesundheitsproblems angewendet werden,

- Comparison, mit welcher Maßnahme soll die Intervention verglichen werden?

- Outcome, was ist der Endpunkt der Studie, also welche gesundheitliche Störung soll beispielsweise behandelt oder gelindert werden?

Soll beispielsweise die Wirksamkeit eines neuen Hörgeräts untersucht werden würde als Patienten Menschen mit andauernden Hörstörungen untersucht werden. Als Intervention das Hörgerät einer neuen Generation, das mit einem herkömmlichen (bewährten) Hörgerät verglichen werden soll (Comparison). Als Outcome könnten Wort- oder Lautverständnis, Orientierung im Alltag oder Teilhabe aufgenommen werden. Die vorgestellte Systematik führt zu folgender Fragestellung:

Verbessert ein Hörgerät der neuen Generation, verglichen mit bewährten Hörgeräten, das Lautverständnis von Patienten mit Schwerhörigkeit?

Zusammenfassend nimmt die Fragestellung das Problem, seine Einflussfaktoren und die erwartete Entwicklung eines Problems auf. Sie orientiert sich am entwickelten theoretischen Rahmen oder (bei qualitativen Studien) an den zu untersuchenden Beobachtungen. Die Fragestellung ist Basis und Orientierung für die Datenerhebung, die Datenanalyse und führt z.T. zur Formulierung von Hypothesen.

4.4 Hypothesen oder die theoriebasierte Suche nach Antworten

Die Annahme oder Vermutung über einen Zusammenhang, eine Veränderung oder einem Zustand wird in Hypothesen formuliert. Forschungshypothesen werden *nicht* ad hoc formuliert und dann untersucht. Die gründliche Auseinandersetzung mit dem Forschungsthema und Veröffentlichungen zum Thema sind Voraussetzungen für die Entwicklung von Hypothesen in der empirischen (quantitativen) Forschung.

Vor der Formulierung von Hypothesen erfolgt eine systematische Literaturanalyse, damit die Spezifizierung des theoretischen Rahmens und darauf aufbauend, die Formulierung geeigneter Forschungsfragen (s. 4.3). Fragestellung und Hypothesen dienen der Eingrenzung des Forschungsthemas. Es wird also i.d.R. eine komplexe Theorie nicht vollumfänglich untersucht, sondern der für das Thema interessierende Ausschnitt einer Theorie. In Hypothesen interessieren einzelne Verknüpfungen von Merkmalen innerhalb von Theorien. Die Formulierung von Hypothesen erfolgt in Aussagesätzen, die (je nach zugrundeliegender Theorie) in eine Konditionalstruktur, also in „Wenn-Dann-Sätze“ überführt werden können. Die meisten Autoren unterscheiden Forschungshypothesen und statistische Hypothesen (Döring et al., 2016, Raithel, 2008, Schäfer, 2016).

4.4.1 Forschungshypothesen

Forschungshypothesen beziehen sich auf *Unterschiede*, *Zusammenhänge* oder *Veränderungen*. In *Unterschiedshypothesen* wird von verschiedenen Merkmalsausprägungen zwischen zwei oder mehr Gruppen ausgegangen und könnten so formuliert sein: „Schwerhörige Patienten mit Hörgeräten der neuen Generation hören besser als schwerhörige Patienten mit herkömmlichen Hörgeräten." Als unabhängige Variable (s. 4.5) geht die Gruppierungsvariable (neues oder altes Hörgerät), als abhängige Variable in unserem Beispiel das „Hören" ein. Auch der Vergleich zwischen mehreren Gruppen erfolgt auf Basis von Unterschiedshypothesen. Beispielsweise könnte die Wirkung unterschiedlicher Behandlungsregimes beim Bluthochdruck untersucht werden (Tabelle 1).

Zusammenhangshypothesen fokussieren auf die begründete gemeinsame Schwankung (Varianz) der Ausprägung zweier Merkmale. Beispielsweise wäre die die Verbindung zwischen Arbeitsanforderungen, Arbeitsbelastungen und Stresssymptomen eine untersuchbare und theoretisch begründete Zusammenhangshypothese (Beispiel für die Darstellung von Zusammenhängen in Tabelle 2).

Veränderungshypothesen setzen die Beobachtung eines Einflussfaktors voraus, auf den sich Veränderungen theoretisch gründen. Im klinischen Kontext wäre die Frage nach der Wirksamkeit von Rauchentwöhnungsprogrammen eine Grundlage für Veränderungshypothesen. Dabei kann bei entsprechender theoretischer Fundierung die Reduktion der Zahl gerauchter Zigaretten angenommen werden.

Tabelle 1: Beispiel für Ergebnisse von Unterschiedshypothesen (hier Mittelwertvergleiche bei abhängigen und unabhängigen Stichproben)

Blutdruck (mm Hg)	t_0 (SD)	t_1 (SD)	t_2 (SD)	F (df1/df2)	Signifikanz p	Effekt ω^2
Kontrollgruppe (N=16)	176,6 (20,5)	172,0 (27,0)	173,5 (32,2)	0,103 (2/30)	0,902	0,00
Blutdrucksenker (N=16)	179,5 (21,8)	115,7 (9,0)	146,1 (12,9)	77,75 (2/30)	<0,001	0,91
Blutdrucksenker & Sport (N=16)	172,4 (21,5)	151,8 (11,1)	126,7 (8,6)	43,70 (2/30)	<0,001	0,84
F (df1/df2)	0,446 (2/45)	64,476 (2/27,38)	24,026 (2/26,27)			
p	0,643	<0,001	<0,001			
Effekt ω^2	0	0,72	0,49			

Anmerkung: $\omega^2 = \frac{f^2}{1+f^2}$, $f^2 = \frac{(F-1)\cdot df_{Zähler}}{n}$, ω^2>0,01, kleiner Effekt, ω^2>0,06, mittlerer Effekt, ω^2>0,14, großer Effekt (Rasch, Friese, Hofmann & Naumann, 2010)

Abhängig von der theoretischen Basis und dem Forschungsstand zu einer Fragestellung werden Hypothesen gerichtet oder ungerichtet formuliert. Eine *gerichtete* Hypothese hebt hervor, welche Gruppe eine größere oder kleinere Ausprägung bei der abhängigen Variable hat (s. 4.5). Die oben formulierte Unterschiedshypothese weist eine Richtung, nämlich schwerhörige Patienten mit einem Hörgerät der neuen Generation hören *besser*. Eine *ungerichtete* Variante dieser Hypothese wäre: „Schwerhörige Patienten mit Hörgeräten der neuen Generation und schwerhörige Patienten mit herkömmlichen Hörgeräten hören unterschiedlich gut.“. Hier wird nicht klargestellt, wer die besseren Ergebnisse erzielt. Bei der Wahl des Signifikanztests spielt die Vorgabe einer Richtung des Zusammenhangs eine Rolle (s. 9.3). Die Hypothesentestung erfolgt bei ungerichteten Hypothesen zwei-, bei gerichteten Hypothesen einseitig (9.1.1, s. auch 4.4.2).

Tabelle 2: Darstellungsbeispiel für Ergebnisse Zusammenhangshypothesen bei einer Studie mit 2 Messzeitpunkten mit Angabe der Signifikanz

		1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.
1.	Kontrolle	**0,51*****	0,31***	0,17***	0,31***	0,27***	0,16***	0,25***	-0,20***	-0,22***	0,18***	0,22***	0,21***	0,24***	0,25***
2.	Gratifikation	0,34***	**0,59*****	0,21***	0,37***	0,34***	0,17***	0,34***	-0,25***	-0,39***	0,23***	0,27***	0,32***	0,29***	0,36***
3.	Teamwork	0,18***	0,27***	**0,52*****	0,23***	0,25***	0,36***	0,19***	-0,18***	-0,27***	0,22***	0,23***	0,22***	0,14***	0,23***
4.	Fairness	0,35***	0,52***	0,30***	**0,57*****	0,29***	0,17***	0,39***	-0,30***	-0,39***	0,20***	0,21***	0,27***	0,21***	0,37***
5.	Vorgesetzter	0,24***	0,41***	0,34***	0,38***	**0,46*****	0,19***	0,31***	-0,15***	-0,23***	0,20***	0,17***	0,22***	0,17***	0,22***
6.	Teamzusammenhalt	0,22***	0,26***	0,49***	0,32***	0,36***	**0,47*****	0,25***	-0,15***	-0,30***	0,23***	0,26***	0,32***	0,27***	0,33***
7.	Kommunikationskultur	0,32***	0,45***	0,29***	0,54***	0,54***	0,35***	**0,53*****	-0,29***	-0,36***	0,24***	0,22***	0,26***	0,22***	0,38***
8.	Erschöpfung	-0,20***	-0,27***	-0,21***	-0,34***	-0,17***	-0,22***	-0,28***	**0,66*****	0,45***	-0,17***	-0,33***	-0,31***	-0,20***	-0,31***
9.	Zynismus	-0,24***	-0,39***	-0,34***	-0,46***	-0,27***	-0,34***	-0,38***	0,59***	**0,67*****	-0,33***	-0,43***	-0,50***	-0,40***	-0,47***
10.	Professionelle Effizienz	0,14***	0,18***	0,18***	0,17***	0,15***	0,26***	0,18***	-0,22***	-0,38***	**0,52*****	0,44***	0,47***	0,37***	0,27***
11.	Vitalität	0,22***	0,28***	0,25***	0,28***	0,22***	0,35***	0,26***	-0,37***	-0,50***	0,67***	**0,64*****	0,54***	0,54***	0,39***
12.	Hingabe	0,22***	0,29***	0,27***	0,29***	0,25***	0,43***	0,28***	-0,34***	-0,59***	0,64***	0,72***	**0,65*****	0,52***	0,44***
13.	Absorbiertheit	0,22***	0,25***	0,15***	0,22***	0,16***	0,34***	0,21***	-0,17***	-0,41***	0,50***	0,65***	0,69***	**0,69*****	0,48***
14.	Commitment	0,30***	0,41***	0,30***	0,46***	0,32***	0,43***	0,42***	-0,38***	-0,57***	0,37***	0,51***	0,55***	0,51***	**0,71*****

Quelle: Beerlage, Arndt, Hering & Springer, 2009, S. 135

Forschungshypothesen werden in einem eigenen Gliederungspunkt einer wissenschaftlichen Arbeit aufgenommen. Sie bilden neben der Fragestellung den Fahrplan für alle Analysen im Forschungsprozess.
An die Formulierung von Forschungshypothesen werden Anforderungen geknüpft, die über die Theoriekonformität hinausgehen. Bortz und Döring (2003) formulieren vier Mindestanforderungen an Forschungshypothesen:

1. Forschungshypothesen beziehen sich auf empirisch untersuchbare reale Sachverhalte. Es können also Gegenstände, die einer objektiven Erfahrung nicht zugänglich sind, nicht untersucht werden: Gottesexistenz, Wiederauferstehung, unbefleckte Empfängnis. Dies deutet nicht auf die Unwahrheit dieser Annahmen hin, sondern vielmehr ist ein Nachweis oder eine Zurückweisung von Dingen, an die Menschen glauben mit wissenschaftlichen Methoden kaum möglich.

2. Formulierungen in Forschungshypothesen umschreiben allgemeingültige Sachverhalte, treffen also für größere Gruppen oder unter bestimmten Bedingungen (zumindest hypothetisch) *immer* zu. Insbesondere Sätze die „es gibt…“ oder „einige …“ enthalten sind keine „All-Sätze“ und somit keine Forschungshypothesen.

3. Forschungshypothesen enthalten in ihrer Aussage Ursache und Wirkung. Sie folgen der Struktur eines Konditionalsatzes und können in eine „Wenn-Dann“ Formulierung übertragen werden.

4. Annahmen in Forschungshypothesen müssen prinzipiell zurückgewiesen werden können, also *falsifizierbar* sein. Die bereits angesprochenen „Es gibt…“ oder „Es kann…“-Sätze sind nicht falsifizierbar, weil jedes potenzielle Ereignis die Hypothese untermauern würde.

In der Forschung werden zusammenfassend auf Basis von theoretischen Annahmen, die sich auf gründlichen Vorüberlegungen, systematischen Recherchen in bestehender Literatur und auf empirischen Befunden gründen, Fragestellungen und Hypothesen formuliert. Theorien werden gebildet von einzelnen Aussagen, die sich auf eine „Ursache-Wirkungsbeziehung“ beziehen. Diese Einzelaussagen sind Hypothesen, die sich auf Unterschiede, Zusammenhänge oder Veränderungen beziehen können. Wissenschaftliche Hypothesen müssen dabei real untersuchbar, empirisch direkt (Einzelbeobachtung) oder durch die begründete Zusammenfassung von Einzelbeobachtungen zugänglich sein. In der Hypothesenformulierung müssen die Allgemeingültigkeit der Aussage sowie Einflussfaktor und Effekt erkennbar werden. Grundsätzlich muss es möglich sein, Hypothesen zurückweisen zu können.

4.4.2 Statistische Hypothesen

In der Formulierung untersuchbarer Hypothesen wird die anstehende Datenanalyse inhaltlich vorgeplant und ihr Ergebnis ausgehend von den entwickelten theoretischen Grundlagen vorweggenommen. In der statistischen Analyse können keine inhaltlichen bzw. qualitativen Aussagen überprüft werden. Daher werden die inhaltlichen Aussagen aus Forschungshypothesen zum Zweck der Datenanalyse in statistische Aussagen übersetzt. Dabei muss zunächst einmal die Frage nach der Messung von Merkmalen geklärt werden (s. 6.2). Die sozial- und gesundheitswissenschaftliche Forschung bedient sich dazu standardisierter Fragebögen, mit denen inhaltliche Merkmale in Zahlen übertragen werden. In der klinischen Forschung werden zudem unterschiedliche Messeinrichtungen zur Diagnostik eingesetzt (Röntgen, CT, Blutuntersuchung, …). Unter der Annahme „Schwerhörige Patienten mit Hörgeräten der neuen Generation hören besser als schwerhörige Patienten mit herkömmlichen Hörgeräten.“ müsste auch geklärt werden, wie verschiedene Grade von Schwerhö-

rigkeit gemessen und woran sich „besser hören“ zeigen soll. Ferner spielt die Art der Hypothese eine Rolle, also ob Unterschiede, Zusammenhänge oder Veränderungen erwartet werden. Die Annahme, schwerhörige erzielen einen höheren Nutzen aus Hörgeräten der neuen Generation impliziert einen Gruppenvergleich, nämlich zwischen schwerhörigen Nutzern von Hörgeräten der alten und Nutzern von Hörgeräten der neuen Generation. Dieser Vergleich beruht auf der Annahme, Nutzer der neuen Hörgerätegeneration hören besser. Das Hörvermögen kann durch standardisiertes Testen des Wortverstehens von abgespielten Wörtern aus unterschiedlichen Lautkombinationen mit verschiedenen Lautstärken erfolgen und auch unter Störgeräuschen getestet werden. Nutzer neuer Hörgeräte müssten im Mittel mehr Worte zweifelsfrei verstehen, als Nutzer von Hörgeräten der alten Generation. Die statistische Aussage wäre hier: $M_{neues\ Hörgerät} > M_{altes\ Hörgerät}$ (lateinische Buchstaben) was bedeutet: der Mittelwert (M) des Wortverstehens ist in der Gruppe der Träger der neuen Hörgerätegeneration höher.

Mit dem Anspruch nach Allgemeingültigkeit von Forschungshypothesen (s. 4.4.1) sollte ein Ergebnis nicht nur für die untersuchte Stichprobe, sondern für alle Patienten mit einer vergleichbaren Schwerhörigkeit gültig sein. Es interessiert bei statistischen Hypothesen, mit denen auf die Wahrscheinlichkeit für bestimmte Ergebnisse in der Stichprobe ausgehend von einer (angenommenen) Situation in der Grundgesamtheit abgehoben wird, im Prinzip also nicht das Ergebnis in der Stichprobe, sondern die Situation in der (nicht untersuchten) Grundgesamtheit: $\mu_{neues\ Hörgerät} > \mu_{altes\ Hörgerät}$ (griechische Buchstaben). Erwartungen in der Stichprobe werden in statistischen Annahmen mit lateinischen, Erwartungen in der Population bzw. der Grundgesamtheit mit griechischen Buchstaben formuliert (Bortz & Döring, 2003).

Neben der statistischen Übersetzung inhaltlicher Annahmen erfolgt im Rahmen der Datenanalyse der Hypothesentest. Wie bereits angesprochen ist es mathematisch nicht möglich, qualitative Aussagen hinsichtlich der Wahrscheinlichkeit ihres Zutreffens in der Grundgesamtheit zu untersuchen. In der Datenanalyse wird ausgehend von der statistischen Annahme ein Hypothesenpaar aus Null- (H_0) und Alternativhypothese (H_A) formuliert (s. 9.1.1). In der Nullhypothese wird von *keinem Unterschied (Zusammenhang …)* in der Grundgesamtheit ausgegangen. Im Blutdruckbeispiel in Tabelle 1 geht die Nullhypothese erstens von keiner Veränderung des Blutdrucks im Zeitverlauf aus und erwartet ebenfalls keine Unterschiede im Behandlungsregime. Damit widerspricht die Nullhypothese den inhaltlichen Annahmen. In der Alternativhypothese wird in der Grundgesamtheit ein Unterschied erwartet oder ein Zusammenhang bzw. es erfolgte eine Veränderung.

Im Rahmen von Forschung wird keine Grundgesamtheit untersucht, sondern eine aus der Grundgesamtheit (möglichst zufällig) gezogene Stichprobe (s. 7.1). Zeigt sich in den Datenanalysen ein Unterschied in der Stichprobe, z.B. im Blutdruck zwischen verschiedenen Behandlungsregimes (Tabelle 1) stellt sich die Frage, ob dieses Ergebnis für die Grundgesamtheit zutreffend ist. Dies wird im Rahmen des Hypothesentestens jedoch nicht berechnet, sondern ob und mit welcher Wahrscheinlichkeit

ein errechneter Unterschied (Zusammenhang) in der Stichprobe möglich ist, wenn in der Grundgesamtheit kein Unterschied (Zusammenhang) besteht (Nullhypothese). Ist diese (Irrtums-) Wahrscheinlichkeit kleiner als 5% ist, wird von statistisch bedeutsamen (signifikanten) Ergebnissen gesprochen (s. 9).

Unter statistischen Hypothesen wird zusammenfassend die Übersetzung einer inhaltlichen Annahme in statistisch untersuchbare Sachverhalte verstanden. Dazu muss für die interessierenden Merkmale ein Messinstrument genutzt werden, mit dem qualitative Aspekte in Zahlen übertragen werden. In der statistischen Analyse interessiert anschließend Wahrscheinlichkeit für einen beobachteten oder extremeren Unterschied (Zusammenhang) in der zufällig gezogenen Stichprobe unter der Annahme, dass in der Grundgesamtheit kein Unterschied (Zusammenhang) vorliegt. Das statistische Hypothesenpaar bildet sich aus Null- (H_0) und Alternativhypothese (H_A), das sich auf die Grundgesamtheit bezieht. Statistisches Hypothesentesten geht von der Nullhypothese in der Grundgesamtheit aus. Ein statistisch bedeutsames Ergebnis liegt vor, wenn ein in der Stichprobe beobachteter Unterschied (Zusammenhang) mit einer Wahrscheinlichkeit von *weniger als 5%* vorliegt, sofern in der Grundgesamtheit kein Unterschied besteht (s. 9.1).

4.5 Variablen

Theorien formulieren grundlegende Verbindungen zwischen Merkmalen. Sie weisen beispielsweise auf das Krebsrisiko des Rauchens hin oder sie verknüpfen Aussagen über Arbeitsbedingungen mit der psychischen Gesundheit von Beschäftigten. Sie lassen die Entwicklung eines Forschungsplans zu, der in der Formulierung von Fragestellungen und Hypothesen skizziert ist. Die sozial- und gesundheitswissenschaftliche sowie die klinische Forschung interessieren „Merkmale“ von Menschen und der Bedingungen unter denen Menschen leben. Eine grundsätzliche Annahme ist dabei: verschiedene Bedingungen bringen unterschiedliche Merkmalsausprägungen hervor. Anstelle des Merkmalsbegriffs wird in der Forschung die Bezeichnung „Variable“ verwendet. Bei einer Variablen handelt es sich um die Menge von Ausprägungen eines spezifischen Merkmals (Bortz & Döring, 2003). Die Variable Schulabschluss liegt beispielsweise in den Ausprägungen „kein Schulabschluss“, „Hauptschulabschluss“, „Realschulabschluss“ und „Abitur“ vor. Die Variable Intelligenz ergibt sich aus der Zusammenfassung von Ergebnissen unterschiedlicher Fragen, die Hinweise auf die Problemlösekompetenz von Menschen geben. Variablenausprägungen werden regelgeleitet bestimmten Zahlen zugeordnet. Im Beispiel Schulabschluss könnten für „kein Schulabschluss“ eine „0“, für den „Hauptschulabschluss“ die „1“ vergeben werden. Realschulabsolventen würden mit „2“, Abiturienten mit „3“ codiert. Auch andere Codierschemen sind denkbar. *Daten* ergeben sich aus allen Merkmalsmessungen (Bortz & Döring, 2003). Der Gegenstand der Untersuchung sind Beschäftigte eines Unternehmens (Abbildung 6).

In der Systematisierung von Variablen gliedern Bortz und Döring (2003) Variablen anhand ihres *Stellenwerts* in Untersuchungen, der *Art der Merkmalsausprägung* sowie entsprechend ihrer *empirischen Zugänglichkeit*.

4.5.1 Stellenwert von Variablen in Untersuchungen

Die Bedeutung und den *Stellenwert* von Variablen in Untersuchungen bestimmt die Theorie. In den theoretischen Aussagen wird deutlich, ob eine Variable Einflussfaktor, Effekt oder Moderator ist. Einflussfaktoren werden als *unabhängige Variablen* (UV) untersucht. Die Ausprägung des als Effekt in die Untersuchung eingehenden Merkmals ist „abhängig" von der Ausprägung der unabhängigen Variablen. Theoretisch begründete Endpunkte einer Untersuchung (Effekte) werden als *abhängige Variablen* (AV) untersucht (Abbildung 6).

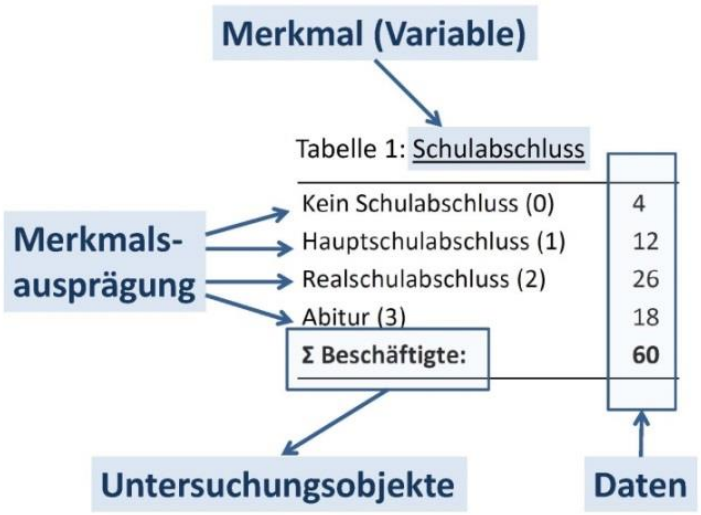


Abbildung 6: Begriffsbestimmung: Merkmal, Variable, Merkmalsausprägung, Untersuchungsobjekte und Daten (eigene Darstellung).

Beispiel für eine theoretisch begründete Verbindung zwischen UV und AV ist die Annahme, Fettleibigkeit erhöht die Herz-Kreislauf-Sterblichkeit. Eine hinreichende theoretische Begründung erfolgt aus der Physiologie des Menschen, aus Annahmen zum Fettstoffwechsel und seinen Auswirkungen auf die Situation der arteriellen Gefäßwände. Dieser Zusammenhang kann, ebenfalls theoretisch fundiert, hinsichtlich seiner Stärke verändert werden. Rauchen kann beispielsweise das Risiko für Herz-Kreislauf-Erkrankungen durch Fettleibigkeit noch verstärken. Regelmäßige sportliche Betätigung kann das Risiko von Herz-Kreislauf-Erkrankungen durch Fettleibigkeit reduzieren. Variablen, die Zusammenhänge modifizieren gehen als *Moderatorvariablen* (MoV) in Untersuchungen ein. In klinischen Studien sind beispielsweise Kontrollvariablen oder (sofern wesentliche Faktoren unberücksichtigt blieben) Störvariablen einflussreiche Moderatorvariablen (Abbildung 7).

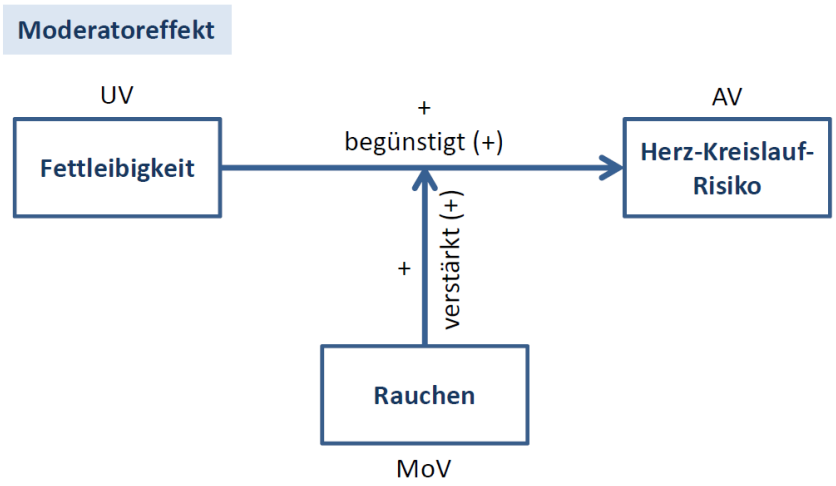


Abbildung 7: Unabhängige, abhängige und Moderatorvariable
UV= unabhängige Variable, AV= abhängige Variable, MoV= Moderatorvariable (eigene Darstellung).

Variablen, durch die ein Zusammenhang zwischen UV und AV erst erkennbar wird, sind *Mediatorvariablen* (MeV). In der sozial- und gesundheitswissenschaftlichen Forschung können „echte" Mediatoreffekte selten beobachtet werden. In den Naturwissenschaften übernehmen Mediatorvariablen Relaisfunktionen. Beispielhaft kann das anhand der Ausschüttung von Hormonen beleuchtet werden. Bestimmte effektorische Hormone aus der Nebennierenrinde werden erst nach einer Aktivierungskaskade ausgeschüttet. Cortisol ist ein effektorisches Hormon und wird aus der Nebennierenrinde nur freigesetzt, wenn ein Releasinghormon (Corticotropes Releasinghormon) vom Hypothalamus ausgeschüttet wird und anschließend vom Hypophysenvorderlappen ein glandotropes Hormon (Adrenocorticotropes Hormon, ACTH) über die Blutbahn zur Nebennierenrinde gelangt (Abbildung 8).

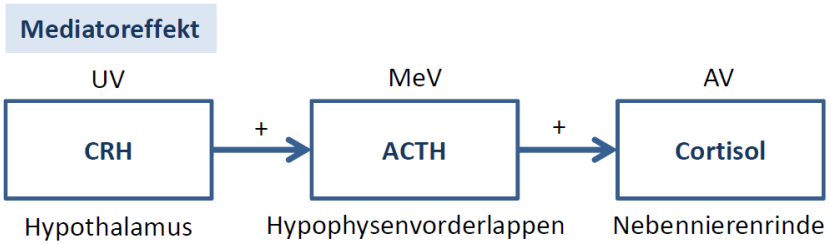


Abbildung 8: Unabhängige, abhängige und Mediatorvariable
UV= unabhängige Variable, AV= abhängige Variable, MeV= Mediatorvariable (eigene Darstellung)

4.5.2 Art der Merkmalsausprägung von Variablen

Variablen liegen entweder als *stetige* (kontinuierliche) oder *diskrete* (diskontinuierliche) Variable vor. Bei stetigen Variablen befinden sich zwischen zwei definierten Merkmalsausprägungen unendlich viele weitere Merkmalsausprägungen. Messwerte sind natürlicher Weise stetige Variablen, zwischen 20 und 21 cm befinden sich unendlich viele Ausprägungen (20.1, 20.01, 20.001, 20.0001, …). Bei diskontinuierlichen bzw. diskreten Variablen liegt eine endliche Anzahl Ausprägungen zwischen zwei Merkmalsausprägungen vor. Zwischen zwei und vier Kindern liegt genau eine weitere Merkmalsausprägung: drei Kinder. Alle kategorialen Merkmale zählen zu den diskreten Variablen: höchster Schulabschluss, Wohnort, Geschlecht, usw. (Bortz & Döring, 2003). Diskrete Variablen können zwei (*dichotom*) oder mehrere Ausprägungen (*polytom*) beinhalten. Die Merkmalsausprägung kann *natürlich* (das Geschlecht) oder *künstlich* diskret (Intervalle des Lebensalters 20 bis 30, 30 bis 40, …, usw.) sein. Stetige Variablen können in diskrete Variablen übertragen werden, in umgekehrte Richtung ist eine Transformation dagegen nicht möglich, d.h. eine diskrete Variable kann nicht in eine stetige Variable umgewandelt werden.

4.5.3 Empirische Zugänglichkeit von Variablen

Nicht zuletzt erfolgt eine Systematisierung von Variablen nach ihrer empirischen Zugänglichkeit. Direkt beobachtbare Merkmale werden als *manifeste Variablen* untersucht. Prinzipiell gehören alle Variablen, die mit *einer* Messung erfasst werden können zu den manifesten Variablen. Körpergröße, Körpermasse, Temperatur oder Blutzuckerspiegel sind manifeste Variablen. *Latente Variablen* sind dagegen nicht direkt beobachtbar, sie werden als hypothetisches Konstrukt untersucht, das durch mehrere manifeste Variablen erklärt wird. Die Formulierung latenter Variablen erfolgt auf der Basis einer theoretischen Annahme und im Prozess der Operationalisierung von Merkmalen (6.2.2). Beispielsweise wird angenommen, das Verhältnis zwischen Körpermasse und Körperoberfläche ist ein aussagekräftiges Konstrukt für die Beschreibung des Gesundheitszustands oder die Diagnose von Krankheiten. Der dahinter stehende Body-Mass-Index (BMI) erschließt sich durch die manifesten Merkmale Körpermasse und Körperoberfläche. Klima ist ein weiteres hypothetisches Konstrukt, das zunächst nicht anhand eines Merkmals direkt beobachtbar ist, sondern erst durch die Zusammenschau von Temperatur- und Niederschlagsverlauf, Sonnenstand, Jahreszeiten und Vegetation definiert werden kann. Die meisten psychologischen und sozialwissenschaftlich interessierenden Merkmale bei Menschen sind latente Merkmale, die sich erst durch mehrere Messungen oder mehrere Fragen erschließen. Auch fallen Krankheitsdiagnosen unter latente Variablen, die durch in der ICD 10 oder im DSM V festgelegte diagnostische Kriterien erfasst werden.

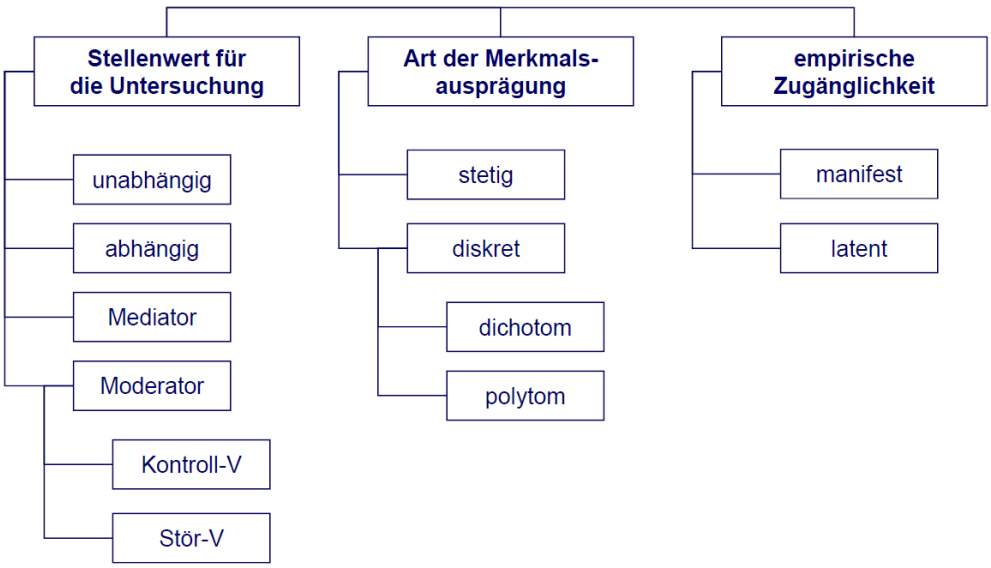


Abbildung 9: Systematisierung von Variablen nach Stellenwert, Merkmalsausprägung und empirischer Zugänglichkeit (Bortz & Döring, 2003)

4.6 Zusammenfassung und Aufgaben

Ausgangspunkt für wissenschaftliche Arbeiten ist die Entwicklung und Formulierung eines geeigneten untersuchbaren Themas. Die Themensuche orientiert sich zunächst stark am Interesse des Forschers und gründet sich auf sein Wissen darüber, welche Themen in einer Disziplin inhaltlich sinnvoll sind. Die Entwicklung einer theoretischen Basis zu einem Thema erfolgt durch systematische Überlegungen die flankiert werden durch Literaturanalysen und die Interpretation empirischer Studien zu einem Thema. Als Theorien werden Aussagensysteme bezeichnet, die den für die Forschung interessierenden Ausschnitt der Realität beleuchten. Jede Aussage in einer sozial- und gesundheitswissenschaftlichen Theorie, die im Rahmen quantitativer Forschungsdesigns untersucht werden soll, muss mit Quellen untersetzt sein. Theorien werden selten insgesamt und häufiger ausschnittsweise untersucht. Sie liefern vorweg die Erklärung für empirische Befunde und sollten so einfach und so komplex formuliert werden, wie es für die Beantwortung einer Forschungsfrage notwendig ist. Aussagen in Theorien sind logisch, widerspruchsfrei, empirisch untersuchbar und liefern zusätzliche Erklärungen.

Fragestellungen gründen sich in quantitativen Studien auf den theoretischen Annahmen, sie nehmen das Problem, seine Einflussfaktoren und die erwartete Entwicklung eines Problems auf. Die Fragestellung ist Basis und Orientierung für Datenerhebung, Datenanalyse und führt zur Formulierung von Hypothesen.

Hypothesen sind Elemente innerhalb des komplexen Aussagesystems in Theorien. Sie werden im Forschungsprozess als Unterschieds-, Zusammenhangs- und Verän-

derungshypothesen formuliert. Wesentliche Anforderungen an Forschungshypothesen sind der Bezug auf untersuchbare Sachverhalte, Allgemeingültigkeit der Aussage, die Aufnahme von Einflussfaktor und Effekt und die Möglichkeit, eine Hypothese auch zu verwerfen (Falsifizierbarkeit). Das statistische Hypothesenpaar umfasst Null- und Alternativhypothese. Beide Hypothesen werden in einer wissenschaftlichen Arbeit nicht verschriftlicht, sie treffen Aussagen für die Grundgesamtheit, deren Eintrittswahrscheinlichkeit auf der Basis von Daten aus Zufallsstichproben analysiert wird. Je unwahrscheinlicher ein Unterschied oder Zusammenhang in der Stichprobe ist, wenn für die Grundgesamtheit kein Unterschied (Nullhypothese) angenommen wird, desto eher wird von statistischer Signifikanz gesprochen.

In Hypothesen wird die Beziehung zwischen Merkmalen beschrieben. Im Forschungskontext werden Merkmale, die in unterschiedlichen Ausprägungen vorliegen können als Variablen untersucht. Variablen können als unabhängige, abhängige Moderator- oder Mediatorvariable in Untersuchungen aufgenommen werden. Die Theorie entscheidet, was Einflussfaktor (UV) und was Effekt (AV) ist. Variablen können stetig oder diskret ausgeprägt sein. Stetig sind Variablen, die zwischen zwei Ausprägungen unendlich viele weitere Ausprägungen umfassen. Diskrete Variablen integrieren eine endliche Zahl weiterer Stufen zwischen zwei Merkmalsausprägungen.

Variablen können entweder direkt mit einer Messung erfasst werden (manifeste Variable) oder sie müssen auf Grundlage systematischer Überlegungen aus mehreren manifesten Variablen entwickelt werden (latente Variablen). Latente Variablen finden sich häufig im sozial- und gesundheitswissenschaftlichen Forschungskontext. Krankheitsdiagnosen oder Persönlichkeitsmerkmale (Engagement, Wohlbefinden, Zufriedenheit) sind latente Variablen, die aus mehreren manifesten Variablen geschätzt werden.

Schlüsselwörter: *abhängige Variable, Alternativhypothese, Analyse, Art der Merkmalsausprägung, Daten, dichotom, diskrete Variable, empirische Zugänglichkeit, falsifizierbar, Forschungshypothesen, gerichtete Hypothese, Hypothesen, Kontrollgruppen, künstlich diskret, latente Variablen, manifeste Variablen, Mediatorvariable, Moderatorvariable, natürlich diskret, Nullhypothese, polytom, stetige Variable, unabhängige Variable, ungerichtete Hypothese, Unterschiedshypothesen, Variablen, Veränderungshypothesen, Zusammenhangshypothesen*

Aufgaben zur Lernkontrolle

1. Welche Komponenten gehören zu Theorien und welche Anforderungen sind an Theorien geknüpft?

2. Worin unterscheiden sich die Fragestellungen und Ihre Ziele bei der qualitativen zur quantitativen Forschung?

3. Wofür steht die Abkürzung PICO?

4. Worin unterscheidet sich eine Forschungshypothese von einer statistischen Hypothese?

5. Nennen Sie die 4 Mindestanforderungen an Hypothesen.

6. Ab wann wird von einem signifikanten Ergebnis gesprochen?

7. Definieren Sie folgende Begriffe und grenzen Sie diese ggf. zu dem anderen Terminus ab:

    a. unabhängige Variable vs. abhängige Variable
    b. Moderatorvariable
    c. Mediatorvariable
    d. stetig vs. diskret
    e. dichotom vs. polyton
    f. manifest vs. latent

Aufgabe zur Berufspraxis

8. Für welche Themen aus der Sprachtherapie interessieren Sie besonders? Würden die Kriterien zur Themenwahlfindung bei Ihnen zutreffen?

Literatur

Antonovsky, A., & Franke, A. (1997). *Salutogenese zur Entmystifizierung der Gesundheit*. Tübingen: DGVT-Verlag.

Atteslander, P., & Cromm, J. (2010). *Methoden der empirischen Sozialforschung* (13., neu bearbeitete und erweiterte Auflage ed.). Berlin: Erich Schmidt Verlag.

Beerlage, I., Arndt, D., Hering, T., & Springer, S. (2009). *Arbeitsbedingungen und Organisationsprofile als Determinanten von Gesundheit, Einsatzfähigkeit sowie haupt- und ehrenamtlichem Engagement bei Einsatzkräften in Einsatzorganisationen des Bevölkerungsschutzes. Abschlussbericht, September 2009*. Magdeburg & Bonn: Hochschule Magdeburg-Stendal.

Bengel, J., Strittmatter, R., & Willmann, H. (2006). *Was erhält Menschen gesund? Antonovskys Modell der Salutogenese - Diskussionsstand und Stellenwert; eine Expertise* (9., erw. Neuaufl ed.). Köln: BZgA.

Bortz, J., & Döring, N. (2003). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler* (3., überarb. Aufl., Nachdr. ed.). Berlin [u.a.]: Springer.

Döring, N., Bortz, J., & Pöschl, S. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften* (5. vollständig überarbeitete, aktualisierte und erweiterte Auflage ed.). Berlin Heidelberg: Springer.

Fawcett, J., & Erckenbrecht, I. (1999). *Spezifische Theorien der Pflege im Überblick*. Bern Göttingen Toronto Seattle: Verlag Hans Huber.

Raithel, J. (2008). *Quantitative Forschung ein Praxiskurs* (2., durchgesehene Auflage ed.). Wiesbaden: VS Verlag für Sozialwissenschaften.

Rasch, B., Friese, M., & Hofmann, W. (2010). *Quantitative Methoden : Einführung in die Statistik für Psychologen und Sozialwissenschaftler Bd. 2* (3., erw. Aufl. ed.). Berlin [u.a.]: Springer.

Schäfer, T. (2016). *Methodenlehre und Statistik Einführung in Datenerhebung, deskriptive Statistik und Inferenzstatistik*. Wiesbaden: Springer.

5 Studiendesigns in der sozial-, gesundheits- und therapiewissenschaftlichen Forschung

Lernergebnisse:

- Studiendesigns können nach zeitlicher Dimension (Kapitel 5.2), ausgehend vom Vergleich zwischen Gruppen (Kapitel 5.3) und nach Studienort (Kapitel 0) differenziert und Ziele unterschiedlicher Studiendesigns erläutert werden.

- Deskriptive Studien können hinsichtlich ihrer Ziele von anderen Studienarten abgegrenzt werden. Ziele deskriptiver Studien können beschrieben und ihre Bedeutung für die klinische und politische Entscheidungsfindung diskutiert werden (Kapitel 5.1).

- Besonderheiten von Einzelfallstudien können genannt und ihre Aussagekraft kritisch gewichtet werden (Kapitel 5.5).

- Die Begriffe interne und externe Validität können definiert, Studienarten hinsichtlich der Gütekriterien bewertet und die Bewertung für die Gewichtung von Studienergebnissen diskutiert werden (Kapitel 5.6).

Das Thema steht, wesentliche theoretische Grundlagen sind erarbeitet, Fragestellung und Hypothesen sind im Entwurf formuliert. Dann kann es eigentlich losgehen. Was jetzt fehlt ist die grundsätzliche Planung des Studiendesigns (5.1 bis 5.5), die Rekrutierung und Auswahl einer Stichprobe und die Festlegung geeigneter Messinstrumente (6.2). Von der Festlegung des Studiendesigns hängt wesentlich die Aussagekraft einer wissenschaftlichen Studie ab. Studiendesigns werden in der Fachliteratur zur quantitativen Forschungsmethodik systematisiert nach ihrer zeitlichen Dimension (5.2), also ob lediglich eine Messung oder Messwiederholungen bei derselben Stichprobe erfolgen, danach, ob Einflussfaktoren kontrolliert werden oder nicht, ob zwei oder mehrere Gruppen unterschiedlich behandelt werden, die Gruppeneinteilung auf Zufall beruht und verblindet wurde (5.3) und wie gut die Rahmenbedingungen oder ergebnisrelevante Störungen kontrolliert werden (50). In den Therapiewissenschaften erfolgen zudem aus unterschiedlichen Gründen Untersuchungen bei nur einer Person in Form von Einzelfallstudien (55.5). Von der Wahl des Studiendesigns und der Sorgfalt bei der Durchführung hängen die Aussagekraft der Studienergebnisse und ihre Übertragbarkeit in die Klinik oder die Praxis ab. Als Gütekriterien werden die interne und externe Validität von Studien betrachtet.

Wie bereits in der knappen Vorstellung des Forschungsprozesses angesprochen, ist qualitative Forschung zirkulär angelegt. Offenheit und Flexibilität im Verlauf einer Untersuchung gelten als wichtige Kennzeichen qualitativer Forschung. Trotzdem ist auch in qualitativen Studien eine reflektierte und begründete Auswahl des Forschungsdesigns von grundlegender Bedeutung. In der Literatur werden verschiedene Vorschläge zur Systematisierung qualitativer Designs diskutiert. Beispielsweise orientiert sich Flick (2000, S. 257) in seinem Vorschlag zur Klassifikation qualitativer Basisdesigns an der zeitlichen Dimension sowie an der Größe der Stichprobe. Mayring (2010, S. 231) unterscheidet – vergleichbar zum quantitativen Forschungsstil – explorative und deskriptive Designs sowie Kausal- und Zusammenhangsanalysen. Verbreiteter ist in der qualitativen Forschungslandschaft die Unterscheidung verschiedener qualitativer Forschungsansätze oder -traditionen. Creswell (2007) nennt in einem einflussreichen Lehrbuch beispielsweise Grounded-Theory-Designs, Phänomenologische Studien, Ethnographische Studien, Narrative Studien und Qualitative Einzelfallstudien als zentrale qualitative Ansätze. Die Konversations- oder Gesprächsanalyse und die Partizipative Aktionsforschung werden in der Gesundheitsforschung vielfach als weitere grundlegende Traditionen genannt. Alle Ansätze formulieren zentrale methodologische und methodische Ausgangspunkte. Einige werden im nachfolgenden kommentierten Reader zur qualitativen Forschung knapp vorgestellt. Die folgenden Kapitel fokussieren Studiendesign des quantitativen Forschungsstils. Das gilt auch für die Vorstellung von Gütekriterien im Kapitel 5.6. Kriterien zur Bewertung qualitativer Forschung werden im qualitativen Reader thematisiert.

5.1 Deskriptive Studien

Fragestellungen und Probleme, denen auf der Basis einer substanzielle Beschreibung einer Population nachgegangen wird, werden mit deskriptiven Studien erforscht (Döring, et al., 2016). Bund, Länder und Kommunen bedienen sich deskriptiver Studien zur Planung der gesundheitlichen Versorgung der Bevölkerung (Gesundheitsberichterstattung), zur Schulplanung (Demografieberichte) oder zur Planung des Öffentlichen Personennahverkehrs. Es existieren Gesetze, die die Datenerfassung und -aufbereitung regeln, wie beispielsweise die Gesundheitsdienstgesetze der Länder. Auch die Wirtschaft nutzt Daten aus populationsweiten Konsumentenbefragungen. Das regelmäßig aktualisierte Sozioökonomische Panel (SOEP) stellt beispielsweise Daten und Informationen zur wirtschaftlichen und sozialen Lage der Bevölkerung, zu den Einkommens- und Schichtverhältnissen bereit. Seit 1984 werden bei der immer selben Kohorte Daten erhoben. Dabei können Lebensläufe unterschiedlicher Bevölkerungsgruppen vor dem Hintergrund individueller Entscheidungen und Qualifikationen sowie gesellschaftlicher und wirtschaftlicher Entwicklungen nachverfolgt werden. Die Informationen sind einerseits Grundlage für Politikberatung, sind zugleich auch Datenbasis für die Grundlagenforschung in den Sozial-, Wirtschafts- und Geisteswissenschaften (Wagner, Göbel, Krause, Pischner & Sieber, 2008). Vom Robert-Koch-Institut wird mit dem Deutschen Erwachsenen Gesundheitssurvey (DEGS) und der Studie zur Gesundheit von Kindern in Deutschland (KiGGS) ebenfalls in Form deskriptiver Studien, teils über mehrere Messzeitpunkte, eine Datenbasis zur Beschreibung der gesundheitlichen Lage Erwachsener und Kinder erarbeitet (Robert-Koch-Institut, 2012). Ein weiteres aktuelles Beispiel für deskriptive Studien ist die Nationale Kohorte (NAKO), bei der über mehrere Jahre hinweg 200.000 Menschen regelmäßig untersucht werden. Die Daten sollen Hinweise auf Einflussfaktoren und die Entstehung der großen Volkskrankheiten (Herzinfarkte, Diabetes, Krebs) liefern (Ahrens & Jöckel, 2015). Auch regelmäßig durchgeführte Zensusuntersuchungen, zu deren Teilnahme deutsche Staatsbürger verpflichtet sind, sind im Prinzip deskriptive Studien.

Zusammenfassend wird über deskriptive Studien die aktuelle Situation oder die Entwicklung bestimmter Merkmale in Bevölkerungsgruppen nachgezeichnet und dokumentiert. Grundlage dafür sind neben Bundes- und Landesgesetzen (Gesundheitsberichterstattung in den Gesundheitsdienstgesetzen) grundsätzliche und vertiefte Interessen aus Wissenschaft, Wirtschaft, Sozialpolitik usw. Neben der gesundheitlichen Lage werden soziale, politische, wirtschaftliche und individuelle Aspekte in Verbindung mit den untersuchten Merkmalen einer Population berücksichtigt. Deskriptive Studien können in jede der anschließend dargestellten Systematisierungsvarianten wissenschaftlicher Studien aufgenommen werden. Sie werden als Quer- und Längsschnittstudien durchgeführt (5.2), können Gruppenvergleiche beinhalten (5.3) und im Feld oder unter kontrollierten Laborbedingungen erfolgen (0).

5.2 Systematisierung von Studien nach der zeitlichen Dimension – Quer-, Längsschnitt-, Fall-Kontroll-, Kohortenstudien

Wissenschaftliche Studien können zu einem oder zu mehreren Messzeitpunkten durchgeführt werden. *Querschnittsstudien* erheben Daten in einer Stichprobe zu einem Zeitpunkt oder einem kurzen Zeitraum (Raithel, 2008). Daten aus Querschnittstudien können, sofern sie Zufallsstichproben entstammen, einer substanziellen Beschreibung einer Population dienen. Sofern gründlich und nachvollziehbar eine theoretische Basis entwickelt (4.2), wissenschaftliche Fragen und Hypothesen aus der Theorie entwickelt wurden (4.3, 4.4), können mit Daten aus Querschnittstudien auch Hinweise auf Zusammenhänge zwischen Merkmalen aufgedeckt werden. Keinesfalls können Aussagen zu zeitlichen Abläufen oder gar Ursache-Wirkungsaussagen getroffen werden. Ein zeitgleiches Auftreten von Fehlernährung und Atherosklerose kann beispielsweise auf Basis der physiologischen und biochemischen Grundlagen durchaus in eine potenzielle Reihenfolge gebracht werden – also zuerst ernähren sich Menschen einseitig, was langfristig zu Veränderungen in der Gefäßwand bzw. zur Einlagerung von Fettstoffwechselresten führen kann. Einen hinreichenden Beleg für diese Aussagen können Querschnittsstudien nicht liefern. Es bleibt eine vorsichtige Formulierung potenzieller Trends auf Basis gut fundierter theoretischer Annahmen. Querschnittstudien werden ebenfalls im klinischen Kontext und zur Untersuchung relevanter Einflussfaktoren auf Krankheitsverläufe eingesetzt. In diesen *Fall-Kontroll-Studien* werden Menschen mit einer Erkrankung verglichen mit Menschen ohne diese Erkrankung (s. 5.3).

Fall-Kontroll-Studien sind Querschnittstudien im klinischen Kontext. An dem skizzierten Beispiel zum möglichen Zusammenhang zwischen Fehlernährung und Atherosklerose sollten Menschen mit Atherosklerose beispielsweise deutlich häufiger fettleibig sein, sich in deutlich stärkerem Ausmaß fettreich ernähren und in ihrer Familie mehr vergleichbare Fälle von Atherosklerose haben. Die Blickrichtung von Querschnitts- und Fall-Kontrollstudien ist *retrospektiv*. Es wird also nicht untersucht, ob ein bestimmter Einflussfaktor im Zeitverlauf zu einer Erkrankung führt, sondern ob Menschen mit einer bestimmten Erkrankung in der Vergangenheit bestimmte Verhaltensweisen ausgeprägter oder häufiger zeigten als Menschen ohne die Erkrankung oder ob eine bestimmte Maßnahme in der Gruppe mit der Erkrankung häufiger oder seltener durchgeführt wurden.

Längsschnittstudien erheben bei ein und derselben Stichprobe Daten zu mehreren Messzeitpunkten. Auf der Grundlage von Längsschnittstudien können Verlaufsaussagen getroffen und Hinweise auf Ursache-Wirkungs- (Kausal-)beziehungen aufgedeckt werden. Nur auf der Basis von Längsschnittuntersuchungen lassen sich theoretische Annahmen über Zeitverläufe oder Ursache-Wirkungsbeziehungen mit Daten untermauern (Raithel, 2008). Allerdings muss auch in der Ausformulierung der Ergebnisse von Längsschnittuntersuchungen zurückhaltend argumentiert werden. Notwendig ist dies zum einen wegen der Untersuchung in Stichproben, nicht in der Grundgesamtheit – hier können aus ganz unterschiedlichen Gründen Fehler ergeben (s. 7), zum anderen aus theoretischen Gründen. Krankheiten oder andere in

den Sozial- und Gesundheitswissenschaften interessierende Endpunkte von Studien sind häufig nicht auf wenige Einflussfaktoren zurückführbar, sondern Ergebnis eines Wechselwirkungsgeflechts zahlreicher Faktoren. Selbst wenn in Längsschnittuntersuchungen relevante Zusammenhänge gefunden werden, die auf Ursache-Wirkungs-Beziehungen hindeuten, so sind dennoch Unsicherheiten groß und zugleich alternative Erklärungen denkbar.

Im klinischen Kontext werden Längsschnittuntersuchungen in Form *randomisierter kontrollierter Studien* (5.3) und als *Kohortenlängsschnittstudien* durchgeführt. Die Untersuchungsrichtung ist *prospektiv*, sie blickt im Zeitverlauf in die Zukunft.

Kernziel von *Kohortenstudien* ist der Nachweis der Bedeutung bestimmter, theoretisch begründeter Einflussfaktoren auf den Gesundheitszustand bzw. den Krankheitsverlauf von Patienten. Basis von Kohortenstudien sind gesunde Menschen, die sich allerdings hinsichtlich ihrer Verhaltensweisen (Raucher vs. Nichtraucher, schlanke vs. übergewichtige Menschen, Sportler vs. Nichtsportler, …) oder bestimmter theoretisch begründeter Einflussfaktoren auf den Gesundheitszustand unterscheiden (hohe vs. niedrige Lärmexposition, leben an Hauptverkehrsstraßen vs. leben im Grünen, hoher vs. niedriger Bildungsstand, …). Im Zeitverlauf verändert sich der Gesundheitszustand nicht in beiden Gruppen gleich. Langfristig sollten Menschen mit der Exposition von Risikofaktoren beim Endpunkt deutlich häufiger von bestimmten Krankheiten betroffen sein. Raucher sollten, sofern die eingeführten theoretischen und physiologischen Annahmen zutreffen, beispielsweise im Zeitverlauf eine kürzere Lebenserwartung, ein höheres Lungenkrebsrisiko und weniger Ausdauer als Nichtraucher haben.

In einigen Quellen werden auch *retrospektiven Kohortenstudien* beschrieben (u.a. Klug, Bender, Blettner & Lange, 2004). Sie beschreiben retrospektive Kohortenstudien als „Einen Sonderfall (…), (der, Anm.) vor allem in der Arbeitsepidemiologie Anwendung findet." (Klug et al., 2004, S. T7). Ähnlich wie bei Fall-Kontroll-Studien wird der Einfluss bestimmter Faktoren und Expositionen bei Menschen mit einer bestimmten Krankheit zurückverfolgt. Ein relevanter Einfluss einer Exposition läge vor, wenn ein Großteil der von einer bestimmten Krankheit betroffener Menschen dieser Exposition in einem bestimmten Ausmaß ausgesetzt war (Klug et al., 2004).

Auch wenn bei retrospektiven Kohortenstudien prinzipiell ein (zurückliegender) Zeitverlauf betrachtet wird, handelt es sich genau genommen nicht um Längsschnittuntersuchungen, sondern um Datenerhebungen zu einem Messzeitpunkt, bei der Informationen aus der Vergangenheit ausgewertet werden. Auf retrospektiven Studien beruhende Aussagen sind insgesamt vorsichtiger zu formulieren, als Ergebnisse prospektiver Studien. Auch wenn Dokumentationen zur Expositionen, wie beispielsweise Raucher- oder Trinktagebücher ausgewertet werden, können Manipulationen oder das Auslassen relevanter Informationen nicht ausgeschlossen werden. Ergebnisse retrospektiver Studien dienen der Untermauerung theoretischer Annahmen, sie können demnach auch Annahmen zu potenziellen Ursache-Wirkungs-

Beziehungen nachgehen. Sie sind allerdings keinesfalls ein Beleg für eine Zeitabfolge oder für Kausalbeziehungen.

Bei einer Systematisierung nach Zeitverlauf kann zusammenfassend in Quer- und Längsschnittstudien unterschieden werden. Querschnittstudien erheben Daten zu einem definierten Messzeitpunkt oder in einem engen Zeitraum. Die Daten ermöglichen Trendaussagen, sie weisen auf (theoretisch fundierte) Zusammenhänge hin, können jedoch keine Kausalbeziehungen aufdecken. Im klinischen Kontext werden Fall-Kontroll-Studien und sog. retrospektive Kohortenstudien zu einem Messzeitpunkt durchgeführt. Längsschnittstudien erheben Daten zu mehreren, inhaltlich begründeten Messzeitpunkten. Auf der Grundlage von Daten aus Längsschnittuntersuchungen sind Verlaufsaussagen möglich. Auch Hinweise auf Ursache-Wirkungs-Beziehungen finden sich in Längsschnittdaten. Kohortenstudien sind in die Zukunft gerichtet, also prospektive Längsschnittstudien.

5.3 Systematisierung von Studien ausgehend von der Kontrolle von Einflussfaktoren – experimentelle und quasi-experimentelle Designs

In der wissenschaftlichen Forschung werden unter Experimenten Untersuchungen subsummiert, in denen Einflussfaktoren in inhaltlich begründeter Weise variiert werden. Der Forscher interveniert und verändert aus gutem Grund bestimmte Rahmenbedingungen, von denen er Einfluss auf Endpunkte erwartet. Dies können Krankheiten, Beeinträchtigungen oder positive Folgen wie Lebenszufriedenheit sein. Es interessiert, ob eine abhängige Variable durch die Variation des Einflussfaktors verändert wird. Studien die Einflussfaktoren variieren und kontrollieren werden je nach methodischem Vorgehen zu *experimentellen*, *quasi-experimentellen* und *nicht-experimentellen* Untersuchungen gezählt (Döring et al., 2016). Experimente im Rahmen klinischer Studien untersuchen die Wirksamkeit von Interventionen. Dabei werden mindestens zwei Gruppen betrachtet: eine Interventions- und eine Kontrollgruppe. Interventionsgruppen erhalten die Intervention, von der ein theoretisch begründeter (Zusatz-)Nutzen erwartet wird. Studienteilnehmer in der Kontrollgruppe erhalten dagegen keine Intervention (unüblich), ein Placebo (das sind wirkstofffreie Präparate, die in Aussehen, Geruch und Geschmack dem Präparat mit Wirkstoff täuschend ähnlich sind) oder die bis dato am besten wirkende Therapie. Klinische experimentelle Studien erhalten zumeist nur dann ein positives Ethikvotum (s. 1.4), wenn die Kontrollgruppe ebenfalls behandelt wird und das am besten wirksame Therapieverfahren erhält. Bei der Untersuchung eines neuen Präparats oder eines neuen Verfahrens muss daher einen begründeter substanzieller Zusatznutzen für Patienten zu erwarten sein.

Experimentelle Studien untersuchen Interventions- und Kontrollgruppen. Die Zuordnung von Patienten erfolgt hier per Zufall, d.h. randomisiert. Die interne Validität dieser Studien erhöht sich, wenn die Studienteilnehmer die persönliche Zuordnung zu den Gruppen *nicht* kennen (also selber nicht wissen, ob sie in der Interventions- oder in der Kontrollgruppe sind) – man spricht in diesem Fall von *Einfachverblindung*. Wenn auch die Untersucher nicht wissen, ob sie ein Verfahren bei der Interventions- oder der Kontrollgruppe anwenden, wird von *Doppelblindstudien* gesprochen. *Dreifachverblindung* heißt, neben Studienteilnehmern und Untersuchern kennen auch Datenauswerter die Gruppenzuordnung nicht (Döring et al., 2016). Randomisierte und mehrfach verblindete klinische Studien haben eine hohe Aussagekraft und erlauben Kausalaussagen bei gründlicher theoretischer Fundierung. Im klinischen Kontext untersuchen *randomisierte, kontrollierte klinische Studien* (RCT) Menschen mit einer bestimmten Erkrankung. Ziel von RCT ist es, die Wirksamkeit und/oder Überlegenheit beispielsweise neuer Therapieverfahren nachzuweisen. Die Stichprobe wird auf Zufallsbasis (randomisiert) in eine Interventions- und eine Kontrollgruppe unterteilt. Die Interventionsgruppe wird mit einem Medikament/einer Intervention behandelt, von dem/der ein günstiger Einfluss auf den Gesundheitszustand – im besten Fall Genesung – erwartet wird. Patienten in der Kontrollgruppe erhalten entweder ein Placebo oder das nach aktuellem Stand der Forschung am besten wirksame Medikament. Der Gesundheitszustand des Patienten wird i.d.R. vor der Intervention (Baselineerhebung oder t_0), kurze Zeit nach der Intervention (t_1) und einige Zeit nach der Intervention (Follow-up, t_2) gemessen. Je nach Fragestellung oder Gesundheitsproblem sind Modifikationen der Gruppenzuweisung und der Messzeitpunkte denkbar. In Tabelle 1 (s. 4.4.1) wird das Ergebnis einer experimentellen klinischen Studie dargestellt.

Bei *quasi-experimentellen* Untersuchungen erfolgt die Gruppenzuordnung nicht zufallsgesteuert. Eine experimentelle Variation erfolgt dennoch, beide Gruppen werden geplant unterschiedlich behandelt. Die Aussagekraft und interne Validität quasi-experimenteller Untersuchungen ist kleiner als bei experimentellen Studien (Döring et al., 2016).

In nicht-experimentellen Untersuchungen erfolgt keine bewusste Gruppeneinteilung, sondern ein Vergleich vorgefundener Gruppen, die sich hinsichtlich der Einflussfaktoren (oder des Behandlungsregimes) unterscheiden. Die Aussagekraft und interne Validität nicht-experimenteller Untersuchungen sind klein (Döring et al., 2016).

Studiendesigns können zusammenfassend danach untergliedert werden, ob bei der unabhängigen Variablen gezielt Variationen (Interventionen) erfolgen oder nicht (Abbildung 10). Experimentelle Studien nutzen in mindestens zwei Gruppen unterschiedliche Interventionen, wobei in der Interventionsgruppe ein Verfahren eingesetzt wird, von dem ein Zusatznutzen erwartet wird. Die Gruppeneinteilung erfolgt auf Zufallsbasis (randomisiert). Auch quasi-experimentelle Studien variieren die Intervention in zumindest zwei Gruppen.

Im Unterschied zu experimentellen Studien erfolgt die Gruppenzuteilung nicht per Zufall. Nicht-experimentelle Studien vergleichen vorgefundene Gruppen, die sich hinsichtlich der Einflussfaktoren bzw. Interventionen unterscheiden. Eine gezielte Einteilung in Gruppen zu Studienzwecken erfolgt hier nicht. Die Aussagekraft und die interne Validität sind bei experimentellen Studien verglichen mit quasi- und nicht-experimentellen Studien hoch.

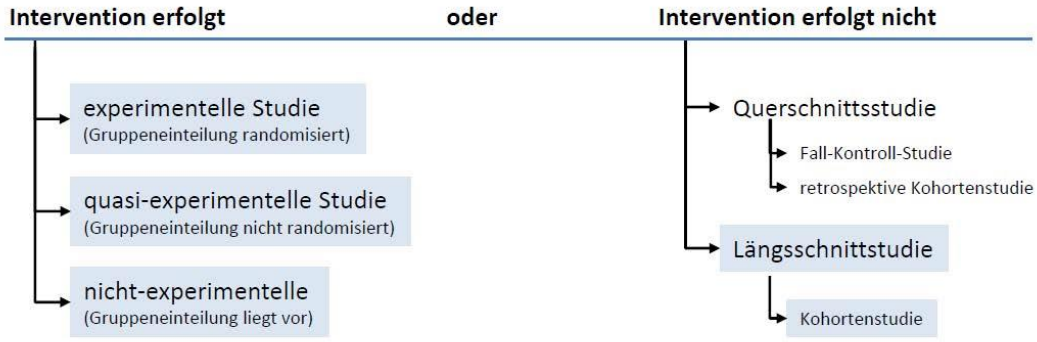


Abbildung 10: Systematisierung von Studiendesigns (modifiziert nach Schäfer, 2007, S. 102). Grau hinterlegte Felder sind als Längsschnittstudien angelegt (eigene Darstellung)

5.4 Studienort (Labor oder Feld)

Auswirkungen auf die Aussagekraft hat neben der Art der Gruppenzuteilung der Untersuchungsort. Hier wird unterschieden, ob eine Untersuchung unter Laborbedingungen oder im Feld durchgeführt wird. In *Laboruntersuchungen* werden die Rahmenbedingungen, unter denen die Studie durchgeführt wird, kontrolliert. Der Einfluss potenzieller Störvariablen, wie Geräusche, die Anwesenheit anderer Personen, Temperatur, Lichtverhältnisse, (…) wird beabsichtigt so klein wie möglich gehalten (zu unabhängigen, abhängigen, Moderator- und Störvariablen s. 4.5). Laborumgebungen entsprechen nachvollziehbarer Weise häufig nicht den natürlichen Umgebungsbedingungen von Studienteilnehmern. Aus diesem Grund ist zwar die interne Validität (s. 5.6) von Laborstudien vergleichsweise hoch, allerdings sind die Ergebnisse nicht ohne weiteres übertragbar (Döring et al., 2016).

Felduntersuchungen finden in der alltäglichen Umgebung von Studienteilnehmern statt. Alle Umgebungsfaktoren nehmen unverändert Einfluss und können die Aussagekraft des Studienergebnisses beeinflussen. Ein Ziel von Felduntersuchungen ist eine breite Übertragbarkeit der Ergebnisse. Allerdings geht dies zu Lasten der internen Validität (s. 5.6), weil auch nicht kontrollierte Umgebungsfaktoren die abhängigen Variablen beeinflussen können (Döring et al., 2016).

Ob eine Studie unter kontrollierten Laborbedingungen oder in der alltäglichen Umgebung von Studienteilnehmern durchgeführt wird, hängt einerseits von den zur Verfügung stehenden Mitteln ab. Andererseits davon, ob die Verknüpfung von unabhängiger und abhängiger Variablen durch eine gezielte Kontrolle potenzieller Störvariablen möglichst störungsfrei abgebildet werden soll (interne Validität) oder ob Ergebnisse auf Untersuchungen in der alltäglichen Umgebung beruhen und gut übertragbar sein sollen (externe Validität).

5.5 Sonderfall Einzelfallstudie

Die Anwendung statistischer Verfahren setzt die Streuung von Merkmalsausprägungen, also Variabilität bzw. Varianz voraus (s. 9). Mit der Befragung einer großen Zahl an Studienteilnehmern und unter der Annahme, Merkmalsausprägungen streuen in der Population zwischen Extrempolen, werden Daten generiert, die entsprechende Varianzerwartungen erfüllen sollten. In verschiedenen Disziplinen, der Psychologie z.B. im Zusammenhang mit einer tiefenpsychologisch fundierten Psychotherapie, in den Therapieberufen, in denen bei einzelnen Patienten die Kombination ganz bestimmter Therapieverfahren notwendig ist oder auch in der Medizin, wenn es um die Untersuchung von Patienten mit seltenen Erkrankungen geht, werden Daten oft nur bei einzelnen Menschen erhoben.

Bei Einzelfallstudien wird die Entwicklung von Merkmalen, z.B. von Gesundheitsparametern, bei einer Person im Zeitverlauf verfolgt. Da bei Einzelmessungen zu den jeweiligen Messzeitpunkten keine auf Varianz basierenden Verfahren genutzt werden könnten, werden mehrere Messungen vor (Baseline), unmittelbar danach und langfristig später (Follow-up) durchgeführt. Die daraus resultierenden Einzelwerte werden dann wie Werte mehrerer Personen für statistische Analysen genutzt. Abbildung 11 stellt beispielhaft eine Einzelfallstudie aus der Physiotherapie dar, bei der durch den therapeutischen Eingriff (Session) Auswirkungen auf die Ganggeschwindigkeit (Gangg) untersucht wird.

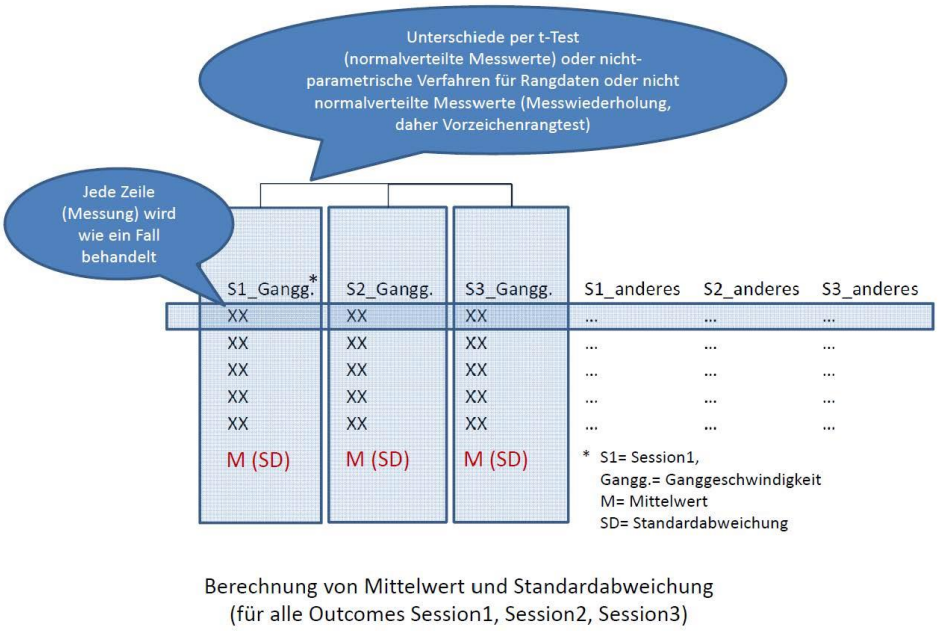


Abbildung 11: Einzelfallstudie nach jeweiligen Therapieeinheiten (Sessions) und die Veränderung der Ganggeschwindigkeit (eigene Darstellung)

Im Unterschied zu Studien bei Stichproben, bei denen jede Datenzeile einen Fall/eine konkrete Person repräsentiert (s. 8.1), sind die Datenzeilen bei Einzelfallstudien konkreten Messungen vorbehalten. Zu jedem Zeitpunkt (Session) werden also mehrere Messungen (>10) der Ganggeschwindigkeit vorgenommen, die aller Wahrscheinlichkeit nach unterschiedliche Messergebnisse hervorbringen – also eine bestimmte Varianz haben. Mit diesen Daten können prinzipiell alle statistischen Verfahren gerechnet werden. Allerdings lässt hier ein statistisch bedeutsames Ergebnis keinen Schluss auf die Grundgesamtheit zu. Ebenso wenig zulässig ist die kausale Bewertung der Ergebnisse (Köhler, 2012).

5.6 Gütekriterien wissenschaftlicher Studien

Grundlegende Gütekriterien von Forschung wurden in 1.3. angesprochen. Ausgehend vom Forschungsdesign, also ob zu einem oder zu mehreren Messzeitpunkten gemessen wurde, ob Interventions- und Kontrollgruppen verglichen werden und ob Studienteilnehmer auf Zufallsbasis Interventions- und Kontrollgruppe zugeordnet wurden, kann die Aussagekraft von Forschungsergebnissen bewertet werden. Die *interne Validität* beschreibt, wie gut Studienergebnisse kausal interpretierbar sind. Sofern Veränderungen bei der abhängigen Variable (also beim Effekt) zweifelsfrei mit der Variation der unabhängigen Variable (also des Einflussfaktors) erklärt werden kann, liegt eine hohe interne Validität vor. Intern valide sind experimentelle Untersuchungen, deren Gruppenzuordnung verblindet ist und die unter der Kontrolle potenzieller Störfaktoren durchgeführt wurde (Laborstudie). Bei der *externen Validität* interessiert, ob Studienergebnisse übertragbar sind, also ein vergleichba-

res Ergebnis auch außerhalb der Untersuchungsbedingungen denkbar ist. Übertragbar sind Studienergebnisse experimenteller Untersuchungen (Kontrollgruppendesign), die unter Feldbedingungen durchgeführt wurden. Die Annahme einer Übertragbarkeit der Studienergebnisse gründet sich auf die natürlichen, nicht oder wenig kontrollierten Rahmenbedingungen der Studie (Döring et al., 2016).

Mangold (2013) stellt adaptiert auf klinische Untersuchungen „Leitfragen zur Beurteilung der Validität“ vor (S. 69). Dazu zählen (ebd., S. 69):

1. Randomisierung der Gruppenzuordnung
2. Geheimhalten der Randomisierungsliste
3. niedrige Abbruchraten
4. Intention to Treat-Analyse – das bedeutet, auch bei denjenigen, die die Intervention vorzeitig beendeten werden weitere Messung durchgeführt
5. Dauer der follow-up-Periode
6. Übereinstimmung der Zielparameter vor Studienbeginn zwischen Interventions- und Kontrollgruppe
7. abgesehen von der Interventions- und Kontrollbehandlung werden Interventions- und Kontrollgruppe gleich behandelt
8. die eingesetzten Messinstrumente sind zuverlässig und gültig, die Messung objektiv
9. bei allen Patienten werden dieselben Messinstrumente eingesetzt
10. Studienteilnehmer und Therapeuten kennen die Gruppeneinteilung nicht (Verblindung).

Nicht alle der genannten Kriterien können für klinische Studien der Gesundheitsberufe ohne weiteres übertragen werden. Schwierigkeiten bestehen insbesondere in der Verblindung von Studienteilnehmern und Therapeuten. Bei der Untersuchung von Intervention in der Logopädie oder anderen therapeutischen und Pflegeberufen ist es schlicht unmöglich, konkrete, auf besonderes Fachwissen beruhende Interventionen, die auf Interaktion mit Patienten beruhen, zu verblinden. Sowohl Therapeut, als auch Patienten werden Unterschiede bemerken. Ein Therapeut kann zudem nicht fachgerecht intervenieren, wenn er nicht weiß, was er eigentlich macht. Ob Medikamente verteilt werden sollen, deren Beschaffenheit in Aussehen, Geschmack und Größe in Interventions- und Kontrollgruppe übereinstimmt oder ob eine spezifische logopädische, therapeutische oder pflegerische Intervention durch-

geführt und untersucht werden soll, macht einen Unterschied. Im letztgenannten Verfahren ist eine Verblindung nicht möglich. Künftig wäre eine kritische methodische Auseinandersetzung mit dem „Evidenzbegriff“ in den Gesundheitsberufen erforderlich und zudem die Definition von Kriterien evidenter Interventionen in den Gesundheitsberufen.

Der Validitätsbegriff wird zusammenfassend für verschiedene Gütebewertungen wissenschaftlicher Untersuchungen verwendet. Fragt man nach der Güte von Studienergebnissen insgesamt interessieren die interne und externe Validität. Bei intern validen Studien können Veränderungen der abhängigen Variablen auf die Variation der unabhängigen Variablen zurückgeführt werden. Studien müssen: 1. mit einem Interventions- und Kontrollgruppendesign durchgeführt werden, bei dem 2. Störfaktoren weitgehend ausgeschaltet bzw. kontrolliert werden. Extern valide Studienergebnisse erlauben die Übertragbarkeit der Ergebnisse auf Situationen außerhalb des Studiendesigns. Experimentelle Untersuchungen im Feld haben eine hohe externe Validität (Döring et al., 2016).

5.7 Zusammenfassung und Aufgaben

Wissenschaftlichen Fragestellungen wird in verschiedenen Studiendesigns nachgegangen. In der Literatur wird überwiegend differenziert, nach Studien zur Beschreibung von Populationen (deskriptive Studien, s. 5.1), nach Anzahl der Messzeitpunkte (5.2), nach der Untergliederung der Studienpopulation im Rahmen experimenteller, quasi-experimenteller oder nicht-experimenteller Designs (5.3), nach den Umgebungsbedingungen in Labor oder Feld (0) und letztlich nach der empirischen Untersuchungen bei einzelnen Menschen oder in Gruppen (5.5). Ziel wissenschaftlicher Studien sind aussagekräftige Ergebnisse, die stark vom geplanten Studiendesign abhängen. Wie gut Effekte auf die untersuchten Einflussfaktoren zurückgeführt werden können, wird anhand der internen Validität bemessen. Ob Studienergebnisse auch außerhalb des Untersuchungsumfelds übertragen werden können, wird mit externer Validität umschrieben.

Deskriptive Studien beschreiben die aktuelle Situation (Querschnittsstudie) oder die Entwicklung (Längsschnittstudie) bestimmter Merkmale in Populationen. Neben Bundes- und Landesgesetzen (Gesundheitsberichterstattung in den Gesundheitsdienstgesetzen) sind Interessen aus Wissenschaft, Wirtschaft, Sozialpolitik forschungsleitend. Beschrieben wird neben gesundheitlichen Aspekten auch die soziale, politische oder wirtschaftliche Situation.

Studiendesigns werden nach Zeitverlauf in Quer- und Längsschnittstudien unterschieden. Die Datenerhebung zu einem Messzeitpunkt erfolgt in Querschnittstudien. In Längsschnittstudien werden Daten zu mehreren Messzeitpunkten erhoben.

Nur anhand von Längsschnittstudien sind Verlaufsaussagen möglich bzw. können Kausalbeziehungen aufgedeckt werden. Im klinischen Kontext zählen Fall-Kontroll- und retrospektive Kohortenstudien zu Querschnittsstudien. Experimentelle Studien und Kohortenstudien sind Längsschnittstudien.

In experimentelle Studien erfolgt eine gezielte Variation der unabhängigen Variablen. In zwei (oder mehr) Gruppen werden unterschiedliche Maßnahmen angewendet. In der Interventionsgruppe wird ein neues Verfahren mit potenziellem Zusatznutzen eingesetzt, Kontrollgruppen erhalten die bewährte Maßnahme. Experimentelle Untersuchungen basieren auf eine Gruppeneinteilung nach dem Zufallsprinzip. Quasi-experimentelle Studien nehmen ebenfalls eine gezielte Gruppenbildung vor, nur erfolgt diese nicht zufallsbasiert. Nicht-experimentelle Untersuchungen nutzen bereits vorab gebildete Gruppen.

Eine Studie kann unter Laborbedingungen oder im Feld durchgeführt werden. Ergebnisse von Felduntersuchungen sind gut übertragbar auf Situationen außerhalb der Studienbedingungen, allerdings können hier unkontrollierte Störfaktoren das Studienergebnis beeinflussen. Laboruntersuchungen sind weniger gut übertragbar, Störvariablen werden dagegen kontrolliert. Die interne Validität von Laborstudien ist daher höher einzuschätzen als in Felduntersuchungen.

Einzelfallstudien sind Sonderfälle der quantitativen empirischen Forschung. Bei der Datenerhebung in Stichproben sind die betrachteten Merkmalsausprägungen zwischen den Studienteilnehmern verschieden. Diese Varianz zwischen Untersuchungsteilnehmern existiert nicht in Einzelfallstudien. Für die statistische Analyse werden bei einzelnen Probanden daher mehrere Messwerte generiert, die anschließend im Zeitverlauf analysiert werden können. Einzelfallstudien werden durchgeführt, wenn die Wirksamkeit sehr komplexer, individueller Interventionen untersucht werden soll. Ergebnisse von Einzelfallstudien erlauben allenfalls Trendaussagen, sind nicht ohne weiteres übertragbar und erlauben keine Kausalinterpretation.

Die Güte wissenschaftlicher Untersuchungen bemisst sich an der internen und externen Validität. Bei intern validen Studien können Veränderungen der abhängigen Variablen auf die Variation der unabhängigen Variablen zurückgeführt werden. Studien müssen: 1. mit einem Interventions- und Kontrollgruppendesign durchgeführt werden, bei dem 2. Störfaktoren weitgehend kontrolliert werden. Extern valide Studienergebnisse ermöglichen die Übertragbarkeit von Ergebnissen auf Situationen außerhalb des Studiendesigns. Experimentelle Untersuchungen im Feld haben eine hohe externe Validität (Döring et al., 2016).

Schlüsselwörter: *abhängige Variable, deskriptiven Studien, Doppelblindstudien, Dreifachverblindung, Einfachverblindung, experimentell, experimentelle Studien, externe Validität, externe Validität, Fall-Kontroll-Studien, Fall(Case-)Studie, Felduntersuchungen, interne Validität, interne Validität, interne Validität, Kohorten-*

längsschnittstudien, Kohortenstudien, Kontrollgruppe, Laboruntersuchungen, Längsschnittstudien, nicht-experimentell, prospektiv Längsschnittstudie, quasi-experimentell, Querschnittsstudien, randomisierte kontrollierte klinische Studien (RCT), randomisierter kontrollierter Studien, retrospektiv, retrospektive Kohortenstudien, unabhängigen Variablen, Varianz

Aufgaben zur Lernkontrolle

1. Grenzen Sie die Begrifflichkeiten „Querschnittsstudie", „Fall-Kontroll-Studie", „Längsschnittstudie" und „Kohortenstudie" voneinander ab.

2. Welchen Unterschied gibt es zwischen Einfach-, Doppelt- und Dreifach-verblindung?

3. Nennen Sie Vor- und Nachteile für Labor- und Felduntersuchungen.

4. Welche 10 Leitfragen zur Beurteilung der Validität stellt Mangold (2013) vor?

Aufgabe zur Berufspraxis

5. Nennen Sie mögliche Interventions- und Kontrollgruppen im sprachtherapeutischen Setting.

Literatur

Creswell, J. W. (2007). *Qualitative inquiry & research design choosing among five approaches* (2. ed.). Thousand Oaks, Calif. [u.a.]: Sage Publ.

Döring, N., Bortz, J., & Pöschl, S. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften* (5. vollständig überarbeitete, aktualisierte und erweiterte Auflage ed.). Berlin Heidelberg: Springer.

Flick, U. (2000). Design und Prozess qualitativer Forschung. In U. Flick, E. v. Kardorff, & I. Steinke (Eds.), *Qualitative Forschung. Ein Handbuch* (pp. 252-265). Reinbek bei Hamburg: Rowohlt.

Klug, S. J., Bender, R., Blettner, M., & Lange, S. (2004). Wichtige epidemiologische Studientypen. [Common epidemiologic study types]. *Dtsch med Wochenschr, 129*(S 3), T7-T10. doi: 10.1055/s-2004-836076

Köhler, T. (2012). Inferenzstatistischer Nachweis intraindividueller Unterschiede im Rahmen von Einzelfallanalysen. *Empirische Sonderpädagogik, 4*(3/4), 265-274.

Mangold, S. (2013). *Evidenzbasiertes Arbeiten in der Physio- und Ergotherapie Reflektiert - systematisch - wissenschaftlich fundiert* (2., aktualisierte Aufl. ed.). Berlin [u.a.]: Springer.

Mayring, P. (2010). Design. In G. Mey & K. Mruck (Eds.), *Handbuch Qualitative Forschung in der Psychologie* (pp. 225-237). Wiesbaden: VS Verlag für Sozialwissenschaften.

Raithel, J. (2008). *Quantitative Forschung ein Praxiskurs* (2., durchgesehene Auflage ed.). Wiesbaden: VS Verlag für Sozialwissenschaften.

Schäfer, S. (2007). Kritische Bewertung von Studien zur Ätiologie. In R. Kunz, G. Ollenschläger, H. Raspe, G. Jonitz, & N. Donner-Banzhoff (Eds.), *Lehrbuch Evidenzbasierte Medizin in Klinik und Praxis* (pp. 101-114). Köln: Deutscher Ärzteverlag.

Wagner, G. G., Göbel, J., Krause, P., Pischner, R., & Sieber, I. (2008). Das Sozio-oekonomische Panel (SOEP): Multidisziplinäres Haushaltspanel und Kohortenstudie für Deutschland – Eine Einführung (für neue Datennutzer) mit einem Ausblick (für erfahrene Anwender). *AStA Wirtschafts- und Sozialstatistisches Archiv, 2*(4), 301-328. doi: 10.1007/s11943-008-0050-y

6 Methoden der Datenerhebung

Lernergebnisse:

- Methoden der Datenerhebung, ihr Einsatzfeld und Probleme bei der Interpretation von Ergebnissen unterschiedlicher Erhebungsarten können erläutert und differenziert werden. Studiendesigns mit Datenerhebung durch Befragen, Beobachten und Experimentieren können entwickelt und in Bezug zur Aussagekraft und Gütediskussion gebracht werden (6.1, 5.3).

- Die Bedeutung des Messens in empirischen Studien kann erläutert werden. Probleme im Zusammenhang mit dem Messen von Merkmalen können beschrieben und Lösungsmöglichkeiten erörtert werden (6.2).

- Messniveaus können beschrieben und hinsichtlich ihrer Bedeutung für die statistische Analyse erläutert werden (6.2.1, 9.3).

- Der Prozess des Operationalisierens von Merkmalen kann anhand eines Forschungsgegenstands beschrieben und Probleme identifiziert werden (6.2).

- Gütekriterien von Messinstrumenten: Objektivität, Reliabilität und Validität können definiert und ihre Bedeutung Zusammenhang diskutiert werden (6.3).

Die Ziele empirischer Forschung: Erklären, Vorhersagen und Verändern (s. 1.1) können auf verschiedenen Wegen erreicht werden. Im Anschluss an die Identifikation der theoretischen Basis und der darauf begründeten Wahl des Forschungsparadigmas – also qualitativ oder quantitativ – steht die Klärung der Art der Datenerhebung und der Rekrutierung der Stichprobe. Aussagen empirischer Forschung, im qualitativen und quantitativen Forschungsansatz, gründen sich auf systematisch erhobenen Daten aus Stichproben. Qualitativer und quantitativer Forschungsansatz unterscheiden sich dabei zum einen in ihrer wissenschaftstheoretischen Fundierung (s. 2.2, 2.3), zum anderen wie systematisch und standardisiert Daten erhoben werden. Die anschließenden Ausführungen beziehen sich dabei zunächst auf Datenerhebungen unter einem quantitativen Paradigma.

Datenerhebung kann auf verschiedenen Wegen erfolgen. Unterschieden wird die Datenerhebung auf der Basis von Beobachtungen, Befragungen und Experimenten (s. 6.1). Ein grundsätzliches Problem in der quantitativen sozial- und gesundheitswissenschaftlichen Forschung beruht auf der Notwendigkeit, qualitative Merkmale und Merkmalsausprägungen in Zahlen übersetzen zu müssen. Diese Übersetzung erfolgt im Prozess der Operationalisierung (6.2), die in der Erarbeitung und Festlegung eines Messverfahrens mündet. Im Rahmen der Messung erfolgt die Übertragung des qualitativen Merkmals (empirisches relativ) in eine Zahl – einen Messwert (numerisches Relativ). Dieser Prozess ist in sozial- und gesundheitswissenschaftlichen Studien deutlich komplizierter und fehleranfälliger als bei Messungen in den Naturwissenschaften (6.2). Grundsätzlich kann kritisch hinterfragt werden, ob ein qualitatives Merkmal in Zahlen übersetzt werden kann und ob eine Interpretation des qualitativen Merkmals dann noch möglich ist. Auch die Adressaten von Forschung, die Studienteilnehmer, können Fragen in Messinstrumenten anders verstehen als vom Forscher intendiert. Die Entwicklung valider und zuverlässiger Messinstrumente gelingt nicht „nebenher“. Für Anwendungsforschung in den Gesundheitsberufen sollten auf bereits vorhandene Messinstrumente zurückgegriffen werden, die über Onlinedatenbanken merkmalsspezifisch recherchiert werden können. Gute Messinstrumente bilden das interessierende Merkmal ab (Validität), geben bei gleicher Merkmalsausprägung gleiche Messwerte aus (Reliabilität) und messen unabhängig vom Untersucher (Objektivität).

Die erhobenen Daten liegen in unterschiedlicher Messgüte vor. Je besser Messwerte bestimmten mathematischen Eigenschaften folgen, desto eher dürfen aussagekräftige statistische Verfahren angewendet werden. Unterschieden wird danach, ob Messwerte auf unterschiedliche Merkmalsausprägungen hinweisen (Nominalskalenniveau), ob Messwerte in eine Rangfolge (besser-schlechter, stärker-schwächer) nahelegen (Ordinalskalenniveau) und ob der Abstand zwischen zwei Merkmalsausprägungen auf der gesamten Skalenbreite gleich ist.

6.1 Befragen, Beobachten und Experimentieren

6.1.1 Befragen

Befragungen sind Kommunikationsprozesse zwischen Menschen, bei denen verbale oder schriftliche Stimuli seitens des Befragers, Reaktionen, d.h. Antworten, auf Seiten der Befragten auslösen (Atteslander & Cromm, 2010). Dabei können Befragungen schriftlich oder mündlich erfolgen, sie können stark oder wenig strukturiert und standardisiert sein, sich an einzelne Menschen oder an Gruppen richten sowie einmal oder mehrfach erfolgen (Quer- und Längsschnittstudie, s. 5.2) (Atteslander & Cromm, 2010). Atteslander und Cromm (2010) systematisieren in ihrem Lehrbuch Befragungen. Die folgenden Ausführungen gründen sich auf die dort veröffentlichte Einteilung.

Mündliche Befragungen erfolgen in direkter oder telefonischer Kommunikation zwischen Befrager und Befragten. Telefonische Befragungen werden häufig im Rahmen von Marktanalysen bei Konsumentenbefragungen eingesetzt.

Fragebögen werden im Rahmen *schriftlicher Befragungen* eingesetzt. Mündliche und schriftliche Befragungen könne in stark und schwach strukturiertem Befragungsrahmen durchgeführt werden.

In *schwach strukturierten* Befragungen sind Befrager (und Befragte) nicht eng an ein Befragungsschema gebunden. Die Formulierungen der Fragen kann dabei ebenso variiert werden, wie die Reihenfolge der Fragen. Ferner sind Ergänzungen möglich, mit denen Befrager das Gespräch steuern können.

Schwach strukturierte Befragungen werden im Rahmen der Erkundung des Forschungsfelds oder zur Präzisierung des Forschungsproblems eingesetzt, sie finden oft unter einem qualitativen Forschungsparadigma Anwendung (s. 2.2, 2.3). Im Unterschied dazu sind Befrager bei stark strukturierten Befragungen an ein vorher ausgearbeitetes Befragungsschema gebunden. Sowohl die Formulierung und die Anzahl der Fragen, das Antwortformat, als auch z.T. die Reihenfolge von Fragen sind festgelegt. Zuverlässige Ergebnisse liefern strukturierte Befragungen, wenn Fragen in der geplanten Form allen Studienteilnehmern gestellt werden. Als Erhebungsinstrumente werden strukturierte Interviewleitfäden oder Fragebögen eingesetzt. Stark strukturierte Befragungen werden im Rahmen theorieüberprüfender Forschung eingesetzt, die einem quantitativen Forschungsparadigma folgen (s. 2.2, 2.3).

Standardisierte Befragungen geben Antworten vor. Es werden hier geschlossene Fragen genutzt. Der Studienteilnehmer kann aus verschiedenen Antwortalternativen die für ihn zutreffende auswählen. Mit standardisierten Befragungen werden vergleichbare Ergebnisse in der Stichprobe erzielt. Erwartet wird außerdem von unterschiedlichen Ergebnissen auf unterschiedliche Merkmalsausprägungen bei den Studienteilnehmern schließen zu können. Allerdings ist trifft diese Annahme nicht

grundsätzlich zu und hängt davon ab, wie Studienteilnehmer Antwortvorgaben interpretieren (s. Urteilsfehler, 6.2).

Wenig (oder nicht-) standardisierte, offene Fragen geben keine Antwortkategorien vor. Die Kategorienbildung erfolgt nach der Befragung. In quantitativen Studien werden standardisiert (geschlossene) und nicht-standardisierte Fragen häufig kombiniert, wobei hier offene Fragen eher sparsam eingesetzt werden.

Bei der Mehrzahl quantitativer (Fragebogen-) Studien handelt es sich um *Individualbefragungen*, jeder Studienteilnehmer wird dabei einzeln befragt.

Die folgenden Aussagen beinhalten mögliche Tätigkeitsmerkmale und Rahmenbedingungen der Arbeit in verschiedenen Behörden/ Organisationen in der polizeilichen und nicht-polizeilichen Gefahrenabwehr. Sie beziehen sich auf Ihre hauptamtliche Tätigkeit **in der Behörde/Organisation der Gefahrenabwehr**, in der Sie den Fragebogen ausgehändigt bekamen. Bitte bewerten Sie die folgenden Aussagen danach, wie häufig Sie sie im letzten Jahr erlebt haben. **Bitte kreuzen Sie nur die Antwort an, die für Sie am ehesten zutrifft.** Beantworten Sie bitte jede der folgenden Fragen.		nie	einige Male im Jahr oder seltener	1x im Monat oder seltener	mehrmals im Monat	1x in der Woche	einige Male in der Woche	täglich
		0	**1**	**2**	**3**	**4**	**5**	**6**
79.	Ich war an einem Einsatz in einer Großschadenslage tätig.	☐	☐	☐	☐	☐	☐	☐
80.	Während eines Einsatzes war die Einsatzstelle schwer zu erreichen.	☐	☐	☐	☐	☐	☐	☐
81.	Während einer Einsatzfahrt gab es Behinderungen durch andere Verkehrsteilnehmer.	☐	☐	☐	☐	☐	☐	☐
82.	Während eines Einsatzes wurde ich verbal provoziert.	☐	☐	☐	☐	☐	☐	☐
83.	Während eines Einsatzes wurde ich körperlich angegriffen.	☐	☐	☐	☐	☐	☐	☐
84.	Während eines Einsatzes, an dem ich beteiligt war, wurde ein(e) Kolleg(in)e/ Kamerad(in) verletzt.	☐	☐	☐	☐	☐	☐	☐

Abbildung 12: Beispiel für stark strukturierte und standardisierte Fragen aus einer Studie zu Belastungen im Rettungsdienst (Beerlage, Arndt, Hering & Springer, 2007)

Gruppenbefragungen und Gruppendiskussionen werden in qualitativen Forschungsdesigns angewendet. Sie stellen besondere Herausforderungen an die Befrager, weil häufig eine oder mehrere Teilnehmer die Diskussion dominieren können und relevante Informationen zurückhaltender Teilnehmer nicht erhoben werden. Abbildung 12 und Abbildung 13 zeigen Beispiele für Erhebungsschemen aus quantitativen und qualitativen Erhebungen.

Teil 1: Allgemeine Auslösung von PSNV und Anforderung an den Ort?	
Hintergrund	**Beispielfragen**
1. **Alarmierung** wann wie durch wen? (Konkrete Schilderung)	• Können Sie sich noch erinnern, was Sie gerade taten, als sie von dem Ereignis erfuhren? Wo waren Sie? • Was ist in der Folge passiert, was haben sie getan? • Was hat sie schließlich veranlasst, sich an den Ort des Geschehens zu bewegen?
2. **Indikatoren** für Interventionsnotwendigkeit: Stressorenprofil, Belastungsbegriff, abgeleitete Indikation. Warum war das ein **Fall von** "Notwendigkeit für PSNV"?	• Wer hatte denn die Idee, dass dort eine Art von [psychologischer/ psychosozialer/ seelsorgerlicher] Arbeit geschehen musste? • Welche Hinweise gab es, dass hier ein Bedarf vorlag? • Können Sie sich noch daran erinnern, in welchen Worten Ihnen das gesagt wurde, oder was Sie selber dachten?
3. Wer **definierte, dass irgendie-**	• Wer gab schließlich den Ausschlag dafür, dass

Abbildung 13: Beispiel für schwach strukturierte und wenig standardisierte Fragen aus einer Studie zur Koordination psychosozialer Notfallversorgung in Großschadenslagen (unveröffentlichter Leitfaden aus der Studie von Beerlage, Hering & Nörenberg, 2006)

6.1.2 Beobachten

Beobachtungen dienen ebenfalls der Datenerhebung in der empirischen Forschung. Unter wissenschaftlichen Beobachtungen wird ein systematischer und nicht kommunikativer Prozess verstanden, bei dem Merkmale mit einem nachvollziehbar entwickelten Beobachtungsplan erfasst werden (Atteslander & Cromm, 2010, Hussy et al., 2008). Beobachtungen werden in der empirischen Forschung eingesetzt, wenn Daten aus der Selbstauskunft der Studienteilnehmer verfälscht sein können (Beller, 2008). Der *Beobachtungsplan* sollte offenlegen, welche Merkmale beobachtet werden sollen, welche Ausprägungen relevant und nicht relevant sind, wie groß der Interpretationsspielraum des Beobachters sein soll, Dauer, Umfang und Ort der Beobachtung und ob Beobachtungen offen oder verdeckt erfolgen sollen (Hussy et al., 2008). Beobachtungen werden in teilnehmende und nicht teilnehmende, Selbst- und Fremdbeobachtung sowie offener und verdeckter Beobachtung unterschieden.

In *teilnehmenden Beobachtungen* ist der Beobachter am Geschehen beteiligt. Das Protokollieren von Sitzungen ist beispielsweise eine teilnehmende Beobachtung. Beobachter, die zugleich dem Geschehen folgen müssen und beteiligt sind, können überfordert werden oder das Geschehen nicht mit der notwendigen Aufmerksamkeit verfolgen.

Im Unterschied dazu ermöglichen *nicht-teilnehmende Beobachtungen* eine stärkere Konzentration auf den Beobachtungsgegenstand (Beller, 2008). Sowohl teilnehmende als auch nicht-teilnehmende Beobachtungen werden in qualitativen und quantitativen Forschungsdesigns angewendet. Im Rahmen quantitativer Untersu-

chungen ist der Beobachtungsplan strukturiert und standardisiert. Wissenschaftliche Beobachtungen werden ferner in Fremd- und Eigenbeobachtung unterschieden.

Fremdbeobachtungen zielen auf die Dokumentation und die Analyse fremden Verhaltens ab. Eigenbeobachtungen dienen der Dokumentation und Analyse eigenen Erlebens und Verhaltens. Um nachvollziehbare und gültige Schlussfolgerungen aus Eigenbeobachtungen ziehen zu können bedarf es hinreichender Erfahrungen. Als problematisch wird im Rahmen von Eigenbeobachtungen gewertet, dass sie Verhalten ändern. Intensives Nachdenken und Reflektieren kann zu veränderten Schlussfolgerungen führen, mit Einfluss auf die Zuverlässigkeit von Beobachtungsergebnissen.

In der letzten Einteilung von Beobachtungen wird gefragt, ob die beobachteten Studienteilnehmer über den Zeitpunkt und den Zeitraum der Beobachtung aufgeklärt sind – dies sind *offene Beobachtungen* – oder ob sie zwar grundsätzlich als Studienteilnehmer über die die Form der Datenerhebung aufgeklärt sind, jedoch nicht über den Zeitpunkt und den Zeitraum der Beobachtung – dabei handelt es sich um *verdeckte Beobachtungen* (Beller, 2008).

Beobachtungsstudien bergen (wie alle Datenerhebungsformen) Potenzial für Verzerrung der Ergebnisse, die z.T. bereits angesprochen wurden. Als problematisch wird dabei gesehen (Beller, 2008, S. 35):

- die potenzielle Überforderung der Beobachter, insbesondere bei teilnehmenden Beobachtungen
- die eventuell unzureichende Trennung von dem beobachteten Verhalten und der Interpretation des Verhaltens
- eine geringe Vertrautheit der Beobachter mit den Besonderheiten der Probandengruppe, die auf unterschiedliche Norm- und Wertvorstellungen zwischen beiden gegründet sein können
- unzureichende oder unvollständige Beobachtungspläne, in denen untersuchungsrelevante Verhaltensweisen und Situationen nicht aufgenommen werden
- strukturelle und inhaltliche Abweichungen zwischen Beobachtungsprotokollen verschiedener Beobachter
- alle Studienteilnehmer müssen über die Erhebungsart Beobachtung aufgeklärt sein, dies und der Einsatz offener Beobachtungen sowie von Eigenbeobachtungen können Einfluss haben auf das Verhalten.

6.1.3 Experimentieren

Daten werden ferner im Rahmen von **Experimenten** gewonnen und für wissenschaftliche Zwecke genutzt. In Experimenten werden Einflussfaktoren gezielt variiert und die Auswirkungen dieser Variationen innerhalb zwischen einer Interventions- und einer Kontrollgruppe verglichen. Unterschieden wird in experimentelle, quasi-experimentelle und nicht-experimentelle Designs, die unter kontrollierten Laborbedingungen oder im Feld erfolgen können (s. dazu im Detail 5.3, 0, Döring et al., 2016).

Studien basieren zusammenfassend auf der Erhebung von Daten, die im Rahmen von Befragungen, Beobachtungen und mit Experimenten gewonnen werden können. Befragungen sind ein kommunikativer Prozess zwischen Befrager und Befragtem, sie werden mündlich oder schriftlich durchgeführt, sind stark oder wenig strukturiert und haben offene oder geschlossene Antwortformate. Sie können in Gruppen oder bei Einzelpersonen durchgeführt werden. Befragungen werden in qualitativen und quantitativen Forschungsdesigns genutzt, wobei der Grad der Strukturierung und Standardisierung in quantitativen Studien höher ist als in qualitativen. Beobachtungen erfolgen auf nicht kommunikativer Basis. Sie werden entweder in Form teilnehmender oder nicht-teilnehmender Beobachtungen, Eigen- oder Fremdbeobachtungen, offen oder verdeckt durchgeführt. Damit aus Beobachtungen zuverlässige Daten resultieren müssen Beobachtungsgegenstand, seine potenziellen Ausprägungen, Ort, Zeitpunkt und Dauer der Beobachtung in einem Beobachtungsplan beschrieben werden. In Experimenten wird über die Variation der Ausprägung von Einflussfaktoren ein Effekt in der Interventionsgruppe, verglichen mit der Kontrollgruppe, erwartet. Die Aussagekraft von experimentellen Studien hängt davon ab, ob die Zuordnung von Studienteilnehmern zu Interventions- und Kontrollgruppe auf Zufall beruht und wie gut Störfaktoren kontrolliert werden (s. dazu im Detail 5.3, 0).

6.2 Merkmale messen und operationalisieren

6.2.1 Messen, Messtheorie, Maßsystem und Messprobleme

Empirische Studien erheben systematisch Daten von Studienteilnehmern, die aus einer bestimmten Population (möglichst auf Zufallsbasis) ausgewählt wurden. Daten ergeben sich aus allen in der Untersuchung durchgeführten Messungen, jede Messung erhebt die Ausprägung *einer* manifesten Variablen. In eine Variable gehen alle Ausprägungen eines Merkmals ein, sie muss jedoch mindestens zwei Ausprägungen umfassen (Sedlmeier & Renkewitz, 2013). Beispielsweise enthält die Variable Geschlecht mindestens die beiden Ausprägungen „männlich“ und „weiblich“, die Variable Schulabschluss z.B. die in Abbildung 6 skizzierten Ausprägungen. Im Rahmen

der Operationalisierung, also der Festlegung auf das Messverfahren zur Abbildung der Merkmale, wird vorab auf Basis der Theorie und der Definition des interessierenden Merkmals die Verknüpfung manifester Variablen zu latenten Variablen beschrieben (4.5). Im Rahmen von Messungen wird die Ausprägung einer bestimmten Variablen (eines Merkmals), bei konkreten Personen zu einem definierten Zeitpunkt abgebildet. Das Messergebnis ist in der quantitativen Forschung eine Zahl.

Messtheorie. Messen umschreibt das Verfahren, regelhaft eine Merkmalsausprägung in einer Zahl abzubilden (Sedlmeier & Renkewitz, 2013, Beller, 2008). In der Messtheorie werden Funktionen für die Abbildung eines *empirischen Relativs* in ein *numerisches Relativ* differenziert. Das *empirische Relativ* umfasst einzelne Objekte in einer bestimmten Relation zueinander (gleich, verschieden, größer, kleiner …). Objekte des empirischen Relativs sollten unterschiedlich sein und mindestens zwei Merkmalsausprägungen umfassen (z.B. Männer und Frauen). Dieselben Merkmalsgruppen (z.B. Männer) haben innerhalb des empirischen Relativs dieselben Merkmalsausprägungen. Verschiedene Merkmalsgruppen (Männer und Frauen) innerhalb des empirischen Relativs unterscheiden sich dagegen hinsichtlich ihrer Merkmalsausprägung (*Äquivalenzrelation*). Können Merkmalsausprägungen neben der Ebene Gleich- und Verschiedenheit auch auf verschiedenen Rängen abgebildet werden (z.B. größer oder kleiner) wird von der *Ordnungsrelation* gesprochen. Steyer und Eid (2001, zitiert in Beller, 2008, S. 26) sprechen drei Kernfragen der Messtheorie an, nämlich:

- Können empirische Beziehungen numerisch abgebildet werden (Repräsentationsproblem)?

- Kann trotz Änderung der Skala eine eindeutige Skala beibehalten werden (Eindeutigkeitsproblem?

- Haben numerische Abbildungen empirischer Beziehungen auch eine empirische Bedeutsamkeit (Bedeutsamkeitsproblem)?

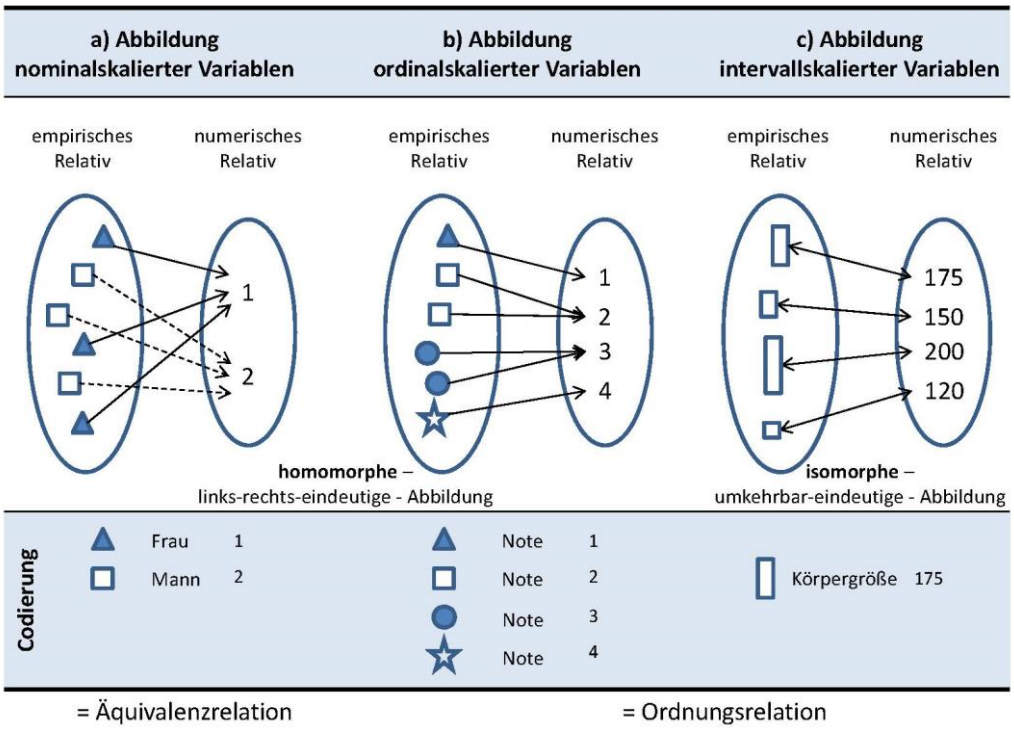


Abbildung 14: Messtheoretische Grundlagen (eigene Darstellung)

Zunächst einmal werden Möglichkeiten der Abbildung eines empirischen Relativs in ein numerisches Relativ beispielhaft erläutert. In Abbildung 14 sind im Bereich a) zwei Objekttypen (Geschlecht männlich und weiblich) im empirischen Relativ unterscheidbar. Im numerischen Relativ erfolgt die Abbildung der Objekttypen mit zwei Zahlen (1, 2). Jedem Objekt im empirischen Relativ wurde eine Zahl zugeordnet, jeweils drei Objekte werden der „1“, und drei weitere der „2“ zugeordnet. Die Abbildung ist links-rechts-eindeutig, also homomorph. Mit diesem Messergebnis werden Gemeinsamkeiten und Unterschiede abgebildet, es handelt es sich um eine Äquivalenzrelation.

Im Bereich b) sind im empirischen Relativ ebenfalls sechs Objekte enthalten, die vier Objekttypen zugeordnet werden können. Die Abbildung erfolgt im numerischen Relativ mit vier Zahlen (Schulnoten) von 1 bis 4. Es wurde ebenfalls jedem Objekt eine Zahl zugeordnet, teilweise „teilen“ sich zwei Objekte eine Zahl (2 und 3). Auch hier ist die Abbildung homomorph links-rechts-eindeutig. Es werden Gemeinsamkeiten und Unterschiede abgebildet. Zudem kann z.B. über das Merkmal „Schulnoten“ eine Rangfolge der Objekte erstellt werden, es handelt sich dabei um eine Ordnungsrelation.

Der Bereich c) enthält im empirischen Relativ vier Objekte, die vier Objekttypen zugeordnet werden können. Die Abbildung erfolgt im numerischen Relativ mit vier Zahlen (Körpergrößen): 120, 150, 175, 200 cm. Es wurde jedem Objekt genau eine Zahl zugeordnet. Keine einzige Zahl ist doppelt vergeben, so dass über die vorlie-

gende *isomorphe Abbildung* auch eine umgekehrt eindeutige Abbildung der Merkmale erfolgen kann. Jeder Zahl kann also ein konkreter Objekttyp zugeordnet werden, was in einer homomorphen, links-rechts-eindeutigen Abbildung nicht möglich ist. In der Messung werden Gemeinsamkeiten, Unterschiede und die Rangfolge abgebildet, es handelt sich ebenfalls um eine Ordnungsrelation.

Ein *numerisches Relativ* setzt sich aus einer Menge Zahlen zusammen, die in einer Relation zueinander stehen, also gleich und verschieden, größer und kleiner sind. Das Ziel von Messungen ist es, zumindest eine *homomorphe*– eine eindeutige – *Abbildung* zu erzielen. Dies bedeutet, von jedem empirischen Relativ (also von jeder Person, die untersucht wird), geht ein Pfeil aus, jeder Person wird eine Zahl zugewiesen. Menschen mit derselben Merkmalsausprägung erhalten dieselbe Zahl. Die Abbildung des empirischen im numerischen Relativ erfolgt hier homomorph, also „links-rechts-eindeutig". Es ist nicht zwingend notwendig, ausgehend von einem numerischen Relativ, also einer Zahl, umgekehrt eine konkrete Person zu identifizieren, wie dies bei einer *isomorphen* – also umkehrbar eindeutigen – *Abbildung* – der Fall wäre.

Auch wenn die Überführung qualitativer Merkmale, z.B. die Studierende M. ist klug, neuen Forschungsthemen gegenüber aufgeschlossen und kann Problemstellungen gut erfassen, in Zahlen im Feld der Sozial- und Gesundheitswissenschaften kritisiert wird, bieten Messungen die Möglichkeit, Personen oder Personengruppen zu vergleichen und Schlussfolgerungen für weitere Forschungsfragen oder für ganz konkrete Maßnahmen zu ziehen. Eine Schlussfolgerung könnte sein, besonders engagierte und begabte Studierende in Begabtenförderprogrammen gezielt auf eine wissenschaftliche Karriere vorzubereiten.

Maßsystem. Messungen liefern Daten unterschiedlicher Qualität. Sie können auf Unterschiede (Kategorien, Nominalskala), zusätzlich auf eine konkrete Rangfolge der Messwerte (Rangdaten, Ordinalskala) und ferner auf die Bestimmung der Relation von Unterschieden zwischen Personen (Messwerte, Intervallskala) fokussieren. Am Beispiel der Studierenden könnte zunächst die Information des höchsten Schulabschlusses gemessen werden. Fachabiturienten erhielten in der Messung beispielsweise eine „1", Abiturienten eine „2". Da der Zahlenwert in diesem Fall (noch) keine Rangfolge unterstellt, ist es gleichgültig, mit welcher Zahl die Codierung des jeweiligen Schulabschlusses erfolgt, das Abitur könnte genauso gut mit einer 34, das Fachabitur mit einer 67 in die Studie eingehen. Alle Abiturienten müssen dieselbe Zahl erhalten, wie auch alle Fachabiturienten ihrerseits mit derselben Zahl kodiert werden müssen, während sich die Zahl zwischen „Abitur" und „Fachabitur" unterscheiden muss. Die aus diesem Beispiel resultierenden Messwerte sind nominalskaliert, sie bilden Kategorien ab und lassen auf Gemeinsamkeiten und Unterschiede von Studienteilnehmern schließen (Äquivalenzrelation) (Abbildung 15).

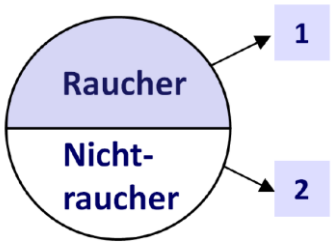


Abbildung 15: Nominalskalierte Daten. Merkmal (Variable) = empirisches Relativ: Rauchstatus, Merkmalsausprägung: Raucher, Nichtraucher, numerisches relativ: 1, 2 (eigene Darstellung)

Bei der Erfassung der „Klugheit“ interessiert nicht nur, ob Menschen gleich oder verschieden klug sind, sondern auch wer mehr oder weniger klug ist, also eine Rangfolge. Bei der Messung sollte eine weniger kluge Person einen kleineren Messwert erreichen als eine klügere Person. Auch hier ist die konkrete Zahl wie bei den kategorialen Variablen nur insofern von Bedeutung als dass kluge Menschen höhere Zahlen erhalten müssen als weniger kluge Menschen. Ob allerdings der Abstand zwischen den Messwerten zweier Personen genauso groß ist, wie ein zahlenwertig gleicher Abstand zwischen zwei anderen unterschiedlich klugen Personen, kann aus dem *Ordinalskalenniveau* nicht geschlossen werden (Abbildung 16).

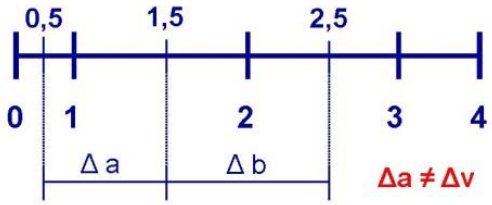


Abbildung 16: Ordinalskalierte Daten. Merkmal (Variable) = empirisches Relativ: Klugheit, Merkmalsausprägung: sehr wenig klug, wenig klug, klug, ziemlich klug, sehr klug, numerisches relativ: 1, 2, 3, 4. Die qualitativen Abstände zwischen gleichen Abständen sind unterschiedlich groß oder unbekannt (eigene Darstellung)

Merkmale, wie die Uhrzeit oder die Temperatur auf der Celsiusskala, können in Messungen vergleichsweise problemlos mit einer über Gleichheit und Verschiedenheit sowie die Rangfolge abbildendem Maß hinaus erfasst werden. Ein bestimmter zahlenwertiger Unterschied zwischen zwei Menschen ist dabei genauso groß wie derselbe Unterschied zwischen zwei anderen Menschen. Der Abstand zwischen 1,71 und 1,76 ist (auch qualitativ) genauso groß, wie der zwischen 1,85 und 1,90. Auf der gesamten möglichen Skalenbreite – also dem gesamten Wertebereich – weist ein gleicher Zahlenunterschied demnach auf den gleichen qualitativen Unterschied hin. Diese *intervallskalierten* Daten können im Rahmen statistischer Analysen genauer untersucht werden als nominal- und ordinalskalierte Daten (Abbildung 17).

Intervallskalierte Daten haben einen willkürlich gesetzten Nullpunkt. Bei der Uhrzeit wurde der Sonnenstand bei der Festlegung berücksichtigt, bei der Temperaturskala nach Celsius der Gefrierpunkt von Wasser (es hätte auch auf den Gefrier- und Siedepunkt eines jeden anderen Stoffs Bezug genommen werden können). Auf *Verhältnisskalenniveau* werden Merkmale mit einem natürlichen Nullpunkt gemessen. Beispiele sind Körpergröße und Körpermasse sowie die Temperaturskala nach Kelvin. Der Nullpunkt dieser Merkmale ist merkmalsimmanent, er bezieht sich nicht auf ein anderes Merkmal. Dies ist bei Körpergröße und -masse einfach nachvollziehbar, der Nullpunkt der Kelvinskala kennzeichnet den (derzeit bekannten) Temperaturnullpunkt, den Stillstand der Atome aller Stoffe.

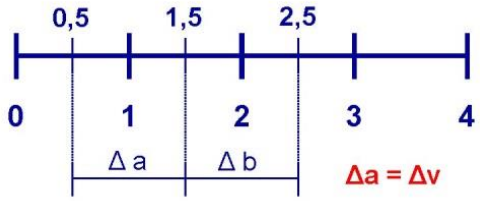


Abbildung 17: Intervallskalierte Daten. Merkmal (Variable) = empirisches Relativ: Uhrzeit, Merkmalsausprägung: 0-24 Uhr, numerisches relativ: 1, 2, (…) 24. Die qualitativen Abstände zwischen gleichen Abständen sind gleich groß (eigene Darstellung)

Messprobleme. In Messungen soll das empirische Relativ in einem numerischen Relativ abgebildet werden. Eine systematische Veränderung einer Skala des numerischen Relativs soll nicht zulasten der Abbildung empirischer Beziehungen gehen und eine numerische Aussage soll empirische Bedeutung haben. Aus der Beantwortung dieser Fragen für die Messung können sich ein Repräsentations-, Eindeutigkeits- und Bedeutsamkeitsproblem ergeben.

Beim *Repräsentationsproblem* stellt sich die Frage, ob ein empirisches Merkmal überhaupt in Zahlen übertragen werden darf und ob die numerische Abbildung des Merkmals alle Facetten des empirischen Merkmals aufnimmt. Im Beispiel könnte Klugheit auf unterschiedlichen Wegen erfasst werden: mit zwei Objekttypen (unklug und klug), auf einer Skala von 1 bis 7 oder unter Nutzung von Verhältnissen, wie beim Intelligenzquotienten. Eine Messung sollte kluge von weniger klugen Menschen unterscheiden können. Im numerischen Relativ müssen ferner auch extrem unkluge und sehr kluge Menschen Abbildung finden. Im Repräsentationsproblem interessiert demnach, ob ein Merkmal in seiner möglichen Ausprägung angemessen in Zahlen übertragen werden kann (Beller, 2008; Sedlmeier & Renkewitz, 2013).

Das *Eindeutigkeitsproblem* bezieht sich auf die Möglichkeit der (mathematischen) Transformation numerischer Relative, bei der die Informationen einer homomorphen Abbildung erhalten bleiben. Beispielsweise können Zensuren auf einer Skala

von „1“ bis „6“ mit jeder beliebigen Zahl addiert werden, ohne dass die in der Relation der Zensuren kodierten Informationen verloren gehen: niedrigere Zahlen umfassen nach wie vor bessere Leistungen als höhere Zahlen (Beller, 2008, Sedlmeier & Renkewitz, 2013).

Der Frage, welche mathematischen Operationen sinnvoll sind, wird im *Bedeutungsproblem* nachgegangen. Im Zensurenbeispiel wäre eine Addition der Werte einer Person, die eine „1“ und einer Person, die eine „4“ erreicht hat in der Summe „5“ ergeben. Bedeutet dies, Einser und Vierer sind genau so wenig leistungsfähig, wie der Fünfer? Die Antwort darauf lautet selbstverständlich nein, diese Berechnung hat keine empirische Bedeutung. Allerdings könnten die Zensuren von männlichen und weiblichen Studierenden verglichen werden, indem der Medianwert (s. 8.4.1) beider Gruppen berechnet wird. Dabei wird die empirische Bedeutung der in Berechnungen genutzten Zensuren unter allen zulässigen Transformationen (s. Eindeutigkeitsproblem) erhalten. Im Mittel lässt sich abschätzen, ob männliche oder weibliche Studierende bessere Zensuren erreichten (Beller, 2008, Sedlmeier & Renkewitz, 2013).

Beim Messen geht es zusammenfassend darum, Merkmale des empirischen Relativs in Zahlen – also das numerische Relativ – zu übertragen. Die quantitative Abbildung der qualitativ beschreibbaren Merkmalsausprägungen in Zahlen sollte auf Gemeinsamkeiten und Unterschiede bei den Merkmalsausprägungen in einer Stichprobe hinweisen – Männer sollten bei der Variable Geschlecht also dieselben Messwerte erhalten. Alle Frauen erhalten ebenfalls dieselben Messwerte, allerdings andere als die Männer. Wenn ausgehend von Werten des numerischen Relativs auf Personen geschlossen werden kann, die bestimmte Eigenschaften aufweisen, wird von einer homomorphen Abbildung gesprochen. Beim Messniveau wird in erster Linie nach dem Informationsgehalt des numerischen Relativs gefragt. Nominalskalierte Daten weisen auf Gemeinsamkeiten und Unterscheide zwischen Objekten hin. Bei ordinalskalierten Daten wird diese Information um eine Rangfolge ergänzt (weniger, mehr, besser, schlechter usw.). Intervallskalierte Daten weisen bei denselben zahlenwertigen Unterschieden auch dieselben qualitativen Unterschiede im gesamten möglichen Wertebereich auf. Vom Messniveau ist ferner abhängig, welche statistischen Analysen und welche Signifikanztests mit den Daten durchgeführt werden dürfen. Messprobleme ergeben sich aus der Übertragung qualitativer Merkmale in Zahlen (Eindeutigkeitsproblem), hinsichtlich der informationserhaltenden Transformation (Eindeutigkeitsproblem) und den Möglichkeiten, empirisch bedeutende Berechnungen mit den Zahlen durchführen zu können (Bedeutungsproblem).

6.2.2 Operationalisieren

Operationalisieren bedeutet, theoretischen Begriffen manifeste, also einer Messung zugängliche, Variablen zuzuordnen (4.5). Basis der Operationalisierung sind der *theoretische Rahmen* – also wie ein Begriff in der existierenden Literatur verwendet wird und die *Begriffsdefinition* – also mit welchen Kriterien ein Begriff beschrieben wird. Die beschreibenden Kriterien bilden den Rahmen für die Formulierung von Items, die in Befragungen eingesetzt werden. Dabei sind mehrere Varianten von Operationalisierung denkbar. Die Art und die Ziele der Untersuchung bestimmen, welche Operationalisierungsvariante der Vorzug gegeben werden sollte (Steyer & Eid, 2001, zit. in Beller, 2008, S. 29f):

1. Vergleichbarkeit mit anderen Studien: sollen eigene Studienergebnisse mit denen anderer Studien vergleichbar sein, ist die Nutzung desselben Messverfahrens nötig. Ein direkter Vergleich ist bereits dann nicht mehr aussagekräftig, wenn einige Fragen des ursprünglichen Messverfahrens weggelassen werden. Daher sollte also auf bereits vorhandene Messverfahren und Fragebögen zurückgegriffen werden.

2. Bei der unabhängigen Variablen sollen für Interventionsstudien hinreichend viele Variationsmöglichkeiten bestehen, weil mit der Modifikation der unabhängigen Variablen Veränderungen bei der abhängigen Variablen erwartet werden. Sollen beispielsweise blutdrucksenkende Verfahren verglichen werden, sollten Verfahren mit unterschiedlichen Wirkmechanismen zur Verfügung stehen.

3. Bei der Messung der abhängigen Variablen ist auf eine den erwarteten Veränderungen entsprechende Sensibilität des Messverfahrens zu achten. Änderungen im Millimeterbereich lassen sich mit einer grob messenden Zentimeterskala nicht abbilden (und umgekehrt).

Die Art der Operationalisierung hat Einfluss auf die Messung, auf das Messniveau und die erlaubten statistischen Verfahren. Wird das Lebensalter nicht als Messwert (konkrete Jahresangabe oder Angabe des Geburtsdatums) sondern auf einem Intervall erhoben („Sind Sie zwischen 20 und 25 (weitere Intervalle) Jahren alt?“), also als nominalskalierter Wert, können nur Häufigkeitsvergleiche berechnet werden. Bei der Operationalisierung sind folgende Schritte empfehlenswert:

1. Identifikation theoretischer Grundlagen des abzubildenden Gegenstands (4.2)

2. Definition des interessierenden Merkmals

3. definitorische Anker, also beschreibende Worte aus der Definition extrahieren

4. anhand der Anker Items formulieren

5. Urteilsdimensionen und verbale Anker festlegen (6.2.3).

Bei der Operationalisierung wird zusammenfassend ein Messverfahren für die unabhängigen und abhängigen Variablen erarbeitet. Basis für die Operationalisierung ist der theoretische Rahmen für die Verbindung zwischen Merkmalen und die jeweilige (ebenfalls theoretisch begründete) Definition der abzubildenden Begriffe. In den Sozialwissenschaften sind häufig mehrere, teils konkurrierende Definitionen und demnach auch unterschiedliche Operationalisierungsvarianten denkbar. Welche Operationalisierungsvariante genutzt wird hängt ab von Anforderungen an die Vergleichbarkeit mit anderen Studien, von den notwendigen Modifikationsstufen der unabhängigen Variablen für Veränderungen bei der abhängigen Variablen und von der erforderlichen Sensibilität bei der Erfassung der abhängigen Variablen.

6.2.3 Beurteilungsvarianten in Befragungen

Das erarbeitete oder bereits vorhandene Messverfahren in sozial- und gesundheitswissenschaftlichen Studien stützt sich überwiegend auf die subjektive Bewertung und Einschätzung von Aussagen und weniger auf das klassisch „naturwissenschaftliche Messen“ von Merkmalsausprägungen. Die auf Einschätzung und die Bewertung von Aussagen beruhenden Verfahren stützen sich auf die Urteilsfähigkeit der Studienteilnehmer. Mit der Nutzung subjektiver Bewertungen sind Anforderungen an die Gestaltung von Beurteilungsverfahren verknüpft. Dies bedeutet, Urteile verschiedener Menschen mit derselben Merkmalsausprägung müssen gleich sein. Zudem müssten mehrere Messungen bei ein und derselben Person, deren Merkmalsausprägung unverändert blieb, dieselben Ergebnisse liefern (Zuverlässigkeit). Schließlich dürfen Beurteilungsergebnisse nicht vom Untersucher beeinflusst sein (Beller, 2008).

Für die Einschätzung und Beurteilung von Aussagen kommen Ratingskalen zum Einsatz. Ratingskalen erheben Merkmalsausprägungen innerhalb eines vorgegebenen qualitativen und quantitativen Wertebereichs mit festgelegten und markierten Abständen = *Skalenverankerung*. Studienteilnehmer beurteilen in Befragungen, welcher Wert für sie zutrifft.

Beller (2008) zufolge ist der Einsatz von Ratingskalen in den Sozial- und Gesundheitswissenschaften mit zwei Annahmen verknüpft: (1) Die Abstände zwischen den einzelnen Abstufungen werden von den Befragten als gleich groß gewertet, *es wird also vom Intervallskalenniveau ausgegangen*, (2) Befragte verstehen unter den er-

klärenden Begriffen dasselbe. Es wird vom Zutreffen beider Annahmen ausgegangen, i.d.R. ohne dass diese Annahme überprüft wird (Beller, 2008).

Je nach Einsatzgebiet und Operationalisierung werden verschiedene *Skalenpolungen*, *Skalenverankerungen*, *Stufenzahlen* und *Urteilsdimensionen* unterschieden (Beller, 2008). Bei der *Skalenpolung* wird in uni- und bipolare Polung unterschieden. Bei der unipolaren Polung bewegt sich der Wertebereich des numerischen Relativs entweder ausschließlich im positiven oder im negativen Bereich. Die bipolare Polung umfasst in einem Extrem negative und im anderen Extrem positive Werte (Tabelle 3).

Tabelle 3: Differenzierung der Skalenpolung (Beispiele nach Beller, 2008, S. 36f)

unipolar	**bipolar**
Durch meine Arbeit fühle ich mich **nicht belastet** **hoch belastet** 0 1 2 3 4 5 6	Durch meine Arbeit fühle ich mich **nicht belastet** **hoch belastet** -3 -2 -1 0 1 2 3

Die Skalenverankerung, kann mit Zahlen, Worten oder grafischen Symbolen erfolgen (Tabelle 4). Ratingskalen erheben Daten auf drei bis (zumeist) sieben Stufen. Die Anzahl der Stufen bemisst sich daran, wie genau abgestuft eine Merkmalsausprägung abgebildet werden soll und wie gut Studienteilnehmer die Stufenzahl differenzieren können. Je mehr Stufen eine Skala umfasst, desto weniger Differenzierungskompetenzen werden unterstellt. Eine weitere Diskussion geht darum, ob Daten auf einer *geraden* oder *ungeraden* Stufenzahl erhoben werden. Bei gerader Stufenzahl fehlt die „natürliche Mitte“. Studienteilnehmer legen sich also bei der Beurteilung tendenziell auf eine Richtung fest. Dagegen haben Skalen mit einer ungeraden Stufenzahl eine natürliche Mitte. Bei der Beantwortung von Fragen/Items, bei denen sich Befragte unsicher sind, könnte ausgeprägter in der Mitte beurteilt werden, als dies bei einer geraden Stufenzahl möglich ist (s. Urteilsfehler) (Beller, 2008).

Tabelle 4: Differenzierung der Skalenverankerung (Beispiele nach Beller, 2008, S. 36f)

Skalenverankerung	**Beispiel**
Zahlen	Durch meine Arbeit fühle ich mich **nicht belastet** **hoch belastet** 0 1 2 3 4 5 6
Worte	Durch meine Arbeit fühle ich mich ___ **nicht belastet** (-1) ___ **etwas belastet** (0) ___ **hoch belastet** (1)
grafische Symbole	Meine Arbeit macht mir **keinen Spaß** **viel Spaß** 

Die Operationalisierung kann ferner mit unterschiedlichen *Urteilsdimensionen* erfolgen. Entscheidend sind hierbei: der Gegenstand, seine Definition und theoretische Fundierung, die Ausprägung des empirischen Relativs, die Differenzierungsfähigkeit der Studienteilnehmer und die notwendige Abstufung von unabhängiger und abhängiger Variable.

Dabei sollte die Abstufung der Urteilsdimensionen widerspruchs- und überschneidungsfrei sein sowie die Merkmalsausprägung in ihrer gesamten Variationsbreite abbilden. Bei den in Tabelle 5 beispielhaft zusammengefassten Urteilsdimensionen wird von Überschneidungs- und Widerspruchsfreiheit ausgegangen, i.d.R. ebenfalls ohne dass diese überprüft wird. Zudem ist die getroffene Intervallskalenannahme kritisch zu werten, weil die inhaltlichen Abstände zwischen den verbalen Markern prinzipiell nicht festgelegt und ebenso nicht auf der gesamten Skalenbreite gleich groß sein müssen – auch wenn dies die numerische Abstufung unterstellt (Beller, 2008).

Tabelle 5: Urteilsdimensionen (Beispiele nach Beller, 2008, S. 37)

Urteilsdimension	**verbale Anker (Beispiele)**
Häufigkeit	nie-selten-gelegentlich-manchmal-häufig-immer
Intensität	gar nicht-kaum-mittelmäßig-ziemlich-außerordentlich
Wahrscheinlichkeit	keinesfalls-wahrscheinlich nicht-vielleicht-ziemlich wahrscheinlich-sicher
Bewertung	völlig falsch-falsch-unentschieden-richtig-völlig richtig
Zustimmung	stimme gar nicht-wenig-teilweise-ziemlich-absolut … zu

Eine Operationalisierung kann zusammenfassend auf der Basis unterschiedlicher Beurteilungsvarianten erfolgen. Zunächst wird unterschieden, ob Skalen bipolar – die Extremwerte befinden sich am negativen und positiven Ende des Wertebereichs – oder unipolar *gepolt* sind – die Extremwerte befinden sich dann im negativen *oder* im positiven Wertebereich. Die Skalenverankerung, also die Differenzierung der Stufen, erfolgt durch Zahlen, Worte oder ist grafisch untersetzt. Als Urteilsdimensionen können Häufigkeiten, die Intensität, der Grad der Zustimmung oder die Wahrscheinlichkeit für Aussagen in Items unterschieden werden.

6.2.4 Beurteilungsfehler in Befragungen

Eine Beurteilung durch verschiedene Studienteilnehmer birgt nicht nur messtheoretische Risiken, sie ist an sich fehleranfällig. Befragte können evtl. nicht viel mit der Aussage im Item/in der Frage anfangen, die Skalenverankerung oder -polung sind den Befragten unklar oder die Urteilsdimensionen passen aus der Sicht der Befragten nicht zum erhobenen Gegenstand. Bei der skizzierten Abbildung von Merkmalsausprägung auf Ratingskalen werden als Urteilsfehler *Boden- oder Deckeneffekt*, die *Tendenz zur Mitte*, *Reihenfolgeneffekte* sowie Fehler bei der *Einschätzung von Menschen* diskutiert (Beller, 2008).

Decken- und *Bodeneffekte*, *die Tendenz zur Mitte* und eine *geringe Variationsbreite* bei der Beantwortung von Items können entstehen, wenn Befragte das Item und die Urteilsdimensionen (Tabelle 5) nicht ausreichend differenzieren können. Aus diesem Problem resultieren Werte entweder am unteren (Bodeneffekt), oberen (Deckeneffekt) oder mittleren (Tendenz zur Mitte) Wertebereich bei zugleich sehr geringer Streuung der Werte (s. 8.3.2). Dem Problem kann mit einer genauen und mit Beispielen für die Extrempunkte untersetzten Instruktion zu den Fragen begegnet werden.

Reihenfolgeeffekte entstehen, wenn Fragen in Fragebögen in ungeeigneter Form sortiert werden. Wenn zunächst viele Fragen mit großer Variationsbreite des Merkmals (z.B. die Körpergröße in cm- oder m) und anschließend viele Fragen zu Merkmalen mit sehr geringer Variationsbreite (z.B. die Größe von Leberflecken in mm) bewertet werden sollen, kann durch „Gewöhnung an die große Merkmalsstreuung“ u.U. der Kontrast des Merkmals mit großer Variationsbreite (Körpergröße) auf Fragen mit geringer Variationsbreite (die Größe des Leberflecks) angewendet werden. Dies kann in einer größeren Streuung der Messwerte von Merkmalen mit geringer Variationsbreite münden, als dies beim empirischen Relativ (also den zu bewertenden Leberfleck) tatsächlich der Fall ist. Als Lösung wird vorgeschlagen, Items mit ähnlicher Variationsbreite nicht en block zu ordnen, sondern Items mit geringer und großer Variationsbreite beim Merkmal zu mischen.

Bei der Einschätzung von Personen können zudem der *Halo-Effekt*, *Milde-Härte-Fehler* und *Rater-Ratee-Interaktion* Quelle von Beurteilungsfehlern sein. Der *Halo-Effekt* umschreibt die Auswirkung der (positiven oder negativen) Bewertung eines Personenmerkmals auf alle anderen Merkmale. Im Alltag können uns Menschen mit bestimmtem Aussehen oder mit bestimmtem Schmuck, Über- oder Untergewicht besonders sympathisch oder unsympathisch sein. Ob diese Menschen aufgeschlossen, empathisch, hilfsbereit sind oder nicht, spielt für die Beurteilung der Person insgesamt dann evtl. nur eine untergeordnete Rolle bzw. werden diese Merkmale dann auch in ähnlicher Weise bewertet, wie das eine, besonders herausstehende Merkmal.

Zudem können Studienteilnehmer auch systematisch zu positiv oder zu negativ bewerten, wobei dann von *Milde-Härte-Fehlern* gesprochen wird.
Bei der *Rater-Ratee-Interaktion* erfolgt vom Befragten (Rater) eine Bewertung einer anderen Person (Ratee) nicht auf Basis der beim Ratee beobachteten Merkmale, sondern in die Beurteilung fließt die Einschätzung der eigenen Ausprägung beim erhobenen Merkmal mit ein (Beller, 2008).

Befragungen, bei denen Aussagen in Items durch die Befragten beurteilt werden, sind hinsichtlich ihrer Aussagekraft kritisch zu bewerten. An die Erarbeitung treffender Items (Operationalisierung), die Festlegung der Skalenpolung, Skalenverankerung und der Urteilsdimensionen werden sehr hohe Anforderungen gestellt. Die Entwicklung eines Fragebogens ist daher keine Aufgabe, die der Bearbeitung einer inhaltlichen Fragestellung untergeordnet ist. Sie ist essenziell um übertragbare und gültige Antworten in Studien zu erhalten. Es ist wenig empfehlenswert für eine Studie ad hoc einen Fragebogen zu entwickeln und ihn ungeprüft anzuwenden. Stattdessen sollte auf etablierte Instrumente zurückgegriffen werden. Dazu bieten für eine Vielzahl sozial- und gesundheitswissenschaftlicher Fragestellungen Datenbanken wie Psyndex ®, Verlage, die sich auf die Veröffentlichung valider Messinstrumente spezialisiert haben (z.B. Hogrefe ®) oder z.B. Bundesoberbehörden (z.B. Bundesanstalt für Arbeitsschutz und Arbeitsmedizin, BAuA) Messinstrumente an, die teils frei, teils kostenpflichtig genutzt werden können.

Beim Einsatz von Fragebögen sind verschiedene Fehlerquellen bekannt, die sich zusammenfassend auf die wahrgenommene Variationsbreite von Merkmalsausprägungen beziehen (Decken- und Bodeneffekte, Tendenz zur Mitte), auf die Reihenfolge, mit der Fragen unterschiedlicher Variationsbreite ihrer Merkmalsausprägung gestellt werden, auf Einflüsse bei der Bewertung von Menschen und die in Halo-Effekten, Milde-Härte-Fehlern oder Rater-Ratee-Interaktion deutlich werden.

6.3 Gütekriterien von Messinstrumenten

Bei der Diskussion um die Güte von Studiendesigns ist mit dem Validitätsbegriff bereits ein Gegenstand vorgestellt worden, der auch bei der Bewertung der Güte von Messinstrumenten bedeutsam ist. Beim Erheben und Abbilden von Merkmalen mit Hilfe von Messinstrumenten (Fragebögen), können verschiedene Probleme die Aussagekraft der Ergebnisse beeinflussen. Probleme bei der Abbildung eines empirischen Relativs in ein numerisches Relativ (6.2.1) sowie Fehlerquellen bei der Beurteilung verschiedener inhaltlicher Abstufungen in Ratingskalen (6.2.4) wurden bereits angesprochen. Ob Messinstrumente gut sind oder nicht hängt also insbesondere von ihrer Gestaltung ab. Ferner müssen sie Ergebnisse liefern, die nicht vom Untersuchungsleiter abhängen, sie sollten zuverlässig dieselben Werte bei denselben Zuständen liefern und den Gegenstand abbilden, den sie vorgeben abzubilden.

Neben rein technischen Voraussetzungen, der Formulierung von Fragen, der korrekten Auswahl der Skalierung und der Urteilsdimensionen spielt hier, wie im gesamten Forschungsprozess, die theoretische Fundierung eine besondere Rolle.

Als Hauptgütekriterien werden *Objektivität*, *Reliabilität (Zuverlässigkeit)* und *Validität (Gültigkeit)* unterschieden, die aufeinander aufbauen: Nur objektive Instrumente könne reliabel, nur reliable Instrumente auch valide sein (Sedlmeier & Renkewitz, 2013).

Messinstrumente sind *objektiv*, wenn sie Ergebnisse ohne den Einfluss einer dritten Person liefern. Sie sind es nicht, wenn ein Ergebnis je nach Testleiter bei ein und demselben Zustand unterschiedlich ist. Bei der Objektivität wird unterschieden zwischen (Sedlmeier & Renkewitz, 2013):

- *Durchführungsobjektivität*: Während der Testdurchführung darf durch unterschiedliche oder verschieden gut verständliche Testinstruktionen verschiedener Untersucher ein Ergebnis nicht beeinflusst werden. Standardisierte Testinstruktionen, die schriftlich gelesen oder mündlich vom Untersucher vorgetragen werden, gewährleisten eine hohe Durchführungsobjektivität.

- *Auswertungsobjektivität*: während der Auswertungsphase gelangen unterschiedliche Auswerter bei Ergebnissen ein und desselben Untersuchungsteilnehmers zum selben Testergebnis. Eine hohe Auswertungsobjektivität wird durch aussagekräftige Auswertungsinstruktionen von Messinstrumenten erzielt.

- *Interpretationsobjektivität*: Testanwender ziehen aus den Ergebnissen der Auswertung dieselben Schlüsse und ein vergleichbares Fazit. Interpretationsobjektivität wird durch die Normierung von Messinstrumenten erzielt. Ein bestimmtes Testergebnis wird mit Normwerten verglichen und Schlüsse z.B. für die therapeutische Intervention gezogen. Beim Fiebermessen sind medizinisch begründete Normwerte und -intervalle festgelegt. Ein Testergebnis unter 37,0 wird als fieberfrei, zwischen 37,0 und 37,9 als erhöhte Temperatur, zwischen 38,0 und 38,5 als leichtes Fieber (…) interpretiert. Auch bei der Interpretation von Testergebnissen sollte die subjektive Bewertung des Untersuchers keine Rolle spielen.

Reliabilität oder Zuverlässigkeit liegt vor, wenn bei ein und demselben Probanden, dessen Zustand sich nicht verändert hat, mit einem Messinstrument bei mehreren Messungen dasselbe Testergebnis erzielt wird. Ein zuverlässiges Thermometer sollte siedendes Wasser auf Meereshöhe stets mit 100 Grad Celsius messen und nicht mal mit 90 und mal mit 110 Grad. Messungen sind stets mit einem Messfehler behaftet, über eine große Anzahl von Messungen beim selben Zustand, sollte der Messfehler

im Mittel „0“ sein. In sozialwissenschaftlichen Untersuchungen wird die Reliabilität durch folgende Verfahren überprüft (Sedlmeier & Renkewitz, 2013):

- *Retest-Reliabilität*, in einer Stichprobe, deren Zustand beim interessierenden Merkmal konstant bleibt, wird derselbe Test zeitversetzt zwei Mal angewendet. Über die Berechnung von Korrelationskoeffizienten (8.4.3) wird überprüft, ob die Messung gleiche (absolute Korrelatio, r = 1) oder sehr ähnlich variierende Werte (r > 0,6) ergab.

- *Paralleltestreliabilität*, einer Stichprobe werden zu zwei Zeitpunkten jeweils zwei Tests vorgelegt, die mit unterschiedlicher Itemformulierung dasselbe Merkmal messen. Auch hier werden Korrelationsanalysen eingesetzt, bei denen sich sowohl zum selben Messzeitpunkt enge Zusammenhänge ergeben sollten, als auch (vergleichbar mit der Re-Test-Reliabilität) zwischen den Messzeitpunkten und den verschiedenen Messinstrumenten.

- *Testhalbierungsmethode*, bei einer Stichprobe wird ein Test durchgeführt. Daran anschließend werden die Testergebnisse einer Hälfte der Stichprobe mit den Ergebnissen der anderen Stichprobenhälfte korreliert. Die Zuordnung zu den Testhälften sollte auf Zufall beruhen.

Tests sind *valide* bzw. die gültig, wenn sie das Merkmal abbilden, was sie vorgeben abzubilden. Eine Waage misst die Masse, ist jedoch ungeeignet die Temperatur zu messen. Ein Thermometer ist valide für die Temperaturmessung in einem bestimmten Bereich. Für die Messung der Körpermasse ist es dagegen ungeeignet. Während bei physikalischen Parametern eine überwiegend eindeutige Zuordnung von zu messenden Konstrukten (empirisches Relativ, Merkmale) zu bestimmten Messinstrumenten (für die Abbildung des empirischen in das numerische Relativ) möglich ist, gelingt dies bei sozial- und gesundheitswissenschaftlichen Untersuchungen häufig nicht. Ob ein Fragebogen tatsächlich die überdauernde Lebenszufriedenheit misst oder einen Ausschnitt davon, z.B. die Arbeitszufriedenheit, bzw. nur auf einen kurzen Zeitraum Bezug nimmt, wird häufig erst nach mehreren Anwendungen und Überprüfungen des Messinstruments erkennbar. Wie ein Instrument inhaltlich gestaltet werden muss, damit ein theoretischer Zielgegenstand korrekt abgebildet wird, bestimmt die Theorie, die darauf beruhende Definition eines Gegenstands und die sich daraus ergebenden definitorischen Anker. Besondere Bedeutung für valide Messinstrumente hat eine sorgfältige Operationalisierung (s. 6.2.2). Validität kann auf folgenden Ebenen betrachtet werden (Sedlmeier & Renkewitz, 2013):

- *Inhaltsvalidität* liegt vor, wenn von allen möglichen Items/Fragen, mit denen ein Merkmal abgebildet werden kann, diejenigen ausgewählt werden, die es hinreichend beschreiben. Dazu müssen alle Items bekannt sein, was bei komplexen Merkmalen, wie Lebenszufriedenheit, Arbeitszufriedenheit oder Depressivität kaum anzunehmen ist. Die Inhaltsvalidität ist bei diesen

Merkmalen mit der Regel: Auswahl relevanter Items aus einem „Itemuniversum“ (Sedlmeier & Renkewitz, 2013, S. 76) nicht bestimmbar. Ob relevante Items ausgewählt wurden wird über subjektive Bewertungen von Experten der jeweiligen Gegenstände bestimmt.

- *Kriteriumsvalidität* prüft ob ein Testergebnis mit dem messenden Kriterium (Gegenstand, Merkmal, empirisches Relativ) zusammenhängt. Da ein Messinstrument ein Kriterium erst misst, erfolgt die Überprüfung der Kriteriumsvalidität über die vergleichende diagnostische Bewertung von Experten oder durch den Einsatz von Messinstrumenten, die ähnliche Gegenstände messen.

- *Konstruktvalidität* untersucht ob ein Messinstrument Ergebnisse liefert, die der theoretischen Diskussion zu einem Gegenstand entsprechen. Ein Messinstrument zur Erfassung der gesundheitlichen Lebensqualität sollte die gesundheitliche Lebensqualität konstruktkonform abbilden. Beispielsweise sollte bei chronisch kranken Menschen eine geringere gesundheitliche Lebensqualität vorliegen als bei gesunden Menschen. Das subjektive körperliche und psychische Wohlbefinden sollte bei einer hohen gesundheitlichen Lebensqualität hoch, Depressivität, Burnout oder hohe Belastung sollten hingegen niedrig ausgeprägt sein.

Die diskutierten Gütekriterien von Messinstrumenten stehen in Beziehung zueinander. Mindestanforderung ist die Objektivität. Nur wenn Instrumente unabhängig vom Testanwender Ergebnisse liefern, messen sie zuverlässig (reliabel). Zuverlässig messende Instrumente liefern Ergebnisse, die eine Interpretation einer Merkmalsausprägung zulassen. Nur zuverlässig messende Instrumente liefern demnach auch valide, also theoretisch konforme Ergebnisse.

Zusammenfassend umschreibt die Objektivität den Grad der Unabhängigkeit von Messergebnissen vom Testanwender. Sie kann über eine standardisierte Instruktion bei der Datenerhebung, Auswertung und Interpretation hergestellt werden. Reliabel sind Messinstrumente, die bei denselben Zuständen auch dieselben Testergebnisse liefern. Valide Messinstrumente bilden ein Merkmal so ab, wie es theoretisch diskutiert wird. Messinstrumente sind valide, wenn sie das messen, was sie vorgeben zu messen.

6.4 Zusammenfassung und Aufgaben

Sozial- und gesundheitswissenschaftliche Studien erheben Daten durch die Befragung und Beobachtungen von Menschen oder im Rahmen von Experimenten (s. auch 5.3). Basis von Befragungen ist Kommunikation zwischen Befrager und Befragtem in mündlicher oder schriftlicher Form, stark oder wenig strukturiert mit offenen

oder geschlossenen Antwortformaten. Befragt werden sowohl Einzelpersonen als auch Gruppen im Rahmen quantitativer und qualitativer Forschungsansätze. In Beobachtungen werden Daten nicht auf der Basis eines kommunikativen Prozesses erhoben. Unterschieden werden teilnehmende und nicht-teilnehmende Beobachtungen, Eigen- oder Fremdbeobachtungen und offene oder verdeckte Beobachtungen. Der Beobachtungsgegenstand, seine potenziellen Ausprägungen, Ort, Zeitpunkt und Dauer der Beobachtung werden vorab in einem Beobachtungsplan festgeschrieben. In Experimenten wird die Ausprägung von Einflussfaktoren variiert und ein bestimmter Effekt beim Endpunkt in der Interventionsgruppe erwartet. Ergebnisse experimenteller Studien sind aussagekräftig, wenn die Zuordnung von Studienteilnehmern zu Interventions- und Kontrollgruppe auf Zufall beruht und Störfaktoren benannt und kontrolliert werden (im Detail s. 5.3, 0).

Quantitative Daten aus Befragungen und Beobachtungen ergeben sich aus Messungen. Beim Messen werden Merkmale von Menschen, die auf physikalischen, biologischen, psychosozialen (…) Informationen beruhen (das empirische Relativ), in Zahlen (numerisches Relativ) übertragen. Die Abbildung des qualitativen Merkmals muss zumindest Gemeinsamkeiten und Unterschiede bei den Merkmalsausprägungen erkennen lassen. Kann anhand der Messwerte auf Personen mit bestimmten Eigenschaften geschlossen werden, handelt es sich um eine homomorphe Abbildung. Messungen erfassen Informationen auf verschiedenen Messniveaus, die sich hinsichtlich des Informationsgehalts unterscheiden: nominalskalierte Daten (Gemeinsamkeiten und Unterscheide), ordinalskalierte Daten (um Rangfolge ergänzt), intervallskalierte Daten (selbe zahlenwertige Unterschiede kennzeichnen dieselben qualitativen Unterschiede im gesamten möglichen Wertebereich). Messprobleme können sich aus der Übertragung qualitativer Merkmale in Zahlen (Eindeutigkeitsproblem), bei der informationserhaltenden Transformation (Eindeutigkeitsproblem) und bei Möglichkeiten, Berechnungen mit den Zahlen durchzuführen (Bedeutungsproblem), ergeben.

Unter Operationalisierung wird die Erarbeitung eines Messverfahrens für unabhängige und abhängige Variablen auf Basis eines theoretischen Rahmens verstanden. Dies erfordert die theoretisch begründete Definition der abzubildenden Begriffe. Für zahlreiche sozial- und gesundheitswissenschaftliche Merkmale werden unterschiedliche, teils konkurrierende Definitionen angewendet, die in unterschiedlichen Operationalisierungsvarianten münden können.

Im Rahmen der Operationalisierung werden ebenfalls Überlegungen zur Beurteilungsvarianten angestellt. Merkmalsausprägungen werden häufig auf Ratingskalen erhoben, auf denen Aussagen in Items bewertet werden. Skalen können bipolar oder unipolar *gepolt* sein. Die Verankerung der qualitativen Aussagen auf der Ratingskala, also die Differenzierung der Stufen, erfolgt durch Zahlen, Worte oder grafisch. Beurteilt werden Häufigkeiten, die Intensität, der Grad der Zustimmung oder die Wahrscheinlichkeit für Aussagen in Items.

Datenerhebung mit Fragebögen bergen Fehlerrisiken, bei der wahrgenommenen Variationsbreite von Merkmalsausprägungen (Decken- und Bodeneffekte, Tendenz zur Mitte), bei der Reihenfolge, mit der Fragen gestellt werden sowie bei Einflüssen im Rahmen der Bewertung durch Menschen, die in Halo-Effekten, Milde-Härte-Fehlern oder Rater-Ratee-Interaktion erkennbar sind.

Messinstrumente sollten ein Merkmal objektiv, reliabel und valide erfassen. Die Objektivität beschreibt den Grad der Unabhängigkeit des Messergebnisses vom Testanwender. Messinstrumente sind reliabel, wenn sie bei denselben Zuständen auch dieselben Testergebnisse liefern. Unter Validität wird die theoriekonforme Abbildung eines Merkmals verstanden. Valide Messinstrumente messen, was sie vorgeben zu messen.

Schlüsselwörter: *Äquivalenzrelation, Auswertungsobjektivität, Bedeutungsproblem, Befragung, Begriffsdefinition, Beobachtung, Beobachtungsplan, Bodeneffekte, Deckeneffekte, Durchführungsobjektivität, Eindeutigkeitsproblem, Einschätzung von Menschen, empirischen Relativs, Experimente, Fremdbeobachtungen, gerade Stufenzahl, geringe Variationsbreite, Gruppenbefragungen, Halo-Effekt, homomorphe Abbildung, Individualbefragungen, Inhaltsvalidität, Interpretationsobjektivität, Intervallskalenniveau, intervallskalierten Daten, isomorphe Abbildung, Konstruktvalidität, Kriteriumsvalidität, Messinstrumente, Messinstrumente Messinstrumente, Milde-Härte-Fehler, mündliche Befragungen, nicht-teilnehmende Beobachtungen, Nominalskalenniveau, numerisches Relativ, Objektivität, offene Beobachtungen, Operationalisieren, Ordinalskalenniveau, Ordinalskalenniveau, Ordnungsrelation, Paralleltestreliabilität, Rater-Ratee-Interaktion, Reihenfolgeneffekte, Reliabilität (Zuverlässigkeit) und Validität (Gültigkeit), Repräsentationsproblem, Retest-Reliabilität, schriftlicher Befragungen, schwach strukturierte Befragungen, Skalenpolung, Skalenverankerung, Standardisierte Befragungen, stark strukturierte Befragungen, teilnehmenden Beobachtungen, Tendenz zur Mitte, Testhalbierungsmethode, ungerade Stufenzahl, Urteilsdimensionen, valide, Variable, verdeckte Beobachtung, Verhältnisskalenniveau, Verzerrung (Bias), wenig (oder nicht-) standardisiert*

Aufgaben zur Lernkontrolle

1. Welche Arten der Befragungen gibt es und wie unterscheiden sie sich voneinander?
2. Wozu dient ein Beobachtungsplan?
3. Welche Probleme können Beobachtungsstudien bergen?
4. Worin besteht der Unterschied zwischen Äquivalenz- und Ordnungsrelation?

5. Was bedeutet homomorph und isomorph? Erläutern Sie dies anhand von Beispielen.

6. Grenzen Sie „Nominalskala“, „Ordinalskala“, „Intervallskala“ und „Verhältnisskala“ voneinander ab.

7. Definieren Sie folgende Beurteilungsfehler:
    a. Decken- und Bodeneffekte
    b. Tendenz zur Mitte
    c. Reihenfolgeeffekte
    d. Halo-Effekt
    e. Milde-Harte-Fehler
    f. Rater-Ratee-Interaktion

Aufgaben zur Berufspraxis

8. Schauen Sie sich relevante sprachtherapeutische Testungen an, die Sie in Ihrer Klinik/Praxis haben (bspw. AAT oder SETK-3-5). Erfüllen diese Testungen die Gütekritierien?

9. Welche Art der Beobachtungen ist Ihrer Meinung nach am wichtigsten für die Sprachtherapie? Warum?

10. Welcher Beurteilungsfehler in Befragungen spielt Ihrer Meinung nach in der Sprachtherapie eine große Rolle? Warum?

Literatur

Atteslander, P., & Cromm, J. (2010). *Methoden der empirischen Sozialforschung* (13., neu bearbeitete und erweiterte Auflage ed.). Berlin: Erich Schmidt Verlag.

Beerlage, I., Arndt, D., & Hering, T. S., Silke. (2007). *Arbeitsbedingungen und Organisationsprofile als Determinanten von Gesundheit, Einsatzfähigkeit sowie haupt- und ehrenamtlichem Engagement bei Einsatzkräften in Einsatzorganisationen des Bevölkerungsschutzes. Erster Zwischenbericht März 2007.* Magdeburg.

Beerlage, I., Hering, T., & Nörenberg, L. (2006). *Entwicklung von Standards und Empfehlungen für eine Netzwerk zur bundesweiten Strukturierung und Organisation psychosozialer Notfallversor¬gung. Schriftenreihe Zivilschutzforschung – Neue Folge Band 57.* Bonn: Bundesamt für Bevölkerungsschutz.

Beller, S. (2008). *Empirisch forschen lernen Konzepte, Methoden, Fallbeispiele, Tipps* (2., überarb. Aufl ed.). Bern: Huber.

Döring, N., Bortz, J., & Pöschl, S. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften* (5. vollständig überarbeitete, aktualisierte und erweiterte Auflage ed.). Berlin Heidelberg: Springer.

Hussy, W., Schreier, M., & Echterhoff, G. (2013). *Forschungsmethoden in Psychologie und Sozialwissenschaften für Bachelor: mit 23 Tabellen* (2., überarb. Aufl. ed.). Berlin u.a.: Springer.

Sedlmeier, P., & Renkewitz, F. (2013). *Forschungsmethoden und Statistik für Psychologen und Sozialwissenschaftler* (2., aktualisierte und erweiterte Auflage ed.). München Harlow Amsterdam: Pearson/Higher Education.

7 Stichprobenauswahl

Lernergebnisse:

- Methoden der zufalls- und nicht zufallsbasierten Stichprobenziehung können differenziert werden. Die Bedeutung der Stichprobenziehung für die Aussagekraft von Ergebnissen kann erörtert werden (7.1, 7.2).

- Der Repräsentativitätsbegriff kann ausgehend von der Art der Stichprobenrekrutierung und der Stichprobengröße erläutert und diskutiert werden.

- Die Bedeutung des Grenzwerttheorems für die Schätzung von Parametern der Grundgesamtheit ausgehend von Stichproben kann erläutert werden (7.3).

Kaum eine Studie untersucht eine Grundgesamtheit (1.2). Die Mehrzahl publizierter wissenschaftlicher Veröffentlichungen unter einem quantitativen Forschungsparadigma beruht auf Studien, die Stichproben untersuchten, die aus einer interessierenden Grundgesamtheit gezogen wurden. Vollerhebungen werden aus inhaltlichen und ökonomischen Gründen selten durchgeführt. Stichproben sollten hinsichtlich der Verteilung von Merkmalen und Merkmalsausprägungen ein Abbild der Grundgesamtheit sein und ihr in wesentlichen Aspekten entsprechen. Zu diesen wesentlichen Aspekten zählen: Geschlechter- und Altersverteilung, Verteilung des Bildungsstands, Einkommen, Familienstand und weitere Merkmale, die Bedeutung bei der Untersuchung wissenschaftlicher Fragestellungen haben können. Insofern hat die Stichprobe Einfluss auf die externe und die interne Validität (5.6). Das Auswahlverfahren hat einen wesentlichen Einfluss, ob der Annahme einer vergleichbaren Verteilung der Merkmale in der Stichprobe gefolgt werden kann oder nicht. Der Untersuchung von Stichproben liegt ferner die Annahme zugrunde, aus Stichprobenergebnissen auf Stimmungen, Merkmale und Urteile in der Grundgesamtheit schließen zu können. Die regelmäßigen Befragungen vor Wahlen oder das Erfragen von Zustimmungsquoten für Politiker ist ein Beispiel dafür. Dass dieser Annahme nicht grundsätzlich gefolgt werden kann, zeigen die nicht seltenen Diskrepanzen zwischen Ergebnissen von Wahlbefragungen und tatsächlichen Wahlergebnissen. Die Initiatoren dieser Befragungen nennen dabei selber Gründe, warum sie bei ihrer Prognose daneben lagen. Häufig wird dabei auf die Stichprobenziehung verwiesen, die bestimmte Wählergruppen nicht berücksichtigen, z.B. Menschen ohne Festnetzanschluss, bestimmte Altersgruppen usw.

Ziel des Auswahlverfahrens ist eine repräsentative Stichprobe. Die landläufige Annahme, Stichproben seien repräsentativ, wenn sie besonders groß sind, ist nicht korrekt. Die Repräsentativität gründet sich auf die Chancen der Mitglieder einer Grundgesamtheit, Bestandteil der Stichprobe zu werden. Nur wenn alle Mitglieder der Grundgesamtheit dieselbe Chance haben für eine Studie ausgewählt zu werden, können Stichproben als repräsentativ bezeichnet werden. Die Anforderungen an die zufällige Stichprobenziehung sind hoch. Sie setzt streng genommen die namentliche Bekanntheit der Grundgesamtheit voraus. Bei der Durchführung und der Interpretation statistischer Analysen werden zumeist Zufallsstichproben vorausgesetzt. Bei der Stichprobenziehung wird in die zufallsbasierte (repräsentative) und nicht zufallsbasierte (nicht repräsentative) Ziehung unterschieden.

Die Fallauswahl im qualitativen Forschungsstil orientiert sich nicht an der Repräsentativität einer Stichprobe, sondern versucht gezielt Fälle auszuwählen, die von besonderem Interesse für den zu untersuchenden Gegenstand sind. Nicht die Größe der Stichprobe wird als ausschlaggebend für die Bedeutsamkeit der Ergebnisse verstanden, sondern der Informationsreichtum der untersuchten Fälle und die Intensität der Analyse der erhobenen Daten. Darüber hinaus wird in verschiedenen Forschungsansätzen eine sukzessive Fallauswahl im Forschungsprozess vorgeschlagen. Verschiedenen Formen der qualitativen Stichprobenziehung diskutiert beispielsweise Patton (2002, S. 230ff.).

7.1 Zufallsbasierte Stichprobenziehung

Voraussetzung für die Gewinnung repräsentativer Stichproben ist eine zufallsbasierte Auswahl der Studienteilnehmer. Zufallsstichproben können in Form *einfacher*, *geschichteter, mehrstufiger Zufallsauswahl* sowie als *Klumpenstichprobe* gewonnen werden.

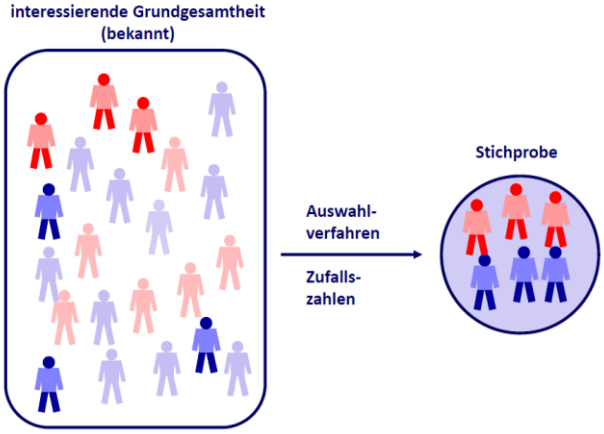


Abbildung 18: Skizze einer einfachen Zufallsauswahl aus einer bekannten Grundgesamtheit (eigene Darstellung)

Bei der *einfachen Zufallsauswahl* werden per Zufallsprinzip Studienteilnehmer aus einer namentlich bekannten Grundgesamtheit gezogen. Das Zufallsprinzip basiert auf der Nutzung von Zufallszahlen, also eher „händisch“, oder es wird die Zufallsoption in Statistikprogrammen genutzt. Eine namentlich bekannte Grundgesamtheit kann über Einwohnermelderegister, Mitgliederlisten von Krankenversicherungen oder Telefonbücher generiert werden (Beller, 2008, Döring et al., 2016) (Abbildung 18).

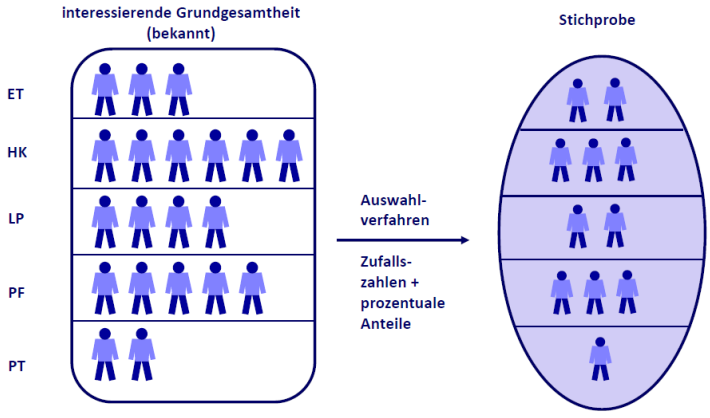


Abbildung 19: Skizze einer geschichteten Zufallsauswahl, die Anteile der Teilpopulationen in der Stichprobe abbildet (ET= Ergotherapie, HK= Hebammenkunde, LP= Logopädie, PF= Pflege, PT= Physiotherapie) (eigene Darstellung)

Bei der *geschichteten Zufallsauswahl* wird die Stichprobe aus ebenfalls namentlich bekannten Teilpopulationen gezogen. Im Unterschied zur einfachen Zufallsauswahl werden hier besondere Merkmale der Grundgesamtheit berücksichtigt. An einer Hochschule haben die einzelnen Studiengänge unterschiedlich viele Studierende. Möchte man die Verteilung der Studierenden in der Stichprobe repräsentiert haben, wird die Stichprobe „geschichtet", also Studiengang für Studiengang, ausgewählt. Dabei wird die Verteilung der Studiengänge über die Größe der Teilstichproben auch in der Stichproben erkennbar (Beller, 2008, Döring et al., 2016) (Abbildung 19).

Die bisherigen Verfahren werden bei einer namentlich bekannten Grundgesamtheit angewendet. In vielen Fällen ist die Grundgesamtheit aus unterschiedlichen Gründen unbekannt. Über die *mehrstufige Zufallsauswahl* oder die Untersuchung einer *Klumpenstichprobe* kann eine Stichprobe zufallsbasiert gezogen werden.

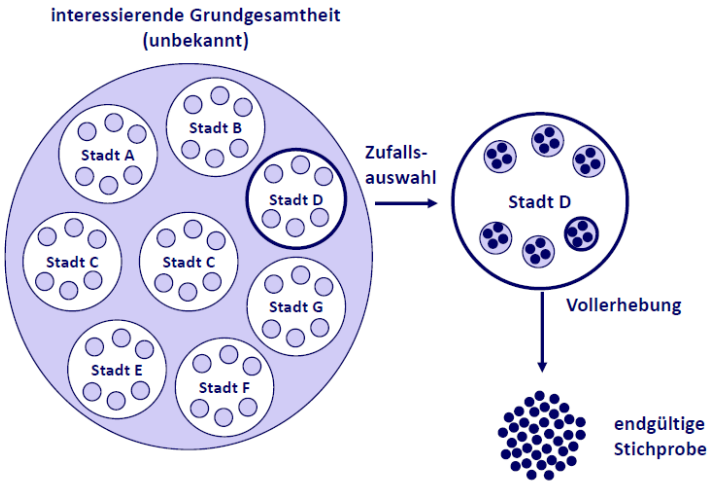


Abbildung 20: Skizze einer Klumpenauswahl (eigene Darstellung)

In der *Klumpenauswahl* erfolgt zunächst die Identifikation der Einheiten, in denen die Grundgesamtheit erkennbar wird. Diese Einheiten können Städte, Hochschulen, Betriebe und andere Institutionen sein. Die für eine Zufallsstichprobe notwendige vollständige Liste Grundgesamtheit wird hier also aus Klumpen, d.h. Städten, Betrieben, Hochschulen, gebildet, in denen die Grundgesamtheit arbeitet, lebt und sich bewegt. Aus der vollständigen Liste werden per Zufall Klumpen gezogen. In diesen Klumpen werden alle Mitglieder befragt (Beller, 2008, Döring et al., 2016) (Abbildung 20).

Die *mehrstufige Zufallsauswahl* erweitert die Klumpenauswahl um eine weitere Zufallsziehung aus dem Klumpen. Je nach Differenzierungsgrad des Klumpens, z.B. in Suborganisationen, Abteilungen und Unterabteilungen bei großen Organisationen, sind mehr als zwei Auswahlstufen denkbar (Döring et al., 2016) (Abbildung 21).

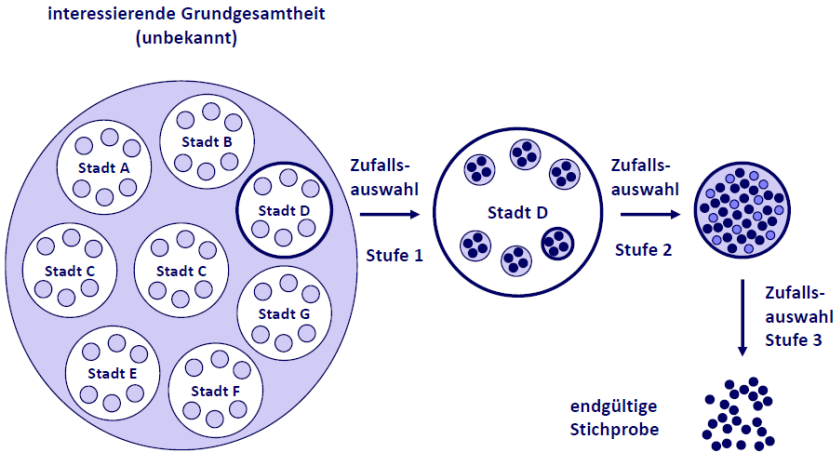


Abbildung 21: Skizze einer mehrstufigen Zufallsauswahl (eigene Darstellung)

7.2 Nicht zufallsbasierte Stichprobenziehung

Für empirische Studien werden Zufallsstichproben angestrebt. Teilweise ist es jedoch nicht möglich und nicht zielführend, Zufallsstichproben zu ziehen. Insbesondere in qualitative Studien kann eine Zufallsauswahl und damit eine Merkmalsstreuung analog zur Grundgesamtheit, kontraindiziert sein. Ferner existieren heikle Stichproben, in Armee, Polizei und anderen Einrichtungen, die auf Geheimhaltung der Identität ihrer Mitglieder angewiesen sind und für die auch für Forschungszwecke keine „Liste der Grundgesamtheit" zur Verfügung gestellt wird. In wenigen quantitativen Studien wird zudem nicht auf eine Verknüpfung von Stichprobenergebnissen mit Annahmen zur Grundgesamtheit abgehoben, wie dies in Pretteststudien der Fall sein kann. Zu den nicht zufallsbasierten Auswahlverfahren zählen *Quotenauswahl*, *Ad-hoc-Auswahl*, *theoriegeleitete Auswahl* und eine Auswahl über das *Schneeballsystem*.

Bei der *Quotenauswahl* wird eine Stichprobe entwickelt, die hinsichtlich der relativen Häufigkeitsverteilung der Grundgesamtheit entspricht. Um Unterschied zur geschichteten Zufallsauswahl erfolgt die Auswahl aus den einzelnen Merkmalsgruppen nicht auf Basis von Zufallszahlen, sondern willkürlich. In der Stichprobe ist das interessierende Merkmal, beispielsweise die Altersgruppen, analog verteilt, wie in der Grundgesamtheit (Döring et al., 2016).

Die *Ad-hoc-Auswahl* ergibt eine Gelegenheitsstichprobe. Dabei werden beispielsweise die ersten 100 Personen befragt, die einen Bahnhof verlassen oder in ein Kino gehen. Auch diese Stichprobenrekrutierung basiert nicht auf Zufall sondern ist willkürlich (Döring et al., 2016).

Eine *theoriegeleitete Auswahl* erfolgt häufig im Rahmen qualitativer Studien. Dabei werden besonders typische oder untypische Fälle ausgewählt. Soll die individuelle Auseinandersetzung alleinerziehender Frauen mit Arbeits- und Alltagsstress unter-

sucht werden, müssen Personen ausgewählt, die diesen Eigenschaften genau entsprechen: Frau, alleinerziehend, arbeitet, ist gestresst (Döring et al., 2016).

In der Auswahl nach dem *Schneeballsystem* benennt eine befragte Person eine oder mehrere weitere Menschen aus ihrem Umfeld bzw. ihrem sozialen Netzwerk. Das Schneeballsystem zur Stichprobenrekrutierung wird auch in qualitativen Studien z.B. zur sozialen Netzwerkanalyse angewendet. Auch eine Auswahl nach dem Schneeball ist willkürlich, erzeugt keine repräsentativen Stichproben und ist nicht geeignet, um mit statistischen Verfahren von der Stichprobe auf die Grundgesamtheit zu schließen (s.9).

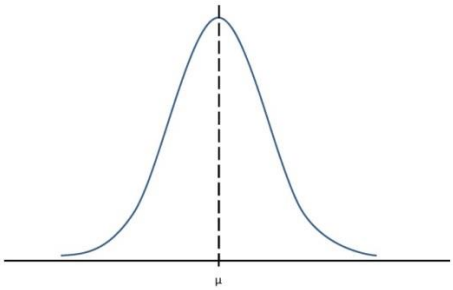
7.3 Stichprobengröße, Genauigkeit von Schätzungen und Grenzwerttheorem

In den bisherigen Ausführungen wurde der Repräsentativitätsbegriff stets in Verbindung gebracht mit der zufälligen Auswahl der Stichprobenauswahl in quantitativen Studien. Wenn alle Mitglieder der Grundgesamtheit dieselbe Chance haben, in eine Stichprobe gezogen zu werden, so die Annahme, nähern sich die Messwerte bei hinreichend großen Stichproben den Werten in der Population an. Auch eine auf Zufall beruhende Stichprobenziehung kann jedoch zufällig Personen ziehen, die im Mittel Messwerte haben, die vom Populationswert deutlich abweichen. Tabelle 6 zeigt unter 1. die Verteilung eines Messwerts in der Grundgesamtheit. Die Nummern 2. bis 5. zeigen die Kennwerteverteilung desselben Merkmals in Stichproben, die aus der Grundgesamtheit gezogen wurden. Sowohl die Verteilung als auch die mittleren Kennwerte weichen von den Parametern in der Grundgesamtheit ab.

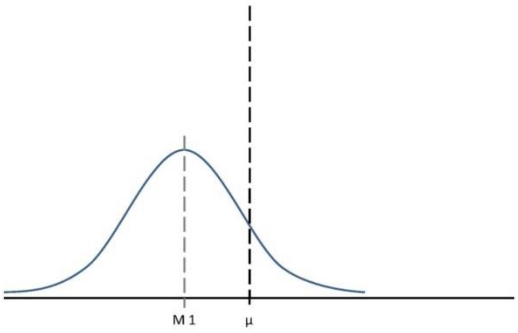
Ein erstes Fazit zeigt: Die Kennwerte in Stichproben, auch in auf Zufallsbasis gezogenen Stichproben stimmen nicht zwingend mit den Kennwerten in der Population überein.

Tabelle 6: Verteilung von Messwerten in der Grundgesamtheit und in daraus auf Zufallsbasis gezogenen Stichproben (Beispiele nach Beller, 2008, S. 91)

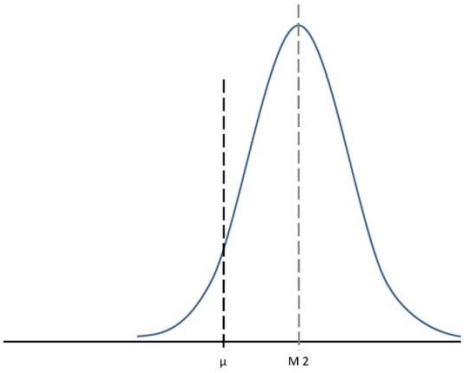
1. Verteilung der Werte in der Grundgesamtheit



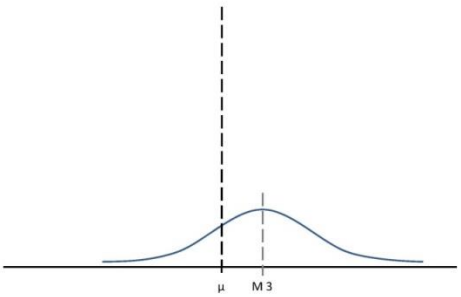
2. Stichprobe 1



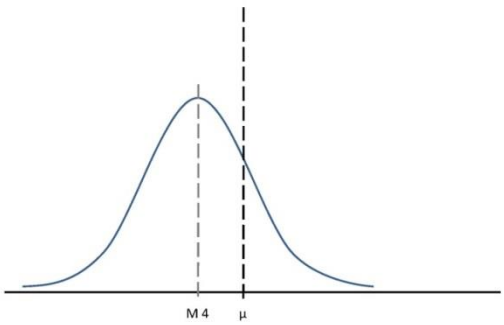
3. Stichprobe 2



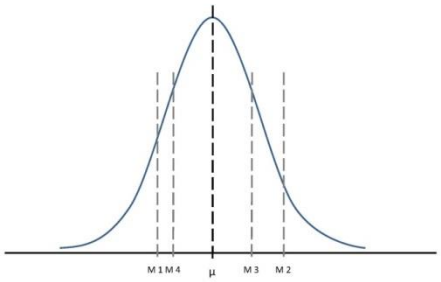
4. Stichprobe 3



5. Stichprobe 4



6. Verteilung der Stichprobenkennwerte im Vergleich mit dem Populationskennwert



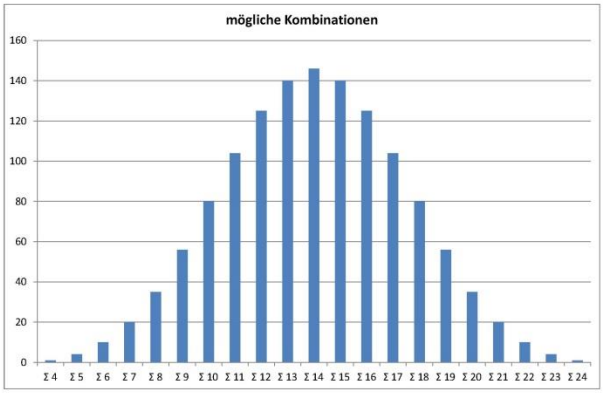
Anmerkung: μ= Mittelwert in der Grundgesamtheit
M1 bis M4= Mittelwert in den jeweiligen Zufallsstichproben

Das Risiko für ungenaue Schätzungen von durchschnittlichen Merkmalsausprägungen und Verteilungen in Stichproben wird kleiner, wenn große Stichproben gezogen werden. Eine für die Berechnung von Maßen der deskriptiven und der Inferenzstatistik zentrale Annahme wird im *zentralen Grenzwerttheorem* beschrieben. Danach nähert sich die Verteilung in Zufallsstichproben bei zunehmender Stichprobengröße einer Normalverteilung bzw. der wahren Verteilung an und der Mittelwert dem Mittelwert in der Grundgesamtheit (Beller, 2008). Beller (2008) zufolge kann bereits bei einer Stichprobengröße der Zufallsstichprobe von n > 30 von annähernd normalverteilten Kennwerten ausgegangen werden. Grundlage für die Anwendung des Grenzwerttheorems, insbesondere im Hinblick auf die Forderung nach normalverteilten Messwerten bei parametrischen Inferenzstatistischen Tests (s. 9.3, z.B. t-Test), ist die zufallsbasierte Ziehung einer Stichprobe. Für nicht zufallsbasierte Stichprobenziehungen ist das Grenzwerttheorem nicht anwendbar, die Überprüfung der Normalverteilungsannahme ist daher bei willkürlichen Stichprobenziehungen erforderlich.

Beispielhaft soll die Annahme des Grenzwerttheorems anhand von Augensummen beim viermaligen Würfeln *eines* Würfels dargestellt werden. Die Augensummen bewegen sich zwischen 4 (viermaliges Würfeln einer 1) und 24 (viermaliges Würfeln einer 6). Insgesamt sind 6^4= 6 x 6 x 6 x6= 1.296 verschiedene Reihenfolgen der Würfelaugen möglich.

Tabelle 7: Wahrscheinlichkeit verschiedener Augensummen bei viermaligem Würfeln eines Würfels mit Darstellung der Verteilung – Grundgesamtheit (N= 1.296)

Würfelsummen	Mögliche Kombinationen	Wahrscheinlichkeit
Σ 4	1	0,0008 (0,08%)
Σ 5	4	0,0031 (0,31%)
Σ 6	10	0,0077 (0,77%)
Σ 7	20	0,0154 (1,54%)
Σ 8	35	0,0270 (2,70%)
Σ 9	56	0,0432 (4,32%)
Σ 10	80	0,0617 (6,17%)
Σ 11	104	0,0802 (8,02%)
Σ 12	125	0,0965 (9,65%)
Σ 13	140	0,1080 (10,8%)
Σ 14	146	0,1127 (11,3%)
Σ 15	140	0,1080 (10,8%)
Σ 16	125	0,0965 (9,65%)
Σ 17	104	0,0802 (8,02%)
Σ 18	80	0,0617 (6,17%)
Σ 19	56	0,0432 (4,32%)
Σ 20	35	0,0270 (2,70%)
Σ 21	20	0,0154 (1,54%)
Σ 22	10	0,0077 (0,77%)
Σ 23	4	0,0031 (0,31%)
Σ 24	1	0,0008 (0,08%)
Summen	**1.296**	**1,0000 (100%)**



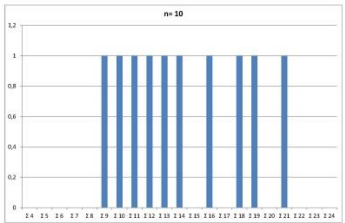
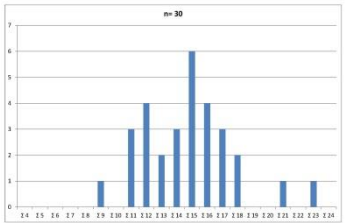
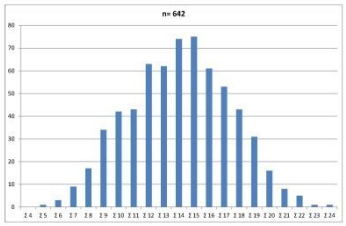
Der Versuch des viermaligen Würfelns eines Würfels wird beispielhaft n = 10, n = 30, n = 50, n = 100, n= 642 mal wiederholt. Tabelle 8 zeigt die Ergebnisse dieser Versuche. Während bei zehnmaliger Wiederholung erwartungsgemäß einige wenige

Augensummen und keine einzige Augensumme mehr als einmal auftraten, zeigt sich bei 30 Wiederholungen eine Häufung bei den wahrscheinlichen Augensummen in der Mitte der Verteilung. Bei 642 Wiederholungen ähnelt die Verteilung einer Normalverteilung (Tabelle 9) bzw. der Verteilung in Tabelle 7.

Tabelle 8: Verteilung der Augensummen bei viermaligem Würfeln eines Würfels mit n= 10, n=30, n= 50, n= 100 und n= 642 Wiederholungen

Würfelsummen	n= 10	n= 30	n= 50	n= 100	n= 642
Σ 4	0	0	0	0	0
Σ 5	0	0	0	2	1
Σ 6	0	0	0	2	3
Σ 7	0	0	0	3	9
Σ 8	0	0	2	5	17
Σ 9	1	1	1	1	34
Σ 10	1	0	1	4	42
Σ 11	1	3	0	5	43
Σ 12	1	4	5	13	63
Σ 13	1	2	6	11	62
Σ 14	1	3	6	9	74
Σ 15	0	6	8	13	75
Σ 16	1	4	8	8	61
Σ 17	0	3	5	5	53
Σ 18	1	2	2	12	43
Σ 19	1	0	3	3	31
Σ 20	0	0	3	2	16
Σ 21	1	1	0	2	8
Σ 22	0	0	0	0	5
Σ 23	0	1	0	0	1
Σ 24	0	0	0	0	1
Summen	**10**	**30**	**50**	**100**	**642**

Tabelle 9: Grafische Darstellung des Versuchs aus Tabelle 8 bei n= 10, n=30 und n= 642 Wiederholungen

n= 10	n= 30	n= 642
		

Die Stichprobenauswahl und die Planung der Stichprobengröße sind zusammenfassend entscheidend dafür, ob die Kennwerte in der Stichprobe die Kennwerte in der Grundgesamtheit repräsentieren. Bei einer Zufallsauswahl wird angenommen, die Kennwerte der Stichprobe folgen den Werten in der Grundgesamtheit hinsichtlich mittlerer Ausprägung und Verteilung. Dennoch können auch Zufallsstichproben die Kennwerte und die Verteilung der Kennwerte verzerrt darstellen. Je größer eine Zufallsstichprobe, desto näher ist der Stichprobenkennwert am wahren Wert in der Grundgesamtheit und desto stärker ähnelt die Verteilung in der Stichprobe einer Normalverteilung. Im Grenzwerttheorem wird dies bereits bei einer Stichprobengröße von n> 30 angenommen (Beller, 2008).

7.4 Zusammenfassung und Aufgaben

Die Stichprobenrekrutierung hat Auswirkungen auf die Aussagekraft wissenschaftlicher Studien und auf die Einsatzmöglichkeiten statistischer Analyseverfahren. Repräsentative Stichproben werden in zufallsbasierten Auswahlverfahren erzeugt. Für die Ziehung einer Zufallsstichprobe muss die Grundgesamtheit namentlich bekannt sein. Aus der Namensliste werden auf Zufallsbasis, „per Hand“ oder mit statistischer Analysesoftware, Studienteilnehmer ausgewählt. Eine einfache Zufallsauswahl wählt aus einer bekannten Grundgesamtheit eine Stichprobe. Die geschichtete Zufallsauswahl ergibt eine Merkmalsverteilung in der Stichprobe, beispielsweise von Wohnort, Altersgruppe, Geschlecht, die (weitgehend) der Verteilung in der Grundgesamtheit entspricht. In jeder Merkmalsgruppe (Schicht) erfolgt die Zufallsauswahl aus einer namentlich bekannten Grundgesamtheit. Klumpen- und mehrstufige Auswahl sind zufallsbasierte Verfahren für eine namentlich nicht bekannte Grundgesamtheit. Die Zufallsauswahl erfolgt ausgehend von organisatorischen Einheiten (Klumpen) wie Städten, Hochschulen, Betrieben, die bekannt sind. Zufallsbasierte Verfahren sind Voraussetzung für viele statistische Analyseverfahren. Eine nicht zufallsgesteuerte Auswahl erfolgt, wenn in qualitativen Studien theoretisch begründet bestimmte Merkmale interessieren oder wenn die Studie nicht auf Aussagen für eine Grundgesamtheit abzielt.

Nur Zufallsstichproben ergeben in quantitativen Studien repräsentative Stichproben, die gemessene Kennwerte im Durchschnitt und Verteilung nah an den wahren werten in der Grundgesamtheit wiedergeben. Bei Stichprobengrößen über n = 30 wird bereits von einer starken Ähnlichkeit von Kennwert und Verteilung zwischen Stichprobe und Grundgesamtheit ausgegangen.

Schlüsselwörter: *Ad-hoc-Auswahl, einfache Zufallsauswahl, geschichtete Zufallsauswahl, Häufigkeitsverteilung, Klumpenstichprobe, Quotenauswahl, mehrstufige*

Zufallsauswahl, Schneeballsystem, theoriegeleitete Auswahl, zentralen Grenzwerttheorem, Zufallsstichprobe

Aufgaben zur Lernkontrolle

1. Ab wann kann eine Stichprobe als repräsentativ gelten?

2. Ordnen Sie die Stichprobenziehungen richtig zu:
   Quotenauswahl, einfache Zufallsstichprobe, geschichtete Zufallsstichprobe, Ad-hoc-Auswahl, Klumpenauswahl, Auswahl über das Schneeballsystem, theoriegeleitete Auswahl, mehrstufige Zufallsauswahl

Zufallsbasierte Stichprobenziehung	Nicht zufallsbasierte Stichprobenziehung

3. Was besagt das zentrale Grenzwertheorem?

4. Ab welcher Stichprobengröße kann bei der Zufallsstichprobe annähernd von einer Normalverteilung ausgegangen werden?

5. Ist das zentrale Grenzwerttheorem für alle Stichprobenziehungen anwendbar?

Literatur

Beller, S. (2008). *Empirisch forschen lernen Konzepte, Methoden, Fallbeispiele, Tipps* (2., überarb. Aufl ed.). Bern: Huber.

Döring, N., Bortz, J., & Pöschl, S. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften* (5. vollständig überarbeitete, aktualisierte und erweiterte Auflage ed.). Berlin Heidelberg: Springer.

Patton, M. Q. (2002). *Qualitative research & evaluation methods* (3. ed.). Thousand Oaks, Calif. [u.a.]: Sage.

8 Daten systematisieren und auswerten

Lernergebnisse:

- Aufbau und Bestandteile der Datentabelle, Spalten- und Zeilenvektoren sowie ihre Inhalte können beschrieben, ihre Bedeutung für die Analyse von Studienergebnissen kann diskutiert werden (8.1).
- Verfahren der univariaten deskriptiven Statistik können ausgehend vom Messniveau differenziert, in Beispieldatensätzen angewendet, tabellarisch zusammengestellt und Ergebnisse schriftlich berichtet werden werden (8.3, 8.4).
- Lage- und Streuungsmaße für verschiedene Messniveaus können differenziert werden. Die hinter Interquartilsbereich und Standardabweichung stehenden Informationen können erläutert und ihre Aussagekraft diskutiert werden (8.4).
- Formen von Kennwerteverteilungen können differenziert, die Lagemaße verschiedener Verteilungsformen können verglichen werden (8.4.2).
- Die Bedeutung standardisierter Werte für den Vergleich von Kennwerten und die Schätzung von Kennwertebereichen können erläutert werden (8.4.4).
- Die Normalverteilung kann beschrieben, die Wahrscheinlichkeit von Kennwertebereichen ausgehend vom arithmetischen Mitte und der Standardabweichung kann berechnet werden (8.4.5).
- Kennwerte der bivariaten Statistik können ausgehend von unterschiedlichen Messniveaus differenziert werden (8.5).
- Kreuztabellen und bivariate Korrelationskoeffizienten können anhand eines Beispieldatensatzes berechnet werden, die dahinter stehende Aussage formuliert, Schlussfolgerungen diskutiert und eine geeignete Form der tabellarischen Ergebnisdarstellung in Studien gewählt werden (8.5).

Den Fragestellungen in der empirischen Forschung wird mit Daten nachgegangen, die in Stichproben, also einem relevanten Ausschnitt der Grundgesamtheit, erhoben werden. Die Erhebung von Daten erfolgt auf der Basis gründlicher Vorüberlegungen. Erhoben werden bestimmte Merkmale von Untersuchungsobjekten, die in bestimmter Form ausgeprägt sein können (s. 4.5). Welche Art von Fragen mit erhobenen Daten beantwortet werden kann, ist neben der Repräsentativität der Stichprobe abhängig vom Studiendesign. Ursache-Wirkungs-beziehungen lassen sich nur in Studien untersuchen, in denen bei ein und derselben Stichprobe mindestens zwei Mal im Zeitverlauf Daten erhoben wurden. Es handelt sich dabei um experimentelle und Längsschnittuntersuchungen (s. 5.2). Im Rahmen der Datenerhebung in quantitativen Studien werden qualitative Merkmale und Eigenschaften in messbar gemacht. Der Prozess des Operationalisierens (also wie kann ein Merkmal mit seinen wesentlichen Facetten abgebildet werden) und des Messens (wie können qualitativen Kriterien Zahlen zugeordnet werden, so dass ein Merkmal in seiner gesamten Ausprägungsbreite erfasst werden kann?) werden quantitative Daten, also Zahlen, generiert, die anschließend ausgewertet werden.

Der quantitative und qualitative Forschungsprozess muss nachvollziehbar und transparent erfolgen. Gewonnene Daten müssen bestimmten Standards folgend ausgewertet und als Ergebnis präsentiert werden. Datenerhebung und Darstellung der Ergebnisse sind kein Selbstzweck, sondern liefern im Idealfall Antworten auf die entwickelten Fragestellungen. Die Zahlen in den Ergebnissen sind Indikatoren auf deren Grundlage Fragestellungen beantwortet werden. Zahlen zu generieren, diese Zahlen nach bestimmten Standards auszuwerten und Ergebnisse in einer, für den Laien zunächst häufig schwer durchschaubaren Logik, zu präsentieren ist nicht Ziel von Forschung. Forschung möchte Antworten auf inhaltliche Fragen geben, die in qualitativen Aussagen münden. In der quantitativen Forschung besteht daher eine besondere sprachliche Herausforderung, Zahlen und quantitative Ergebnisse (wieder) inhaltliche Bedeutung zu verleihen. Die zunächst in der Messung in quantitative Merkmalsausprägungen überführten qualitativen Eigenschaften von Merkmalsträgern müssen durch den quantitativ Forschenden in nachvollziehbare, theoretisch begründbare Aussagen (zurück-) übersetzt werden. Jedes noch so komplexe und aufwändige statistische Rechenverfahren verfolgt nur einen Zweck: Antworten auf die theoretisch entwickelte Fragestellung zu liefern. Die methodische Kompetenz eines quantitativ Forschenden zeigt sich nicht allein in der Beherrschung statistischer Methoden und ihren mathematischen Grundlagen. Sie wird erkennbar in seiner Fähigkeit, Ergebnisse auch sehr komplexer Analyseverfahren einfach und nachvollziehbar darzustellen und zu erläutern.

Die Datenauswertung beginnt bei einer Datentabelle, die in bestimmter Weise aufgebaut ist (8.1). Uni- und bivariate deskriptive (beschreibende) Statistik dienen der Beschreibung der Stichprobe ausgehend von einer inhaltlichen Fragestellung. Univariat bedeutet, es werden einzelne Variablen/ Merkmale beschrieben. Je nach Messniveau (s. 6.2.1) erfolgt die Beschreibung durch absolute und relative Häufigkeiten (8.3) oder durch Lage- und Streuungsmaße (8.4). Bi- (und multi-) variat

heißt, zwei (bi-) oder mehrere (multi-) Variablen werden ausgehend von theoretisch begründeten Fragestellungen im Zusammenhang betrachtet. Dieser Studienbrief beleuchtet von den bivariate Verfahren Kreuztabellen zur Darstellung von Zusammenhängen zwischen nominalskalierten Daten und die bivariate Korrelation, zur Ermittlung von Zusammenhängen bei intervallskaliertem Datenmaterial (8.5).

8.1 Die Datentabelle – Basis weitergehender Analysen

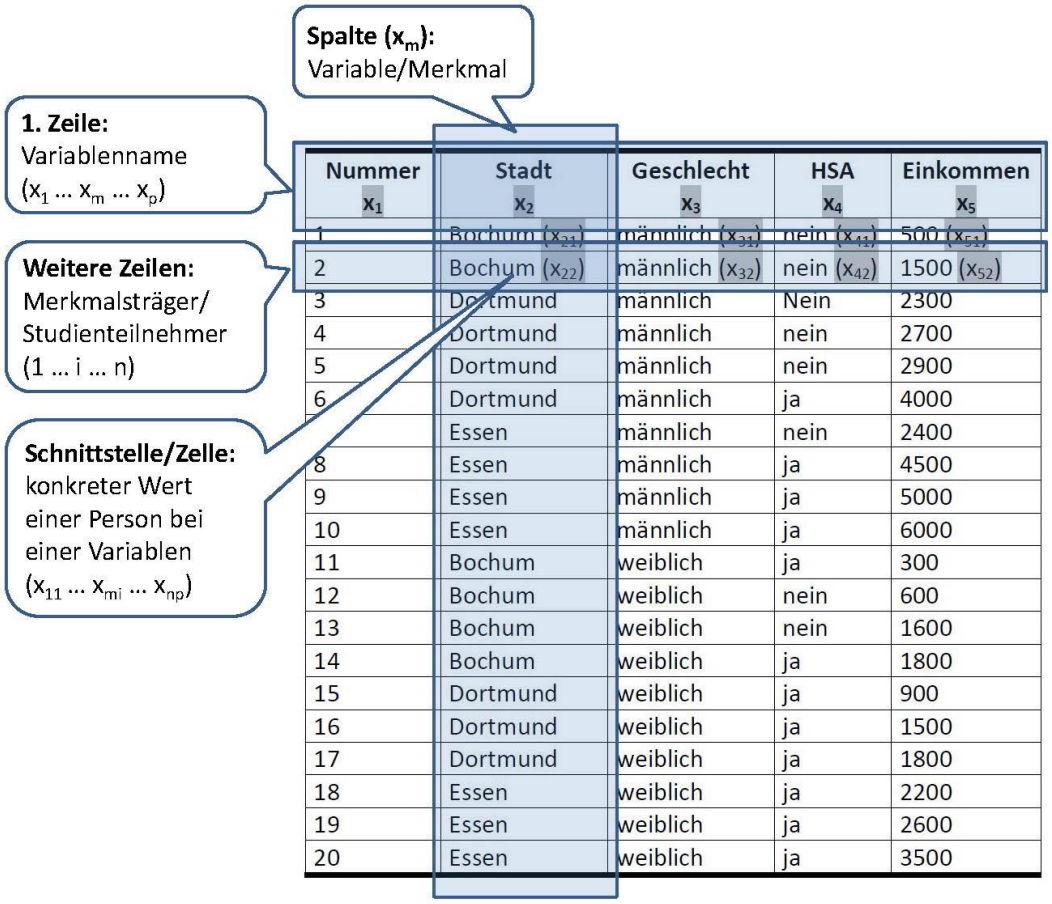
In der Datentabelle werden alle Merkmale/Variablen mit ihren unterschiedlichen Ausprägungen bestimmten Merkmalsträgern (Studienteilnehmern) zugeordnet. Sie hat die Funktion die Rohwerte nachvollziehbar darzustellen und liefert die Basis für statistische Analysen per Hand oder per Statistiksoftware.

Datentabellen nehmen Merkmalsträger in den Zeilen (n) und Variablen in den Spalten (p) auf. Eine Datentabelle enthält demnach n · p (n mal p) Werte. Die erste Zeile einer Datentabelle ist für den Variablennamen reserviert. Jeder Proband bzw. Studienteilnehmer hat genau die Anzahl an Werten, wie es Spalten, also Variablen, in einer Datentabelle gibt. Alle Werte einer Spalte, d.h. Daten aller Merkmalsträger zu einer Variablen, werden als *Spaltenvektor*, alle Werte einer Zeile, d.h. alle Daten eines Merkmalsträger, als *Zeilenvektor* bezeichnet (Eid, Gollwitzer & Schmitt, 2011).

Abbildung 22 stellt eine typische Datentabelle dar. Die qualitativen Merkmale Herkunft, Geschlecht und Hochschulabschluss werden als verbale Daten dargestellt. Für die Berechnung in Statistikprogrammen müssen diese empirischen Relative in numerische Relative übersetzt werden (s. 6.2). Die Variable Stadt könnte mit: Bochum = 1, Dortmund = 2 und Essen = 3, das Geschlecht mit: weiblich = 1, männlich = 2 und der Hochschulabschluss (HSA) mit ja = 1, nein = 2 codiert werden. Die erste Zeile umfasst den Variablennamen. Jede weitere Zeile nimmt die Daten des jeweiligen Merkmalsträgers zu den einzelnen Variablen auf (n Zeilenvektoren = Anzahl der Studienteilnehmer). Jede Spalte nimmt die Daten aller Merkmalsträger zum jeweiligen Merkmal auf.

In Abbildung 22 finden sich unter dem Variablennamen und neben einigen Merkmalsausprägungen Buchstaben-Zahlen-Kombinationen (z.B. x_{21}). Dabei handelt es sich um Indexbezeichnungen, die im weiteren Studienbrief noch häufiger Bedeutung haben werden. Sie ermöglichen eine Zuordnung von Einzelwerten zu Variablen und Merkmalsträgern. Jeder Wert eines Merkmalsträgers kann damit einer konkreten Variable zugeordnet werden – und umgekehrt. Der Index hat ebenfalls Bedeutung bei der Auflösung statistischer Formeln. Variablen und Merkmalsträger werden in den Formeln mit ihrem Index aufgenommen. Als Index für eine Variable wird der Buchstabe „x_m“ genutzt. Die zugehörige Zahl umschreibt die Variablennummer. Am konkreten Beispiel bezeichnet x_2 die Variable Stadt. Merkmalsträger werden mit dem Index „i“ beschrieben. Eine konkrete Merkmalsausprägung wird mit dem Index „x_{mi}“ beschrieben. Im konkreten Beispiel erhält die Merkmalsausprägung des 2. Merkmalsträgers bei der Variable Stadt den Index „x_{22}“. Da in einer Datentabelle p

Variablen (Spaltenvektoren) und n Merkmalsträger (Zeilenvektoren) enthalten sind, existieren insgesamt „x_{np}“ Messwerte (Abbildung 22).



Nummer x_1	Stadt x_2	Geschlecht x_3	HSA x_4	Einkommen x_5
1	Bochum (x_{21})	männlich (x_{31})	nein (x_{41})	500 (x_{51})
2	Bochum (x_{22})	männlich (x_{32})	nein (x_{42})	1500 (x_{52})
3	Dortmund	männlich	Nein	2300
4	Dortmund	männlich	nein	2700
5	Dortmund	männlich	nein	2900
6	Dortmund	männlich	ja	4000
7	Essen	männlich	nein	2400
8	Essen	männlich	ja	4500
9	Essen	männlich	ja	5000
10	Essen	männlich	ja	6000
11	Bochum	weiblich	ja	300
12	Bochum	weiblich	nein	600
13	Bochum	weiblich	nein	1600
14	Bochum	weiblich	ja	1800
15	Dortmund	weiblich	ja	900
16	Dortmund	weiblich	ja	1500
17	Dortmund	weiblich	ja	1800
18	Essen	weiblich	ja	2200
19	Essen	weiblich	ja	2600
20	Essen	weiblich	ja	3500

Abbildung 22: Skizze einer mehrstufigen Zufallsauswahl (angelehnt an Eid et al., 2011, S. 100).

In der Datentabelle werden zusammenfassend alle Merkmalsausprägungen von Merkmalsträgern (Studienteilnehmern) dargestellt. Sie enthält in den Spalten (p) die Variablen/Merkmale (x_m), in den Zeilen (n) die Merkmalsträger (i). Der Index einer konkreten Merkmalsausprägung ergibt sich aus der Kombination von Spalten- und Zeilenindex (x_{mi}). Eine Datentabelle enthält $n \cdot p$ Werte. Alle Statistikprogramme nutzen Datentabellen in der beschriebenen Form für Analysen.

8.2 Fiktive Studie: Schulabschluss, Glück und Einkommen

Für die Beispielanalysen und die Herleitung der statistischen Verfahren werden Daten einer fiktiven Beispielstudie herangezogen. Die Studie geht der Frage nach, ob Schulabschluss und Glück in Verbindung stehen. Es sollen ferner einige Nebenfragen

beantwortet werden. Dazu zählen u.a.: Haben Frauen häufiger einen Hochschulabschluss? Verdienen Frauen weniger als Männer? Gibt es einen Zusammenhang zwischen dem Wohnort und Einkommen? Verdienen Hochschulabsolventen (HSA) mehr als Menschen ohne Hochschulabschluss? Tabelle 10 enthält die Rohdatenliste, auf der alle nachfolgenden Analysen aufbauen.

Tabelle 10: Datenmatrix der Studie „Schulabschluss, Glück und Einkommen“

Nummer	**Stadt**	**Geschlecht**	**HSA**	**Note**	**Glück**	**Einkommen**
1	Bochum	weiblich	ja	4	2	300
2	Bochum	männlich	nein	4	3	500
3	Bochum	weiblich	nein	3	4	600
4	Dortmund	weiblich	ja	4	2	900
5	Dortmund	weiblich	ja	4	1	1400
6	Bochum	männlich	nein	2	5	1500
7	Bochum	weiblich	nein	2	4	1600
8	Dortmund	weiblich	ja	2	4	1700
9	Bochum	weiblich	ja	3	2	1800
10	Essen	weiblich	ja	4	4	2200
11	Dortmund	männlich	nein	2	3	2300
12	Essen	männlich	nein	3	2	2400
13	Essen	weiblich	ja	2	4	2600
14	Dortmund	männlich	nein	1	5	2700
15	Dortmund	männlich	nein	4	2	2900
16	Essen	weiblich	ja	3	3	3500
17	Dortmund	männlich	ja	2	5	4000
18	Essen	männlich	ja	2	4	4500
19	Essen	männlich	ja	1	6	5000
20	Essen	männlich	ja	2	3	6000
	Summe					**48400**

Die untersuchten Merkmale werden auf den folgenden Messniveaus erfasst (s. 6.2.1) und sind folgendermaßen codiert:

- **Nominalskalenniveau**
  - **Stadt: Bochum = 1, Dortmund = 2, Essen = 3**
  - **Geschlecht: weiblich = 1, männlich = 2**
  - **HSA: ja = 1, nein = 2 (Hochschulabsolvent)**
  - **Note: sehr gut = 1, gut = 2, befriedigend = 3, genügend = 4**
  - **Glück: Wertebereich 1 bis 6 auf einer Ratingskala von sehr unglücklich = 1 bis sehr glücklich = 6**
- **Intervallskalenniveau**
  - **Einkommen**

Zunächst werden die deskriptiven Statistiken zur Studie ausgegeben. Dabei erfolgt die Darstellung der sinnvollen Maßzahlen zu den einzelnen Messniveaus (Tabelle 11).

Tabelle 11: Messniveau und sinnvolle Maßzahlen

	Häufigkeit		Lagemaße			Streuungsmaße		
Messniveau	**n**	**h**	**Mo**	**Md**	**M**	**Range**	**IQR**	**SD**
Nominalskalenniveau	X	X	X					
Ordinalskalenniveau	X	X	X	X		X	X	
Intervallskalenniveau	X	X	X	X	X	X	X	X

Anmerkungen: n = absolute Häufigkeit, h = relative Häufigkeit
Mo = Modalwert, Md = Medianwert, M = arithmetisches Mittel
Range = absoluter Wertebereich, IQR = Interquartile Range, SD = Standardab weichung

8.3 Univariate deskriptive Statistik I – Häufigkeiten

Mit univariater deskriptiver Statistik wird die Häufigkeit einzelner Merkmalsausprägungen oder die Ausprägung einzelner Merkmale in der Stichprobe bzw. in Teilstichproben beschrieben. Die Häufigkeitsverteilungen von Merkmalen werden aus der Rohliste der Daten entwickelt. Es interessiert also, wie häufig bestimmte Merkmalsausprägungen in der Stichprobe auftreten. Prinzipiell lassen sich Häufigkeiten mit Daten jedes Messniveaus darstellen. Wirklich sinnvoll ist dies i.d.R. mit nominalskalierten, also kategorialen oder allenfalls ordinalskalierten, Daten (s. 6.2.1). Unterschieden wird die Darstellung absoluter, relativer und prozentualer sowie kumulierter Häufigkeiten. Sofern es inhaltlich notwendig ist können auch intervallskalierte Merkmale über Häufigkeiten beschrieben werden (8.3.1). Dazu erfolgt eine Zusammenfassung einzelner Werte zu Wertebereichen, die unterschiedlich häufig besetzt sein können (8.3.2).

8.3.1 Absolute, relative und kumulative Häufigkeiten bei nominalskalierten Variablen

Als *absolute Häufigkeit* (n) wird die Gesamtzahl aller Nennungen zu einer Kategorie berichtet. *Relative Häufigkeit* (h) setzt die Häufigkeit in einer Kategorie in *Relation* zur Gesamthäufigkeit, sie kann Werte zwischen 0 und 1 annehmen. Die relative Häufigkeit wird aus dem Quotienten von absoluter Häufigkeit in der Kategorie und der absoluten Gesamthäufigkeit gebildet (F. 1). Sofern von 10 Studienteilnehmern 4 Männer und 6 Frauen sind, treten Männern demnach mit einer relativen Häufigkeit von 0.4 auf. Anstelle der relativen Häufigkeit wird häufig die *prozentuale Häufigkeit* (h%) berichtet (F. 2). Es wird empfohlen entweder ausschließlich die prozentuale

Häufigkeit zu berichten oder relative und prozentuale Häufigkeit gemeinsam. Sofern prozentuale oder relative Häufigkeiten berichtet werden, *müssen* Bezugsdaten, also die absolute Häufigkeit in der jeweiligen Kategorie/ Merkmalsausprägung oder die Gesamthäufigkeit genannt werden.

Kumulierte Häufigkeiten liefern bei *ordinalskaliertem Datenmaterial* Hinweise auf die Verteilung einzelner Bereiche von Merkmalsausprägungen. Beispielsweise lässt sich erkennen, in welcher Kategorie die Mitte einer Stichprobe liegt (eine Hälfte hat einen niedrigeren, die andere Hälfte einen höheren Wert) oder welcher Wertebereich beim größten Teil der Stichprobe auftritt.

(F. 1)	$h = \frac{n_j}{n}$	h = relative Häufigkeit n_j = absolute Häufigkeit in der Kategorie n= Gesamthäufigkeit/Stichprobengröße
(F. 2)	$h\% = \frac{n_j}{n} \cdot 100$	

Formeln: Leonhart (2013)

Häufigkeiten können zu allen Merkmalen/Variablen der fiktiven Beispielstudie berichtet werden. Weniger sinnvoll aber möglich ist die Häufigkeitsdarstellung mit intervallskalierten Daten, in der Beispielstudie ist dies das Merkmal Einkommen. Hier kommt jede spezifische Merkmalsausprägung nur jeweils einmal vor. Allerdings ist es möglich einzelne Werte zu Wertebereichen zusammenzufassen, also eine intervallskalierte Variable in eine nominalskalierte überführt wird. Zunächst erfolgt daher die Darstellung der Häufigkeitsmaße nominal- und ordinalskalierter Merkmale. Anschließend wird ein Verfahren eingeführt, mit dem auch Häufigkeiten intervallskalierte Daten sinnvoll dargestellt werden können (8.3.2).

Eine mögliche Darstellung der Häufigkeiten ausgehend von der Rohdatentabelle (Tabelle 10) erfolgt in Tabelle 12. Dabei wird auf die Darstellung der relativen Häufigkeit verzichtet und entsprechende Prozentwerte angegeben. Die Gestaltung von Tabellen in diesem Studienbrief orientiert sich an den Richtlinien zur Manuskriptgestaltung der Deutschen Gesellschaft für Psychologie (DGPs) (2007).

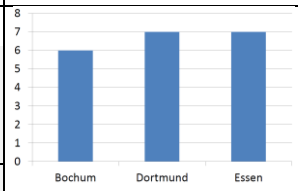
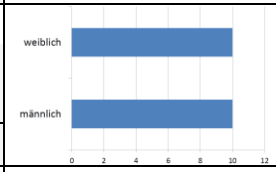
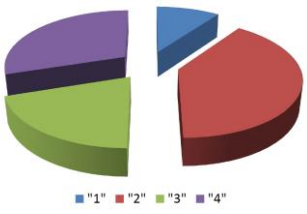
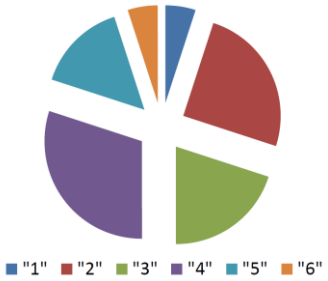
Ausgehend von den Ergebnissen in Tabelle 12 kann die Stichprobe in einer wissenschaftlichen Arbeit wie folgt beschrieben werden:

> In Bochum leben weniger Studienteilnehmer als in Essen und Dortmund. Männer und Frauen sind in gleich großer Zahl vertreten. Ein Hochschulstudium schlossen 60 Prozent (12) Teilnehmer ab.
>
> Bei den Merkmalen Abschlussnote und Glück werden neben den absoluten und prozentualen Häufigkeiten auch die kumulierten Häufigkeiten berichtet. Die kumulierten Häufigkeiten zeigen den Wertebereich, den ein großer bzw. ein kleiner Teil der Stichprobe einnimmt. Bei den Abschlussnoten überwiegt die Note 2 gefolgt von der 4, 3 und 1. Die guten und sehr guten sowie befrie-

digende und genügende Noten kommen in der Stichprobe in etwa gleich oft vor. Etwa die Hälfte der Stichprobe weist Glückswerte von mehr als 3 auf, ist also verhältnismäßig glücklich. Sehr glücklich und sehr unglücklich erleben sich nur wenige Teilnehmer.

Neben der tabellarischen Darstellung absoluter und relativer Häufigkeiten kann auch die grafische Abbildung der Ergebnisse erfolgen. Für Häufigkeiten eignen sich Balken-, Kreis- oder Tortendiagramme (Tabelle 12). Abbildungen sollten generell sparsam und zur Darstellung zentraler Antworten auf Fragestellungen verwendet werden.

Tabelle 12: Häufigkeiten von Stadt, Geschlecht, Hochschulabsolvent, Abschlussnote und Glück

Merkmal	**n (%)**	**kum n (kum %)**	**Abbildung**
Stadt			
Bochum	6 (30)		
Dortmund	7 (35)		
Essen	7 (35)		
Summe	20 (100)		
Geschlecht			
weiblich	10 (50)		
männlich	10 (50)		
Summe	20 (100)		
Hochschulabsolvent			
ja	12 (60)		
nein	8 (40)		
Summe	20		
Schulnote			
1	2 (10)	2 (10)	
2	8 (40)	10 (50)	
3	4 (20)	14 (70)	
4	6 (30)	20 (100)	
Summe	20 (100)		
Glück			
1	1 (5)	1 (5)	
2	5 (25)	6 (30)	
3	4 (20)	10 (50)	
4	6 (30)	16 (80)	
5	3 (15)	19 (95)	
6	1 (5)	20 (100)	
Summe	20 (100)		

8.3.2 Absolute, relative und kumulative Häufigkeiten bei intervallskalierten Variablen

Die Angabe der Häufigkeit in einzelnen Merkmalsausprägungen bei intervallskalierten Merkmalen ist ausgehend von Rohwerten wenig erkenntnisbringend, anschaulich und sinnvoll. Zwischen einzelnen Messwerten sind definitorisch unendlich viele Werte möglich. Bei Messwerten finden sich, je nach Stichprobengröße, wahrscheinlich nur wenige Merkmalsausprägungen, in denen mehr als eine Angabe enthalten ist. Im Beispieldatensatz (Tabelle 10) ist beispielsweise kein einziger Einkommenswert mehr als einmal vorhanden. Sofern es inhaltlich sinnvoll ist und zur Beantwortung aufgeworfener Forschungsfragen beiträgt, können auch von intervallskalierten Variablen Häufigkeiten berichtet werden. Dazu erfolgt eine Zuordnung einzelner Werte zur Wertebereichen. Dies kann intuitiv und willkürlich erfolgen oder über Faustregeln, mit denen die geeignete Anzahl Kategorien ausgehend von der Anzahl der Werte in der Stichprobe bestimmt werden können. Der Prozess der Umwandlung intervallskalierter Merkmale in kategoriale kann in folgenden Schritten erfolgen:

1. Bestimmung der geeigneten Kategorienanzahl (intuitiv oder über Faustformel, s. (F. 3). Das Ergebnis ist zumeist keine ganze Zahl, es muss daher auf- oder abgerundet werden.

2. Bestimmung des maximalen Wertebereichs der erhobenen Daten (Range).

3. Bestimmung des Wertebereichs in jeder Kategorie – der Kategorienbreite. Dazu erfolgt die Division von maximalem Wertebereich (s. Punkt 2 dieser Aufzählung) durch die errechnete (oder intuitiv bestimmte) Kategorienzahl. Dieses Ergebnis ist ebenfalls zumeist keine ganze Zahl, so dass auch hier gerundet werden muss.

4. Festlegen der Kategorien anhand der Wertebereiche (3.), Auszählung der Häufigkeiten.

Alle Kategorien müssen dieselbe Breite haben, also denselben Wertebereich umfassen und sie müssen überschneidungsfrei sein. Eine Kategorie darf also keine Werte einer anderen beinhalten. Das Vorgehen wird exemplarisch am Merkmal Einkommen des Musterdatensatzes beschrieben.

Einkommen (EUR), sortiert anhand Tabelle 10**:**

300, 500, 600, 900, 1400, 1500, 1600, 1700, 1800, 2200, 2300, 2400, 2600, 2700, 2900, 3500, 4000, 4500, 5000, 6000

1. Bestimmung der geeigneten Kategorienzahl.

(F. 3) $m = 1 + 3.32 \cdot \log(n)$

$m = 1 + 3.32 \cdot \log(20)$

$m = 1 + 3.32 \cdot 1.30$

$m = 1 + 4.32$

$m \sim 5$

m = Kategorienzahl
log = dekadischer Logarithmus
n = Gesamthäufigkeit/ Stichprobengröße

Ausgehend von 20 Werten beim Merkmal Einkommen erscheint die Bildung von fünf Kategorien sinnvoll.

2. Bestimmung des maximalen Wertebereichs (Range).

(F. 4) $Range = x_{i\,max} - x_{i\,min} + 1$

$x_{i\,max}$ = höchster Messwert
$x_{i\,min}$ = niedrigster Messwert

$Range = 6000 - 300 + 1$

$Range = 5701$

Der maximale Wertebereich umfasst 5.701 mögliche Werte.

3. Bestimmung Kategorienbreite.

Zur Bestimmung der Kategorienbreite wird der Range durch die errechnete Kategorienzahl geteilt. In einer Kategorie könnten sich demnach 1.140,20 Werte befinden. Diese Zahl ist ungeeignet, eine Rundung auf 1.200 als Kategorienbreite ist empfehlenswert. Bei dieser Kategorienbreite würden sich über 5 Kategorien 6.000 mögliche Werte ergeben, mehr als im Range enthalten sind. Daher werden Unter- und/oder Obergrenze des Range erweitert. Dies ist erlaubt, weil so nach wie vor alle Messwerte im Range enthalten sind. Möglich wäre es in diesem Fall den Wertebereich bei 200 zu beginnen und bei 6.199 zu beenden.

4. Festlegung der Kategorien und der zugeordneten Wertebereiche, Auszählung der Häufigkeiten.

Formeln: Leonhart (2013)

Tabelle 13 zeigt das Ergebnis der Umwandlung eines intervallskalierten Merkmals in ein nominalskaliertes Merkmal. Für wissenschaftliche Arbeiten wird die folgende Formulierung empfohlen:

Mehr als die Hälfte der Studienteilnehmer verdient weniger als 2.600 EUR im Monat. Spitzenverdienste zwischen 3.800 und 6.000 EUR erzielt lediglich ein Fünftel. Ebenso viele Studienteilnehmer erhalten weniger als 1.400 EUR im Monat.

Tabelle 13: Häufigkeiten von Einkommenskategorien

Merkmal	**n (%)**	**kum n (kum %)**
Einkommen (EUR)		
200-1.399	4 (20)	4 (20)
1.400-2.599	8 (40)	12 (60)
2.600-3.799	4 (20)	16 (80)
3.800-4.999	2 (10)	18 (90)
5.000-6.199	2 (10)	20 (100)
Summe	**20 (100)**	

8.4 Univariate deskriptive Statistik II – Lage und Streuung

Bei ordinal- und intervallskalierten Daten erfolgt die Beschreibung über Maße der zentralen Tendenz (Lagemaße) (8.4.1) und Streuungsmaße (Variabilität) (88.4.3) (Beller, 2008).

8.4.1 Lagemaße: Modus, Median, arithmetisches Mittel

Lagemaße geben einen Durchschnittswert wieder, der aus der Verteilung von Merkmalsausprägungen errechnet wird (Beller, 2008). Unterschieden werden drei Lagemaße (Beller, 2008; Schäfer, 2016):

- Modalwert (Modus, Mo): Der Modalwert ist diejenige Merkmalsausprägung, die innerhalb eines Spaltenvektors (Merkmal) am häufigsten gemessen wird. Er ist stabil gegenüber Ausreißerwerten und kann für alle Messniveaus angewendet werden.

- Medianwert (Median, Md): Der Median teilt die Stichprobe in zwei gleichgroße Hälften. Eine Hälfte nimmt Merkmalsausprägungen unter, die andere Hälfte über dem Median ein. Er ist stabil gegenüber Ausreißerwerten und kann für ordinal- und intervallskaliertes Datenmaterial verwendet werden.

- Arithmetisches Mittel (Mittelwert, M): Das aritmetische Mittel umschreibt die Summe aller Merkmalswerte dividiert durch die Anzahl der Merkmals-

träger (n). Es ist empfindlich gegenüber Ausreißerwerten und ist anwendbar für Daten mit hohem Messniveau (Intervall- und Verhältnisskalenniveau).

Der *Modalwert* wird von der am häufigsten besetzten Kategorie beschrieben. Tabelle 14 enthält die Häufigkeit der verschiedenen Abschlussnoten. Im Beispiel liegt der Modalwert bei der Note „2“, die Mehrzahl der Studienteilnehmer hat also beim Schulabschluss eine „2“ erzielt. Beim Modalwert spielt keine Rolle, ob Extremwerte hinzukommen, beispielsweise Studienteilnehmer, die mit 5 oder 6 abschlossen. Ein Modalwert ist nur bestimmbar, wenn genau eine Kategorie am häufigsten besetzt ist. Beller (2008) schlägt bei zwei *nebeneinander liegenden*, gleich häufig mehrheitlich besetzten Kategorien vor, als Modalwert die Mitte zwischen beiden Kategorien als Modalwert zu berichten. Dieses Vorgehen ist nicht möglich, wenn Moduskategorien nicht nebeneinander liegen. In der Beispielstudie wären das beispielsweise „2“ und „4“. Wenn, wie in Tabelle 13, Kategorien durch Wertebereiche beschrieben werden, ist Beller (2008) zufolge der Modalwert in der Mitte der am häufigsten besetzten Kategorie. Im Beispiel ist dies der Bereich zwischen 1.400 und 2.599 EUR, die Mitte liegt hier bei 2.000 EUR.

Tabelle 14: Modalwert des Merkmals Schulnote

Merkmal	**n (%)**	**kum n (kum %)**
Schulnote		
1	2 (10)	2 (10)
2	**8 (40)**	**10 (50)**
3	4 (20)	14 (70)
4	6 (30)	20 (100)
Summe (N)	**20 (100)**	

Der *Medianwert* kennzeichnet die Mitte einer Verteilung, indem eine nach Werten geordnete Stichprobe genau in zwei Hälften geteilt wird. In der Beispielstudie liegen für das Einkommen genau 20 Werte vor. Die „Mitte“ der Verteilung liegt demnach bei den Messwerten zwischen dem 10. und 11. Platz der Verteilung. Er errechnet sich bei einer geraden Anzahl von Werten (durch „2“ teilbare Anzahl) aus dem arithmetischen Mittel der benachbarten Werte: (2.200 + 2.300)/2 = 2.250 (F. 5) (Abbildung 23).

(F. 5)	**Median gerades N** $Md = \frac{x_{\frac{N}{2}} + x_{\frac{N}{2}+1}}{2}$	$x_{\frac{N}{2}}$ = Wert an der Hälfte der Verteilung $x_{\frac{N}{2}+1}$ = Wert, der dem Wert an der Hälfte der Verteilung folgt
(F. 6)	**Median ungerades N** $Md = x_{\frac{N+1}{2}}$	$x_{\frac{N+1}{2}}$ = bei ungeraden Stichproben wird dem N ein Wert addiert (aus 19 wird 20), der Median liegt dann an der Hälfte der Verteilung (am Beispiel beim 10. Wert

Formeln: Leonhart (2013)

Voraussetzung für die Berechnung des Median ist eine nach Größe des Messwerts geordnete Verteilung ordinal- oder intervallskalierter Daten. Dabei spielt es keine Rolle, ob auf- oder absteigend sortiert wurde. Statistikprogramme berechnen den Median auch aus einer ungeordneten Reihung der Werte, hier muss also keine „händische“ Sortierung erfolgen.

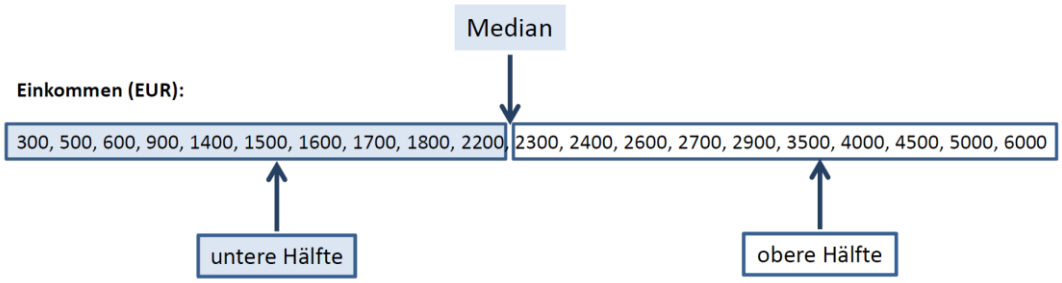


Abbildung 23: Grafische Darstellung des Medianwerts der Variable Einkommen (eigene Darstellung)

Der Median ist robust und unsensibel gegenüber Ausreißern am unteren und oberen Ende einer Verteilung. Im Beispiel würde der Median konstant bleiben, auch wenn der höchste Wert nicht 6.000 EUR wäre, sondern 600.000 EUR.

Das *arithmetische Mittel* oder der Durchschnittswert ist ein häufig auch im Alltag angewendetes Lagemaß. Es wird aus der Summe aller Merkmalsausprägungen geteilt durch die Anzahl der Merkmalsausprägungen berechnet (F. 7). Die Daten müssen mindestens auf Intervallskalenniveau vorliegen.

(F. 7) **Arithmetisches Mittel**

$$M = \frac{\Sigma_{i=1}^{N} x_i}{N} = \Sigma_{i=1}^{N} x_i \cdot \frac{1}{N}$$

$\Sigma_{i=1}^{N} x_i$ = Summe der Messwerte einer Variablen

N = Anzahl der Werte

Formel: Leonhart (2013)

Am Einkommensbeispiel liegt das arithmetische Mittel bei 48.400 (Tabelle 10) geteilt durch 20 (Anzahl der Messwerte), dies entspricht 2.420 EUR. Im Durchschnitt verdienen die Studienteilnehmer 2.420 EUR. Das arithmetische Mittel ist empfindlich gegenüber Ausreißerwerten. Würde in der Beispielstudie der höchste Wert nicht 6.000, sondern 60.000 EUR betragen, läge das arithmetische Mittel bei 5.120 EUR.

Informationen zu den für den Bericht von Lagemaßen erforderlichen Messniveaus liefert Tabelle 11. Alle Lagemaße sind nur aussagekräftig, wenn auch die Streuung der Werte (s. 8.4) berichtet wird. Ein arithmetisches Mittel von Einkommenswerten zwischen 1.200 und 1.600 EUR spiegelt daher eher die reale Verteilung des Merkmals in der Stichprobe, als eines, bei dem sich das Einkommen in der Stichprobe zwischen 200 und 60.000 EUR bewegt. Streuungsmaße geben Hinweise auf die Breite der Streuung von Merkmalen. Breite Streuungen und große Wertebereiche lassen sich nicht verhindern, sie schränken jedoch die Aussagekraft der Lagemaße und aller weiteren statistischen Analysen ein. **Neben den Lagemaßen muss die Streuung und die damit in Verbindung stehenden möglicherweise kritischen Schlussfolgerungen für die Aussagekraft einer Studie berichtet werden** (8.4.3).

8.4.2 Beziehung zwischen den Lagemaßen und Art der Verteilung

Anhand von Unterschieden oder der Gleichheit von Modus, Median und arithmetischen Mittel kann bestimmt werden, ob eine Verteilung „schief“ oder symmetrisch ist. Schief verteilt sind ungleich bzw. ungerecht verteilte Güter, wie Einkommen oder Vermögen. Beim Einkommen beispielsweise verdient die übergroße Mehrheit vergleichsweise kleine Summen, während wenige Menschen sehr hohe Einkommen erzielen.

Eine Verteilung ist symmetrisch, wenn rechts und links des arithmetischen Mittels dieselbe Anzahl an Studienteilnehmern Werte im selben Abstand vom arithmetischen Mittel hat, wobei die Mehrzahl eng um den Mittelwert verteilt sein sollte (Abbildung 24, links, siehe auch 8.4.5).

Nicht symmetrische Verteilungen können entweder linkssteil/rechtsschief oder rechtssteil/linksschief sein. Bei symmetrischen Verteilungen nehmen arithmetisches Mittel, Median und Modus denselben Wert ein. Bei linkssteilen/rechtsschiefen Verteilungen liegen Ausreißer „nach oben“ vor, die Mehrzahl hat dagegen niedrige und mittlere Merkmalsausprägungen. Beim Gehaltsbeispiel (Tabelle 11) erhielten zahlreiche Menschen mittlere und niedrige Einkommen, nur wenige dagegen hohe und

sehr hohe Einkommen. Der Modus, also der am häufigsten vorkommende Wert, hat dabei von allen Lagemaßen die niedrigste Merkmalsausprägung, das arithmetische Mittel die höchste.

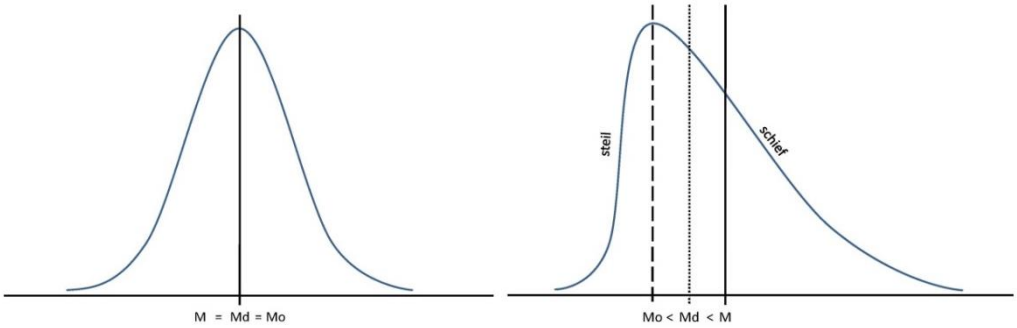


Abbildung 24: Symmetrische (links) und linkssteil/rechtsschiefe (rechts) Verteilung mit Verhältnissen der Lagemaße (eigene Darstellung)

Bei rechtssteil/linksschiefen Verteilungen ist die Mehrzahl der Werte hoch oder sehr hoch ausgeprägt, wenige gering oder auf mittlerem Niveau (Abbildung 25). Im Gehaltsbeispiel hätten viele Menschen ein hohes und sehr hohes Einkommen, wenige mittlere und niedrige. Der Modus ist bei einer linksschiefen Verteilung der höchste, das arithmetische Mittel das am niedrigsten ausgeprägte Lagemaß.

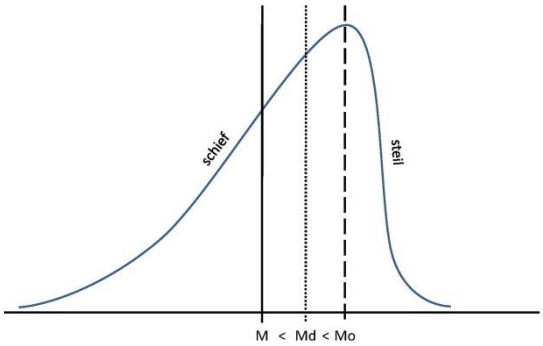


Abbildung 25: Rechtssteil/linksschiefe Verteilung mit Verhältnissen der Lagemaße (eigene Darstellung)

8.4.3 Streuungsmaße: Range, Interquartilsbereich, Standardabweichung

Die Streuung der Merkmalsausprägung ist ein wesentlicher Parameter bei der Bewertung der Lagemaße. Bei ein und derselben zentralen Tendenz können Werte unterschiedlich breit streuen. Je breiter die Streuung von Merkmalsausprägungen ist, desto weniger sagt ein Mittelwert über die Verteilung von Merkmalsausprägungen aus. Abbildung 26 stellt anhand von Daten zum täglichen Zigarettenkonsum die Daten von jeweils 10 Probanden aus den Jahren 1995 und 2016 dar. In beiden Jahren wurden täglich durchschnittlich 4,2 Zigaretten geraucht. 2016 ist verglichen mit 1995 nicht nur der Wertebereich größer (Range 2016: 2 bis 7; 1995: 3 bis 6), sondern auch eine markante Häufung von Merkmalsausprägungen in der Mitte findet

sich nicht mehr. Durch die größere Streuung der Merkmalsausprägungen ist die Genauigkeit und Aussagekraft des arithmetischen Mittels von 2016 geringer als 1995.

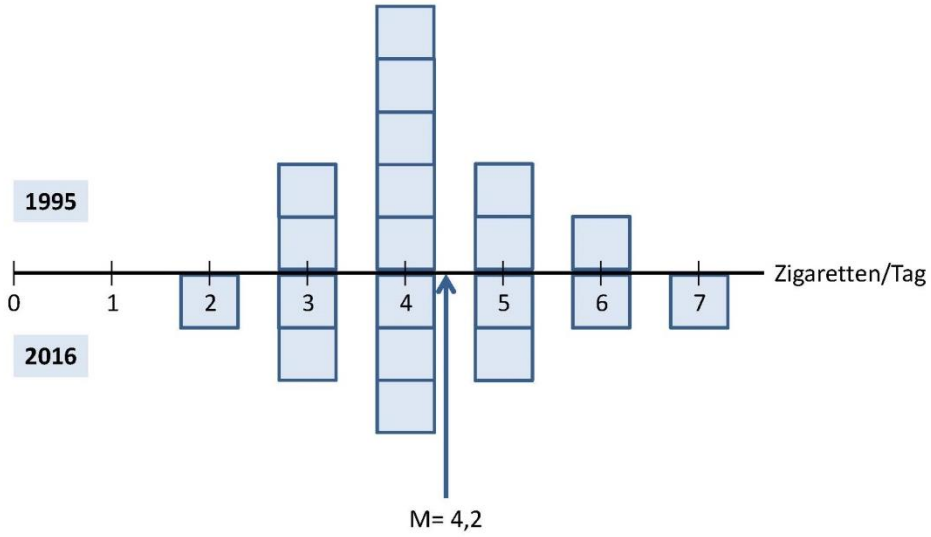


Abbildung 26: Verschiedene Verteilungen bei identischem arithmetischen Mittel (angelehnt an Beller, 2008, S. 69).

Folgende Streuungsmaße werden berichtet:

- **der Range zum Modus, (Median und arithmetischen Mittel)**
- **der Interquartilsbereich zum Median und**
- **die Standardabweichung zum arithmetischen Mittel.**

Der *Range* umfasst den gesamten Bereich den Merkmalsausprägungen in der Stichprobe einnehmen. Er liefert Informationen zum Modus, Median und arithmetischem Mittel.

Der *Interquartilsbereich* zum Medianwert grenzt die Werte zwischen dem 25 und 75 Prozent Perzentil ein. 50 Prozent der Studienteilnehmer haben Werte innerhalb des Interquartilsbereichs, 25 Prozent liegen mit ihrer Merkmalsausprägung unterhalb, 25 Prozent oberhalb. Die Grenzen des Interquartilsbereichs lassen sich mit den Formeln (F. 5 (F. 6 berechnen, indem jeweils der Median der unteren und der oberen Hälfte berechnet wird. Ausgehend von der Beispielstudie zum Einkommen lässt sich der Interquartilsbereich grafisch veranschaulichen (Abbildung 27).

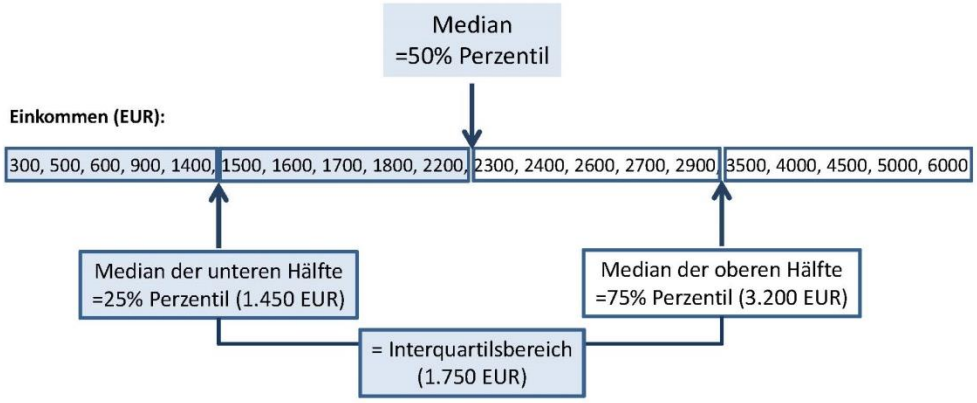


Abbildung 27: Streuung um den Median: Grafische Veranschaulichung des Interquartilsbereichs (eigene Darstellung)

Eine übliche Darstellungsform von Median und Interquartilsbereich erfolgt in Boxplotdiagrammen. Die „Box“ stellt dabei den Interquartilsbereich zwischen dem 25 und 75 Prozent-Perzentil dar. Der fette Balken in der Box zeigt den Median bzw. das 50 Prozent-Perzentil. Die oben und unten erkennbaren „Antennen“ werden als Whisker-Plots bezeichnet. Sie umfassen entweder den gesamten Wertebereich, den Bereich zwischen 2,5 und 97,5 Prozent-Perzentil oder den Bereich des 1,5-fachen Interquartilsbereichs. Sofern der Whisker-Plot nicht den gesamten Wertebereich umschließt, werden Ausreißerwerte in die Darstellung aufgenommen. Sie erscheinen als Punkte, Sterne, kleine Kreise mit Zahlen. Die Zahlen weisen auf den Zeilenvektor (also den Fall) hin, der diesen Wert einnimmt (Abbildung 28).

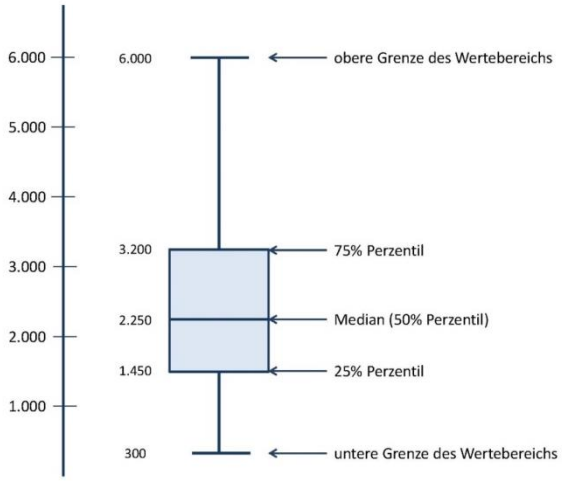


Abbildung 28: Beschriftetes Boxplotdiagramm auf Basis des Merkmals Einkommen in der Beispielsstudie (eigene Darstellung)

Die *Standardabweichung* (s) ist das Streuungsmaß des arithmetischen Mittels. Etwa 68% der Studienteilnehmer haben Messwerte, die eine Standardabweichung unter und über dem berechneten arithmetischen Mittel liegen. Je größer der Wert der Standardabweichung im Verhältnis zum arithmetischen Mittel ist, desto größer ist

die Streuung der Messwerte in der Stichprobe und desto weniger Aussagekraft hat das arithmetische Mittel. Voraussetzung für die Berechnung der Standardabweichung sind wie beim arithmetischen Mittel intervallskalierte Daten (Beller, 2008) (F. 9). Basis für die Berechnung der Standardabweichung ist die Bestimmung der *Varianz* (s^2) (F. 8). Unter Varianz wird die Summe der quadrierten Abweichungen der einzelnen Messwerte von arithmetischen Mittel verstanden. Durch die Quadrierung haben große Abweichungen von Messwert und arithmetischem Mittel einen größeren Einfluss auf die Varianz als kleine. Die Varianz hat durch die Quadrierung der Abweichungen nicht mehr dasselbe Maß, wie die Messwerte, sie bezieht sich auf quadrierte Messwerte und das arithmetische Mittel aus den quadrierten Messwerten. Erst durch das Ziehen der Quadratwurzel aus dem Wert der Varianz sind Rückschlüsse auf die Streuung der Messwerte möglich (Beller, 2008). Die Wurzel aus der Varianz ist die Standardabweichung.

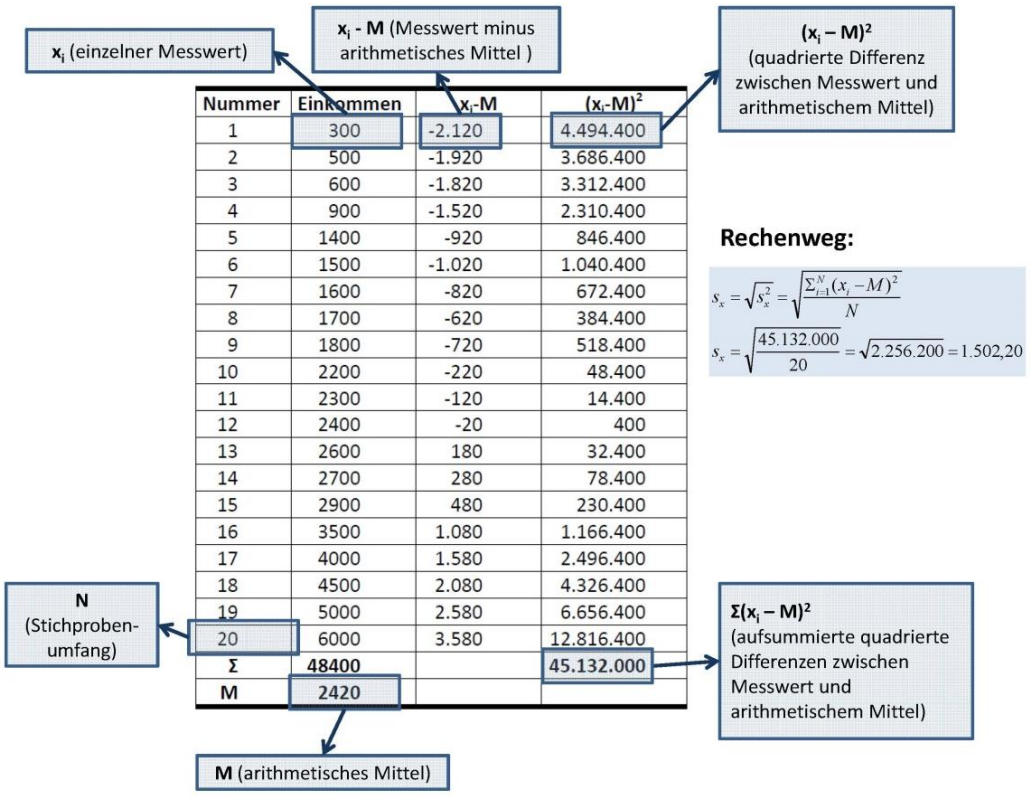
(F. 8)	**Varianz** $s_x^2 = \frac{\sum_{i=1}^{N}(x_i - M)^2}{N}$	s^2 = Summe quadrierter Abweichungen bezogen auf N (Varianz) x_i = Messwert M = arithmetisches Mittel N = Stichprobengröße
(F. 9)	**Standardabweichung** $s_x = \sqrt{s_x^2} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - M)^2}{N}}$	s = Standardabweichung

Formeln: Leonhart (2013)

Anhand der Formeln (F. 8) und (F. 9) wird die Standardabweichung errechnet. Exemplarisch erfolgt dies für die intervallskalierte Variable Einkommen der Beispielstudie (Auszug aus Tabelle 10). Zunächst wird für jeden Messwert die Differenz zum Mittelwert berechnet [x_i – M]. Anschließend wird jede errechnete Differenz quadriert [$(x_i - M)^2$] und die Ergebnisse aufsummiert [$\sum_{i=1}^{N}(x_i - M)^2$]. Diese Summe (45.132.000) wird durch die Anzahl der Messwerte (N=20) dividiert, das Ergebnis ist die Varianz (s^2): 2.256.200. Die Quadratwurzel der Varianz ergibt schließlich die Standardabweichung (s): 1.502,20 (Abbildung 29).

Grundsätzlich dürfen arithmetisches Mittel und Standardabweichung nur bei intervallskaliertem Datenmaterial berechnet werden. In der Beispielsstudie sind die Note und die Variable Glück als ordinalskaliert angegeben. Als Lagemaß für beide Merkmale darf der Median, als Streuungsmaß der Interquartilsbereich berechnet werden (s. Tabelle 11). In der Psychologie und den Sozialwissenschaften wird, nicht ohne Kritik, vom Intervallskalenniveau der Daten ausgegangen, die mit Ratingskalen erhoben werden (6.2.2). „Glück“ in der Beispielstudie ist eine Variable, die per Fragebogen mit Ratingskalen abgebildet wird. Somit kann auch für das Glück die Berechnung von arithmetischem Mittel und Standardabweichung erfolgen. In sozialwissenschaftlichen Veröffentlichungen wird dieser Annahme überwiegend gefolgt und für

mit Ratingskalen erhobene Merkmale arithmetisches Mittel und Standardabweichung berechnet.



Nummer	Einkommen	x_i-M	$(x_i-M)^2$
1	300	-2.120	4.494.400
2	500	-1.920	3.686.400
3	600	-1.820	3.312.400
4	900	-1.520	2.310.400
5	1400	-920	846.400
6	1500	-1.020	1.040.400
7	1600	-820	672.400
8	1700	-620	384.400
9	1800	-720	518.400
10	2200	-220	48.400
11	2300	-120	14.400
12	2400	-20	400
13	2600	180	32.400
14	2700	280	78.400
15	2900	480	230.400
16	3500	1.080	1.166.400
17	4000	1.580	2.496.400
18	4500	2.080	4.326.400
19	5000	2.580	6.656.400
20	6000	3.580	12.816.400
Σ	48400		45.132.000
M	2420		

Abbildung 29: Berechnung der Standardabweichung mit Auflösung der Formel (eigene Darstellung)

Lage- und Streuungsmaße werden in Studien tabellarisch berichtet (Tabelle 15). Eine grafische Darstellung ist auch hier per Balkendiagramm oder Boxplot möglich, ist aber nur sinnvoll, wenn mit der Grafik wesentliche Fragestellungen beantwortet werden.

Für wissenschaftliche Arbeiten dient der folgende Text der Beschreibung der Merkmale:

> Das mittlere Einkommen liegt bei 2.420 EUR bei mäßiger Streuung der Werte und unterhalb der Mitte des Wertebereichs in der Stichprobe. Untere und mittlere Einkommen dominieren demnach. Mehrheitlich erzielten die Befragten eine gute Abschlussnote, die Hälfte der Befragten hat dabei eine bessere Note als 2,5 erzielt. Gemessen am möglichen Wertebereich sind die Studienteilnehmer vergleichsweise glücklich. Die Mehrheit gibt mit „4“ (ziemlich glücklich) einen oberhalb der Mitte des Wertebereichs an.

Tabelle 15: Deskriptive Statistik der Merkmale Einkommen, Abschlussnote und Glück

Merkmal	M (SD)	Md (IQR)	Mo (Range)
Einkommen	2.420,00 (1.502,20)	2.250,00 (3.050,00)	- (300-6.000)
Abschlussnote	- (-)	2,50 (2,00)	2 (1-4)
Glück	3,40 (1,28)	3,50 (4,00)	4 (1-6)

Anmerkungen: M = arithmetisches Mittel, SD = Standardabweichung, Md = Median, IQR = Interquartilsbereich, Mo = Modalwert

8.4.4 z-Standardisierung

Merkmalsausprägungen von Variablen mit verschiedenen Wertebereichen sind anhand ihrer Rohwerte nicht vergleichbar. Die z-Standardisierung von Werten auf Grundlage von Mittelwert und Standardabweichung ermöglicht den Vergleich auch sehr unterschiedlicher Merkmale. Dabei gehen nicht mehr die ursprünglichen Messwerte in die Betrachtung ein, sondern ihre Relation zu arithmetischem Mittel und Standardabweichung. *Ein z-Wert drückt aus, wie viele Standardabweichungen ein Messwert vom Mittelwert entfernt ist.* Ein Wert, der genau dem arithmetischen Mittel entspricht erhält dabei eine „0“, Messwerte, die genau eine Standardabweichung unter dem arithmetischen Mittel liegen, erhalten eine „-1“. Messwerte genau eine Standardabweichung über dem arithmetischen Mittel haben einen z-Wert von „1“.

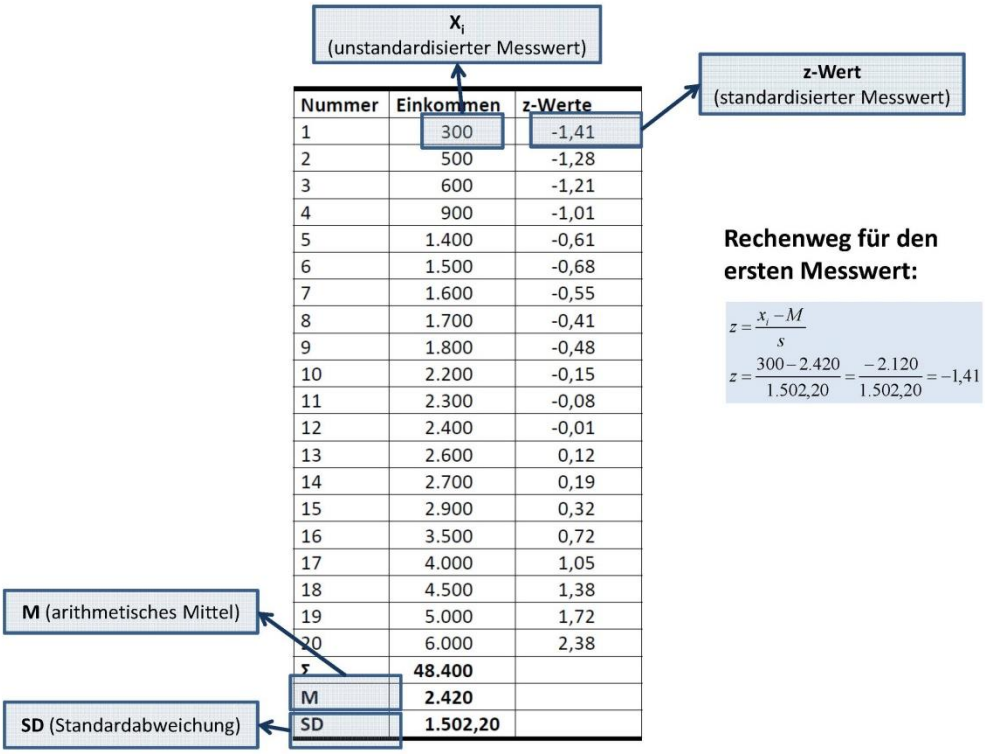
Der z-Wert ist das Verhältnis aus der Differenz von Messwert und arithmetischem Mittel in Relation zur Standardabweichung (F. 10). Abbildung 30 zeigt eine Beispielrechnung der z-Werte des Einkommens.

(F. 10) **z-Standardisierung**

$$z = \frac{x_i - M}{s}$$

x_i = Messwert
M = arithmetisches Mittel
s = Standardabweichung

Formel: Leonhart (2013)



Nummer	Einkommen	z-Werte
1	300	-1,41
2	500	-1,28
3	600	-1,21
4	900	-1,01
5	1.400	-0,61
6	1.500	-0,68
7	1.600	-0,55
8	1.700	-0,41
9	1.800	-0,48
10	2.200	-0,15
11	2.300	-0,08
12	2.400	-0,01
13	2.600	0,12
14	2.700	0,19
15	2.900	0,32
16	3.500	0,72
17	4.000	1,05
18	4.500	1,38
19	5.000	1,72
20	6.000	2,38
Σ	**48.400**	
M	**2.420**	
SD	**1.502,20**	

Abbildung 30: Berechnung des z-Werts (eigene Darstellung)

8.4.5 Normalverteilung und Wahrscheinlichkeit von Wertebereichen normalverteilter Daten

Auf der Basis von standardisierten z-Werten wird nicht nur eine indirekte Vergleichbarkeit zwischen verschiedenen Wertebereichen möglich. Über die Bestimmung von Flächenanteilen unterhalb der Normalverteilungskurve, die bestimmte Wertebereiche einnehmen, kann die Wahrscheinlichkeit jedes beliebigen Wertebereichs bestimmt werden – vorausgesetzt die Normalverteilung kann angenommen werden.

Die *Normalverteilung* hat die Form einer Glocke und kann für alle natürlichen Merkmale (Körpergröße, -gewicht usw.) in der Grundgesamtheit angenommen werden. Auf der x-Achse (untere waagerechte Achse des Diagramms) sind die Werte eines Merkmals eingetragen. Der Wertebereich unterhalb der Normalverteilungskurve ist unendlich groß. Die Normalverteilungskurve verläuft asymptotisch, d.h. sie nähert sich an ihrem rechten und linken Ende der x-Achse an, erreicht sie aber nicht. Im mittleren Bereich ist die Normalverteilungskurve aufgewölbt. Die zugehörigen Werte kommen besonders häufig vor. Die Dichte der Normalverteilung ist in der Mitte des Wertebereichs am größten, die Wahrscheinlichkeit einen Wert im mittleren Ausprägungsbereich zu erreichen ist entsprechend größer als einen am Rand. Die Dichte ist an den Rändern am geringsten. Der deutsche Mathematiker Gauß (1777 bis 1855) beschrieb die Dichtefunktion der Normalverteilung zuerst.

Auch für psychologische und sozialwissenschaftliche Merkmale ist es sinnvoll, die Normalverteilung in der Grundgesamtheit anzunehmen (Eid et al., 2011), auch wenn diese Annahme durchaus kontrovers diskutiert wird und nicht alle psychologischen und sozialwissenschaftlichen Merkmale normalverteilt sind (z.B. Lebensqualität, Döring et al., 2016, S. 476).

Der Flächenanteil, den ein bestimmter Wertebereich eines Merkmals unter der Normalverteilungskurve einnimmt, wird über die Dichtefunktion der Normalverteilung errechnet (Eid et al., 2011, S. 183). Bei normalverteilten Daten nimmt die Fläche unter der Normalverteilungskurve (i.w.S. der Stichprobenanteil) im Bereich zwischen einer Standardabweichung unter und über dem arithmetischen Mittel 68,26 Prozent ein (Abbildung 31). Über die z-Tabelle können die meisten Flächenanteile für Wertebereiche bestimmt werden (Anhang 1). Die Normalverteilung ist eine stetige Verteilung, zwischen jedem Werteintervall sind unendlich viele Werte möglich. Die Bestimmung der Wahrscheinlichkeit eines bestimmten Wertes ist daher nicht möglich, nur die von Wertebereichen (Rasch, Friese, Hofmann & Naumann, 2004).

Für jeden Bereich links oder rechts eines z-Werts oder zwischen zwei z-Werten kann der Flächenanteil unter der Normalverteilungskurve bestimmt werden.

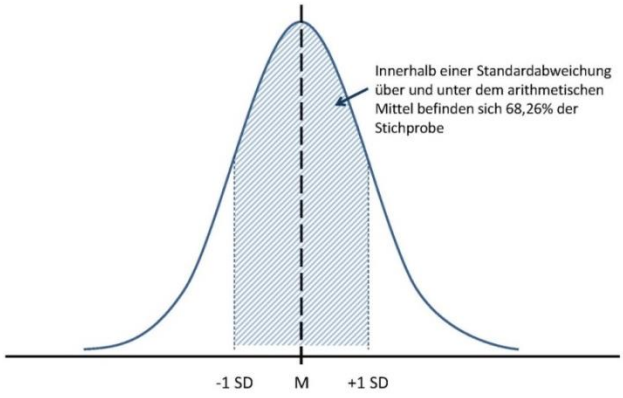


Abbildung 31: Normalverteilungskurve mit dem Flächenanteil innerhalb einer Standardabweichung über und unter dem arithmetischen Mittel (eigene Darstellung)

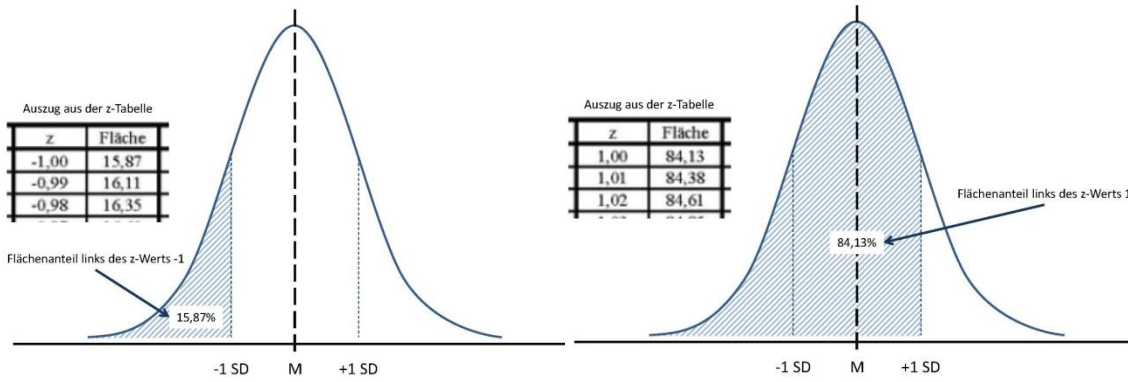


Abbildung 32: Visualisierung der Flächenanteile jeweils links der z-Werte „-1“ (linke Seite) und „1“ (rechte Seite) mit den jeweiligen Auszug aus der z-Tabelle (eigene Darstellung)

Abbildung 32 visualisiert die Berechnung von Wahrscheinlichkeiten unter der Normalverteilungskurve und zeigt Auszüge der z-Tabelle. Auf der linken Seite ist mit z „-1“ der Wert eine Standardabweichung unter dem Mittelwert eingetragen. Der Flächenanteil unter der Normalverteilungskurve links dieses Werts beträgt 15,87 Prozent. Der rechte Tabellenausschnitt zeigt den Flächenanteil unter der Normalverteilungskurve links des z-Werts „1“, er beträgt 84,13 Prozent. Die Differenz dieser Flächenanteile, also der Bereich zwischen einer Standardabweichung unter dem arithmetischen Mittel und einer Standardabweichung darüber, beträgt 68,26 Prozent (Abbildung 31). Das bedeutet: einen Messwert im Bereich einer Standardabweichung unter und einer Standardabweichung über dem Mittelwert zu erreichen ist zu 68,26% wahrscheinlich.

Analog zum skizzierten Vorgehen kann für jedes Werteintervall der Flächenanteil und die Wahrscheinlichkeit ermittelt werden. Überwiegend gelingt dies über das Ablesen der jeweiligen Flächenwerte zu den entsprechenden z-Werten aus der z-Tabelle (Anhang 1). In der z-Tabelle sind alle Flächenanteile *links des jeweiligen z-Werts* dargestellt. Der abgelesene Wert drückt also den Flächenanteil für den abgelesenen z-Wert und kleinere Werte aus. Soll die Wahrscheinlichkeit rechts des z-Werts, also für einen z-Wert und größere Werte, bestimmt werden, muss der abgelesene Wert von 100 Prozent subtrahiert werden.

Sollen Wahrscheinlichkeiten für konkrete Messwertebereiche, also beispielsweise die Wahrscheinlichkeit, 5.000 EUR oder mehr zu verdienen, berechnet werden, wird folgendes Vorgehen vorgeschlagen:

1. Bestimmung des z-Werts zum konkreten Messwert (F. 10)

2. Ablesen des Flächenanteils links des z-Werts in der z-Tabelle (Anhang 1)

3. Berechnung des Flächenanteils rechts des z-Werts (100 – p_z, p_z umschreibt die Wahrscheinlichkeit/den Flächenanteil links des jeweiligen z-Werts)

Das Vorgehen bei der Bestimmung konkreter Wahrscheinlichkeiten für Wertebereiche anhand der Variable Einkommen wird anschließend beispielhaft dargestellt. Folgende Wahrscheinlichkeiten werden gesucht:

1. 4.500 EUR oder *mehr* (erfordert die Bestimmung des Flächenanteils *rechts* des z-Werts)

2. zwischen 950 und 1.400 EUR (Bestimmung eines Wertebereichs)

3. 10 EUR oder weniger zu verdienen (einfaches Ablesen der z-Tabelle).

Übung1:

4.500 EUR oder mehr.

1. **Berechnung des z-Werts (Anzahl der Standardabweichungen, die der Wert vom Mittelwert entfernt ist) (F. 10):**
   $z_{4.500}$ = (4.500 – 2.420) / 1.502,20 = 1,38

2. **Flächenanteil links des z-Werts laut z-Tabelle ($p_{z\ 4.500}$): 91,62%**
   das heißt, die Wahrscheinlichkeit 4.500 EUR oder *weniger* zu verdienen liegt bei 91,62%.

3. **Berechnung der Gegenwahrscheinlichkeit: 100 – p_z = 100% - 91,62% = 8,38%**

Die Wahrscheinlichkeit 4.500 EUR oder mehr zu verdienen beträgt ausgehend von den erhobenen Einkommenswerten in der Stichprobe 8,38 Prozent.

Übung 2:

zwischen 950 und 1.400 EUR verdienen

1. **Berechnung der z-Werte beider Einkommen(F. 10):**
   z_{950} = (950 – 2.420) / 1.502,20 = -0,98
   $z_{1.400}$ = (1.400 – 2.420) / 1.502,20 = -0,68

2. **Flächenanteil links des z-Werts laut z-Tabelle ($p_{z\ 950}$): 16,35%**
   das heißt, die Wahrscheinlichkeit 950 EUR oder *weniger* zu verdienen liegt bei 16,35%.
   Flächenanteil links des z-Werts laut z-Tabelle ($p_{z\ 1.400}$): 24,83%
   das heißt, die Wahrscheinlichkeit 1.400 EUR oder *weniger* zu verdienen liegt bei 24,83%

3. **Berechnung der Wahrscheinlichkeit durch Bestimmung der Differenz zwischen beiden Wahrscheinlichkeiten: $p_{z\ 1.400} – p_{z\ 950}$ = 24,83% – 16,35% = 8,48%**

Die Wahrscheinlichkeit zwischen 950 und 1.400 EUR zu verdienen beträgt ausgehend von den erhobenen Einkommenswerten in der Stichprobe 8,48 Prozent.

Übung 3:

10 EUR oder weniger verdienen

1. **Berechnung des z-Werts von 400 EUR (F. 10):**
   z_{10} = (10 – 2.420) / 1.502,20 = -1,60

2. **Flächenanteil links des z-Werts laut z-Tabelle ($p_{z\ 10}$): 5,48%**
   das heißt, die Wahrscheinlichkeit 10 EUR oder *weniger* zu verdienen liegt bei 5,48%.

Die Wahrscheinlichkeit 10 EUR oder weniger zu verdienen beträgt ausgehend von den erhobenen Einkommenswerten in der Stichprobe 5,48 Prozent.

Im letzten Beispiel ist die Wahrscheinlichkeit nahe der 5-Prozent-Schwelle, die im Kapitel „Signifikanz und Relevanz“ Bedeutung bekommt (s. 9). Werte unterhalb der 5- bzw. oberhalb der 95-Prozent-Schwelle sind so wenig wahrscheinlich, dass ihr Auftreten nicht mehr mit Zufall erklärt werden kann. Diese Werte können ebenso gut zu einer anderen Verteilung, einer anderen Grundgesamtheit, gehören, in der sie wahrscheinlicher sind. Beim Einkommensbeispiel kommt ein Verdienst von 10 EUR oder weniger vermutlich nicht in der Grundgesamtheit der Arbeitnehmer vor, weil er ausgehend von arithmetischem Mittel und Standardabweichung weniger als 5% wahrscheinlich ist. So wenig Einkommen erzielt vermutlich die Grundgesamtheit Schüler bzw. nicht erwerbstätige Hausmännern oder -frauen. Über diese Annahmen werden Aussagen zur statistischen Signifikanz abgeleitet.

Zur Schätzung der Verteilung der Messwerte in der Grundgesamtheit werden zu den Maßen der zentralen Tendenz und Streuung auch die Konfidenzintervalle angegeben. Ein weiteres Maß zur Schätzung der Verteilung der Messwerte in der Grundgesamtheit ist der Standardfehler des arithmetischen Mittels (9.4).

8.5 Bivariate deskriptive Statistik – Koinzidenz, Korrelation, Kausalität – Kreuztabelle, Rangkorrelation und Punkt-Moment-Korrelation

Ziel der bivariaten deskriptiven Statistik ist es, Zusammenhängen zwischen zwei Merkmalen in der Stichprobe aufzudecken. Verfahren der bivariaten Statistik werden häufig zur Beantwortung von Forschungsfragen genutzt. Sie finden Muster gemeinsam auftretender Veränderungen der Werte zweier Variablen in der Stichprobe. Drei Begriffe werden unterschieden:

1. *Koinzidenz* meint das zufällige gemeinsame Auftreten zweier Merkmale, ohne dass dafür eine inhaltlich plausible oder theoretisch untermauerte Erklärung existiert.

2. *Korrelation* tritt auf, wenn zwei Merkmale sich in der Stichprobe in ähnlicher Weise verändern. Dieses Muster zeigt sich beispielsweise, wenn in einer Stichprobe kleine Menschen (Variable Körpergröße) leicht (Variable Gewicht) und große Menschen schwer sind. Eine Korrelation muss sich theoretisch untermauern lassen. Für die Untersuchung von Korrelationen sind fundierte theoretische Vorüberlegungen erforderlich (s. 4.2). Korrelation bezieht sich auf Zusammenhänge in Querschnittsstudien, Ursache-Wirkungsbeziehungen decken Korrelationsanalysen in Querschnittstudien nicht auf, nur ein zeitgleich variables Auftreten zweier Merkmale.

3. *Kausalität* (Ursache-Wirkungs-Beziehung) bedeutet wie die Korrelation auch die gemeinsame Varianz zweier Merkmale. Im Unterschied zum zeitgleichen Auftreten erlaubt Kausalität eine Verlaufsbeschreibung, in dem ein Merkmal vor dem anderen auftreten muss. Kausalität kann in Studien untersucht

werden, die ein und dieselbe Stichprobe mindestens zwei Mal nacheinander untersuchen (Längsschnittstudie, s. 5.2, 5.3).

Die häufig eingesetzten menügesteuerte Statistiksoftwareprogramme, wie beispielsweise SPSS, verleiten dazu, unfundiert Korrelationen zwischen allen möglichen Variablen zu berechnen. Dies ist nicht sinnvoll, bringt keinen Erkenntnisgewinn und gleicht „Kaffeesatzleserei“. Walter Krämer (2013) stellt ein paar Ergebnisse theoretisch nicht fundierter Korrelationsanalysen vor, die von ihm als „Nonsenskorrelationen“ (S. 172) bezeichnet werden. Beispielsweise werden Zusammenhänge zwischen Kalziumgehalt im Knochen und der Zahl unverheirateter Tanten, Schuhgröße und Lesbarkeit der Handschrift genannt, die sich sämtlich durch nicht berücksichtigte Dritt- (oder Stör-)Variablen erklären lassen. Die Zeit (2015/13, 26. März 2015) stellt Scheinkorrelationen beispielsweise zwischen der geernteten Menge Endiviensalats und der Anzahl gestorbener Menschen durch Parasiten und Infektionskrankheiten vor. Auch der Zusammenhang zwischen der Anzahl Störche und der Geburtenrate fällt unter Scheinkorrelationen. Diese Zusammenhänge bestehen und lassen sich berechnen. Sie sind genauso zufällig, wie ein recht unwahrscheinlicher Sechser im Lotto getippt werden kann. Wie bei alle anderen quantitativen Analysen sind auch bei Korrelationsanalysen die gründliche theoretische Auseinandersetzung und die Beschreibung eines Modells notwendig, in dem auch die Richtung des Zusammenhangs (positiv oder negativ) beschrieben wird.

In Abbildung 33 werden Korrelationen verschiedener Richtungen visualisiert. Dabei ist jeder Punkt das Messwertepaar eines Studienteilnehmers zweier Variablen (hier Körpergröße und Körpergewicht). Von links unten nach rechts oben aufsteigende Punktwolken decken positive Korrelationen auf. Negative Korrelationen zeigen von links oben nach rechts unten absteigende Punktwolken. Je schmaler die Ellipse um die Punktwolke ist, desto enger ist ein Zusammenhang. Sofern alle Punkte auf einer diagonal verlaufenden Linie liegen, wird von einer absoluten Korrelation gesprochen. Eine Punktwolke, um die ein Kreis gezeichnet werden kann, zeigt keine Korrelation. Das linke Bild in Abbildung 33 zeigt eine positive Korrelation. Menschen die groß sind, sind auch schwer, kleine Menschen dagegen sind leicht. In der Bildmitte liegt eine negative Korrelation vor. Hier sind große Menschen leicht und kleine Menschen schwer. Auf der rechten Abbildungsseite liegt kein Zusammenhang zwischen den Variablen Körpergewicht und Körpergröße vor.

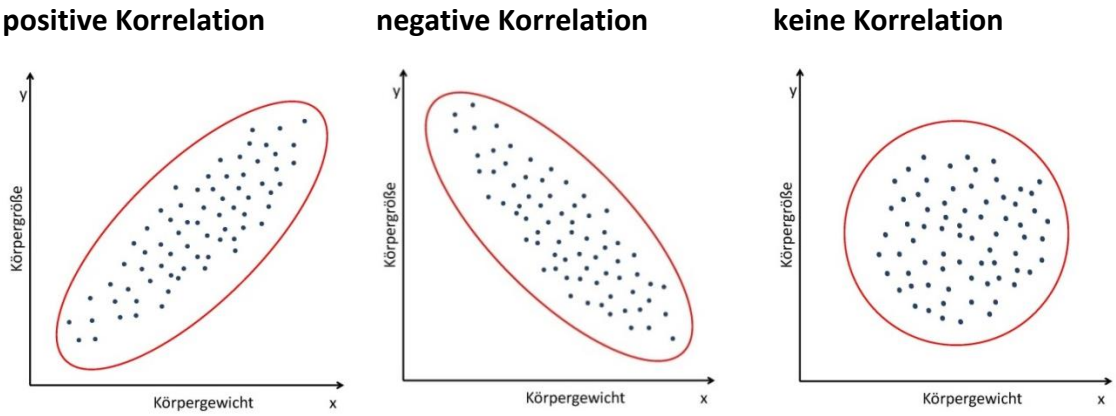


Abbildung 33: Visualisierung von Korrelationen unterschiedlicher Richtungen (eigene Darstellung)

Korrelationen können mit allen Messniveaus (s. 6.2.1) berechnet werden. Bei nominalskalierten Datenmaterial werden Zusammenhänge mit Hilfe von Kreuztabellen dargestellt. Die *Produkt-Moment-Korrelation* wird zwischen intervallskalierten und normalverteilten Daten berechnet (Pearson-Korrelation). Der *Rangkorrelationskoeffizient* (Spearman-Korrelation) kann berechnet werden, wenn eines der Merkmale oder beide nicht intervallskaliert und normalverteilt ist, zumindest aber auf Ordinalskalenniveau vorliegt (Tabelle 16).

Tabelle 16: Messniveau und Berechnung von Zusammenhängen.

Messniveau	**Kreuztabelle**	**Spearman**	**Pearson**
Nominalskalenniveau	X		
Ordinalskalenniveau	X	X	
Intervallskalenniveau	X	X	X

Korrelationskoeffizienten werden standardisiert berichtet, sie können Werte zwischen „-1“ und „+1 einnehmen. Wobei „0“ keine Korrelation bedeutet, je näher der Korrelationskoeffizient an „-1“ oder „+1“ ist, desto enger ist der Zusammenhang. Bei der Bewertung von Korrelationskoeffizienten lehnen sich Döring et al. (2016) an Cohen (1988) an (Cohen, 1988, zit. in Döring et al., 2016, S. 820). **Danach weisen Korrelationskoeffizienten ab 0,10 auf kleine, ab 0,3 auf mittlere und ab 0,5 auf enge Zusammenhänge hin.**

In diesem Studienbrief wird die Berechnung der Korrelation zwischen nominalskalierten Merkmalen (8.5.1) und intervallskalierten Merkmalen näher betrachtet. Die Interpretation des Spearman-Korrelationskoeffizienten für ordinalskaliertes Datenmaterial unterscheidet sich nicht vom Pearson-Korrelationskoeffizienten. In der

sozialwissenschaftlichen Forschung werden die Anforderungen an die Berechnung der Produkt-Moment- bzw. Pearson-Korrelation teils nicht erfüllt, weil Daten nicht normalverteilt vorliegen oder die Verteilung in der Grundgesamtheit unbekannt ist (s. 88.4.5). In diesem Fall muss mit den entsprechenden Statistikprogrammen eine Berechnung mit dem Rangkorrelationskoeffizienten nach Spearman erfolgen. Unterschiede fallen bei der Berechnung beider Korrelationskoeffizienten auf: Bei der Pearson-Korrelation werden Unterschiede zwischen Merkmalsausprägung und Mittelwert des jeweiligen Merkmals (Varianz, s. 8.4.3) in die Formel aufgenommen (F. 13). In die Berechnung der Spearman-Korrelation fließen die Rangplätze der Messwerte und die Unterschiede *zwischen den Rangplätzen* der korrelierten Merkmale in die Berechnung ein.

8.5.1 Korrelation zwischen nominalskalierten Merkmale

Zusammenhänge zwischen nominalskalierten Merkmale mit zwei Ausprägungen (dichotome Merkmale) werden über Kreuztabellen dargestellt. Die Berechnung des standardisierten Korrelationskoeffizient über den ϕ-Koeffizienten oder über Chancenverhältnisse. Kreuztabellen, Vierfeldertafeln oder 2 x 2-Tabellen enthalten vier Zellen mit Werten, die mit Buchstaben, alternativ mit Buchstaben-Zahlenkombination bezeichnet werden (Tabelle 17). Die Kenntnis der Zellbezeichnung ist für die Nutzung von Formeln zur Berechnung von Zusammenhängen zwischen nominalskalierten Variablen hilfreich (F. 11, (F. 12). Die Merkmalsausprägungen müssen in einer hypothesenkonformen Weise geordnet werden. In der ersten Spalte sollte die interessierende Ausprägung der abhängigen Variable aufgenommen werden, in der ersten Zeile die Ausprägung *der* unabhängigen Variable, von der ein Einfluss auf die abhängige Variable erwartet wird.

Tabelle 17: Gestaltung einer Kreuztabelle und Beschriftung der Zellen

Merkmal	**Merkmal 2** *Ausprägung 1*	**Merkmal 2** *Ausprägung 2*	**Zeilensumme**
Merkmal 1 *Ausprägung 1*	**a (n_{11})**	**b (n_{12})**	
Merkmal 1 *Ausprägung 2*	**c (n_{21})**	**d (n_{21})**	
Spaltensumme			

In der Beispielstudie (Tabelle 10) werden Verbindungen zwischen dem Geschlecht und dem Hochschulabschluss untersucht. Angenommen wird, Frauen haben häufiger einen Hochschulabschluss als Männer. Zunächst werden zur Überprüfung der Annahme absolute und relative Häufigkeiten der Merkmale in eine Kreuztabelle (Tabelle 18) übertragen. In der Tabelle können Zeilen- oder Spaltenprozente angegeben werden. Im Beispiel werden relative Häufigkeiten auf Basis der Zeilensummen aufgenommen. Der Bericht relativer Häufigkeiten lässt Rückschlüsse auf die Verteilung von Hochschulabsolventen in der Gruppe der Männer und Frauen zu.

Zeilenprozente stellen im Beispiel dar, wie groß der Anteil Hochschulabsolventen bei Männern und Frauen ist. Spaltenprozente, die nicht in Tabelle 18 aufgenommen wurden, zeigen den Anteil Frauen (und Männer) in beiden Ausprägungen des Hochschulabschlusses. Anteilig haben mehr Frauen als Männer einen Hochschulabschluss.

Tabelle 18: Kreuztabelle Geschlecht und Hochschulabschluss.

	Hochschulabschluss		
Geschlecht	**ja (%)**	**nein (%)**	**Zeilensumme**
weiblich (%)	8 (80)	2 (20)	10 (100)
männlich (%)	4 (40)	6 (60)	10 (100)
Spaltensumme	12 (60)	8 (40)	20

Der Zusammenhang zwischen den nominalskalierten Merkmalen Geschlecht und Hochschulabschluss wird über den ϕ-Koeffizienten oder das Odds-Ratio (F. 12) berechnet.

(F. 11) **ϕ-Koeffizient**

$$\hat{\varphi} = \frac{n_{11} \cdot n_{22} - n_{12} \cdot n_{21}}{\sqrt{(n_{11} \cdot n_{21}) + (n_{12} \cdot n_{22}) + (n_{11} \cdot n_{12}) + (n_{21} \cdot n_{22})}}$$

alternativ

$$\hat{\varphi} = \frac{a \cdot d - b \cdot c}{\sqrt{(a \cdot c) + (b \cdot d) + (a \cdot b) + (c \cdot d)}}$$

n_{xy} = absolute Häufigkeit in der Zelle mit Nummer

a-d = alternative Bezeichnung der Zellennummer

Formel: Eid et al. (2013)

Der ϕ-Koeffizient kann wie alle standardisierten Korrelationskoeffizienten Werte zwischen -1 und +1 einnehmen. Allerdings kann das Vorzeichen nicht analog den in Abbildung 33 dargestellten Richtungen interpretiert werden. Das Vorzeichen ist beim ϕ-Koeffizienten abhängig von der Codierung des Merkmals, also ob Frauen in der ersten oder zweiten Zeile, bzw. ob der Hochschulabschluss in der ersten oder zweiten Spalte aufgenommen wird und hat keine inhaltliche Bedeutung (Eid et al., 2011).

Das Ergebnis 0,48 weist auf eine mittelhohe (s. 8.5, Döring et al., 2016) systematische Verbindung zwischen den Merkmalen Geschlecht und Hochschulabschluss hin. Männer und Frauen haben in der Stichprobe demnach unterschiedlich häufig einen Hochschulabschluss, Frauen häufiger als Männer (Abbildung 34).

	Hochschulabschluss				
Geschlecht	**ja (%)**		**nein (%)**		**Zeilensumme**
weiblich (%)	8 (80)	n_{11} (a)	2 (20)	n_{12} (b)	10 (100)
männlich (%)	4 (40)	n_{21} (c)	6 (60)	n_{22} (d)	10 (100)
Spaltensumme	12 (60)		8 (40)		20

$$\widehat{\varphi} = \frac{n_{11} \cdot n_{22} - n_{12} \cdot n_{21}}{\sqrt{(n_{11} \cdot n_{21}) + (n_{12} \cdot n_{22}) + (n_{11} \cdot n_{12}) + (n_{21} \cdot n_{22})}}$$

$$\widehat{\varphi} = \frac{8 \cdot 6 - 2 \cdot 4}{\sqrt{(8 \cdot 4) + (2 \cdot 6) + (8 \cdot 2) + (4 \cdot 6)}}$$

$$\widehat{\varphi} = \frac{40}{84} = \mathbf{0,48}$$

Abbildung 34: Berechnung des ϕ-Koeffizient zwischen zwei nominalskalierten Merkmalen (eigene Darstellung)

Eine weitere Möglichkeit, den Zusammenhang dichotomer nominalskalierter Variablen zu bestimmen, ist die Berechnung von Chancenverhältnissen (F. 12, Abbildung 35).

Beim Odds-Ratio (Chancenverhältnis) wird zunächst die Chance jeweils für Männer und Frauen berechnet, einen Hochschulabschluss zu haben. Vier Mal mehr Frauen haben einen Hochschulabschluss als keinen. Bei den Männern ist ein Hochschulabschluss seltener, die Chance einen Hochschulabschluss zu haben beträgt hier nur 0,67. Das Chancenverhältnis von 6 drückt aus, dass Frauen eine sechs Mal größere Chance auf einen Hochschulabschluss haben als Männer.

(F. 12) **Odds-Ratio**

$$OR = \frac{n_{11} \div n_{12}}{n_{21} \div n_{22}} = \frac{a \div b}{c \div d}$$

n_{xy} = absolute Häufigkeit in der Zelle mit Nummer

a-d = alternative Bezeichnung der Zellennummer

	Hochschulabschluss		
Geschlecht	**ja (%)**	**nein (%)**	**Zeilensumme**
weiblich (%)	8 (80) n_{11} (a)	2 (20) n_{12} (b)	10 (100)
männlich (%)	4 (40) n_{21} (c)	6 (60) n_{22} (d)	10 (100)
Spaltensumme	12 (60)	8 (40)	20

$$OR = \frac{n_{11} \div n_{12}}{n_{21} \div n_{22}} = \frac{a \div b}{c \div d}$$

$$OR = \frac{8 \div 2}{4 \div 6} = \frac{4}{0{,}67} = \mathbf{6}$$

Abbildung 35: Berechnung des Odds-Ratios zwischen zwei nominalskalierten Merkmalen (eigene Darstellung)

8.5.2 Korrelation zwischen intervallskalierten Merkmalen

Zusammenhänge zwischen intervallskalierten Daten werden als Pearson-Korrelationskoeffizient (auch Produkt-Moment-Korrelation) angegeben. Wie alle standardisierten Korrelationskoeffizienten bewegt er sich in einem Wertebereich zwischen -1 und +1. Grundlage der Berechnung nach Pearson sind die Kovarianz und die Standardabweichung der Merkmalsausprägungen jedes Merkmals (F. 13).

(F. 13) **Produkt-Moment-Korrelation**

$$r_{xy} = \frac{cov_{xy}}{s_x \cdot s_y}$$

aufgelöst so:

$$r_{xy} = \frac{1}{N} \cdot \frac{\sum_{i=1}^{N}(x_i - M) \cdot (y_i - M)}{\sqrt{\frac{\sum_{i=1}^{N}(x_i - M_x)^2}{n_x} \cdot \frac{\sum_{i=1}^{N}(y_i - M_y)^2}{n_y}}}$$

vereinfacht:

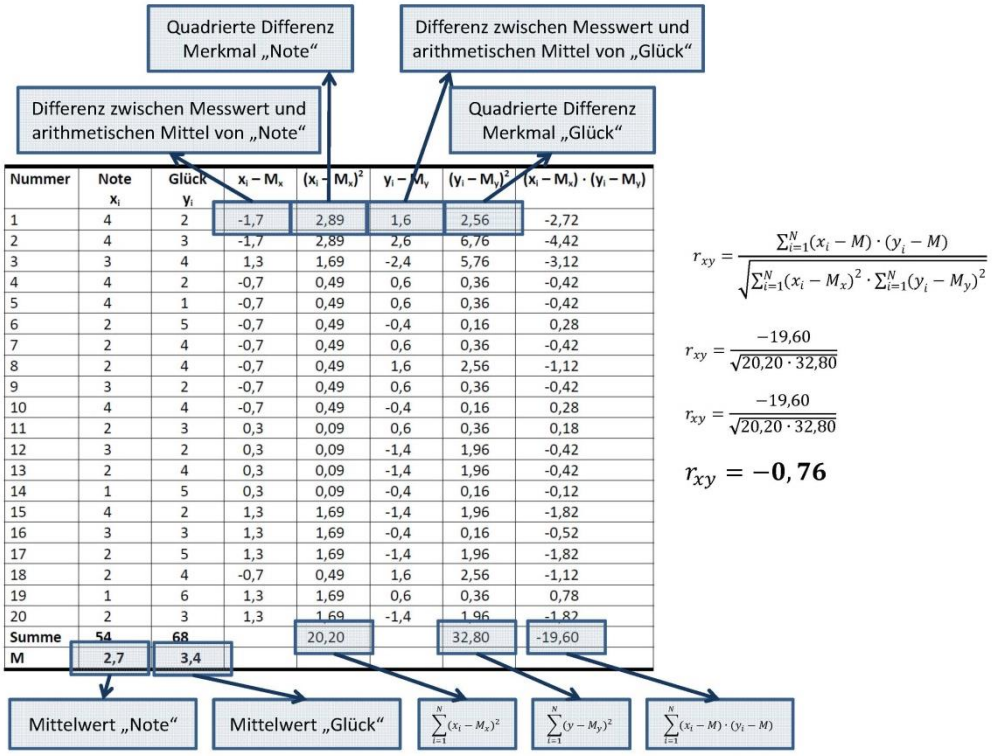
$$r_{xy} = \frac{\sum_{i=1}^{N}(x_i - M) \cdot (y_i - M)}{\sqrt{\sum_{i=1}^{N}(x_i - M_x)^2 \cdot \sum_{i=1}^{N}(y_i - M_y)^2}}$$

cov_{xy} = Summe des Produkts der Differenzen zwischen Mess- und Mittelwert der Merkmale (Kovarianz)

s_x = Standardabweichung des Merkmals

Formel: Leonhart (2013)

Zwischen der Abschlussnote und der Ausprägung von Glück besteht ein enger negativer Zusammenhang (Abbildung 36). Je höher der Messwert bei der Abschlussnote – also je schlechter die Note – desto kleiner ist die Ausprägung beim Glück. Menschen mit schlechter Abschlussnote fühlen sich also unglücklicher.



Nummer	Note x_i	Glück y_i	$x_i - M_x$	$(x_i - M_x)^2$	$y_i - M_y$	$(y_i - M_y)^2$	$(x_i - M_x) \cdot (y_i - M_y)$
1	4	2	-1,7	2,89	1,6	2,56	-2,72
2	4	3	-1,7	2,89	2,6	6,76	-4,42
3	3	4	1,3	1,69	-2,4	5,76	-3,12
4	4	2	-0,7	0,49	0,6	0,36	-0,42
5	4	1	-0,7	0,49	0,6	0,36	-0,42
6	2	5	-0,7	0,49	-0,4	0,16	0,28
7	2	4	-0,7	0,49	0,6	0,36	-0,42
8	2	4	-0,7	0,49	1,6	2,56	-1,12
9	3	2	-0,7	0,49	0,6	0,36	-0,42
10	4	4	-0,7	0,49	-0,4	0,16	0,28
11	2	3	0,3	0,09	0,6	0,36	0,18
12	3	2	0,3	0,09	-1,4	1,96	-0,42
13	2	4	0,3	0,09	-1,4	1,96	-0,42
14	1	5	0,3	0,09	-0,4	0,16	-0,12
15	4	2	1,3	1,69	-1,4	1,96	-1,82
16	3	3	1,3	1,69	-0,4	0,16	-0,52
17	2	5	1,3	1,69	-1,4	1,96	-1,82
18	2	4	-0,7	0,49	1,6	2,56	-1,12
19	1	6	1,3	1,69	0,6	0,36	0,78
20	2	3	1,3	1,69	-1,4	1,96	-1,82
Summe	**54**	**68**		20,20		32,80	-19,60
M	**2,7**	**3,4**					

Abbildung 36: Berechnung der Produkt-Moment-Korrelation (eigene Darstellung)

8.6 Zusammenfassung und Aufgaben

Die Darstellung und Beschreibung von Daten dient Transparenzzwecken. Die Ergebnisse deskriptiver Statistik fließen in die Beschreibung der Stichprobe und in die Analyse der entwickelten Fragestellungen ein. Dazu werden zunächst alle Merkmalsausprägungen in eine Datentabelle aufgenommen. Datentabellen für weitere statistische Analysen enthalten in den Spalten (x_m) die untersuchten Merkmale und in den Zeilen die jeweiligen Studienteilnehmer. Die erste Zeile der Datentabelle nimmt den Variablennamen auf. Eine Spalte mit allen Daten zu einem Merkmal wird als Spaltenvektor, eine Zeile die Merkmalsausprägungen eines Studienteilnehmers zu allen Merkmalen enthält, Zeilenvektor bezeichnet. Eine so gestaltete Datentabelle kann in allen Statistikprogrammen für die Datenanalyse genutzt werden.

Die nachfolgenden Rechenschritte erfolgen mit Daten einer fiktiven Studie, in der als Merkmale Monatseinkommen, Wohnort, Geschlecht, Hochschulabschluss, Abschlussnote und Glück untersucht wurden (8.2).

In der deskriptiven Statistik werden Verfahren der uni-, bi- und multivariaten Statistik unterschieden. Univariate deskriptivstatistische Verfahren dienen der Darstellung absoluter, relativer und prozentualer Häufigkeiten. Die Häufigkeit von Merkmalsausprägungen kann für alle Messniveaus abgebildet werden, ist aber insbesondere bei kleinen Stichproben für nominalskaliertes Datenmaterial sinnvoll (8.3). Lage- und Streuungsmaße werden zur Beschreibung ordinal- und intervallskalierter

Daten verwendet. Je nach Messniveau werden als Lagemaß Modal- oder Medianwert bzw. das arithmetische Mittel berichtet. Ein Lagemaß ist umso aussagekräftiger, je kleiner die Streuung der Merkmalsausprägung ist. Daher muss neben dem Lagemaß auch die Streuung berichtet werden. Zu den wichtigen Streuungsmaßen gehören Range (Modalwert), Interquartilsbereich (Median), Varianz und Standardabweichung (arithmetisches Mittel) (8.4).

Bivariate deskriptivstatistische Verfahren decken Zusammenhänge zwischen zwei Merkmalen auf, die theoretisch begründet sein müssen. Je nach Messniveau werden bivariate Statistiken in Kreuztabellen (nominalskalierte Daten), als Rang- (ordinal- oder nicht normalverteilte intervallskalierte Daten) oder Produkt-Moment-Korrelationskoeffizient (normalverteilte intervallskalierte Daten) berichtet. Standardisierte Korrelationskoeffizienten nehmen Werte zwischen -1 und +1 ein. Je näher der Wert an der „1“ ist, desto enger, je näher an „0“ desto schwächer ist der Zusammenhang. Grundsätzlich dürfen Korrelationen nur berechnet werden, wenn ihr Auftreten theoretisch erklärbar ist, also ein empirisch begründetes Modell in der wissenschaftlichen Arbeit entwickelt wurde (8.5).

Alle Ergebnisse der deskriptiven Statistik sind für die Stichprobe, von der die Daten stammen gültig. Das Interesse von Forschung geht über die Beschreibung von Stichprobenergebnissen hinaus. Daher werden durch inferenzstatistische Verfahren Schlüsse von Stichproben auf die jeweilige Grundgesamtheit gezogen. Auf die Bedeutung und die Aussage von Signifikanztests wird anschließend näher eingegangen.

Schlüsselwörter: *absolute Häufigkeit, Datentabellen, deskriptive (beschreibende) Statistik, dichotome Merkmale/Variable, Häufigkeitsverteilung, Interquartilsbereich, Kausalität, Koinzidenz, Konfidenzintervall, Korrelation, Korrelationskoeffizient, Kumulierte Häufigkeiten, Medianwert/Median, Mittelwert/arithmetische Mittel, Modalwert/Modus, n, Normalverteilung, ordinalskaliertes Datenmaterial, Produkt-Moment-Korrelation, prozentuale Häufigkeit, Range, Rangkorrelations-koeffizient, Relation, relative Häufigkeit, Spaltenvektor, Standard-abweichung, Stör-(Confounding) Variable, Variable, Varianz, z-Wert, Zeilenvektor*

Aufgaben zur Lernkontrolle

1. Definieren Sie Modal- und Medianwert sowie das arithmetische Mittel.
2. Wie ist ein Boxplot aufgebaut? Verdeutlichen Sie dies anhand einer Skizze.
3. Definieren Sie die Begriffe Varianz und Standardabweichung.
4. Wie ist eine Datentabelle aufgebaut? Fertigen Sie hierzu am besten eine Skizze an.
5. Welche Darstellungsformen für Häufigkeiten gibt es?
6. Was wird mit einer z-Standardisierung erreicht?

Literatur

Beller, S. (2008). *Empirisch forschen lernen Konzepte, Methoden, Fallbeispiele, Tipps* (2., überarb. Aufl ed.). Bern: Huber.

Deutsche Gesellschaft für Psychologie. (2007). *Richtlinien zur Manuskriptgestaltung* (3., überarbeitete und erweiterte Auflage ed.). Göttingen [u.a.]: Hogrefe.

Döring, N., Bortz, J., & Pöschl, S. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften* (5. vollständig überarbeitete, aktualisierte und erweiterte Auflage ed.). Berlin Heidelberg: Springer.

Eid, M., Gollwitzer, M., & Schmitt, M. (2011). *Statistik und Forschungsmethoden. Lehrbuch* (2. korrigierte Auflage). Weinheim u.a.: Beltz.

Krämer, W. (2013). *So lügt man mit Statistik* (Überarb. Neuausg. der ungekürzten Taschenbuchausg., 4. Aufl. ed.). München [u.a.]: Piper.

Leonhart, R. (2013). *Lehrbuch Statistik Einstieg und Vertiefung* (3., überarbeitete und erweiterte Auflage ed.). Bern: Verlag Hans Huber.

Rasch, B., Friese, M., Hofmann, W., & Naumann, E. (2004). *Quantitative Methoden Band 1*. Berlin u.a.: Springer.

Schäfer, T. (2016). *Methodenlehre und Statistik Einführung in Datenerhebung, deskriptive Statistik und Inferenzstatistik*. Wiesbaden: Springer.

9 Inferenzstatistik

Lernergebnisse:

- Der Signifikanzbegriff kann definiert und die Aussagekraft signifikanter Ergebnisse diskutiert werden (9.1).
- Statistische Hypothesen und Fehler beim Hypthesentesten können erläutert und hinsichtlich ihrer Bedeutung für signifikante Ergebnisse gewichtet werden (9.1).
- Signifikanz und Relevanz , Kriterien zur Entscheidung über die Relevanz signifikanter Ergebnisse können exemplarisch erläutert werden: Woran erkennt man relevante Ergebnisse? (9.1).
- Der Signifikanzbegriff kann ausgehend von der Wahrscheinlichkeit des Auftretens einer Merkmalsausprägung erläutert werden. Dazu können Bernoulli-Experimente angewendet werden (9.2).
- Signifikanztests für Nominal-, Ordinal- und Intervalldaten können ausgehend von inhaltlichen Hypothesen systematisiert werden (9.3).
- Signifikanztests für Unterschieds- und Zusammenhangshypothesen können angewendet, die Ergebnisse transparent in Tabellenform und schriftlich berichtet werden(9.3).
- Der Begriff Konfidenzintervall kann definiert werden. Die Bedeutung von Konfidenzintervallen zur Einschätzung der Signifikanz von Unterschieden und Zusammenhängen können erläutert werden (9.4).
- Ausgehend von veröffentlichten Studienergebnissen können signifikante und nicht-signifikante Ergebnisse berichtet und ihre Relevanz diskutiert werden.

Forschung geht Fragestellungen von übergreifendem Interesse nach. Sie untersucht sie aber nicht in der Grundgesamtheit, sondern in Stichproben, die aus der Grundgesamtheit gezogen werden (s. 0). Ergebnisse aus Stichproben sind grundsätzlich nur für die Stichprobe gültig. Durch die Nutzung von wahrscheinlichkeitstheoretisch fundierten Verfahren ist es möglich von Stichprobenergebnissen auf die Lage in der Grundgesamtheit zu schließen. Die Aussagekraft dieser Schlüsse ist jedoch kleiner als die starke Betonung „signifikanter Ergebnisse“ in einschlägigen Veröffentlichungen vermuten lässt.

Innerhalb der Inferenzstatistik werden Signifikanztests auf der Basis von Annahmen überprüft, die für die Grundgesamtheit formuliert werden. Dabei wird überprüft, mit welcher Wahrscheinlichkeit das aufgefundene oder ein noch extremeres Ergebnis in der Stichprobe möglich ist, wenn die formulierte Annahme in der Grundgesamtheit zutreffen sollte. Die Überprüfung von Signifikanz erfolgt also mit (1) der Formulierung einer Annahme zur „Lage in der Grundgesamtheit“ (s. statistische Hypothesen, 4.4.2). Die Mehrzahl der Signifikanztests geht der Gültigkeit der Nullhypothese nach. Darin wird angenommen, in der Grundgesamtheit treffen *keine* der theoretisch hergeleiteten, in der Fragestellung und den Hypothesen formulierten Annahmen zu (Nullhypothese). (2) werden diese Annahmen mit den Daten der Stichprobe überprüft. Es wird nicht berechnet, wie wahrscheinlich ein Ergebnis auch für die Grundgesamtheit zutrifft, sondern wie wahrscheinlich ein bestimmtes Ergebnis in der Stichprobe ist, wenn in der Grundgesamtheit die formulierte Annahme, zumeist die Nullhypothese, zutreffend ist (9.1).

Eine Annäherung an den Signifikanzbegriff erfolgt zunächst über die Überprüfung von Wahrscheinlichkeiten für beliebige Verteilungen von Merkmalsausprägungen, wenn für die Grundgesamtheit eine ganz bestimmte Verteilung zutrifft (Bernoulli-Experimente, 9.1.4).

Ob ein Ergebnis statistisch signifikant ist oder nicht kann für Unterschieds-, (Veränderungs-) und Zusammenhangshypothesen untersucht werden. Je nach Messniveau der zugrundeliegenden Daten können verschiedene Verfahren zur statistischen Überprüfung von Hypothesen eingesetzt werden. Betrachtet werden der chi-Quadrattest für absolute und relative Häufigkeiten und der t-Test zur Überprüfung von Mittelwertunterschieden bei intervallskaliertem und normalverteilten Daten. Weitere Testverfahren werden definiert und ihr Anwendungsbereich beschrieben (9.3).

Über die Berechnung von Konfidenzintervallen, die für alle Lagemaße, relative und prozentuale Häufigkeiten sowie für jede andere statistische Kennziffer bestimmt werden können, kann der wahre Wert in der Population geschätzt werden. Über den Vergleich von 95% oder 99% Konfidenzintervallen sind zudem Rückschlüsse auf die Signifikanz von Unterschieden und Zusammenhängen möglich, sofern Konfidenzintervalle überschneidungsfrei sind (9.4) bzw. die „0“ oder „1“ nicht beinhalten.

9.1 Vorstellungen von Signifikanz

9.1.1 Signifikanz ist die Wahrscheinlichkeit von Stichprobenwerten Signifikanztestung ist Hypothesentestung

Mit Signifikanztests wird die Wahrscheinlichkeit für Ergebnisse in der Stichprobe bestimmt, unter der Annahme, dass ein bestimmter Zustand in der Grundgesamtheit vorliegt. Diese Annahme ist mathematisch in verschiedenen Wahrscheinlichkeitsverteilungen begründet. Beispiele für Wahrscheinlichkeitsverteilungen, auf deren Grundlage Signifikanztests erfolgen, sind die Normalverteilung, die t-, chi-Quadrat oder F-Verteilung. Ein Signifikanztest wird von folgenden Annahmen geleitet:

1. ein Merkmal hat einen bestimmten Kennwert (z.B. arithmetisches Mittel) in der Grundgesamtheit

2. Kennwerte von zufällig gezogenen Stichproben streuen um den wahren (unbekannten) Kennwert in der Population (s. 7.3)

3. sind Kennwerte zweier Zufallsstichproben sehr ähnlich, entstammen sie wahrscheinlich aus derselben Grundgesamtheit

4. Kennwerte, die sich zwischen zwei Zufallsstichproben deutlich unterscheiden, stammen eventuell nicht aus derselben Grundgesamtheit

5. die Wahrscheinlichkeit des Auftretens einer Differenz zwischen Stichprobenkennwerten wird mit Hilfe einer Wahrscheinlichkeitsverteilung (z.B. t-Verteilung) unter der Annahme berechnet, dass in der Grundgesamtheit kein Unterschied besteht (Nullhypothese)

6. ist die Wahrscheinlichkeit für die beobachtete Differenz kleiner als 5 Prozent, so ist dieser Unterschied nicht mit der Nullhypothese vereinbar. Der Unterschied kam wahrscheinlich nicht zufällig zustande.

Wahrscheinlichkeitsverteilungen, die für Signifikanztests herangezogen werden, sind durch die Anzahl der Freiheitsgrade bestimmt. Je mehr Freiheitsgrade vorhanden sind, d.h. je größer die untersuchte Stichprobe ist, desto schmaler sind Wahrscheinlichkeitsverteilungen. Freiheitsgrade sind diejenigen Werte einer Verteilung, die jeden beliebigen Wert annehmen können, ohne das Ergebnis zu verändern. Bei großen Stichproben, d.h. einer großen Zahl an Freiheitsgraden, werden bereits kleine Unterschiede signifikant, wogegen bei kleinen Stichproben Unterschiede der Kennwerte zwischen zwei Gruppen groß sein müssen, um ein signifikantes Ergebnis zu erzielen. Ein signifikantes Ergebnis ist daher nicht zwingend auch inhaltlich relevant. Der beobachtete oder ein noch größerer Unterschied ist lediglich wenig wahr-

scheinlich, wenn in der Grundgesamtheit kein Unterschied vorliegt. In diesem Fall gilt die Nullhypothese.

Hypothesentesten

Mit dem Signifikanztest wird ***nicht*** die Wahrscheinlichkeit für bestimmte Unterschiede in der Grundgesamtheit geschätzt, sondern die Wahrscheinlichkeit für ein Ergebnis (z.B. einen Unterschied) in der Stichprobe, wenn in der Grundgesamtheit die Nullhypothese gilt. Ein Unterschied oder ein Zusammenhang ist signifikant, wenn die Wahrscheinlichkeit für ein Ergebnis in der Stichprobe ausgehend von der Nullhypothese in der Grundgesamtheit kleiner als 5% ist.

In der *Nullhypothese* (H_0) werden keine Unterschiede zwischen Kennwerten zweier Stichproben in der Grundgesamtheit erwartet bzw. keine Zusammenhänge zwischen Merkmalen. Diese Annahme widerspricht eigentlich der Fragestellung und in den (inhaltlichen) Hypothesen einer wissenschaftlichen Arbeit. Das Ziel des Forschers ist es, die Nullhypothese zurückzuweisen und die *Alternativhypothese* (H_A) anzunehmen (s. 4.4.2). Die Nullhypothese wird zurückgewiesen, wenn ein Kennwerteintervall ausgehend von der zugrundeliegenden Wahrscheinlichkeitsverteilung (z.B. Normalverteilung) weniger als 5% wahrscheinlich ist.

Zwei- oder einseitiges Testen?

Hypothesentests werden ein- oder zweiseitig berechnet. Ein *einseitiger Test* erfordert eine gerichtete Hypothese (s. 4.4). Gerichtete Hypothesen geben die Richtung des Unterschieds bzw. des Zusammenhangs vor (Männer sind *größer als* Frauen, Körpergröße und Körpergewicht stehen in *positivem* Zusammenhang – wer größer ist, ist schwerer).

Das *zweiseitige Testen* erfolgt, wenn in der inhaltlichen Hypothese keine Richtung des Unterschieds oder des Zusammenhangs formuliert wurde (Männer und Frauen sind unterschiedlich groß, Körpergröße und Körpergewicht hängen zusammen).

Forschungshypothesen sollten prinzipiell gerichtet formuliert werden. Bei einer sehr schwacher theoretischer Fundierung kann es jedoch notwendig sein, Annahmen explorativ zu überprüfen – eine Richtungsannahme also erst nach der Untersuchung zu formulieren. In diesem Fall muss der Unterschied oder der Zusammenhang höher sein als beim einseitigen Testen, damit ein Unterschied signifikant wird (s. 9.3).

Abbildung 37 stellt Annahmen und Prinzipien statistischer Signifikanztests dar. Die linke Seite stellt die angenommene Verteilung in der Grundgesamtheit dar. Innerhalb dieser Verteilung sind bestimmte Kennwerte mehr oder weniger wahrscheinlich (s. dazu auch 8.4.5). Die dunkel gezeichneten Flächen der linken Verteilung markieren den Ablehnungsbereich der Nullhypothese, Werte in diesem Bereich sind weniger als 5% wahrscheinlich. Ein Kennwert in der Stichprobe, der in diesen Be-

reich fällt, könnte ebenso gut mit einer größeren Wahrscheinlichkeit in einer anderen Grundgesamtheit/Verteilung auftreten (rechte Verteilung in Abbildung 37).

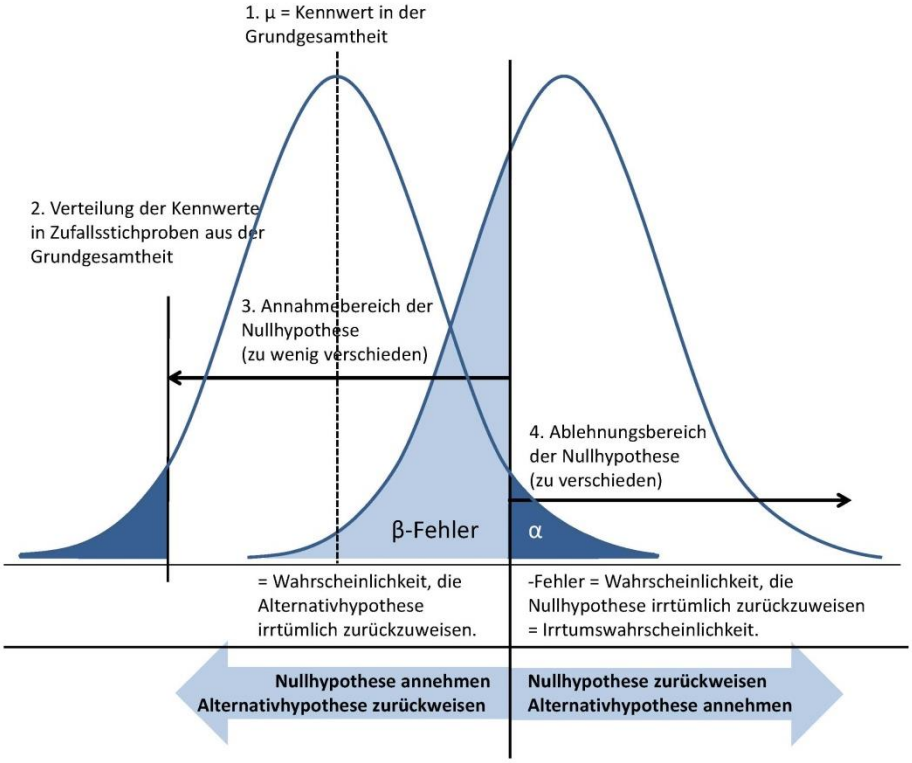


Abbildung 37: Darstellung des Konzepts Signifikanz mit α- und β-Fehlerwahrscheinlichkeit (eigene Darstellung)

Aussagekraft von Signifikanztests (s. auch 9.1.4)

Signifikanztests testen die Gültigkeit der Nullhypothese, also der Wahrscheinlichkeit von Kennwerten in der Stichprobe, sofern die Nullhypothese in der Grundgesamtheit gilt. Diese Überprüfung erfolgt auf Basis von Wahrscheinlichkeitsverteilungen, deren Anwendung für bestimmte Merkmale und Hypothesen in den folgenden Gliederungspunkten beleuchtet wird. Das Ergebnis eines Signifikanztests erlaubt ausschließlich die folgende Aussage: *„Der vorliegende oder ein noch extremerer Unterschied von Kennwerten (oder Zusammenhängen) in der Stichprobe, ist mit der Irrtumswahrscheinlichkeit „p“ möglich, wenn in der Grundgesamtheit kein Unterschied/Zusammenhang vorliegt, d.h. die Nullhypothese gilt.“*

Es ist weder ein Rückschluss auf die wahren Unterschiede in der Grundgesamtheit möglich, noch auf eine Wahrscheinlichkeit, mit der errechnete Unterschiede oder Zusammenhänge in der Grundgesamtheit vorliegen. Die Irrtumswahrscheinlichkeit „p“ weist nicht auf die Wahrscheinlichkeit des Zutreffens oder Nichtzutreffens der Nullhypothese hin, sondern nur auf die Wahrscheinlichkeit, mit der ein Zurückweisen der Nullhypothese fälschlicher Weise erfolgt (s. 9.1.2).

9.1.2 Wer schlussfolgert, kann irren – α- und ß-Fehler beim Hypothesentesten

Die Schlussfolgerung aus dem Signifikanztest kann fehlerbehaftet sein und unterliegt einer *Irrtumswahrscheinlichkeit* bei der Annahme oder dem Zurückweisen von Null- oder Alternativhypothese. Unterschieden wird in den Fehler erster (α-Fehler) und zweiter Art (β-Fehler).

Der *α-Fehler*, auch als Irrtumswahrscheinlichkeit bezeichnet, bedeutet, auch wenn ein Kennwert in einer Verteilung weniger als 5% Wahrscheinlich ist, kann er dennoch zu dieser Verteilung gehören. Unterhalb der Normalverteilungskurve (und jeder anderen Verteilungskurve auch) ist prinzipiell jeder Wert möglich. Extreme Werte werden jedoch zunehmend unwahrscheinlich. Ein Kennwert, der in einer Verteilung weniger als 5% wahrscheinlich ist, könnte in einer anderen Verteilung mit größerer Wahrscheinlichkeit auftreten (s. 9.1.1). Diese Annahme begründet das Zurückweisen der Nullhypothese, auch wenn diese Entscheidung mit einer Wahrscheinlichkeit von 5 % oder weniger (hängt ab von der Ausprägung des Kennwerts) irrtümlich erfolgt sein kann. Die Irrtumswahrscheinlichkeit eines Signifikanztests wird in Studien berichtet und erhält den Buchstaben „p“ (Tabelle 19).

Tabelle 19: α- und β-Fehler bei der Entscheidung für H_0 oder H_A (Beller, 2008, S. 104)

Entscheidung in der Stichprobe für	In der Population gilt	
	H_0	H_A
H_0	korrekt	**β-Fehler**
H_A	**α -Fehler**	korrekt

Als *ß-Fehler* wird das *irrtümliche Zurückweisen der Alternativhypothese* bezeichnet. Wenn ein Kennwert in der angenommenen Wahrscheinlichkeitsverteilung der Grundgesamtheit liegt (linke Verteilung in Abbildung 37) und darin zu mehr als 5% wahrscheinlich ist, könnte er dennoch mit einer bestimmten Wahrscheinlichkeit zu einer anderen Verteilung gehören (rechte Seite in Abbildung 37).

α- und β-Fehler beschreiben die Wahrscheinlichkeit von Fehlschlüssen, die bei der Bestimmung der Wahrscheinlichkeit eines Kennwerts in der Stichprobe unter Annahme einer Verteilung in der Grundgesamtheit entstehen (Tabelle 19).

9.1.3 Welcher konkrete Wert fällt in den Ablehnungsbereich der Nullhypothese? Beispielrechnung auf Basis der Beispielstudie (8.2, 88.4)

Die Wahrscheinlichkeit für bestimmte Intervalle von Messwerten eines Merkmals kann auf der Basis von Wahrscheinlichkeitsverteilungen berechnet werden. Am Bei-

spiel Einkommen aus der Beispielstudie wird der untere und obere Ablehnungsbereich der Nullhypothese auf Basis der Normalverteilung durch Umstellung der Formel zur Berechnung der z-Werte berechnet (F. 10). Zunächst werden dazu anhand der z-Tabelle (Anhang 1) die z-Werte für den 5% Flächenanteil (links vom arithmetischen Mittel) unter der Normalverteilungskurve, anschließend der z-Wert für den 95%-Anteil (bleibt ein Rest von 5% Flächenanteil rechts), bestimmt. Für die linke Seite der Verteilung ist dies ein z-Wert von -1,64, (5% Flächenanteil) für die rechte Seite ein z von 1,64 (95% Flächenanteil, Rest 5%). Weniger als 5% wahrscheinlich sind im unteren Teil der Verteilung -43,60 EUR, im oberen Teil der Verteilung (rechts) 4.883,60 EUR. Einkommenswerte jenseits dieser Grenzen gehören evtl. zu einer anderen Grundgesamtheit, in der sie wahrscheinlicher sind (Abbildung 38).

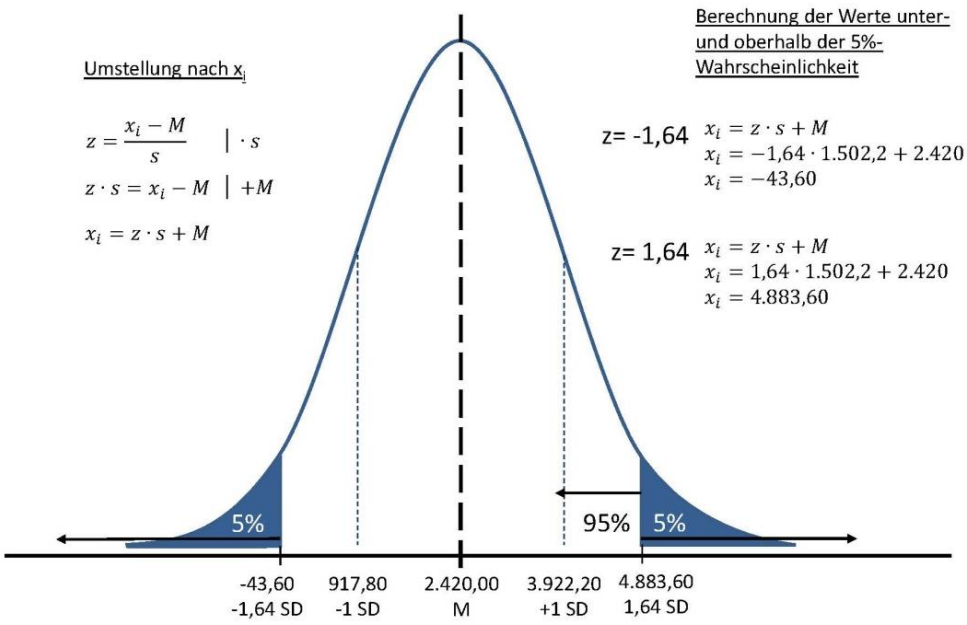


Abbildung 38: Beispiel der Berechnung „unwahrscheinlicher Werte“ am Beispiel der Variable Einkommen (eigene Darstellung)

9.1.4 Signifikanz, Relevanz und Effekt

Ob ein Unterschied oder Zusammenhang signifikant wird, ist nicht nur von der Größe des Unterschieds oder der Stärke des Zusammenhangs abhängig. Daneben ist die Stichprobengröße ein entscheidender Einflussfaktor auf die Signifikanz. Wahrscheinlichkeitsverteilungen auf deren Basis eine Einschätzung der Irrtumswahrscheinlichkeit erfolgt, sind durch die Freiheitsgrade determiniert. Je mehr Freiheitsgrade – d.h. überwiegend, je größer die untersuchte Stichprobe ist, desto schmalgipfliger sind Wahrscheinlichkeitsverteilungen. In der Mitte der Verteilung ist dann eine größere Zahl an Fällen als an ihren Rändern. Unterschiede müssen dann nicht mehr sehr groß sein, damit ein Stichprobenkennwert in den Ablehnungsbereich der Nullhypothese fällt.

Anhand eines Signifikanztests kann daher *nicht auf die inhaltliche Bedeutung* eines Kennwerteunterschieds oder eines Zusammenhangs geschlossen werden, sondern nur ob dieser Unterschied oder der gefundene Zusammenhang unter der Annahme einer bestimmten Verteilung in der Grundgesamtheit wahrscheinlich ist. Es ist Aufgabe des Forscherteams die inhaltliche Relevanz signifikanter Ergebnisse einzuschätzen. Basis für eine Bewertung signifikanter Ergebnisse können in der Theorie formulierte Überlegungen oder klinische Normwerte sein.

Daneben wird empfohlen, Effektgrößen signifikanter Ergebnisse zu berichten. Eine Effektgröße wird bestimmt von der Größe des Unterschieds zwischen zwei Kennwerten, der Varianz der Kennwerteverteilung und aus der Enge eines Zusammenhangs (s. 8.5). Döring et al. (2016, S. 821) geben einen Überblick über verschiedene Effektgrößen zu den wichtigsten Signifikanztests. Die Berechnung von Effektgrößen wird in diesem Studienbrief nicht vertieft, es wird jedoch auf die Notwendigkeit der Effektgrößenbestimmung insbesondere bei der Berechnung von Signifikanztests in großen Stichproben verwiesen.

Signifikanz allein ist zusammenfassend kein hinreichendes Kriterium zur Einschätzung der inhaltlichen und klinischen Bedeutung von Studienergebnissen.

9.2 Bernoulli-Experimente

Die Aussage von Signifikanztests kann anhand eines weiteren Beispiels skizziert werden. Wir erinnern uns: ein Signifikanztest überprüft Hypothesen, i.d.R. unter der Annahme, dass in der Grundgesamtheit kein Unterschied zwischen verschiedenen Gruppen erkennbar wird (Nullhypothese). Basis ist also die Annahme einer Verteilung der Kennwerte in der Grundgesamtheit, die zufällig zu unterschiedlichen Ergebnissen in Stichproben führen kann (7.3). Allerdings sollten diese Unterschiede nur zufällig auftreten, also so klein sein, dass die Nullhypothese beibehalten werden kann. Nur wenn sich systematische Unterschiede finden, also Werte, die in den Ablehnungsbereich der Nullhypothese fallen (linker und rechter 5%-Flächenteil der Wahrscheinlichkeitsverteilung, Abbildung 37), wird von einem signifikanten Ergebnis gesprochen. Bernoulli Experimente ermöglichen die Bestimmung der Wahrscheinlichkeit von Verteilungen bei maximal zwei möglichen Ereignissen bzw. Ausprägungen eines Merkmals (z.B. Männer vs. Frauen). Auf Basis der Annahme, Männer und Frauen kommen gleich häufig in der Grundgesamtheit vor, kann überprüft werden, ob von zehn Führungspositionen nur zwei mit Frauen besetzt sind auf Zufall beruht, also mit der Annahme einer 50:50-Verteilung vereinbar ist, oder ob der Unterschied wenig wahrscheinlich ist, wenn in der Grundgesamtheit eine 50:50-Verteilung vorliegt.

In der Hochschulmedizin werden 10% der Lehrstuhlprofessuren von Frauen besetzt (academics.de, o.A.). Dagegen sind ca. zwei Drittel der Studienanfänger in der Medizin weiblich (Hibbeler & Korzilius, 2008). Sofern die Leistungen von Medizinstudentinnen denen ihrer männlichen Kollegen entsprechen, sie ähnliche Ziele in ihrer beruflichen Karriere verfolgen, wissenschaftlich ähnlich interessiert sind, ebenso zielstrebig eine wissenschaftliche Laufbahn verfolgen und dieselben Chancen haben, wie ihre männlichen Kommilitonen, sollten zwei Drittel der Lehrstühle in der Medizin mit Frauen besetzt sein. Dass dies nicht so ist, zeigen die veröffentlichten Zahlen. Aus denen könnten verschiedene Schlüsse gezogen werden. Einer dieser Schlüsse könnte sein, Frauen werden systematisch benachteiligt oder haben schlichtweg kein Interesse an einer wissenschaftlichen Laufbahn in der Medizin. Ebenso gut kann die stark von den Studienanfängerzahlen abweichende Geschlechterverteilung bei den Lehrstuhlinhabern Zufall sein. Mit Bernoulli-Experimenten kann getestet werden, ob die Wahrscheinlichkeit für lediglich eine Frau auf zehn Medizinlehrstühle wahrscheinlich ist, wenn zwei Drittel der Studienanfänger in der Medizin weiblich ist (p = 0,66) (F. 14).

In (F. 14) wird mit $\binom{n}{k}$ die Anzahl möglicher Kombinationen berechnet, wenn aus einer bestimmten Anzahl (n) von Versuchen eine Anzahl (k) Treffer erzielt werden soll. Der Term beruht auf Annahmen der Wahrscheinlichkeitstheorie zur „Kombinatorik ohne Zurücklegen“, die an dieser Stelle nicht vertieft wird. Es ist dennoch empfehlenswert, sich mit Grundlagen der Wahrscheinlichkeitstheorie zu befassen, die über die Ausführungen dieses Studienbriefs hinausgehen, nicht nur um die (Un-) Wahrscheinlichkeit eines Sechser im 6 aus 49 Lotto zu berechnen oder beim Black Jack Karten zu zählen. In jedem Lehrbuch, in dem Statistik vertieft behandelt wird, ist ein ausführlicher Abschnitt zur Wahrscheinlichkeitstheorie vorhanden. Empfehlenswert sind die Bücher von Leonhart (2013) sowie von Rumsey, Engel, Baum und Kühnel (2012), die einen Einstieg auch mit mathematischen Grundkenntnissen ermöglichen.

(F. 14)	**Bernoulli-Experimente**	n = Versuchsanzahl
		k = Trefferzahl
	$p(k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$	p = Wahrscheinlichkeit
		1 – p = Gegenwahrscheinlichkeit
	$p(k) = \frac{n!}{k! \cdot (n-k)!} \cdot p^k \cdot (1-p)^{n-k}$	n! = n-Fakultät: n · (n - 1) · (n - 2) ·… · 2 · 1
		k! = k-Fakultät: k · (k - 1) · (k - 2) ·… · 2 · 1

Formel: Leonhart (2013)

Die konkrete Berechnung der Wahrscheinlichkeit für eine oder weniger als eine Frau unter zehn Medizinerlehrstühlen wird in drei Schritten berechnet (Leonhart, 2013):

1. Berechnung der Wahrscheinlichkeit, dass *keine* Frau unter den Medizinerlehrstuhlinhabern zu finden ist [p(0)]

2. Berechnung der Wahrscheinlichkeit, dass *eine* Frau unter den Medizinerlehrstuhlinhabern zu finden ist [p(1)]
3. Addieren der in 1. und 2. berechneten Wahrscheinlichkeiten [p(0) + p(1)]

Ist die kumulierte Wahrscheinlichkeit p(0) + p(1) kleiner als 5% kann nicht von einer zufälligen Abweichung gesprochen werden. Es entscheiden sich dann signifikant weniger Frauen für eine wissenschaftliche Laufbahn in der Medizin, als die Studienanfängerzahlen erwarten lassen – bzw. Frauen sind systematisch benachteiligt – je nach zugrundeliegender Theorie.

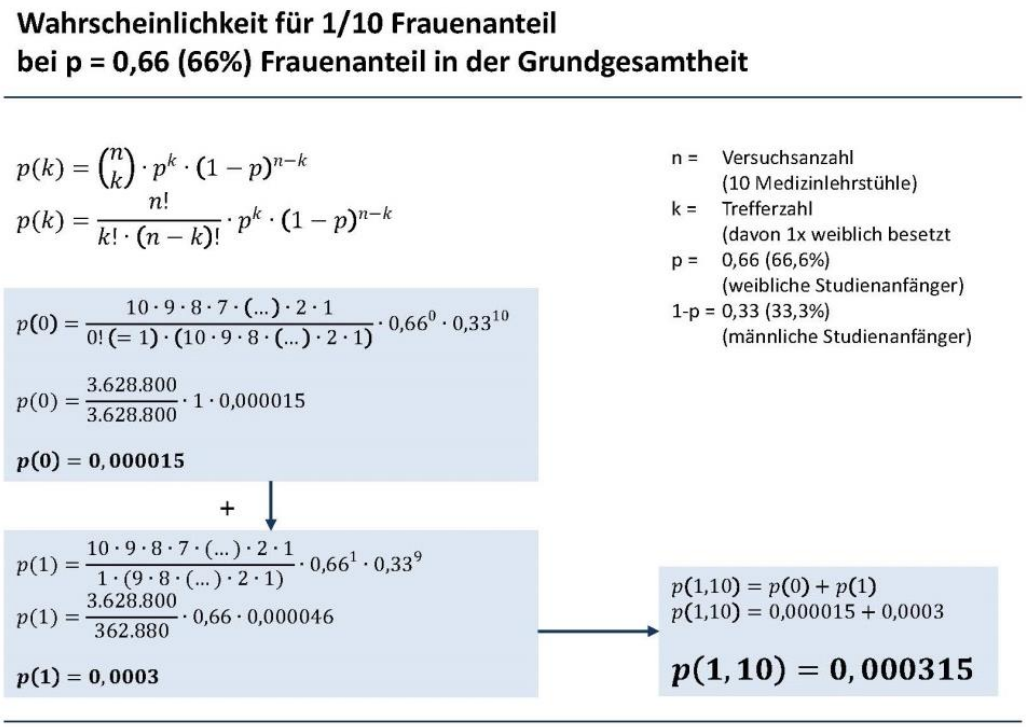


Abbildung 39: Berechnung der Wahrscheinlichkeit für 1/10 weiblichen Lehrstuhlinhabern bei 66,6% weiblichen Studienanfängern (eigene Darstellung)

Vorausgesetzt das Interesse an einer wissenschaftlichen Laufbahn in der Medizin, unterscheidet sich nicht zwischen Männern und Frauen und Frauen schließen ihr Medizinstudium ebenso häufig erfolgreich ab wie Männer, beruht bei 66,6% weiblichen Studienanfängern nur eine Lehrstuhlinhaberin auf zehn Lehrstühle in der Medizin wahrscheinlich nicht auf Zufall und ist weniger als 5% wahrscheinlich (p = 0,000315; 0,0315%) und fällt somit in den Ablehnungsbereich der Nullhypothese (9.1.1). Dies könnte auf fundamental unterschiedliche Interessen oder eine systematische Benachteiligung von Frauen in der Medizin zurückzuführen sein (Abbildung 39).

9.3 Signifikanztests bei Häufigkeiten, Rang- und Intervalldaten

Je nach inhaltlichem Interesse und Messniveau kommen verschiedene Signifikanztests infrage. Näher vorgestellt werden anschließend Signifikanztests für Unterschieds- bzw. Veränderungshypothesen (die auch von Unterschieden ausgehen) (9.3.1) und für Zusammenhangshypothesen (9.3.2). Auf unterschiedliche Verfahren ausgehend vom Messniveau (6.2.1) wird jeweils Bezug genommen.

Tabelle 20 systematisiert die häufig angewendeten Signifikanztests für Unterschiedshypothesen, die zugrundeliegende Wahrscheinlichkeitsverteilung und die Voraussetzungen ausgehend vom Messniveau und der Verteilung der Stichprobenkennwerte.

Tabelle 20: Voraussetzung für inferenzstatistische Standardverfahren mit der zugrundeliegenden Wahrscheinlichkeitsverteilung

	Testverfahren	**Voraussetzung**
1 Mittelwert	*einfacher t-Test* (t-Verteilung)	Intervallskala UND Normalverteilung UND Zufallstichprobe
2 Mittelwerte	*doppelter t-Test* (unabhängige und verbundene Stichproben) (t-Verteilung)	Intervallskala UND Normalverteilung UND Zufallstichprobe
	U-Test (Paardifferenztest) *Wilcoxon-Test* (Vorzeichen-Rangtest)	Ordinal- ODER Intervallskala
>2 Mittelwerte	Einfaktorielle Varianzanalyse (ANOVA) (F-Verteilung)	Intervallskala UND Normalverteilung UND Zufallstichprobe
	Kruskal-Wallis-Test	Ordinal- ODER Intervallskala
Häufigkeiten	chi-Quadrat-Test (chi-Quadrat-Verteilung)	Nominal- Ordinal- ODER Intervallskala

9.3.1 Unterschieds- (Veränderungs-) Hypothesen

In Unterschiedshypothesen werden verschieden hohe Merkmalsausprägungen (Ordinal- und Intervallskalenniveau) oder Häufigkeiten (Nominalskalenniveau) bestimmter Merkmalsausprägungen in zwei oder mehreren Gruppen erwartet (s. 4.4).

Signifikanztest bei Häufigkeiten

Die Beschreibung von Häufigkeiten unterschiedlicher Merkmalsausprägungen kann für jedes Messniveau erfolgen, insbesondere für nominalskaliertes Datenmaterial (s. 8.3). Wenn in zwei oder mehr Gruppen bei einem Merkmal verschieden große Merkmalsausprägungen erwartet werden, erfolgt eine Darstellung dieser Merkmalsverteilung in Kreuztabellen (s. 8.5.1). Von einer unterschiedlichen Häufigkeitsverteilung kann auf Zusammenhänge zwischen zwei nominalskalierten Merkmalen geschlossen werden. Ob gemessene Unterschiede zufällig auftreten oder systematisch (also signifikant) sind, wird mit dem chi-Quadrat-Test überprüft. Die chi-Quadrat-Verteilung ist die der Berechnung des chi-Quadrat-Tests zugrundeliegende Wahrscheinlichkeitsverteilung (s. 9.1.1, Anhang 2).

(F. 15) **Freiheitsgrade beim chi-Quadrat-Test**

$$df = (p-1) \cdot (q-1)$$

p= Anzahl der Zeilen
q= Anzahl der Spalten

(F. 16) **Berechnung erwarteter Zellenhäufigkeiten**

$$f_{e(j,k)} = \frac{Zeilensumme_j \cdot Spaltensumme_k}{N}$$

(F. 17) **Chi-Quadrat-Test**

$$\chi^2 = \sum_{j=1}^{p} \sum_{k=1}^{q} \frac{(f_{b(j,k)} - f_{e(j,k)})^2}{f_{e(j,k)}}$$

$f_{b(j,k)}$ = beobachtete Zellenhäufigkeit
$f_{e(j,k)}$ = erwartete Zellenhäufigkeit

signifikant bei chi-Quadrat > kritisches chi-Quadrat

Formeln: Leonhart (2013)

Der chi-Quadrat-Test wird in den folgenden Schritten berechnet (F. 15), (F. 16), (F. 17) (Leonhart, 2013):

1. Berechnung der Freiheitsgrade (2 x 2-Tabellen haben einen Freiheitsgrad) (F. 15)

2. Bestimmung des kritischen chi-Werts aus der chi-Verteilungsfunktion (Anhang 2)

3. Berechnung der erwarteten Zellenhäufigkeiten (F. 16)

4. Berechnung des chi-Quadrat-Tests (F. 17).

5. Bestimmung der Effektgröße

Die Beispielrechnung in Abbildung 40 untersucht ausgehend von den Ergebnissen in Tabelle 18 anhand des chi-Quadrat-Tests, ob die beobachteten Häufigkeiten beim Hochschulabschluss von Männern und Frauen signifikant von dem abweichen, was rechnerisch zu erwarten wäre. Die zugrundeliegende inhaltliche Hypothese kann gerichtet oder ungerichtet formuliert sein. Die Beispielrechnung geht von einer gerichteten Hypothese aus, es wird erwartet, dass Frauen häufiger einen Hochschulabschluss haben als Männer. Der kritische chi-Wert bei *einem* Freiheitsgrad liegt bei 3,841. Damit das Ergebnis als signifikant eingestuft werden kann, muss der chi-Quadrat-Test einen größeren Wert als das kritische chi-Quadrat aufweisen. Dies ist im vorliegenden Beispiel nicht der Fall. Männer und Frauen haben in der Stichprobe unterschiedlich häufig einen Hochschulabschluss. Dieser Unterschied ist nicht signifikant.

1. + 2. Bestimmung der Freiheitsgrade und kritisches chi-Quadrat

$df = (p-1)\cdot(q-1)$
df=1

bei einer Irrtumswahrscheinlichkeit <0,05 muss bei 0,95 (1-seitig) oder 0,975 (2-seitig) abgelesen werden

1-α

ν \ λ	0,005	0,010	0,025	0,050	0,100	0,900	0,950	0,975	0,990	0,995
1	0,000	0,000	0,001	0,004	0,016	2,706	3,841	5,024	6,635	7,879

einseitiger Test — zweiseitiger Test

3. Berechnung erwarteter Zellenhäufigkeiten

$$f_{e(j,k)} = \frac{Zeilensumme_j \cdot Spaltensumme_k}{N} = \frac{10 \cdot 12}{20} = \mathbf{6}$$ → Für die Zelle „Frauen mit Hochschulabschluss“

4. Berechnung und Aufsummierung der quadrierten Differenzen zwischen beobachtetem und erwartetem Wert

	Hochschulabschluss ja (%) $f_{b(j,k)}$	ja (%) $f_{e(j,k)}$		Hochschulabschluss nein (%) $f_{b(j,k)}$	nein $f_{e(j,k)}$		Zeilensumme
Geschlecht							
weiblich (%)	8 (80)	6	0,66	2 (20)	4	0,66	10 (100)
männlich (%)	4 (40)		0,66	6 (60)	4	0,66	10 (100)
Spaltensumme	12 (60)			8 (40)			20

$$\chi^2 = \sum_{j=1}^{p}\sum_{k=1}^{q}\frac{(f_{b(j,k)} - f_{e(j,k)})^2}{f_{e(j,k)}}$$

$$\chi^2 = \frac{(8-6)^2}{6} + \frac{(4-6)^2}{6} + \frac{(2-4)^2}{4} + \frac{(6-4)^2}{4}$$

χ^2 =**2,67**

Abbildung 40: Berechnung des chi-Quadrat-Test bei einer 2 x 2-Tabelle (eigene Darstellung)

Die Ergebnisse eines chi-Quadrat-Tests werden in die Ergebnistabelle aufgenommen, ein Beispiel zeigt Tabelle 21. Berichtet wird neben dem empirischen chi-Quadrat, das mit Statistikprogrammen ausgegeben wird die Anzahl der Freiheitsgrade (df) (F. 16) und die Irrtumswahrscheinlichkeit (p).

Tabelle 21: Ergebnisdarstellung von chi-Quadrat-Tests in wissenschaftlichen Arbeiten

	Männlich (%)	**Weiblich (%)**	**Summe**	**chi² (df) p**
Verein ja	4 (66,7)	6 (42,9)	10	0,904 (1) 0,66
Verein nein	2 (33,3)	8 (57,1)	10	
Summe	6	14	20	
Bochum	3 (50,0)	5 (35,7)	8	1,236 (2) 0,46
Dortmund	1 (16,7)	6 (52,9)	7	
Essen	2 (33,3)	3 (21,4)	5	
Summe	6	14	20	

Signifikanztest bei Lagemaßen – arithmetisches Mittel in zwei Gruppen

Das arithmetische Mittel kann für intervallskalierte Daten berechnet werden. Unterschiede beim arithmetischen Mittel zwischen *zwei Gruppen* werden mit dem t-Test berechnet, sofern die Voraussetzungen laut Tabelle 20 erfüllt sind (Intervallskalenniveau und normalverteilte Daten). Beide Gruppen können als unabhängige oder abhängige Stichproben vorliegen. Bei *unabhängigen Stichproben* kann *jeder* Studienteilnehmer *einer* bestimmten Gruppe zugeordnet werden. Sofern dieselben Studienteilnehmer zwei Mal untersucht werden, ergeben die Messungen *abhängige Stichproben*. Sollen Unterschiede von arithmetischen Mitteln zwischen unabhängigen Stichproben untersucht werden, erfolgt dies mit dem t-Test für unabhängige Stichproben. Unterschiede im arithmetischen Mittel zwischen abhängigen Stichproben werden mit dem t-Test für abhängige Stichproben auf Signifikanz getestet. In der Beispielstudie wurde das Einkommen als intervallskalierte Variable erfasst, als Gruppierungsvariable liegen Geschlecht und Hochschulabschluss vor. Mit einem t-Test für unabhängige Stichproben wird ausgehend von den Daten der Beispielstudie in Tabelle 10 untersucht, ob Männer und Frauen signifikant unterschiedlich verdienen.

(F. 18) **Freiheitsgrade beim t-Test**

$$df = n_1 + n_2 - 2$$

(F. 19) **t-Test**

$$t = \frac{M_1 - M_2}{\hat{\sigma}_{\bar{x}_1 - \bar{x}_2}}$$

$$\hat{\sigma}_{M_1 - M_2} = \sqrt{\frac{(n_1 - 1) \cdot {s_1}^2 + (n_2 - 1) \cdot {s_2}^2}{n_1 + n_2 - 2} \cdot \left(\frac{1}{n_1} + \frac{1}{n_1}\right)}$$

n_1= Stichprobengröße Gruppe 1
n_2= Stichprobengröße Gruppe 2
M_1 = Mittelwert Gruppe 1
M_2 = Mittelwert Gruppe 2
${s_1}^2$ = Varianz Gruppe 1
${s_2}^2$ = Varianz Gruppe 1

signifikant bei t > kritisches t

Formeln: Leonhart (2013)

Folgende Schritte führen zum Ergebnis beim t-Test für unabhängige Stichproben (Leonhart, 2013):

- Berechnung der Freiheitsgrade (F. 18)
- Bestimmung des kritischen t anhand der Verteilungstabelle der t-Verteilung (Anhang 3)
- Berechnung von arithmetischem Mittel und Varianz für jede Gruppe (s. 8.4)
- Prüfung auf Homogenität der Varianzen zwischen beiden Stichproben. Dies wird in der folgenden Beispielrechnung nicht durchgeführt. Der dazu empfohlene Levene-Test auf Varianzhomogenität wird routinemäßig in Statistikprogrammen durchgeführt.
- Berechnung des t-Test (F. 19).
- Berechnung der Effektgröße (nicht in dieser Beispielrechnung, nähere Informationen zur Effektgröße s. 9.1.4 sowie im Detail bei Döring et al., 2016, S. 821).

Der t-Tests bei unabhängigen Stichproben wird beispielhaft in Abbildung 41 veranschaulicht. Der kritische t-Wert liegt bei einseitiger Testung (gerichtete Unterschiedshypothese) bei 1,734 (Auszug aus Anhang 3). Der empirische (d.h. berechnete) t-Wert liegt bei 5,25 und damit deutlich über dem kritischen Wert. Männer und Frauen unterscheiden sich hinsichtlich ihres Einkommens, wobei Männer signifikant besser verdienen als Frauen.

1. + 2. Bestimmung der Freiheitsgrade und kritisches t

$df = n_1 + n_2 - 2$

df=18

bei einer Irrtumswahrscheinlichkeit <0,05 muss bei 0,95 (1-seitig) oder 0,975 (2-seitig) abgelesen werden

1-α

λ / ν	0,9000	0,9500	0,9750	0,9900	0,9950	0,9995
18	1,330	1,734	2,101	2,552	2,878	3,922

einseitiger Test — zweiseitiger Test

3. + 4. Arithmetisches Mittel und Varianz für jede Gruppe und t-Test

	Einkommen	
	männlich	weiblich
	500,00	300,00
	1500,00	600,00
	2300,00	900,00
	2400,00	1400,00
	2700,00	1600,00
	2900,00	1700,00
	4000,00	1800,00
	4500,00	2200,00
	5000,00	2600,00
	6000,00	3500,00
M	3180,00	1660,00
s	1592,98	905,76
s^2	2537600,00	820400,00

$$\hat{\sigma}_{M_1-M_2} = \sqrt{\frac{(n_1-1)\cdot {s_1}^2 + (n_2-1)\cdot {s_2}^2}{n_1+n_2-2} \cdot \left(\frac{1}{n_1}+\frac{1}{n_1}\right)}$$

$$\hat{\sigma}_{M_1-M_2} = \sqrt{\frac{9\cdot 2.537.600 + 9\cdot 820.400}{18} \cdot \left(\frac{1}{10}+\frac{1}{10}\right)}$$

$$\hat{\sigma}_{M_1-M_2} = \mathbf{289,74}$$

$$t = \frac{M_1 - M_2}{\hat{\sigma}_{\bar{x}_1-\bar{x}_2}}$$

$$t = \frac{3.180 - 1.660}{\mathbf{289,74}}$$

$$t = \mathbf{5,25}$$

Abbildung 41: Berechnung des t-Tests bei unabhängigen Stichproben (eigene Darstellung)

Tabelle 22: Ergebnisdarstellung von t-Tests bei unabhängigen Stichproben in wissenschaftlichen Arbeiten

Variable	**Verein ja (SD)**	**Verein nein (SD)**	t (df) p
Puls t_0	62 (5,6)	74 (8,4)	2,367 (18) 0,02
Puls t_1	96 (10,3)	128 (14,6)	6,32 (18) <0,001
Puls t_2	70 (8,6)	82 (6,4)	3,281 (18) 0,002

Anmerkungen: SD = Standardabweichung, t_0 bis t_2 = Messzeitpunkte

Im Ergebnisbericht zum t-Tests für unabhängige und abhängige Stichproben werden der empirische t-Wert, die Freiheitsgrade (df) und die genaue Irrtumswahrscheinlichkeit (p) berichtet (Beispiel in Tabelle 22, Tabelle 23).

Tabelle 23: Ergebnisdarstellung von t- Tests bei abhängigen Stichproben in wissenschaftlichen Arbeiten

Variable	M (SD) [95% CI]	t (df)	p
Puls t_0	72 (5,6) [67,4;76,6]	3,767	(18)
Puls t_1	116 (10,3) [108,8;123,2]	<0,001	
Puls t_1	116 (10,3) [108,8;123,2]	3,281 (18)	0,002
Puls t_2	76 (7,8) [71,3;80,7]		
Puls t_0	72 (5,6) [67,4;76,6]	1,153 (18)	0,13
Puls t_2	76 (7,8) [71,3;80,7]		

Anmerkungen: SD = Standardabweichung, t_0 bis t_2 = Messzeitpunkte, [95% CI] = 95% Konfidenzintervall (s. 9.4)

Signifikanztest bei Lagemaßen – Median oder nicht normalverteilte Daten

Alternativ zum t-Test für unabhängige Stichproben, der explizit nur bei intervallskalierten und normalverteilten Daten berechnet werden sollte (Grenzwerttheorem beachten, 7.3), können sog. nicht parametrische (verteilungsfreie) Tests berechnet werden und so die Signifikanz von Unterschieden von Lagemaßen (z.B. Median) bestimmt werden. Nicht parametrisch bedeutet, dass beim Datenmaterial die Abstände zwischen gleichen Wertebereichen unterschiedlich sein können (Ordinalskalenniveau, s. 6.2.1) oder die Daten nicht normalverteilt sind (s. 8.4.5).

Das alternative Testverfahren zum t-Test für unabhängige Stichproben ist der U-Test nach Mann und Whitney (auch Rangdifferenztest). Die Werte beider Gruppen gehen nicht als Rohwerte, beispielsweise als konkrete Einkommenswerte, sondern als Rangwerte in die Untersuchung ein. Die Messwerte werden in Rangplätze überführt, derjenige mit dem höchsten Einkommen erhält den ersten Rang (Platz 1), das niedrigste Einkommen erhält den schlechtesten Rang (im Beispiel Platz 20). Die Rangplätze werden für die Gesamtstichprobe vergeben, anschließend erfolgt eine Trennung der Gruppen.

Eine weitere häufig genutzte nichtparametrische Alternative zum t-Test für abhängige Stichproben ist der Wilcoxon-Test (auch Vorzeichenrangtest). Hier wird wie beim U-Test ebenfalls mit den Rangplätzen und nicht den Rohwerten gerechnet.

Signifikanztest bei Lagemaßen mehr als zwei Gruppen

Die einfaktorielle Varianzanalyse (ANOVA) ist ein parametrisches Testverfahren zur Berechnung der Signifikanz von Unterschieden bei Lagemaßen zwischen mehr als zwei Gruppen. Die nicht parametrische Alternative ist der Kruskal-Wallis-Test, der ebenfalls Rangplätzen in die Berechnung aufnimmt (Tabelle 20).

Bei der Untersuchung von Unterschiedshypothesen werden zusammenfassend Verfahren für nominalskaliertes Datenmaterial (chi-Quadrat-Test), ordinalskaliertes und nicht normalverteiltes intervallskaliertes Datenmaterial (U-, Wilcoxon-Test und Kruskal-Wallis-Test) sowie für normalverteiltes intervallskaliertes Datenmaterial (t-Test für abhängige und unabhängige Stichproben, ANOVA) unterschieden. Dabei werden zunächst inhaltliche und statistische Hypothesen formuliert, Freiheitsgrade bestimmt, die zugrundeliegende Wahrscheinlichkeitsverteilung definiert und überprüft, mit welcher Wahrscheinlichkeit Unterschiede bei Gültigkeit der Nullhypothese auftreten. Signifikante Unterschiede sind unter Annahme der Nullhypothese weniger als 5% wahrscheinlich. Nach Berechnung der statistischen Bedeutsamkeit muss durch den Forscher die inhaltliche Relevanz auf theoretischer Grundlage oder anhand klinischer Normwerte bestimmt. Daneben geben Effektmaße Hinweise auf die Relevanz von Unterschieden.

9.3.2 Zusammenhangshypothesen

In 8.5 wurden bivariate Korrelationskoeffizienten vorgestellt. Anhand des t-Test für Produkt-Moment-Korrelationen wird beispielhaft die statistische Bedeutsamkeit des berechneten Korrelationskoeffizienten bestimmt. Die Nullhypothese beim Signifikanztest von Korrelationskoeffizienten erwartet keinen Zusammenhang zwischen zwei Merkmalen. Signifikante Korrelationen sind so hoch ausgeprägt, dass sie mit der Annahme, in der Grundgesamtheit liegt kein Zusammenhang vor, nicht vereinbar sind.

Die Signifikanztestung von Produkt-Moment-Korrelationen verläuft in folgenden Schritten (Leonhart, 2013):

- Berechnung der Freiheitsgrade analog zu den Freiheitsgraden beim t-Test für unabhängige Stichproben (F. 18)
- Bestimmung des kritischen t anhand der Verteilungstabelle der t-Verteilung (Anhang 3)
- Berechnung des t-Test für Produkt-Moment-Korrelationen (F. 20) und Vergleich zwischen kritischem und empirischen t-Wert ($t_{krit} < t_{empirisch}$)
- Der Korrelationskoeffizient ist zugleich die Effektgröße (s. 8.5).

(F. 20) **t-Test für Korrelationskoeffizienten gegen Null**

N= Stichprobengröße
r= Korrelationskoeffizient

$$t_{N-2} = \frac{r \cdot \sqrt{N-2}}{\sqrt{1-r^2}}$$

Formel: Leonhart (2013)

Ausgehend vom Zusammenhang zwischen Glück und Abschlussnote in 8.5.2 (S. 169) (r = -0,76) ergibt sich folgender Rechenweg (Leonhart, 2013):

Der Analyse erfolgt mit 18 Freiheitsgraden (s. 9.3.1) (N – 2 = df; 20 – 2 = 18)

Der kritische t-Wert liegt bei einseitigem Testen (Richtung des Zusamme hangs wird in den Hypothesen formuliert) bei 1,734, bei einer ungerichteten Hypothese beträgt der kritische t-Wert 2,101 (s. Anhang 3, Abbildung 41).

$$\boldsymbol{t_{N-2} = \frac{r \cdot \sqrt{N-2}}{\sqrt{1-r^2}}}$$

$$\boldsymbol{t_{N-2} = \frac{-0,76 \cdot \sqrt{18}}{\sqrt{1-0,58}}}$$

$$\boldsymbol{t_{N-2} = -4,98}$$

Der empirische t-Wert ist *größer* als die kritischen t-Werte bei gerichteten und ungerichteten Hypothesen. Der beobachtete bzw. ein noch extremerer Zusammenhang in der Stichprobe ist unwahrscheinlich, wenn in der Grundgesamtheit kein Zusammenhang besteht.

Bei einem Korrelationskoeffizienten von -0,76 ist ein enger, aussagekräftiger Zusammenhang zwischen Abschlussnote und Glück. Menschen mit einer guten Abschlussnote fühlen sich insgesamt deutlich glücklicher als Menschen mit einer schlechten Abschlussnote (s. 8.5).

Ergebnisse von Signifikanztests bei Korrelationskoeffizienten werden im Unterschied zu den bisher vorgestellten Testverfahren nicht im Detail, also mit empirischem t-Wert, Freiheitsgraden und Irrtumswahrscheinlichkeit angegeben (s. Tabelle 21 bis Tabelle 23), sondern stark vereinfacht berichtet. Zu jedem Korrelationskoeffizienten wird der Bereich, in dem sich die Irrtumswahrscheinlichkeit bewegt, mit Sternchen (*) gekennzeichnet. Eine Irrtumswahrscheinlichkeit über 0,05 erhält keine Kennzeichnung, 0,05 und weniger 1 Stern (*), 0,01 und weniger 2 Sterne (**), 0,001 oder weniger 3 Sterne (***) (Tabelle 24).

Tabelle 24: Ergebnisdarstellung von t- Tests bei Korrelationskoeffizienten mit Angabe von arithmetischem Mittel und Standardabweichung in wissenschaftlichen Arbeiten

Puls	**M (SD)**	**1**	**2**	**3**	**4**	**5**
1. t_0 Eigen-	74,9 (8,7)	1				
2. t_0 Fremd-	73,4 (10,1)	0,80**	1			
3. t_1 Eigen-	105,4 (25,3)	0,54*	0,51	1		
4. t_1 Fremd-	102,3 (24,1)	0,50	0,54*	0,97***	1	
5. t_2 Eigen-	73,4 (11,6)	0,84***	0,76**	0,35	0,41	1
6. t_2 Fremd-	72,0 (11,3)	0,66*	0,63*	0,55*	0,61*	0,80**

* p < 0,05, ** p < 0,01, *** p < 0,001

Signifikanztests bei Korrelationskoeffizienten werden zusammenfassend ebenfalls auf Basis einer Nullhypothese berechnet, wobei in der Grundgesamtheit kein Zusammenhang erwartet wird. Für Produkt-Moment-Korrelationskoeffizienten werden ebenfalls t-Werte berechnet und mit einem kritischen t-Wert verglichen (Anhang 3). Sofern ein signifikanter Zusammenhang vorliegt bedeutet dies, ausgehend von der Annahme, in der Grundgesamtheit liegt kein Zusammenhang zwischen zwei Merkmalen vor, ist die Wahrscheinlichkeit für den empirischen Zusammenhang in der Stichprobe kleiner als 5%. Der Korrelationskoeffizient ist zugleich das Effektmaß einer bivariaten Korrelation (8.5, Döring et al., 2016, S 821).

9.4 Das Konfidenzintervall

Innerhalb des Konfidenzintervalls liegt mit einer bestimmten Wahrscheinlichkeit ein Teil der Stichprobenkennwerte, wenn eine Studie mit unterschiedlichen Zufallsstichproben aus derselben Grundgesamtheit mehrmals wiederholt wird. Die Zuordnung dieses Themas zur Inferenzstatistik ist eigentlich nicht ganz korrekt. Jedoch ist das Konfidenzintervall ein einfaches und anschauliches Maß, die Signifikanz von Unterschieden und bei Zusammenhängen zu schätzen. Häufig werden 95% oder 99% Konfidenzintervalle angegeben. Sofern dieselben Daten bei Zufallsstichproben aus derselben Grundgesamtheit erhoben werden, bewegen sich die ermittelten Stichprobenkennwerte (Lage- oder Zusammenhangsmaße) bei 95 von 100 Studien innerhalb des 95% Konfidenzintervalls. Konfidenzintervalle können für alle deskriptivstatistischen Kennwerte berechnet werden (s. 8). Anhand von 95% oder 99% Konfidenzintervallen sind Rückschlüsse auf die statistische Bedeutsamkeit möglich.

Auf Basis der Ergebnisse der Beispielstudie (Abbildung 41) wird anschließend das 95% Konfidenzintervall des arithmetischen Mittels berechnet. In die Berechnung des Konfidenzintervalls arithmetischer Mittel fließen Standardfehler und der z-Wert für das jeweilige Konfidenzintervall (95% oder 99%) ein (Anhang 1, 8.4.4).

Der *Standardfehler* beschreibt die Streuung ca. 68% der Stichprobenkennwerte um den Populationsmittelwert, wenn gleich große Stichproben aus einer Grundgesamtheit untersucht werden. Es fallen Ähnlichkeiten mit der Standardabweichung auf (8.4.3), die jedoch die Verteilung von Kennwerten in der *Stichprobe* beschreibt und nicht die Verteilung der Kennwerte vieler Zufallsstichproben in der *Grundgesamtheit*. Konfidenzintervalle haben eine Unter- und eine Obergrenze. Der Ablesepunkt des z-Werts für ein 95%-Konfidenzintervall liegt daher nicht beim 95% Flächenanteil in der z-Werte-Tabelle, sondern bei 97,5%. Der 95%-Flächenanteil bzw. das Intervall, in dem 95/100 Stichprobenkennwerte liegen wird um 2,5% nach rechts verschoben. Am linken (unteren) und am rechten (oberen) Ende der Verteilung bleibt jeweils ein Rest von 2,5%. Insgesamt liegen dann 5/100 Stichprobenkennwerte außerhalb des Konfidenzintervalls (Abbildung 42).

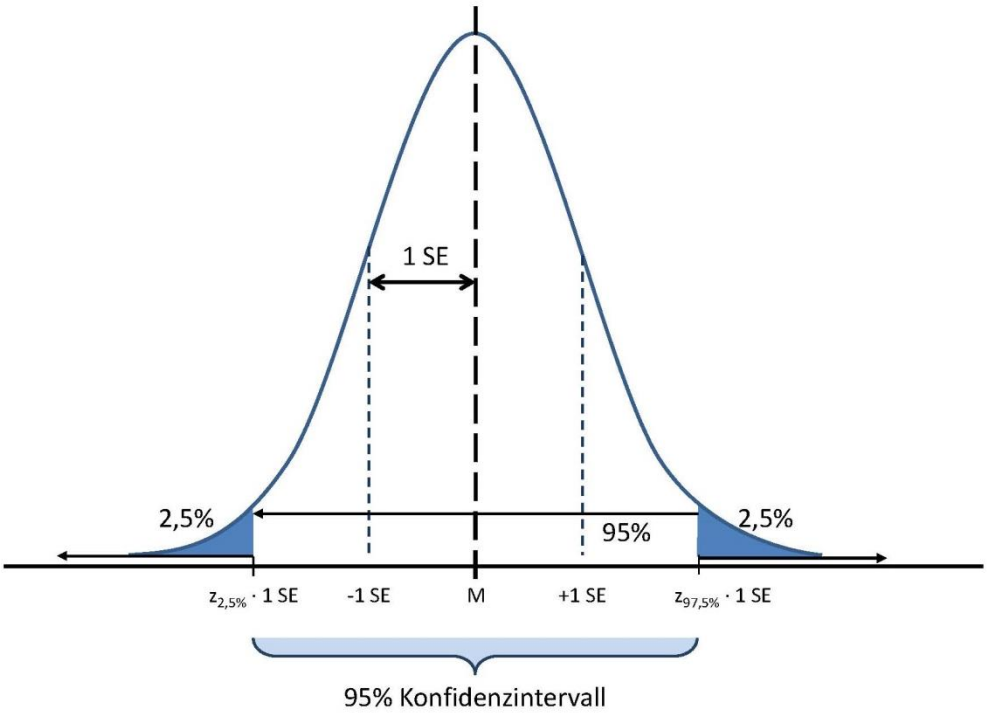


Abbildung 42: Standardfehler (SE) und 95% Konfidenzintervall (modifiziert nach Beller, 2008, S. 96)

(F. 21)	**Standardfehler des arithmetischen Mittels** $SE = \frac{s_x}{\sqrt{N}}$	SE = Standardfehler s_x = Standardabweichung N = Stichprobengröße
(F. 22)	**95% Konfidenzinervall des arithmetischen Mittels** $M \pm z_{97,5\%} \cdot SE$	$z_{97,5\%}$ = z-Wert des 97,5% Flächenanteils

Formeln: Beller (2008)

Bei der Berechnung von 95%-Konfidenzintervallen wird folgendes Vorgehen empfohlen:

1. Bestimmung des z-Werts zum 97,5% Flächenanteil der Normalverteilung (s. 8.4.4, Anhang 1) (1,96)

2. Bestimmung des Standardfehlers des Mittelwerts (F. 21)

3. Berechnung von Unter- und Obergrenze des 95% Konfidenzintervalls des arithmetischen Mittels

Der Rechenweg wird in Abbildung 43 veranschaulicht. Bereits in Abbildung 41 auf Seite 173 werden Unterschiede bei den arithmetischen Mitteln der Einkommen zwischen Männern und Frauen erkennbar, die signifikant sind. Auch anhand der 95%-Konfidenzintervalle lässt sich abschätzen, ob Unterschiede signifikant sind. Empfohlen wird eine Abschätzung jedoch erst bei Stichprobengrößen ab 20 pro Gruppe. Sehr kleine Stichproben sind mit vergleichsweise großen Standardfehlern verbunden. Im Nenner der Formel zur Berechnung des Standardfehlers wird die Quadratwurzel aus der Stichprobengröße gezogen (F. 20). Je größer die Stichprobe ist, desto größer ist der Wert im Nenner der Formel und desto kleiner das Ergebnis der Berechnung des Standardfehlers. Auch eine kleine Streuung der Kennwerte in den Gruppen (der Zähler in der Formel zur Berechnung des Standardfehlers) führt zu kleinen Standardfehlern.

1. Bestimmung des z-Werts 97,5%-Flächenanteils der Normalverteilung

z-Wert des 97,5% Flächenanteils →

1,95	97,44
1,96	97,50
1,97	97,56

→ 1,96

2. Standardfehler des arithmetischen Mittels für Männer und Frauen

Männer: $SE = \frac{s_x}{\sqrt{N}} = \frac{1.592{,}98}{\sqrt{10}} = \mathbf{503{,}74}$

Frauen: $SE = \frac{s_x}{\sqrt{N}} = \frac{905{,}76}{\sqrt{10}} = \mathbf{286{,}43}$

3. 95% Konfidenzintervalle der arithmetischen Mittel bei Männern und Frauen

	Einkommen	
	männlich	weiblich
	500,00	300,00
	1500,00	600,00
	2300,00	900,00
	2400,00	1400,00
	2700,00	1600,00
	2900,00	1700,00
	4000,00	1800,00
	4500,00	2200,00
	5000,00	2600,00
	6000,00	3500,00
M	3180,00	1660,00
s	1592,98	905,76
s^2	2537600,00	820400,00

Männer: $M \pm z_{97{,}5\%} \cdot SE$
$3.180 \pm 1{,}96 \cdot 503{,}74$
$3.180 \pm 987{,}33$

3.180,00 [2.192,67;4.167,33]

Frauen: $M \pm z_{97{,}5\%} \cdot SE$
$1.660 \pm 1{,}96 \cdot 286{,}43$
$1.660 \pm 561{,}40$

1.660,00 [1.098,60;2.221,40]

95%-Konfidenzintervalle

Abbildung 43: Berechnung von Standardfehler und 95% Konfidenzintervall des arithmetischen Mittels bei zwei Gruppen (eigene Darstellung)

Im berechneten Beispiel sind die Standardabweichungen vergleichsweise groß und die Stichprobengröße in den Gruppen mit N = 10 klein. Daher sind die 95% Konfidenzintervalle breit *und überschneiden sich*. Dies deutet auf nicht signifikante Unterschiede der arithmetischen Mittel des Einkommens zwischen Männern und Frauen hin (Abbildung 44).

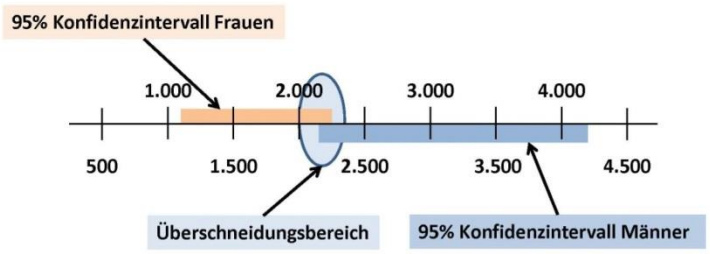


Abbildung 44: Rückschlüsse auf die Signifikanz der Unterschiede arithmetischer Mittel zwischen zwei Gruppen (eigene Darstellung)

Überschneidungsfreie 95%-Konfidenzintervalle erlauben die Annahme, beide Stichproben (hier Männer und Frauen) entstammen nicht derselben Grundgesamtheit. Der Einkommensunterschied zwischen beiden Gruppen wäre dann bei einer Irr-

tumswahrscheinlichkeit von kleiner als 5% signifikant. Diese Annahmen sind für 95% Konfidenzintervalle aller Kennwerte zulässig, die Zufallsstichproben entstammen (Tabelle 25). Anhand von Konfidenzintervallen sind Aussagen zur Plausibilität und zur Relevanz von Unterschieden und Zusammenhängen möglich (Prel, Hommel, Röhrig & Blettner, 2009). Konfidenzintervalle können gemeinsam mit Lage- und Streuungsmaßen berichtet werden. Im wissenschaftlichen Text wird anschließend auf mögliche signifikante Unterschiede und Zusammenhänge verwiesen (Ergänzung der 95%-CI in Tabelle 22, S. 173).

Tabelle 25: Rückschlüsse auf Signifikanz bei verschiedenen Kennwerten ausgehend vom 95% Konfidenzintervall (95% CI)

Kennwert	**Rückschluss auf die Signifikanz von Kennwerten und Kennwertunterschieden ausgehend von 95% CI**
Arithmetisches Mittel	Überschneidungsfreiheit
Relative Häufigkeiten	Überschneidungsfreiheit
Odds-Ratio	im 95% CI ist die „1,0“ nicht enthalten >1: Untergrenze des 95% CI ist **größer als 1,0** <1: Obergrenze des 95% CI ist **kleiner als 1,0** mehrere Odds-ratio: Überschneidungsfreiheit
Korrelationskoeffizient	im 95% CI ist die „0“ nicht enthalten >0: Untergrenze des 95% CI ist **größer als 0** <0: Obergrenze des 95% CI ist **kleiner als 0** mehrere Korrelationskoeffizienten: Überschneidungsfreiheit

Das Konfidenzintervall schätzt die Verteilung der Kennwerte bei mehrfacher Wiederholung einer Studie mit Zufallsstichproben aus derselben Grundgesamtheit. Für alle Kennwerte können Konfidenzintervalle geschätzt werden. Häufig wird das 95% Konfidenzintervall angegeben. Es bedeuten, wird eine Studie mit unterschiedlichen Zufallsstichproben 100 Mal wiederholt, so finden sich in 95 dieser Studien Kennwerte innerhalb des 95% Konfidenzintervall. Über den Vergleich von 95% Konfidenzintervallen sind Rückschlüsse auf die statistische Bedeutsamkeit von Unterschieden und Zusammenhängen möglich. Dabei werden nicht Hypothesen ausgehend von einer Wahrscheinlichkeitsverteilung getestet, sondern wahrscheinliche Kennwerteverteilung in der Grundgesamtheit aus Stichprobenkennwerten geschätzt. Überschneiden sich 95% Konfidenzintervalle zweier Mittelwerte nicht, entstammen die Stichpro-

ben mit einer „Irrtumswahrscheinlichkeit“ von 5% nicht aus derselben Grundgesamtheit.

9.5 Zusammenfassung und Aufgaben

Die Nutzung inferenzstatistikscher Verfahren stellt vereinfacht gesprochen eine Verbindung zwischen Daten einer Zufallsstichprobe und der Grundgesamtheit her. Ziel von Forschung sind Ergebnisse und Aussagen, die über eine Stichprobe hinaus Rückschlüsse für die Grundgesamtheit ermöglichen. Signifikanztests und die Berechnung von Konfidenzintervallen ermöglichen Wahrscheinlichkeitsaussagen ausgehend von Stichprobenkennwerten für die Grundgesamtheit. Signifikante Ergebnisse sind jedoch nicht zwangsläufig von inhaltlicher Bedeutung. Ebenso wenig lassen signifikante Ergebnisse Aussagen über die Gültigkeit oder Ungültigkeit von Aussagen für die Grundgesamtheit zu. Die Aussage signifikanter Ergebnisse beschränkt sich auf die Wahrscheinlichkeit von Unterschieden oder Zusammenhängen in der Stichprobe, wenn für die Grundgesamtheit keinerlei Unterschiede oder Zusammenhänge angenommen werden. Ob signifikante Unterschiede auch inhaltliche Relevanz haben, insbesondere bei sehr großen Stichproben, entscheidet der Forscher auf Basis der theoretischen Diskussion, in klinischen Studien auch anhand von Normwerten biologischer, chemischer und physikalischer Parameter. Ein Signifikanztest liefert eine Entscheidungshilfe, keine ausschließliche Grundlage für Entscheidungen, beispielsweise, ob eine bestimmte Therapie angewendet werden soll. Grundbegriffe in der Inferenzstatistik sind Null- und Alternativhypothese sowie α- und β-Fehler. Der α-Fehler quantifiziert die Irrtumswahrscheinlichkeit, mit der eine Nullhypothese zurückgewiesen wird, obwohl sie zutreffend ist. Die Wahrscheinlichkeit, mit der eine Alternativhypothese zurückgewiesen wird, beschreibt der β-Fehler.

Die Vorstellung von Signifikanz lässt sich über die Überprüfung der Wahrscheinlichkeit z.B. für die Geschlechterverteilung in einer Stichprobe, mit Bernoulli-Experimenten veranschaulichen. Bernoulli Experimente erlauben Rückschlüsse auf die Wahrscheinlichkeit der Verteilung von Merkmalsausprägungen bei Merkmalen mit zwei Ausprägungen (z.B. das Geschlecht) in der Stichprobe, wenn in der Grundgesamtheit eine bestimmte Verteilung beider Merkmalsausprägungen vorliegt. Der Frauenanteil bei Lehrstuhlinhabern in der Medizin ließ sich im Beispiel nicht mit Zufall erklären. Der Grund für den systematischen Unterschied zum Frauenanteil in der Grundgesamtheit findet sich in theoretischen Vorüberlegungen. Plausibel wären ein geringeres Interesse von Frauen an einer wissenschaftlichen Laufbahn oder ihre systematische Benachteiligung.

Signifikanztests gründen sich auf Annahmen über die wahrscheinliche Verteilung von Merkmalsausprägungen in der Grundgesamtheit. Es wird die Gültigkeit statistischer Hypothesen getestet, i. d. R. der Nullhypothese. Ergebnisse werden als signifikant gewertet, wenn die Irrtumswahrscheinlichkeit (p) kleiner als 5% ist. Typische Signifikanztests bei Unterschieden sind der chi-Quadrat-Test bei nominalskalierten

Daten und der t-Test für unabhängige und abhängige Stichproben bei normalverteilten intervallskalierten Merkmalen. Die Signifikanz von Unterschieden kann auch bei ordinalskalierten oder nicht normalverteilten Daten ermittelt werden. Bei unabhängigen Stichproben eignet sich der U-Test, bei abhängigen Stichproben der Wilcoxon-Test.

Konfidenzintervalle sind Schätzer der Verteilung von Stichprobenkennwerten bei mehrfacher Wiederholung einer Studie mit Stichproben aus derselben Grundgesamtheit. Der Vergleich verschiedener 95% Konfidenzintervalle ermöglicht Rückschlüsse darauf, ob Stichprobenkennwerte aus zwei (oder mehrere Gruppen) derselben Grundgesamtheit entstammen – in diesem Fall würden sie sich überschneiden.

Schlüsselwörter: *abhängige Stichproben, Alpha/α-Fehler, Alternativhypothese, arithmetische Mittel, Beta/β-Fehler, chi-Quadrat-Test, doppelter t-Test, Effektgröße/-stärke, einfacher t-Test, einseitiger Test, Freiheitsgrade, Grundgesamtheit, Häufigkeitsverteilung, Irrtumswahrscheinlichkeit, Konfidenzintervall, Korrelationskoeffizienten, Median, n, Nullhypothese, Nullhypothese, Produkt-Moment-Korrelation, Standardfehler, Stichprobe, t-Test, U-Test, unabhängigen Stichproben, Varianz, Varianzanalyse, Wahrscheinlichkeitsverteilungen, Wilcoxon-Test, z-Werte, zweiseitig Testen*

Aufgaben zur Lernkontrolle

1. Worin liegt der Unterschied zwischen einem ein- und zweiseitigem Test?

2. Was ist der α-Fehler, was der β-Fehler?

3. Wozu sind die Freiheitsgrade da?

4. Welche Voraussetzungen müssen für den chi-Quadrat, den einfachen t-Test, den zweifachen t-Test, den U-Test, den Wilcoxon-Test und der einfaktoriellen Varianzanalyse (ANOVA) gegeben sein, damit diese angewandt werden können? Verdeutlichen Sie Ihr Ergebnis in einer Tabelle.

5. Worin liegt der Unterschied zwischen abhängiger und unabhängiger Stichprobe?

6. Was bedeutet nicht parametrische Tests? Welche kennen Sie?

Aufgabe zur Berufspraxis

7. Wählen Sie eine quantitative Studie in Ihrem sprachtherapeutischen Themengebiet aus und schauen Sie sich die Studie genau an.
    a. Wie viele Versuchspersonen nahmen an der Studie teil?
    b. Welche Hypothesenart wurde in dieser Studie genutzt?
    c. Welcher statistische Test wurde verwendet? Erfüllt der verwendete Test die Anwendungsvoraussetzungen?
    d. Wurde das Lagemaß sinnvoll ausgewählt?
    e. Handelt es sich um abhängige oder unabhängige Stichproben?

Literatur

academics.de. (o.A.). Anteil an Medizin-Professorinnen wächst stetig. Retrieved 8. August 2016, from https://www.academics.de/wissenschaft/ professur_medizin_52245.html

Beller, S. (2008). *Empirisch forschen lernen Konzepte, Methoden, Fallbeispiele, Tipps* (2., überarb. Aufl ed.). Bern: Huber.

Döring, N., Bortz, J., & Pöschl, S. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften* (5. vollständig überarbeitete, aktualisierte und erweiterte Auflage ed.). Berlin Heidelberg: Springer.

Hibbeler, B., & Korzilius, H. (2008). Arztberuf: Die Medizin wird weiblich. *Deutsches Ärzteblatt International, 105*(12), 609.

Leonhart, R. (2013). *Lehrbuch Statistik Einstieg und Vertiefung* (3., überarbeitete und erweiterte Auflage ed.). Bern: Verlag Hans Huber.

Prel, J.-B. d., Hommel, G., Röhrig, B., & Blettner, M. (2009). Konfidenzintervall oder p-Wert? Teil 4 der Serie zur Bewertung wissenschaftlicher Publikationen. *Deutsches Ärzteblatt International, 106*(19), 335-339. doi: 10.3238/arztebl.2009.0335

Rumsey, D., Engel, R., Baum, K., & Kühnel, P. (2012). *Wahrscheinlichkeitsrechnung für Dummies* (2., überarb. Aufl. ed.). Weinheim: Wiley-VCH.

10 Präsentation und Veröffentlichung von Studienergebnissen

Lernergebnisse:

- Formen der Veröffentlichung von Forschungsergebnissen können beschrieben werden (10.1).
- Der peer-review-Prozess kann erläutert, die Bedeutung des peer-review-Verfahrens für die wissenschaftliche Reputation kann kritisch diskutiert werden (10.1).
- Aufbau und Inhalt der Gliederungspunkte wissenschaftlicher Arbeiten können differenziert und beschrieben werden. Anhand eines Beispielthemas kann die Anfertigung eines Abstracts mit der empfohlenen Gliederung für wissenschaftliche Arbeiten gelingen (10.2).
- Ausgehend von Beispielergebnissen kann eine knappe, einfache und präzise Beschreibung des theoretischen Hintergrunds eines selbst gewählten Themas und ein differenzierter Ergebnisbericht ausgehend von Beispieldaten gelingen (10.3).

Die Veröffentlichung von Forschungsergebnissen liegt im Interesse des Forschers und der *scientific community* zur Nutzung in weiterführenden Arbeiten. Schlussfolgerungen aus Studien fließen so in andere Studien ein und werden selber Basis für weitere Forschung. Nicht nur für die Durchführung von Studien und Forschung insgesamt bestehen Konventionen und Regeln, sondern auch für das Schreiben wissenschaftlicher Arbeiten. Im wissenschaftlichen Schreiben werden die Gütekriterien von Forschung, u.a. Nachvollziehbarkeit, Transparenz, Objektivität, weiter verfolgt. Daneben werden durch das wissenschaftliche Schreiben und die Veröffentlichung von Studienergebnissen, Forschungsergebnisse einer breiten Öffentlichkeit zugänglich gemacht. Forschende sollten dabei um eine einfache aber korrekte Sprache bemüht sein.

Die Veröffentlichungsliste von Wissenschaftlern ist wichtiges Kriterium für Forschungsleistung und *impact* von Forschung. Einer breiten Öffentlichkeit stehen Forschungsergebnisse allerdings bis heute nicht zur Verfügung, weil veröffentlichte Aufsätze oft nur gegen Geld erworben werden können. Zunehmend werden Arbeiten über Open-Access-Zeitschriften publiziert, die für den Nutzer kostenfrei sind.

Nachdem zunächst Typen von Veröffentlichungen differenziert werden, die sich insbesondere im Grad der Formalisierung und der wissenschaftlichen Qualität unterscheiden (10.1), wird anschließend der Aufbau und Inhalt wissenschaftlicher Arbeiten beschrieben (10.2). Daran schließen sich Formulierungsempfehlungen und sogenannte „no-go's" (10.3) sowie eine Empfehlung zur zeitlichen Planung im Schreibprozess von Abschlussarbeiten (10.4).

10.1 Veröffentlichung wissenschaftlicher Arbeiten

Döring et al. (2016) sehen in wissenschaftlichen Veröffentlichung die „wichtigste ‚Währung' der Wissenschaft" (S. 787). Verschiedene Publikationstypen können unterschieden werden. Grundsätzlich kann jedes Medium, das Forschungsergebnisse enthält, als Veröffentlichung gesehen werden. Der Wert einer Publikation bemisst sich an ihrem „Impact", also ihrer Verbreitung und wie häufig sie zitiert wird. Dieser Impact ist umso größer, je besser die Qualität eines wissenschaftlichen Mediums ist. Für wissenschaftliche Veröffentlichungen ist das Peer-Review-Verfahren ein verbreitetes Instrument zur Qualitätssicherung. Dabei wird nicht allein vom Herausgeber einer Zeitschrift oder vom Verlag entschieden, ob eine Arbeit zur Veröffentlichung angenommen wird, sondern zwei oder mehr im Prozess anonym bleibende Gutachter lesen und bewerten eine zur Veröffentlichung eingereichte Arbeit. Nicht nur die Gutachter bleiben im Peer-Review-Prozess anonym, auch die Autoren werden den Gutachtern nicht genannt. Dieses Verfahren soll Einflussnahme auf wissenschaftliche Publikationen verhindern und die Qualität wissenschaftlicher Veröffentlichungen sichern (Abbildung 45). Zwar nutzen die meisten ernstzunehmenden wissenschaftlichen Medien das Peer-Review-Verfahren, allerdings wird es bereits seit einigen Jahren kritisiert. Kritiker bemängeln ein System der Zensur, das bestimmte Themen und Forschungsfelder ausschließt, und für Vetternwirtschaft und Betrug

anfällig sei. Ob das häufig genutzte verblindete Peer-Review-Verfahren die Qualität wissenschaftlicher Veröffentlichungen sichert, wird außerdem kritisch diskutiert (Fröhlich, 2003). Das verblindete (also anonymisierte) Peer-Review-Verfahren wird in einigen internationalen Zeitschriften durch ein offenes Verfahren ersetzt, indem im Begutachtungsprozess alle Namen bekannt sind und die Rückmeldung der Gutachter ebenfalls offengelegt wird.

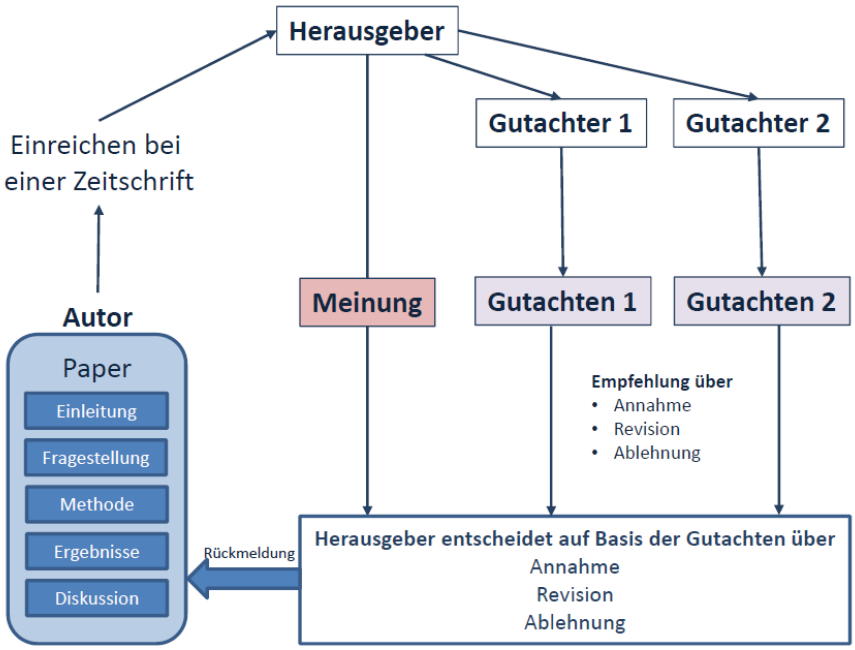


Abbildung 45: Peer-Review-Verfahren (eigene Darstellung)

Alle (wissenschaftlichen) Veröffentlichungen sind in wissenschaftlichen Arbeiten grundsätzlich zitierfähig. Allerdings bemisst sich der Wert einer eigenen Arbeit auch an der Aussagekraft und der Qualität des zitierten Materials. Eine Arbeit, die überwiegend Webseiten zitiert oder graue Schriften, die ohne peer-review-Verfahren veröffentlicht wurden, hat eine geringere Aussagekraft als Arbeiten, in denen peer-reviewte Aufsätze und Bücher zitiert werden. Bereits in Hausarbeiten im Studium legen Lehrende Wert auf eine aussagekräftige Literaturliste, in der klassische Internetseiten nur in sehr kleinem Umfang aufgenommen sein sollten. Peer-reviewte Veröffentlichungen können vorliegen als (u.a. Döring et al., 2016):

- *Monografie*, dies sind größere Arbeiten eines oder mehrerer Autoren, die sich *einem* Forschungsthema widmen. Häufig werden Dissertations- und Habilitationsschriften als Monografie veröffentlicht

- *Beitrag in einem Herausgeberwerk*, Herausgeberwerke nehmen mehrere Aufsätze und Beiträge unterschiedlicher Autoren auf. Der verantwortliche Herausgeber koordiniert die Anfertigung des Werks und kann selber als Autor eines oder mehrerer Beitrags in einem Herausgeberwerk auftreten

- *Aufsatz in einer Fachzeitschrift*, Forschungsergebnisse werden häufig zunächst in Fachzeitschriften publiziert. Für die Erarbeitung des theoretischen Rahmens und des Forschungsstands zu einem Thema sind Aufsätze in Zeitschriften das aktuelle und aussagekräftige Medium

- *Wissenschaftliches Abstract*, nach Fachkongressen und wissenschaftlichen Tagungen werden Zusammenfassungen der Beiträge teils in Fachzeitschriften als Abstractband veröffentlicht. Diese Abstracts sind zitierfähig und haben dieselbe Aussagekraft wie andere peer-reviewte Publikationen

- *Vortrag auf Fachkongressen*, Vorträge, die nach einem Begutachtungsverfahren auf wissenschaftlichen Kongressen gehalten werden, sind ebenfalls aussagekräftige Veröffentlichungen. Sie sind zitierfähig, sollten allerdings nur einen kleinen Anteil des Literaturverzeichnisses einnehmen

- *Posterpräsentation auf Fachkongressen*, Poster werden im Rahmen von Postersessions auf Fachkongressen präsentiert. Sie sind inhaltlich wie wissenschaftliche Arbeiten aufgebaut, nehmen dabei visuelle Elemente in der Präsentation auf.

Die genannten Veröffentlichungen liegen häufig auch nicht „*peer-reviewed*“ vor. Teils ist nicht leicht zu erkennen, ob es sich um eine begutachtete Veröffentlichung handelt oder nicht. Grundsätzlich sollten sowohl *peer-reviewed* als auch nicht begutachtete Veröffentlichungen, die in einer wissenschaftlichen Arbeit zitiert werden, zurückhaltend und kritisch diskutiert werden. Keine Veröffentlichung ist fehlerfrei und kaum eine Studie liefert unverzerrte, fehlerfreie Ergebnisse. In wissenschaftlichen Arbeiten, egal ob *peer reviewed* oder nicht, sollten in einem eigenen Gliederungspunkt am Ende der Arbeit potenzielle Fehlerquellen offengelegt und Ergebnisse ausgehend davon relativiert werden (s. 10.2).

Neben den genannte offiziellen, quasi-qualitätsgesicherten zitierfähigen wissenschaftlichen Arbeiten sind sogenannte *graue Schriften* Arbeiten, bei denen kein formales Begutachtungsverfahren im Sinne eines peer-review erfolgt. Dies bedeutet jedoch nicht, diese Arbeiten sollten unberücksichtigt zu lassen oder sie als nicht zitierfähig zu werten. Beispiele für graue Schriften sind Zwischen- und Abschlussberichte von Forschungsprojekten, die bei öffentlicher Förderung über die durchführende Einrichtung und teilweise auch über den öffentlichen Mittelgeber abgerufen werden können. Dass sie keinem Peer-Review-Verfahren unterliegen heißt nicht, sie werden nicht begutachtet. Allerdings fungiert hier der Mittelgeber oder von ihm beauftragte Forschungsbeiräte als begutachtendes und qualitätssicherndes Gremium. Graue Schriften enthalten Ergebnisse oft in einem höheren Auflösungsgrad als anschließende Zeitschriften- und Abstractveröffentlichungen. Auch wissenschaftliche Gutachten und Fachgutachten werden als graue Schriften bezeichnet und sind grundsätzlich zitierfähig.

10.2 Aufbau und Gliederung

Für alle Disziplinen existieren Richtlinien, die Orientierung und geben für die Veröffentlichung wissenschaftlicher Ergebnisse (u.a. Deutsche Gesellschaft für Psychologie, 2007). Darin wird der Aufbau, die Gliederung, der Bericht von Ergebnissen in Tabellen und Abbildungen sowie die Zitierweise und die Gestaltung des Literaturverzeichnisses detailliert beschrieben. Auch Fachzeitschriften legen Autorenrichtlinien vor.

Eine wissenschaftliche Arbeit setzt sich zusammen aus einem Deckblatt mit Titel, Autoren und ihren Institutionen, Inhalts-, Tabellen- und Abbildungsverzeichnis, einer Zusammenfassung, einer Einleitung, dem Hintergrund mit theoretischem Rahmen, Fragestellung und Hypothesen, der Methodenbeschreibung, dem Ergebnisbericht sowie der Diskussion, Schlussfolgerungen für die Praxis und einer Methodenkritik (Beller, 2008).

Eine *Zusammenfassung* (Abstract) ist die vollständige und knappe Zusammenfassung der wissenschaftlichen Arbeit. Inhaltlich orientiert sich die Gliederung des Abstract an der Struktur wissenschaftlicher Arbeiten. Es umfasst Einleitung/Hintergrund, Ziele und Fragestellung, Material und Methoden, Ergebnisse, Diskussion, Fazit für die Praxis und Methodenkritik. Bei Studienarbeiten sollte ein Abstract einen Umfang von 300 Wörtern haben, bei Bachelor- und Masterarbeiten eine DIN A4-Seite und einen angemessen höheren Umfang bei Dissertationen (Beller, 2008).

Das *Inhaltsverzeichnis* führt alle Gliederungspunkte wissenschaftlicher Arbeit auf. Es umfasst mehrere Gliederungsebenen (1., 1.1, 1.1.1 usw.). Je nach Umfang der Arbeit sollten nicht mehr als vier Gliederungsebenen erstellt werden. Jede Gliederungsebene muss mindestens zwei Gliederungspunkte umfassen. Wenn eine Arbeit beispielsweise einen Gliederungspunkt 1.1 enthält, muss diesem ein Gliederungspunkt 1.2 folgen. Unterkapitel behandeln ein eigenes, dem übergeordneten Kapitel zugeordnetes Thema (Beller, 2008).

Die *Einleitung* ist der Start einer wissenschaftlichen Arbeit. Gelingt er, weckt eine Einleitung Interesse, wird eine Arbeit mit größerer Wahrscheinlichkeit gelesen. In der Einleitung wird das Problem skizziert. Es sollte bereits am Anfang erkennbar werden, warum eine Arbeit überhaupt geschrieben werden sollte. Der theoretische Rahmen wird ebenso umrissen, wie der im Hintergrund näher zu beschreibende Stand der Forschung zu einem Thema (Beller, 2008). Die Einleitung endet mit einem verschriftlichten Inhaltsverzeichnis (s. 1), mit folgendem Formulierungsvorschlag:

Ausgehend von der [theoretischen und empirischen Diskussion] werden im Anschluss Methoden der Arbeit beschrieben [kurz beschreiben welche!]. In den Ergebnissen werden (…) betrachtet. Abschließend wird ausgehend von der einleitend beschriebenen Problemstellung [Nennen!] und den Ergebnissen [Nennen!] ein konkretes, praktisch umsetzbares Praxismodell entwickelt und kritisch diskutiert.

In größeren wissenschaftlichen Arbeiten wird der *Hintergrund* in mehreren Gliederungspunkten entwickelt. Neben den theoretischen Grundlagen wird der aktuelle Stand der Forschung beschrieben. Aus dem Hintergrund muss auch ersichtlich werden, warum ein Thema beforscht werden sollte. Der Hintergrund leitet zu offenen wissenschaftlichen Fragen über (Beller, 2008).

Die *Fragestellung* nimmt die im Hintergrund aufgeworfenen Probleme auf und stellt einzelne, beantwortbare Forschungsfragen (4.3). Die Fragestellung enthält Fragesätze mit einem Fragezeichen am Satzende, keine Aussagesätze. Fragestellungen orientieren sich an den einleitend entwickelten Problemstellung (Beller, 2008).

Hypothesen sind die vorweggenommenen Antworten auf die Fragestellung und gründen sich auf den im Hintergrund entwickelten theoretischen Rahmen (4.4). In einer wissenschaftlichen Arbeit werden inhaltliche Hypothesen (sog. Forschungshypothesen) als Unterschieds-, Zusammenhangs- oder Veränderungshypothesen formuliert. Dass Rahmen der Analyse statistische (Null-)Hypothesen getestet werden ist selbsterklärend und bedarf keiner näheren Erläuterung. Auch die Formulierung von Nullhypothesen ist grundsätzlich nicht erforderlich (Beller, 2008).

Der *Methodenteil* beschreibt alle methodischen Aspekte einer Untersuchung. Auf Grundlage der Beschreibung der Methoden der Arbeit, sollte es anderen Forschern möglich sein, die Arbeit zu wiederholen. Ein Methodenteil enthält (Beller, 2008):

- die Beschreibung der Stichprobe, sofern Menschen befragt wurden (nicht in reinen Literaturarbeiten), die Beschreibung der Grundgesamtheit, das Auswahlverfahren, Aussagen zur Repräsentativität, Nennung und Beschreibung der Stichprobenteilnehmermerkmale (Geschlecht und Alter und weitere soziodemografische Merkmale, die für die Fragestellung und Interpretation der Ergebnisse von Bedeutung sind)

- die Beschreibung der Literaturrecherche, die verwendeten Literaturdatenbanken, Suchbegriffe und die Kombination der Suchbegriffe, Trefferzahl, Anzahl relevanter Treffer, Gründe für den Ausschluss von Arbeiten (z.B. doppelte Treffer, keine inhaltliche Relevanz, Studiendesign, untersuchte Stichprobe, untersuchte Verfahren, verwendete Messinstrumente etc.)

- das Material, mit dem Daten erhoben und mit dem Interventionen durchgeführt wurden

- das Studiendesign mit Nennung unabhängiger, abhängiger, Moderator-, Mediator- sowie potenzieller Störvariablen, die Art der Studie, z.B. Querschnitts- oder Längsschnittstudie, Fall-Kontroll- oder Kohortenstudie, retrospektive oder prospektive Studie, Review, Literaturarbeit

- die Art der Durchführung der Studie, also wie die Materialien angewendet wurden und ggf. welche Instruktionen an Untersucher und Teilnehmer gegeben wurden

Der *Ergebnisteil* präsentiert die Daten in einem für die Fragestellung angemessenen Auflösungsgrad. Ein Absatz im Ergebnisbericht sollte zuerst allgemeine Angaben, dann notwendige Details enthalten. Empfohlen wird inhaltliche Fragen, auf die im Absatz konkret eingegangen wird, in knapper Form zu wiederholen. Daran anschließend wird ausgeführt, wie zentrale Begriffe verstanden werden bevor statistische Details berichtet werden. Dabei sollten im Text die Trends erörtert werden und für Details in Form von Zahlen auf Tabellen verwiesen werden. Abschließend erfolgt eine zusammenfassende Aussage, die zur nächsten Beantwortung der Forschungsfrage überleitet. Details werden in Tabellen berichtet. Abbildungen und Diagramme dienen der Visiualisierung von Antworten auf wesentliche Forschungsfragen (Beller, 2008).

Der *Diskussionsteil* ist ein sehr zentraler und wichtiger Teil wissenschaftlicher Arbeiten. Er greift die Problemstellung im Hintergrund auf. Es werden zudem praktisch relevante Schlussfolgerungen aus den Daten gezogen. Ziel ist keine einfache Zusammenfassung, sondern eine argumentative Auseinandersetzung mit den Ergebnissen, der Bedeutung für die Praxis, mit ihrer Eignung für die Beantwortung der Fragestellung und mit ihrer Bedeutung für den im Hintergrund erarbeiteten theoretischen Rahmen. Eine Diskussion kann folgendermaßen gegliedert werden (Beller, 2008):

- 1. Absatz: Hintergrund der Arbeit und wesentliche Fragestellung, Beschreibung wesentlicher Ergebnisse, Beantwortung der Fragestellung ohne dass bereits Schlussfolgerungen gezogen werden

- 2. Absatz, Interpretation der Ergebnisse, Erläuterung ihrer Bedeutung in Bezug auf die im Hintergrund beschriebenen Theorien

- 3. bis 5. Absatz, Vergleich der Ergebnisse mit dem aktuellen Forschungsstand, der ebenfalls im Hintergrund beschrieben wurde. Quellen, die eigene Annahmen stützen werden ebenso angeführt, wie den eigenen Ergebnissen widersprechende Veröffentlichungen

- 6. Absatz, Limitierungen, Grenzen und Quellen von Verzerrung der eigenen Studie nennen und Beschreiben, Ausführungen zur Übertragbarkeit der Ergebnisse auf andere Situationen

- 7. Absatz, ungelöste Probleme nennen sowie Vorschläge für künftige Forschungsprojekte unterbreiten, offene Forschungsfragen nennen

- 8. Absatz, Schlussfolgerung und die Bedeutung der Ergebnisse für die Praxis formulieren, keine Schlussfolgerungen ziehen, die durch Deine Ergebnisse nicht belegt sind

Im *Literaturverzeichnis* werden alle Quellen der Arbeit genannt. Diese Offenlegung der in wissenschaftlichen Arbeiten verwendeten Quellen hat nicht ausschließlich rechtliche Gründe und schützt vor Plagiatsvorwürfen. Zu zitieren korrespondiert mit den Grundsätzen der Redlichkeit wissenschaftlichen Arbeitens, wie beispielsweise: Transparenz, Überprüfbarkeit, Nachvollziehbarkeit und Ehrlichkeit. Die Angabe der verwendeten Quellen ermöglicht es Lesern und Wissenschaftlern Gedankengänge nachzuvollziehen, sie Denkschulen zuzuordnen und für eigene wissenschaftliche Argumentationen zu nutzen. Für die Quellenangabe existieren Zitierrichtlinien unterschiedlicher Disziplinen, wie beispielsweise die Richtlinien der Deutschen Gesellschaft für Psychologie (DGPs) (2007).

10.3 Sprache der Wissenschaft

Wissenschaftliche Texte werden eher quer als präzise (Wort für Wort) gelesen. An die Sprache und Formulierungen in wissenschaftlichen Arbeiten werden daher Anforderungen gestellt, die klar von denen deutscher Lyrik und Prosa abweichen. Wissenschaftliche Texte sollen (Plümper, 2008):

- einfach und präzise sein sowie

- Fremdworte, sprachliche Neuschöpfungen und Abkürzungen vermeiden,

Im Schreibprozess wird empfohlen, zunächst einen Rohtext zu verfassen und anschließend überflüssige Passagen großzügig zu löschen. Wer nicht bereit ist, ca. 30% des Rohtextes zu löschen (oder abzulegen) wird vermutlich keine stringente Argumentationskette erarbeiten können (Plümper, 2008).

10.3.1 Satzbau

Texte sind eine systematische Folge einzelner Sätzen. Der Satzbau ist wesentlicher Einflussfaktor auf die „Melodie“ eines Textes. Ein Thema kann mit großer Leichtigkeit und Stringenz aber auch chaotisch und additiv erarbeitet werden. Mit einfa-

chen Regeln lassen sich mit etwas Übung gut lesbare Texte verfassen (Plümper, 2008).

Das Prinzip der Einfachheit der wissenschaftlichen Sprache kann bereits durch kurze Sätze erreicht werden. Hauptargumentationen sind in jedem Fall Bestandteil des Hauptsatzes:

> *„Die Kernaussage der vorliegenden Untersuchung ist, dass Kinder aus benachteiligten, bildungsfernen Milieus deutlich seltener ein naturgesundes Gebiss haben, als Kinder aus nicht benachteiligten Milieus."*
> Lösung:
> *„Kinder aus sozial benachteiligten Familien haben verglichen mit Kindern aus nicht benachteiligten Familien seltener ein naturgesundes Gebiss."*
> *Statt „Es ist offensichtlich, dass…" „Offensichtlich …"* schreiben.
> *Statt „Sehr wahrscheinlich ist, dass…" „Wahrscheinlich …"* schreiben.

Eine breite sprachliche Varianz verbessert die Lesbarkeit. Sie kann durch die Kombination kurzer und langer Sätze erreicht werden, ohne ein festes Muster dieser Abfolge im Text (Plümper, 2008).

Schachtelsätze. Ein weiteres Problem sind Sätze mit zahlreichen Satzteilen, wie dieser:

> *„Wenn Kinder bereits frühzeitig Kinderbetreuungs- und Bildungsangebote, selbstverständlich von zertifizierten Anbietern, die ein Akkreditierungsverfahren durchlaufen haben, nutzen, ist die Wahrscheinlichkeit für Schulerfolge, z.B. höhere Abschlüsse, seltene Wiederholung von Klassen, größer, allerdings nur dann, wenn die Eltern auch bereit sind, eine regelmäßige Teilnahme sicherzustellen."*

Mit wenig Aufwand lassen sich Schachtelsätze entwirren und in mehrere Sätze aufteilen. Dabei sollte ebenfalls erwogen werden, überflüssige Gedanken zu entfernen (Plümper, 2008). Das Schachtelsatzbeispiel kann in dieses besser lesbare Fragment übertragen werden:

> *„Eine größere Chance auf höhere Abschlüsse und ein geringeres Risiko für Schulversagen haben Kinder, die bereits frühzeitig Kinderbetreuungs- und Bildungsangebote wahrnehmen konnten, die zertifiziert sind. Die Eltern sind verantwortlich eine regelmäßige Teilnahme ihrer Kinder sicherzustellen."*

Bezug. Werden Substantive durch Pronomen wie *dessen* oder *deren* ersetzt, muss klar bleiben, worauf sich das Pronom bezieht (Plümper, 2008). Beispielhaft für unklare Bezüge ist der folgende Satz:

> *„Zweistündliches Lagern, die Weichlagerung mit Luftkammermatratzen und die Nutzung von Schaffellen, deren Masseverteilungseigenschaften für eine beson-*

ders effektive Entlastung sorgen, kann das Risiko des Wundliegens deutlich reduzieren."

Die Vermutung liegt nahe, *deren* bezieht sich auf die Schaffelle. *Deren* kann ebenso gut auf die anderen genannten Weichlagerungsverfahren bezogen sein, die ebenfalls günstige Masseverteilungseigenschaften aufweisen. Als Lösung könnte das ersetzte Substantiv genutzt werden oder ein Satz mit klarem Bezug formuliert werden:

„Schaffelle sorgen durch ihre Masseverteilungseigenschaften für eine besonders effektive Entlastung Dekubitus gefährdeter Hautareale. Zudem reduzieren zweistündliches Lagern und der Einsatz von Luftkammermatratzen für eine Senkung des Dekubitusrisikos.

Passive Formulierung. Wissenschaftliche Texte nehmen nicht die 1. Person in den Text auf, sondern formulieren passiv. Nicht der Akteur (der Forscher) und sein Handeln interessieren, sondern das Verfahren und seine Resultate. Anstelle des folgenden aktiv formulierten Satzes (Plümper, 2008):

„Zweistündliches Lagern reduzierte das Dekubitusrisiko um 20%.

…können alternativ folgende Formulierungen und Verben genutzt werden:

1. *können*:
   Zweistündliches Lagern konnte das Dekubitusrisiko um 20% reduzieren.
2. *man kann*:
   Durch zweistündliches Lagern konnte man das Dekubitusrisiko um 20% reduzieren.
3. *sich lassen*:
   Durch zweistündliches Lagern ließ sich das Dekubitusrisiko um 20% reduzieren.
4. *Infinitiv + „zu"*:
   Mit zweistündlichem Lagern war es möglich, das Dekubitusrisiko um 20% zu reduzieren.
5. *Adjektiv + „-bar"*:
   Durch zweistündliches Lagern war das Dekubitusrisiko um 20% reduzierbar.

Genitivkonstruktion. Über die Formulierung von Genitivsätzen gelingt ebenfalls eine Verbesserung der Satzrhythmik, wie am folgenden Beispiel lesbar wird (Plümper, 2008):

*„Die Vermittlung **von** berufsrelevanten Fähigkeiten droht im deutschen universitären Alltag verloren zu gehen."* (Plümper, 2008, S. 122)

Das Ersetzen des Worts „von" durch eine Genitivkonstruktion verbessert die Satzrhythmik und die Betonung auf das Substantiv „Fähigkeiten":

„Die Vermittlung berufsrelevanter Fähigkeiten droht im deutschen universitären Alltag verloren zu gehen." (Plümper, 2008, S. 122)

10.3.2 Nutzung von Fachbegriffen

Jede Disziplin hat ihre Fachbegriffe, die in wissenschaftlichen Texten genutzt werden. Wissenschaftliche Texte wenden sich jedoch nicht nur an einen möglicherweise kleinen Kreis einer *scientific community*, sondern an eine breite interessierte Fachöffentlichkeit. Dieser werden nicht alle Fachbegriffe bekannt sein. Fachbegriffe werde daher einerseits nur dort verwendet werden, wo es erforderlich ist und andererseits bei der ersten Nennung definiert. Eingeführte Fachbegriffe werden konsistent in vergleichbaren Kontexten verwendet. In einigen Disziplinen existieren mehrere, teils konkurrierende Definitionen eines Begriffs (s. 4.2, 6.2.2). In diesem Fall ist eine Festlegung auf *eine* Definition erforderlich und der Fachbegriff wird im gesamten Text im definierten Sinn verwendet (Plümper, 2008).

10.3.3 Beispiele einer differenzierten und fachlich korrekten Ausdrucksweise in Diskussionen

Ergebnisse eigener Untersuchungen können auch bei Einhaltung der wesentlichen wissenschaftlichen Standards nur bedingt verallgemeinert und auf die Grundgesamtheit bezogen werden (s. 7, 9.1). Der Schluss von der Stichprobe auf die Grundgesamtheit ist zwar prinzipiell möglich, der Befund in einer Studie, ein oder mehrere Therapieerfolge oder weitere Effekte von Interventionen *beweisen* allerdings niemals die Wirksamkeit einer untersuchten Maßnahme für eine vergleichbare Kohorte. Selbst wenn Ergebnisse in mehreren Studien repliziert werden, stützt dies allenfalls die Evidenz einer Maßnahme – sie beweist ihre Wirksamkeit nicht. Allerdings sind vergleichbare Ergebnisse in mehreren Studien ein guter Hinweis auf die Eignung von Interventionen – aber kein Beweis.

Im Rahmen der Diskussion eigener Ergebnisse ist es, trotz allen berechtigten Stolzes auf das gefundene Ergebnis, wissenschaftlich korrekt und aus statistischen Überlegungen heraus sogar gefordert, zurückhaltend zu interpretieren. Im Ergebnisteil berichtete *Befunde* sind Befunde einer *Stichprobe*, d.h., sie treffen in der gefundenen Weise nur *für die Stichprobe* zu (s. 9.1).

Folgende Formulierungen gelten in sozial- und gesundheitswissenschaftlichen Arbeiten als „no-go's":

- *Die Studie konnte beweisen…*
  (wenn es eine sozial- oder humanwissenschaftliche Studie war, kann sie DAS nicht)
  Besser: *Die Ergebnisse der Studie weisen darauf hin*

- *Damit wurde auch die Annahme belegt…*
  (Brötchen kann man belegen oder Zungen sind belegt)
  Besser: *Dies stützt auch die Annahme, …*

- *Daraus lässt sich ableiten…*
  (Blitze kann man ableiten)
  Besser: *Dies lässt schließen auf …*

- …eigene Ergänzungen?

Zurückhaltende und daher empfehlenswerte Formulierungen für wissenschaftliches Schreiben in den Sozial- und Gesundheitswissenschaften sind u.a.:

- …deuten hin auf

- Mit einiger Wahrscheinlichkeit ist/sind…

- Dieses Ergebnis ist zwar statistisch signifikant, jedoch im Hinblick auf den zahlenmäßigen Unterschied nur bedingt inhaltlich bedeutsam

- Die Ergebnisse stützen die Annahme einer/eines … (Aussage in einer Hypothese)

- Damit wird auch erkennbar, dass…

- Die Studienteilnehmer erlebten sich mehr/weniger/

- Somit finden sich einmal mehr Hinweise auf … (die Wirksamkeit einer Maßnahme)

- …lässt ebenso die Vermutung zu… oder …legt ebenfalls die Vermutung nahe…

- Anhand der Untersuchungsergebnisse können Einflüsse von … auf … angenommen werden oder …sind Einflüsse von … auf … wahrscheinlich…

- Je weniger/je mehr … wahrgenommen wurde, desto weniger/mehr … erlebten sich die Studienteilnehmer…

- Merkmale der (Therapie A) stehen mit weniger/mehr Symptomen in Verbindung… oder ...sind mit weniger/mehr Symptomen verbunden…

- Entgegen den Annahmen fanden sich keine…

- Überraschender Weise fielen trotz … keine wesentlichen Unterschiede zwischen (Therapie A und B/Patientengruppe A/B) auf…

- Offenbar haben Patienten, die mit Therapie A behandelt wurden, eine größere Chance/ein größeres Risiko für…
- Möglicherweise stehen Maßnahmen aus Therapie A nicht nur mit … in Verbindung, sondern wirken sich auch auf … aus.

Die Auseinandersetzung mit sprachlichen Anforderungen ist wesentlich für wissenschaftlich nachvollziehbares und korrektes Schreiben. Kernaussagen und Hauptargumente gehören in einen Hauptsatz am Beginn eines Satzes. Lange und kurze Sätze abwechselnd einzusetzen erhöht die sprachliche Varianz und die Aufmerksamkeit des Lesers. Schachtelsätze werden vermieden und mehrere einfache Hauptsätze formuliert. Satzbezüge müssen eindeutig sein. Es muss klar sein, auf welches Substantiv sich Pronomen wie *deren*, *welche*, *diese* (u.a.) beziehen. Nur die absolut notwendigen Fachbegriffe werden in wissenschaftlichen Arbeiten verwendet, Abkürzungen sparsam eingesetzt. Genitivkonstruktionen unterstützen die Satzrhythmik und helfen, Sätze zu verkürzen, ohne Informationen zu vernachlässigen. Wissenschaftliches Schreiben heißt aus der Distanz, ohne emotionale Verfärbungen, zurückhaltend und differenziert schriftlich zu argumentieren. Sozial- und gesundheitswissenschaftliche Studien können zudem nichts beweisen oder belegen. Sie finden mehr oder weniger deutliche Hinweise, die für oder gegen eine Theorie sprechen.

10.4 Zeitplan für das Abfassen wissenschaftlicher Abschlussarbeiten

Insbesondere größere wissenschaftliche Arbeiten, z.B. Abschlussarbeiten auf Bachelor- und Masterniveau, werden nicht an wenigen Tagen geschrieben. Die Prüfungsordnungen geben einen Zeitraum vor, in dem Arbeiten angefertigt werden. Eine Zeitplanung hilft die Arbeit innerhalb dieser Zeit tatsächlich abzuschließen und eine sprachlich gewandte, wissenschaftlich korrekte Arbeit anzufertigen und einzureichen.

Abbildung 46 stellt einen Musterzeitplan für eine Masterarbeit vor. Für größere Arbeiten und längere Laufzeiten dienen die Zeitanteile als Orientierung. Die Planung sollte einen Zeitpuffer am Ende beinhalten, in dem Korrekturlesungen, die sprachliche Straffung und das Ausdrucken der Arbeit geplant werden.

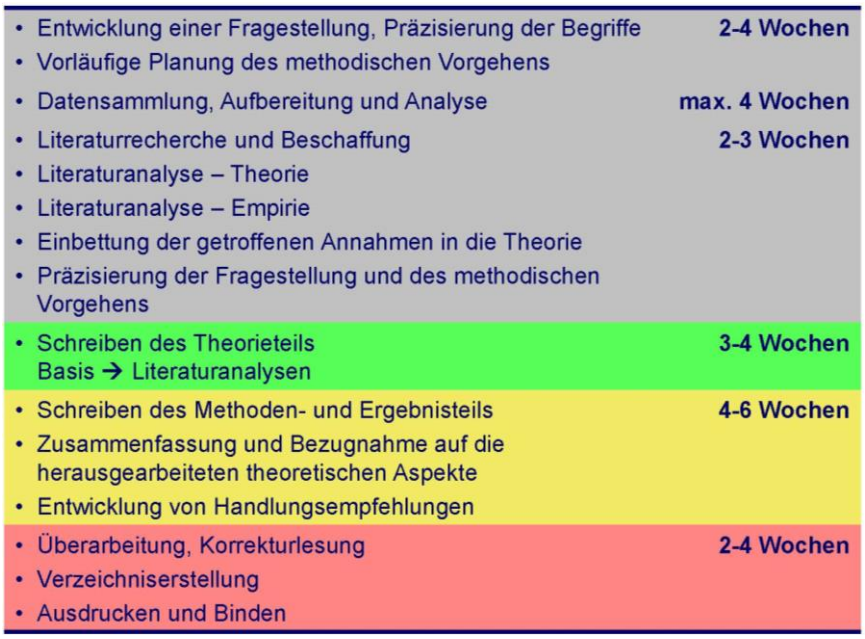


Abbildung 46: Vorschlag für die zeitliche Planung großer wissenschaftlicher Arbeiten

10.5 Zusammenfassung und Aufgaben

Die Veröffentlichung wissenschaftlicher Arbeiten ist eine „Währung der Wissenschaft“ (Döring et al., 2016). An Veröffentlichungen bemisst sich der Ruf von Wissenschaftlern und wissenschaftlicher Institutionen. Die Veröffentlichungsliste ist ferner wesentliches Kriterium bei der Besetzung von Stellen in Forschungseinrichtungen, u.a. von Professuren. Aussagekräftige und „hochwertige Veröffentlichungen“ werden vorab einem Überprüfungsverfahren unterzogen, bei dem mindestens zwei anonym bleibende Gutachter die Arbeit ihrer Kollegen bewerten, die im Prozess ebenfalls anonym bleiben. Dieses Blind-Peer-Review-Verfahren ist ein verbreitetes und aber auch kritisiertes Instrument der wissenschaftlichen Qualitätssicherung.

Die Veröffentlichung von Forschungsergebnissen erfolgt auf verschiedenen Wegen. Sehr aktuell und aussagekräftig sind *peer-reviewed* Aufsätze in Fachzeitschriften. Daneben werden wissenschaftliche Arbeiten als Monografien (häufig Dissertation- oder Habilitationsschriften), als Beitrag in einem Herausgeberwerk, als Vortrag oder Poster auf wissenschaftlichen Kongressen veröffentlicht. Neben den genannten, überwiegend *peer-reviewed* Veröffentlichungen, existieren nicht weniger aussagekräftige Forschungsberichte oder Gutachten, die unter *grauen Schriften* subsummiert werden. Für die theoretische Fundierung eigener wissenschaftlicher Arbeiten sollten in erster Linie Zeitschriftenaufsätze und Monografien herangezogen werden. Graue Schriften sind ebenso geeignet, sollten aber in geringerer Zahl aufgenommen werden als *peer-reviewed* Veröffentlichungen.

Die wissenschaftliche Fachsprache in Veröffentlichungen ist ein Kriterium für den Impact und die Verbreitung einer Arbeit. Wissenschaftliche Texte bedienen sich

einer einfachen und präzisen Sprache. Fachbegriffe zudem einheitlich verwendet und jede Argumentation erfolgt zurückhaltend.

Eine Zeitplanung kann bei größeren Arbeiten Stress vermeiden und eine zeitgerechte Anfertigung der Arbeit sicherstellen.

Schlüsselwörter: *Analyse, Aufsatz in einer Fachzeitschrift, Befunde, Beitrag in einem Herausgeberwerk, Diskussionsteil, Einleitung, Ergebnisteil, Fragestellung, graue Schriften, Hintergrund, Hypothesen, Inhaltsverzeichnis, Literaturverzeichnis, Methodenteil, Monografie, Peer-Review-Verfahren, Posterpräsentation auf Fachkongressen, Stichprobe, Vortrag auf Fachkongressen, Wissenschaftliches Abstract, Zusammenfassung/Abstract*

Aufgaben zur Lernkontrolle

1. Was bedeutet Impact?

2. Beschreiben Sie, wie ein peer-review-Verfahren abläuft?

3. Wie ist eine wissenschaftliche Arbeit gegliedert?

4. Erstellen Sie einen Zeitplan für Ihr Projekt.

Aufgabe zur Berufspraxis

5. **Recherchieren Sie den Impact für folgende, sprachtherapeutische Zeitschriften:**
    a. Aphasiology,
    b. Dysphagia
    c. American Journal of Speech-Language Pathology
    d. Sprache Stimme Gehör
    e. International Journal of Language & Communication Disorders
    f. Journal of Speech, Language and Hearing Research

Literatur

Deutsche Gesellschaft für Psychologie. (2007). *Richtlinien zur Manuskriptgestaltung* (3., überarbeitete und erweiterte Auflage ed.). Göttingen [u.a.]: Hogrefe.

Döring, N., Bortz, J., & Pöschl, S. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften* (5. vollständig überarbeitete, aktualisierte und erweiterte Auflage ed.). Berlin Heidelberg: Springer.

Fröhlich, G. (2003). Anonyme Kritik: Peer Review auf dem Prüfstand der Wissenschaftsforschung. *Medizin Bibliothek Information, 3*(2), 33-39.

Beller, S. (2008). *Empirisch forschen lernen Konzepte, Methoden, Fallbeispiele, Tipps* (2., überarb. Aufl ed.). Bern: Huber.

Plümper, T. (2008). *Effizient schreiben Leitfaden zum Verfassen von Qualifizierungsarbeiten und wissenschaftlichen Texten : [mit ausführlichen Check-Listen]* (2., vollst. überarb. und erw. Aufl. ed.). München: Oldenbourg.

Anhang

I. z-Tabelle

z-TABELLE Prozentuelle Anteile der Standardnormalverteilung links vom tabellierten z-Wert

z	Fläche	z	Fläche	z	Fläche	z	Fläche	z	Fläche	z	Fläche	z	Fläche	z	Fläche	z	Fläche	z	Fläche	z	Fläche
-3,00	0,13	-2,50	0,62	-2,00	2,28	-1,50	6,68	-1,00	15,87	-0,50	30,85	0,00	50,00	0,50	69,15	1,00	84,13	1,50	93,32	2,00	97,72	2,50	99,38
-2,99	0,14	-2,49	0,64	-1,99	2,33	-1,49	6,81	-0,99	16,11	-0,49	31,21	0,01	50,40	0,51	69,50	1,01	84,38	1,51	93,45	2,01	97,78	2,51	99,40
-2,98	0,14	-2,48	0,66	-1,98	2,39	-1,48	6,94	-0,98	16,35	-0,48	31,56	0,02	50,80	0,52	69,85	1,02	84,61	1,52	93,57	2,02	97,83	2,52	99,41
-2,97	0,15	-2,47	0,68	-1,97	2,44	-1,47	7,08	-0,97	16,60	-0,47	31,92	0,03	51,20	0,53	70,19	1,03	84,85	1,53	93,70	2,03	97,88	2,53	99,43
-2,96	0,15	-2,46	0,69	-1,96	2,50	-1,46	7,21	-0,96	16,85	-0,46	32,28	0,04	51,60	0,54	70,54	1,04	85,08	1,54	93,82	2,04	97,93	2,54	99,45
-2,95	0,16	-2,45	0,71	-1,95	2,56	-1,45	7,35	-0,95	17,11	-0,45	32,64	0,05	51,99	0,55	70,88	1,05	85,31	1,55	93,94	2,05	97,98	2,55	99,46
-2,94	0,16	-2,44	0,73	-1,94	2,62	-1,44	7,49	-0,94	17,36	-0,44	33,00	0,06	52,39	0,56	71,23	1,06	85,54	1,56	94,06	2,06	98,03	2,56	99,48
-2,93	0,17	-2,43	0,75	-1,93	2,68	-1,43	7,64	-0,93	17,62	-0,43	33,36	0,07	52,79	0,57	71,57	1,07	85,77	1,57	94,18	2,07	98,08	2,57	99,49
-2,92	0,18	-2,42	0,78	-1,92	2,74	-1,42	7,78	-0,92	17,88	-0,42	33,72	0,08	53,19	0,58	71,90	1,08	85,99	1,58	94,29	2,08	98,12	2,58	99,51
-2,91	0,18	-2,41	0,80	-1,91	2,81	-1,41	7,93	-0,91	18,14	-0,41	34,09	0,09	53,59	0,59	72,24	1,09	86,21	1,59	94,41	2,09	98,17	2,59	99,52
-2,90	0,19	-2,40	0,82	-1,90	2,87	-1,40	8,08	-0,90	18,41	-0,40	34,46	0,10	53,98	0,60	72,57	1,10	86,43	1,60	94,52	2,10	98,21	2,60	99,53
-2,89	0,19	-2,39	0,84	-1,89	2,94	-1,39	8,23	-0,89	18,67	-0,39	34,83	0,11	54,38	0,61	72,91	1,11	86,65	1,61	94,63	2,11	98,26	2,61	99,55
-2,88	0,20	-2,38	0,87	-1,88	3,01	-1,38	8,38	-0,88	18,94	-0,38	35,20	0,12	54,78	0,62	73,24	1,12	86,86	1,62	94,74	2,12	98,30	2,62	99,56
-2,87	0,21	-2,37	0,89	-1,87	3,07	-1,37	8,53	-0,87	19,22	-0,37	35,57	0,13	55,17	0,63	73,57	1,13	87,08	1,63	94,84	2,13	98,34	2,63	99,57
-2,86	0,21	-2,36	0,91	-1,86	3,14	-1,36	8,69	-0,86	19,49	-0,36	35,94	0,14	55,57	0,64	73,89	1,14	87,29	1,64	94,95	2,14	98,38	2,64	99,59
-2,85	0,22	-2,35	0,94	-1,85	3,22	-1,35	8,85	-0,85	19,77	-0,35	36,32	0,15	55,96	0,65	74,22	1,15	87,49	1,65	95,05	2,15	98,42	2,65	99,60
-2,84	0,23	-2,34	0,96	-1,84	3,29	-1,34	9,01	-0,84	20,05	-0,34	36,69	0,16	56,36	0,66	74,54	1,16	87,70	1,66	95,15	2,16	98,46	2,66	99,61
-2,83	0,23	-2,33	0,99	-1,83	3,36	-1,33	9,18	-0,83	20,33	-0,33	37,07	0,17	56,75	0,67	74,86	1,17	87,90	1,67	95,25	2,17	98,50	2,67	99,62
-2,82	0,24	-2,32	1,02	-1,82	3,44	-1,32	9,34	-0,82	20,61	-0,32	37,45	0,18	57,14	0,68	75,17	1,18	88,10	1,68	95,35	2,18	98,54	2,68	99,63
-2,81	0,25	-2,31	1,04	-1,81	3,51	-1,31	9,51	-0,81	20,90	-0,31	37,83	0,19	57,53	0,69	75,49	1,19	88,30	1,69	95,45	2,19	98,57	2,69	99,64
-2,80	0,26	-2,30	1,07	-1,80	3,59	-1,30	9,68	-0,80	21,19	-0,30	38,21	0,20	57,93	0,70	75,80	1,20	88,49	1,70	95,54	2,20	98,61	2,70	99,65
-2,79	0,26	-2,29	1,10	-1,79	3,67	-1,29	9,85	-0,79	21,48	-0,29	38,59	0,21	58,32	0,71	76,11	1,21	88,69	1,71	95,64	2,21	98,64	2,71	99,66
-2,78	0,27	-2,28	1,13	-1,78	3,75	-1,28	10,03	-0,78	21,77	-0,28	38,97	0,22	58,71	0,72	76,42	1,22	88,88	1,72	95,73	2,22	98,68	2,72	99,67
-2,77	0,28	-2,27	1,16	-1,77	3,84	-1,27	10,20	-0,77	22,07	-0,27	39,36	0,23	59,10	0,73	76,73	1,23	89,07	1,73	95,82	2,23	98,71	2,73	99,68
-2,76	0,29	-2,26	1,19	-1,76	3,92	-1,26	10,38	-0,76	22,36	-0,26	39,74	0,24	59,48	0,74	77,03	1,24	89,25	1,74	95,91	2,24	98,75	2,74	99,69
-2,75	0,30	-2,25	1,22	-1,75	4,01	-1,25	10,56	-0,75	22,66	-0,25	40,13	0,25	59,87	0,75	77,34	1,25	89,44	1,75	95,99	2,25	98,78	2,75	99,70
-2,74	0,31	-2,24	1,25	-1,74	4,09	-1,24	10,75	-0,74	22,97	-0,24	40,52	0,26	60,26	0,76	77,64	1,26	89,62	1,76	96,08	2,26	98,81	2,76	99,71
-2,73	0,32	-2,23	1,29	-1,73	4,18	-1,23	10,93	-0,73	23,27	-0,23	40,90	0,27	60,64	0,77	77,93	1,27	89,80	1,77	96,16	2,27	98,84	2,77	99,72
-2,72	0,33	-2,22	1,32	-1,72	4,27	-1,22	11,12	-0,72	23,58	-0,22	41,29	0,28	61,03	0,78	78,23	1,28	89,97	1,78	96,25	2,28	98,87	2,78	99,73
-2,71	0,34	-2,21	1,36	-1,71	4,36	-1,21	11,31	-0,71	23,89	-0,21	41,68	0,29	61,41	0,79	78,52	1,29	90,15	1,79	96,33	2,29	98,90	2,79	99,74
-2,70	0,35	-2,20	1,39	-1,70	4,46	-1,20	11,51	-0,70	24,20	-0,20	42,07	0,30	61,79	0,80	78,81	1,30	90,32	1,80	96,41	2,30	98,93	2,80	99,74
-2,69	0,36	-2,19	1,43	-1,69	4,55	-1,19	11,70	-0,69	24,51	-0,19	42,47	0,31	62,17	0,81	79,10	1,31	90,49	1,81	96,49	2,31	98,96	2,81	99,75
-2,68	0,37	-2,18	1,46	-1,68	4,65	-1,18	11,90	-0,68	24,83	-0,18	42,86	0,32	62,55	0,82	79,39	1,32	90,66	1,82	96,56	2,32	98,98	2,82	99,76
-2,67	0,38	-2,17	1,50	-1,67	4,75	-1,17	12,10	-0,67	25,14	-0,17	43,25	0,33	62,93	0,83	79,67	1,33	90,82	1,83	96,64	2,33	99,01	2,83	99,77
-2,66	0,39	-2,16	1,54	-1,66	4,85	-1,16	12,30	-0,66	25,46	-0,16	43,64	0,34	63,31	0,84	79,95	1,34	90,99	1,84	96,71	2,34	99,04	2,84	99,77
-2,65	0,40	-2,15	1,58	-1,65	4,95	-1,15	12,51	-0,65	25,78	-0,15	44,04	0,35	63,68	0,85	80,23	1,35	91,15	1,85	96,78	2,35	99,06	2,85	99,78
-2,64	0,41	-2,14	1,62	-1,64	5,05	-1,14	12,71	-0,64	26,11	-0,14	44,43	0,36	64,06	0,86	80,51	1,36	91,31	1,86	96,86	2,36	99,09	2,86	99,79
-2,63	0,43	-2,13	1,66	-1,63	5,16	-1,13	12,92	-0,63	26,43	-0,13	44,83	0,37	64,43	0,87	80,78	1,37	91,47	1,87	96,93	2,37	99,11	2,87	99,79
-2,62	0,44	-2,12	1,70	-1,62	5,26	-1,12	13,14	-0,62	26,76	-0,12	45,22	0,38	64,80	0,88	81,06	1,38	91,62	1,88	96,99	2,38	99,13	2,88	99,80
-2,61	0,45	-2,11	1,74	-1,61	5,37	-1,11	13,35	-0,61	27,09	-0,11	45,62	0,39	65,17	0,89	81,33	1,39	91,77	1,89	97,06	2,39	99,16	2,89	99,81
-2,60	0,47	-2,10	1,79	-1,60	5,48	-1,10	13,57	-0,60	27,43	-0,10	46,02	0,40	65,54	0,90	81,59	1,40	91,92	1,90	97,13	2,40	99,18	2,90	99,81
-2,59	0,48	-2,09	1,83	-1,59	5,59	-1,09	13,79	-0,59	27,76	-0,09	46,41	0,41	65,91	0,91	81,86	1,41	92,07	1,91	97,19	2,41	99,20	2,91	99,82
-2,58	0,49	-2,08	1,88	-1,58	5,71	-1,08	14,01	-0,58	28,10	-0,08	46,81	0,42	66,28	0,92	82,12	1,42	92,22	1,92	97,26	2,42	99,22	2,92	99,82
-2,57	0,51	-2,07	1,92	-1,57	5,82	-1,07	14,23	-0,57	28,43	-0,07	47,21	0,43	66,64	0,93	82,38	1,43	92,36	1,93	97,32	2,43	99,25	2,93	99,83
-2,56	0,52	-2,06	1,97	-1,56	5,94	-1,06	14,46	-0,56	28,77	-0,06	47,61	0,44	67,00	0,94	82,64	1,44	92,51	1,94	97,38	2,44	99,27	2,94	99,84
-2,55	0,54	-2,05	2,02	-1,55	6,06	-1,05	14,69	-0,55	29,12	-0,05	48,01	0,45	67,36	0,95	82,89	1,45	92,65	1,95	97,44	2,45	99,29	2,95	99,84
-2,54	0,55	-2,04	2,07	-1,54	6,18	-1,04	14,92	-0,54	29,46	-0,04	48,40	0,46	67,72	0,96	83,15	1,46	92,79	1,96	97,50	2,46	99,31	2,96	99,85
-2,53	0,57	-2,03	2,12	-1,53	6,30	-1,03	15,15	-0,53	29,81	-0,03	48,80	0,47	68,08	0,97	83,40	1,47	92,92	1,97	97,56	2,47	99,32	2,97	99,85
-2,52	0,59	-2,02	2,17	-1,52	6,43	-1,02	15,39	-0,52	30,15	-0,02	49,20	0,48	68,44	0,98	83,65	1,48	93,06	1,98	97,61	2,48	99,34	2,98	99,86
-2,51	0,60	-2,01	2,22	-1,51	6,55	-1,01	15,62	-0,51	30,50	-0,01	49,60	0,49	68,79	0,99	83,89	1,49	93,19	1,99	97,67	2,49	99,36	2,99	99,86
																						3,00	99,87

(Quelle: http://css-kti.tugraz.at/research/cssarchive/courses/fm2/TABELLENWERTE.pdf)

II. Verteilungsfunktion der chi-Quadrat-Verteilung

ν \ λ	0,005	0,010	0,025	0,050	0,100	0,900	0,950	0,975	0,990	0,995
1	0,000	0,000	0,001	0,004	0,016	2,706	3,841	5,024	6,635	7,879
2	0,010	0,020	0,051	0,103	0,211	4,605	5,991	7,378	9,210	10,597
3	0,072	0,115	0,216	0,352	0,584	6,251	7,815	9,348	11,345	12,838
4	0,207	0,297	0,484	0,711	1,064	7,779	9,488	11,143	13,277	14,860
5	0,412	0,554	0,831	1,145	1,610	9,236	11,070	12,832	15,086	16,750
6	0,676	0,872	1,237	1,635	2,204	10,645	12,592	14,449	16,812	18,548
7	0,989	1,239	1,690	2,167	2,833	12,017	14,067	16,013	18,475	20,278
8	1,344	1,647	2,180	2,733	3,490	13,362	15,507	17,535	20,090	21,955
9	1,735	2,088	2,700	3,325	4,168	14,684	16,919	19,023	21,666	23,589
10	2,156	2,558	3,247	3,940	4,865	15,987	18,307	20,483	23,209	25,188
11	2,603	3,053	3,816	4,575	5,578	17,275	19,675	21,920	24,725	26,757
12	3,074	3,571	4,404	5,226	6,304	18,549	21,026	23,337	26,217	28,300
13	3,565	4,107	5,009	5,892	7,041	19,812	22,362	24,736	27,688	29,819
14	4,075	4,660	5,629	6,571	7,790	21,064	23,685	26,119	29,141	31,319
15	4,601	5,229	6,262	7,261	8,547	22,307	24,996	27,488	30,578	32,801
16	5,142	5,812	6,908	7,962	9,312	23,542	26,296	28,845	32,000	34,267
17	5,697	6,408	7,564	8,672	10,085	24,769	27,587	30,191	33,409	35,718
18	6,265	7,015	8,231	9,390	10,865	25,989	28,869	31,526	34,805	37,156
19	6,844	7,633	8,907	10,117	11,651	27,204	30,144	32,852	36,191	38,582
20	7,434	8,260	9,591	10,851	12,443	28,412	31,410	34,170	37,566	39,997
21	8,034	8,897	10,283	11,591	13,240	29,615	32,671	35,479	38,932	41,401
22	8,643	9,542	10,982	12,338	14,041	30,813	33,924	36,781	40,289	42,796
23	9,260	10,196	11,689	13,091	14,848	32,007	35,172	38,076	41,638	44,181
24	9,886	10,856	12,401	13,848	15,659	33,196	36,415	39,364	42,980	45,558
25	10,520	11,524	13,120	14,611	16,473	34,382	37,652	40,646	44,314	46,928
26	11,160	12,198	13,844	15,379	17,292	35,563	38,885	41,923	45,642	48,290
27	11,808	12,878	14,573	16,151	18,114	36,741	40,113	43,195	46,963	49,645
28	12,461	13,565	15,308	16,928	18,939	37,916	41,337	44,461	48,278	50,994
29	13,121	14,256	16,047	17,708	19,768	39,087	42,557	45,722	49,588	52,335
30	13,787	14,953	16,791	18,493	20,599	40,256	43,773	46,979	50,892	53,672
40	20,707	22,164	24,433	26,509	29,051	51,805	55,758	59,342	63,691	66,766
50	27,991	29,707	32,357	34,764	37,689	63,167	67,505	71,420	76,154	79,490
60	35,534	37,485	40,482	43,188	46,459	74,397	79,082	83,298	88,379	91,952
70	43,275	45,442	48,758	51,739	55,329	85,527	90,531	95,023	100,425	104,215
80	51,172	53,540	57,153	60,391	64,278	96,578	101,879	106,629	112,329	116,321
90	59,196	61,754	65,647	69,126	73,291	107,565	113,145	118,136	124,116	128,299
100	67,328	70,065	74,222	77,929	82,358	118,498	124,342	129,561	135,807	140,170

http://www.ivwl.uni-kassel.de/kosfeld/lehre/zeitreihen/Verteilungstabellen2.pdf

III. Verteilungsfunktion der t-Verteilung

ν \ λ	0,9000	0,9500	0,9750	0,9900	0,9950	0,9995
1	3,078	6,314	12,706	31,821	63,657	636,619
2	1,886	2,920	4,303	6,965	9,925	31,598
3	1,638	2,353	3,182	4,541	5,841	12,941
4	1,533	2,132	2,776	3,747	4,604	8,610
5	1,476	2,015	2,571	3,365	4,032	6,859
6	1,440	1,943	2,447	3,143	3,707	5,959
7	1,415	1,895	2,365	2,998	3,499	5,405
8	1,397	1,860	2,306	2,896	3,355	5,041
9	1,383	1,833	2,262	2,821	3,250	4,781
10	1,372	1,812	2,228	2,764	3,169	4,587
11	1,363	1,796	2,201	2,718	3,106	4,437
12	1,356	1,782	2,179	2,681	3,055	4,318
13	1,350	1,771	2,160	2,650	3,012	4,221
14	1,345	1,761	2,145	2,624	2,977	4,140
15	1,341	1,753	2,131	2,602	2,947	4,073
16	1,337	1,746	2,120	2,583	2,921	4,015
17	1,333	1,740	2,110	2,567	2,898	3,965
18	1,330	1,734	2,101	2,552	2,878	3,922
19	1,328	1,729	2,093	2,539	2,861	3,883
20	1,325	1,725	2,086	2,528	2,845	3,850
21	1,323	1,721	2,080	2,518	2,831	3,819
22	1,321	1,717	2,074	2,508	2,819	3,792
23	1,319	1,714	2,069	2,500	2,807	3,767
24	1,318	1,711	2,064	2,492	2,797	3,745
25	1,316	1,708	2,060	2,485	2,787	3,725
26	1,315	1,706	2,056	2,479	2,779	3,707
27	1,314	1,703	2,052	2,473	2,771	3,690
28	1,313	1,701	2,048	2,467	2,763	3,674
29	1,311	1,699	2,045	2,462	2,756	3,659
30	1,310	1,697	2,042	2,457	2,750	3,646
40	1,303	1,684	2,021	2,423	2,704	3,551
60	1,296	1,671	2,000	2,390	2,660	3,460
120	1,289	1,658	1,980	2,358	2,617	3,373
∞	1,282	1,645	1,960	2,326	2,576	3,291

http://www.ivwl.uni-kassel.de/kosfeld/lehre/zeitreihen/Verteilungstabellen2.pdf

IV. Literaturverzeichnis

academics.de. (o.A.). Anteil an Medizin-Professorinnen wächst stetig. Retrieved 8. August 2016, from https://www.academics.de/wissenschaft/ professur_medizin_52245.html

Ahrens, W., & Jöckel, K.-H. (2015). The benefit of large-scale cohort studies for health research: the example of the German National Cohort. *Bundesgesundheitsblatt - Gesundheitsforschung - Gesundheitsschutz, 58*(8), 813-821. doi: 10.1007/s00103-015-2182-x

Antonovsky, A. (1991). *Health, stress, and coping* (1st ed.). San Francisco [u.a.]: Jossey-Bass Publ.

Antonovsky, A., & Franke, A. (1997). *Salutogenese zur Entmystifizierung der Gesundheit*. Tübingen: DGVT-Verlag.

Atteslander, P., & Cromm, J. (2010). *Methoden der empirischen Sozialforschung* (13., neu bearbeitete und erweiterte Auflage ed.). Berlin: Erich Schmidt Verlag.

Balzert, H., Schröder, M., Schäfer, C., & Motte, P. (2011). *Wissenschaftliches Arbeiten Ethik, Inhalt & Form wiss. Arbeiten, Handwerkszeug, Quellen, Projektmanagement, Präsentation* (2., um 50 Prozent erw. und aktualisierte Aufl. ed.). Herdecke [u.a.]: W3L-Verl.

Beerlage, I., Arndt, D., Hering, T., & Springer, S. (2009). *Arbeitsbedingungen und Organisationsprofile als Determinanten von Gesundheit, Einsatzfähigkeit sowie haupt- und ehrenamtlichem Engagement bei Einsatzkräften in Einsatzorganisationen des Bevölkerungsschutzes. Abschlussbericht, September 2009*. Magdeburg & Bonn: Hochschule Magdeburg-Stendal.

Beerlage, I., Arndt, D., & Hering, T. S., Silke. (2007). *Arbeitsbedingungen und Organisationsprofile als Determinanten von Gesundheit, Einsatzfähigkeit sowie haupt- und ehrenamtlichem Engagement bei Einsatzkräften in Einsatzorganisationen des Bevölkerungsschutzes. Erster Zwischenbericht März 2007.* Magdeburg.

Beerlage, I., Hering, T., & Nörenberg, L. (2006). *Entwicklung von Standards und Empfehlungen für eine Netzwerk zur bundesweiten Strukturierung und Organisation psychosozialer Notfallversor¬gung. Schriftenreihe Zivilschutzforschung – Neue Folge Band 57.* Bonn: Bundesamt für Bevölkerungsschutz.

Beller, S. (2008). *Empirisch forschen lernen Konzepte, Methoden, Fallbeispiele, Tipps* (2., überarb. Aufl ed.). Bern: Huber.

Bengel, J., Strittmatter, R., & Willmann, H. (2006). *Was erhält Menschen gesund? Antonovskys Modell der Salutogenese - Diskussionsstand und Stellenwert ; eine Expertise* (9., erw. Neuaufl ed.). Köln: BZgA.

Borgetto, B. (1999). *Berufsbiographie und chronische Krankheit Handlungsrationalität am Beispiel von Patienten nach koronarer Bypassoperation*. Opladen [u.a.]: Westdt. Verl.

Bortz, J., & Döring, N. (2003). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler* (3., überarb. Aufl., Nachdr. ed.). Berlin [u.a.]: Springer.

Chalmers, A. F., Bergemann, N., & Altstötter-Gleich, C. (2007). *Wege der Wissenschaft Einführung in die Wissenschaftstheorie* (6., verb. Aufl. ed.). Berlin [u.a.]: Springer.

Creswell, J. W. (2007). *Qualitative inquiry & research design choosing among five approaches* (2. ed.). Thousand Oaks, Calif. [u.a.]: Sage Publ.

Deutsche Gesellschaft für Psychologie. (2007). *Richtlinien zur Manuskriptgestaltung* (3., überarbeitete und erweiterte Auflage ed.). Göttingen [u.a.]: Hogrefe.

Döring, N., Bortz, J., & Pöschl, S. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften* (5. vollständig überarbeitete, aktualisierte und erweiterte Auflage ed.). Berlin Heidelberg: Springer.

Eid, M., Gollwitzer, M., & Schmitt, M. (2011). *Statistik und Forschungsmethoden. Lehrbuch* (2. korrigierte Auflage). Weinheim u.a.: Beltz.

Fawcett, J., & Erckenbrecht, I. (1999). *Spezifische Theorien der Pflege im Überblick*. Bern Göttingen Toronto Seattle: Verlag Hans Huber.

Fiedler, K. (2016). Empfehlungen der DGPs-Kommission „Qualität der psychologischen Forschung“. *Psychologische Rundschau, 67*(1), 59-74. doi:10.1026/0033-3042/a000316

Flick, U. (2000). Design und Prozess qualitativer Forschung. In U. Flick, E. v. Kardorff, & I. Steinke (Eds.), *Qualitative Forschung. Ein Handbuch* (pp. 252-265). Reinbek bei Hamburg: Rowohlt.

Fröhlich, G. (2003). Anonyme Kritik: Peer Review auf dem Prüfstand der Wissenschaftsforschung. *Medizin Bibliothek Information, 3*(2), 33-39.

Herschel, M. (2013). *Das KliFo-Buch Praxisbuch klinische Forschung* (2., vollst. überarb. und erw. Aufl. ed.). Stuttgart: Schattauer.

Hibbeler, B., & Korzilius, H. (2008). Arztberuf: Die Medizin wird weiblich. *Deutsches Ärzteblatt International, 105*(12), 609.

Hussy, W., Schreier, M., & Echterhoff, G. (2013). *Forschungsmethoden in Psychologie und Sozialwissenschaften für Bachelor: mit 23 Tabellen* (2., überarb. Aufl. ed.). Berlin u.a.: Springer.

Klug, S. J., Bender, R., Blettner, M., & Lange, S. (2004). Wichtige epidemiologische Studientypen. [Common epidemiologic study types]. *Dtsch med Wochenschr, 129*(S 3), T7-T10. doi: 10.1055/s-2004-836076

Köhler, T. (2012). Inferenzstatistischer Nachweis intraindividueller Unterschiede im Rahmen von Einzelfallanalysen. *Empirische Sonderpädagogik, 4*(3/4), 265-274.

Krämer, W. (2013). *So lügt man mit Statistik* (Überarb. Neuausg. der ungekürzten Taschenbuchausg., 4. Aufl. ed.). München [u.a.]: Piper.

Krauss, S., & Wassner, C. (2001). Wie man das Testen von Hypothesen einführen sollte. *Stochastik in der Schule, 21*(1), 29-34.

Leonhart, R. (2013). *Lehrbuch Statistik Einstieg und Vertiefung* (3., überarbeitete und erweiterte Auflage ed.). Bern: Verlag Hans Huber.

Mangold, S. (2013). *Evidenzbasiertes Arbeiten in der Physio- und Ergotherapie Reflektiert - systematisch - wissenschaftlich fundiert* (2., aktualisierte Aufl. ed.). Berlin [u.a.]: Springer.

Mayring, P. (2010). Design. In G. Mey & K. Mruck (Eds.), *Handbuch Qualitative Forschung in der Psychologie* (pp. 225-237). Wiesbaden: VS Verlag für Sozialwissenschaften.

Patton, M. Q. (2002). *Qualitative research & evaluation methods* (3. ed.). Thousand Oaks, Calif. [u.a.]: Sage.

Plümper, T. (2008). *Effizient schreiben Leitfaden zum Verfassen von Qualifizierungsarbeiten und wissenschaftlichen Texten : [mit ausführlichen Check-Listen]* (2., vollst. überarb. und erw. Aufl. ed.). München: Oldenbourg.

Popper, K. R. (1989). *Logik der Forschung* (9., verb. Aufl. ed.). Tübingen: Mohr.

Prel, J.-B. d., Hommel, G., Röhrig, B., & Blettner, M. (2009). Konfidenzintervall oder p-Wert? Teil 4 der Serie zur Bewertung wissenschaftlicher Publikati-

onen. *Deutsches Ärzteblatt International, 106*(19), 335-339. doi: 10.3238/arztebl.2009.0335

Raithel, J. (2008). *Quantitative Forschung ein Praxiskurs* (2., durchgesehene Auflage ed.). Wiesbaden: VS Verlag für Sozialwissenschaften.

Rasch, B., Friese, M., & Hofmann, W. (2010). *Quantitative Methoden : Einführung in die Statistik für Psychologen und Sozialwissenschaftler Bd. 2* (3., erw. Aufl. ed.). Berlin [u.a.]: Springer.

Rasch, B., Friese, M., Hofmann, W., & Naumann, E. (2004). *Quantitative Methoden Band 1*. Berlin u.a.: Springer.

Robert-Koch-Institut. (2012). *Die Gesundheit von Erwachsenen in Deutschland*. Berlin: Eigenverlag.

Rott, C. (1995). Basiskarte: Alltagswissen. In K.-E. Rogge (Ed.), *Methodenatlas* (pp. 12-18). Heidelberg: Springer.

Rumsey, D., Engel, R., Baum, K., & Kühnel, P. (2012). *Wahrscheinlichkeitsrechnung für Dummies* (2., überarb. Aufl. ed.). Weinheim: Wiley-VCH.

Schäfer, S. (2007). Kritische Bewertung von Studien zur Ätiologie. In R. Kunz, G. Ollenschläger, H. Raspe, G. Jonitz, & N. Donner-Banzhoff (Eds.), *Lehrbuch Evidenzbasierte Medizin in Klinik und Praxis* (pp. 101-114). Köln: Deutscher Ärzteverlag.

Schäfer, T. (2016). *Methodenlehre und Statistik Einführung in Datenerhebung, deskriptive Statistik und Inferenzstatistik*. Wiesbaden: Springer.

Sedlmeier, P., & Renkewitz, F. (2013). *Forschungsmethoden und Statistik für Psychologen und Sozialwissenschaftler* (2., aktualisierte und erweiterte Auflage ed.). München Harlow Amsterdam: Pearson/Higher Education.

Steinke, I. (2000). Gütekriterien qualitativer Forschung. In U. Flick, E. v. Kardorff, & I. Steinke (Eds.), *Qualitative Forschung. Ein Handbuch* (pp. 319-331). Reinbek bei Hamburg: Rowohlt.

Wagner, G. G., Göbel, J., Krause, P., Pischner, R., & Sieber, I. (2008). Das Sozio-oekonomische Panel (SOEP): Multidisziplinäres Haushaltspanel und Kohortenstudie für Deutschland – Eine Einführung (für neue Datennutzer) mit einem Ausblick (für erfahrene Anwender). *AStA Wirtschafts- und Sozialstatistisches Archiv, 2*(4), 301-328. doi: 10.1007/s11943-008-0050-y

Wikipedia. (2016). Induktion (Philosophie). Retrieved 18.5.2016, from https://de.wikipedia.org/wiki/Induktion_(Philosophie)

V. Literatur zur Vertiefung

Beller, S. (2008). ***Empirisch forschen lernen Konzepte, Methoden, Fallbeispiele, Tipps*** **(2., überarb. Aufl ed.). Bern: Huber.**

Der Beller ist eine knappe und leicht verständliche Einführung in die quantitativen Forschungsmethoden. Er skizziert den Forschungsprozess und führt in grundlegende Logiken und Berechnungen innerhalb der quantitativen Forschung ein. Zusammen mit dem Leonhart (2013, s.u.) ein unschlagbares Team, mit dem jedes statistische Problem methodisch korrekt gelöst werden kann.

Döring, N., Bortz, J., & Pöschl, S. (2016). ***Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften*** **(5. vollständig überarbeitete, aktualisierte und erweiterte Auflage ed.). Berlin Heidelberg: Springer.**

Döring et al. gelingt seit Jahren diese ausführliche, sehr fundierte und nahezu vollständige Aufbereitung des Themas Forschungsmethoden mit anschaulichen Beispielen. Alle mathematischen Operationen sind verständlich hergeleitet. Dieses Buch ist eine Arbeitsgrundlage für die vertiefte Auseinandersetzung mit quantitativen und qualitativen Forschungsmethoden.

Leonhart, R. (2013). ***Lehrbuch Statistik Einstieg und Vertiefung*** **(3., überarbeitete und erweiterte Auflage ed.). Bern: Verlag Hans Huber.**

Der Leonhart führt vollständig, auch für mathematisch grundgebildete Menschen leicht verständlich und nachvollziehbar in alle Verfahren der deskriptiven und Inferenzstatistik ein. Dieses Buch setzt Standards als Nachschlagewerk und zur selbständigen Vertiefung des ungeliebten Fachs Statistik.

U. Flick, E. v. Kardorff, & I. Steinke (Eds.) (2000). ***Qualitative Forschung. Ein Handbuch*****. Reinbek bei Hamburg: Rowohlt.**

Für qualitativ Forschende ist der Flick et al. ernstzunehmendes ein Standardwerk, das leicht verständlich und vollständig in die verschiedenen qualitative Verfahren einführt.

VI. Schlüsselwörter

Schlüsselwörter	Kapitelangabe
abhängige Stichproben	9.1.3
abhängige Variable	4.5.1/5.3
absolute Häufigkeit	8.3
Ad-hoc-Auswahl	7.2
Alltagswissen	2.1
Alpha/ α-Fehler	9.1.2
Alternativhypothese	4.4.2/9.1.1
Analyse	4.4.2/10.2
Äquivalenzrelation	6.2.1
Art der Merkmalsausprägung	4.5
Aufsatz in einer Fachzeitschrift	10.1
Auswertungsobjektivität	6.3
Bedeutungsproblem	6.2.1
Befragung	6.1.1
Befunde	10.3.1
Begriffsdefinition	6.2.2
Beitrag in einem Herausgeberwerk	10.1
Beobachtung	6.1.2
Beobachtungsplan	6.1.2
Beschreiben	1.1
Beta/ β-Fehler	9.1.2
Bodeneffekte	6.2.4
chi-Quadrat-Test	9.1.3/9.3

VII. Glossar

Begriff	Definition
Abhängige Variable:	Variable deren Ausprägung durch unabhängige Variablen beeinflusst wird bzw. der in einer Studie interessierende Effekt/Outcome/Endpunkt
Abstract:	kurze Zusammenfassung einer wissenschaftlichen Arbeit, gegliedert in Hintergrund, Fragestellung, Methode, Ergebnisse, Diskussion und Fazit für die Praxis (Empfehlung)
Alpha:	Fehler erster Art, in Signifikanztest die Wahrscheinlichkeit, mit der die Nullhypothese zurückgewiesen wird, obwohl sie zutrifft (z.B. Wahrscheinlichkeit mit der ein beobachteter oder ein noch extremerer Unterschied auftritt, wenn in der Grundgesamtheit kein Unterschied besteht bzw. die Nullhypothese (s.) gilt), bzw. (2) Maß der internen Konsistenz (Reliabilität) von Messinstrumenten (s. Cronbachs alpha)
Analyse:	Prozess der Verarbeitung quantitativer und/oder qualitativer Daten mit dem Ziel, die Forschungsfrage(n) zu beantworten
Attrition:	Verlust von Studienteilnehmern während der Studie (Drop-out). Drop-out kann die interne Validität und die Repräsentativität einer Studie beeinträchtigen
Beschreibende (deskriptive) Studie:	Phänomene oder Populationen werden beschrieben. Alle wissenschaftlichen Arbeiten, auch analytische, beschreiben u.a. Stichprobe und Messinstrumente.
Beta:	(1) oder Fehler zweiter Art, in Signifikanztests die Wahrscheinlichkeit mit der die Alternativhypothese zurückgewiesen wird, obwohl sie zutrifft, bzw. (2) in multiplen Regressionsanalysen das standardisierte Gewicht der unabhängigen Variablen auf die abhängige Variable

Bias:	sämtliche Faktoren mit Einfluss auf das Studienergebnis (abhängige Variable), die nicht originärer Gegenstand einer Untersuchung sind
Chi-Quadrat-Test:	nonparametrischer statistischer Test zur Untersuchung des Zusammenhangs zwischen zwei (oder mehreren) nominalskalierten Variablen
Cronbachs alpha:	Kennzahl zur Abschätzung der internen Konsistenz eines Messinstruments/einer Skala, die aus inhaltlich sinnvollen Items (Einzelfragen) zusammengesetzt ist (Reliabilität eines Messverfahrens)
Deskriptive Statistik:	uni- und bivariate Verfahren, die der Beschreibung und Zusammenfassung von Daten in einer Stichprobe dienen. Maße univariater deskriptiver Statistik sind absolute und relative Häufigkeiten, Median, Mittelwert und Standardabweichung. Bivariate Maße sind beispielsweise Rang- und Maßkorrelationskoeffizienten
Dichotome Variable:	Variable die genau zwei Werte annehmen kann
Diskriminante Validität:	Grad des Zusammenhangs zwischen den Ergebnissen von Messverfahren für unterschiedliche Konstrukte bei ein und derselben Stichprobe zu einem Messzeitpunkt. Ein Messinstrument ist diskriminant valide, wenn ein vorher angenommener Zusammenhang (positiv oder negativ) der Messergebnisse unterschiedliche Konstrukte messender Instrumente erkennbar wird (z.B. sollten Lebenszufriedenheit und psychische Belastungen negativ korrelieren) (s. auch konvergente Validität
Effektstärke:	statistische Größe, mit der die Bedeutung von Zusammenhängen/Unterschieden geschätzt wird. Häufig verwendet werden Eta-Quadrat, Cohens d und Kappa
Experimentalgruppe:	in klinischen Studien zufällig gebildete Gruppe aus Studienteilnehmern, bei denen eine Intervention (z.B. Medikamentengabe) erfolgt bzw. ein potenziell wirksames Verfahren angewendet wird (s. auch Kontrollgruppe)

Experimentelle Studie:	Studienart in der unabhängige Variablen kontrolliert verändert werden um beispielsweise die Wirksamkeit von Medikamenten oder therapeutischer/pflegerischer Interventionen zu untersuchen. Voraussetzung für aussagekräftige Ergebnisse ist die Bildung von mindestens zwei Gruppen, in der unterschiedliche Variationen der unabhängigen Variablen erfolgen (Experimental- und Kontrollgruppe). Stichprobenziehung und Zuordnung zu Experimental- und Kontrollgruppe beruht auf Zufall.
Explorative Studie:	Studienart, in der die grundsätzliche Beschaffenheit von Phänomenen beschrieben wird – (qualitative Studien, Theorieentwicklung)
Faktorenanalyse:	statistisches Verfahren zur Datenreduktion. Eine große Anzahl von Variablen wird zu einer kleineren Anzahl Faktoren zusammengeführt. Basis ist nicht die inhaltliche Übereinstimmung, sondern die gemeinsame Varianz von Variablen
Fall(Case-)Studie:	spezifisches Studiendesign das Daten eines Falles/eines Patienten erhebt und qualitativ/quantitativ untersucht
Freiheitsgrade (degrees of freedom, df):	Anzahl der frei schätzbaren, frei variierbaren Parameter in statistischen Tests. In Stichproben sind bis auf einen Parameter alle weiteren variabel, die Freiheitsgrade errechnen sich häufig durch N-1
F-Test:	s. Varianzanalyse
Halo Effect:	Tendenz zur Über- oder Unterschätzung bestimmter Merkmale in Studien aufgrund des (subjektiven) Gesamteindrucks, den Beobachter von einer Person gewonnen haben
Häufigkeitsverteilung:	Darstellung der Häufigkeit bestimmter Ausprägungen einer Variable, geordnet von der kleinsten bis zur größten Ausprägung (Histogramm, Balkendiagramm, Stengel-Blatt-Diagramm)

Hawthorneeffekt:	Änderung des Verhaltens von Studienteilnehmern die wissen, dass sie Gegenstand wissenschaftlicher Studien sind
Histogramm:	grafische Darstellung der Häufigkeit von Nennungen bestimmter Variablenausprägungen in Form benachbarter Balken, deren Höhe der Häufigkeit einer bestimmten Merkmalsausprägung entspricht
Hypothese:	inhaltliche Aussage zum erwarteten Zusammenhang zwischen zwei oder mehr Variablen. Hypothesen gründen sich auf theoretischen Vorüberlegungen bzw. auf den aktuellen Forschungsstand
Interne Konsistenz:	drückt aus, in welchem Ausmaß alle Items einer Skala dieselbe Dimension eines Konstrukts messen
Interne Validität:	Maß der Studienqualität das bestimmt, in welchem Maß Veränderungen der abhängigen Variable auf die Variation der unabhängigen Variable zurückzuführen sind
Interne Validität:	Veränderungen der abhängigen können zweifelsfrei auf die Variation der unabhängigen Variablen zurückgeführt werden bzw. existiert keine plausible Alternativerklärung für Veränderungen der abhängigen Variable
Intervallskalenniveau:	Messung von Daten in Rangordnung mit gleichen Abständen zwischen den Merkmalsausprägungen (z.B. Temperatur in Grad Celsius)
Kausalzusammenhang:	ursächlicher Zusammenhang, d.h., Veränderungen abhängigen Variable beruhen auf Variationen der unabhängigen Variable. Zur Untersuchung von Kausalzusammenhängen muss ein und dieselbe Population zu mindestens zwei Messzeitpunkten untersucht werden (s. 8.5 Längsschnittstudie)
Konfidenzintervall:	Vertrauensintervall (95- oder 99% Konfidenzintervall), ist der Bereich um einen Stichprobenkennwert (Mittelwert, prozentuale Verteilung, Risikoschätzer u.a.), in dem der

	Stichprobenkennwert in 95 (99) Studien liegen würde, wenn die Studie 100 mal wiederholt würde. Nur in 5 (1) der Studien würde ein Ergebnis außerhalb des Konfidenzintervalls erzielt.
Konstruktvalidität:	Zuverlässigkeit mit dem ein Messinstrument ein Konstrukt abbildet/misst.
Kontrollgruppe:	in klinischen Studien zufällig gebildete Gruppe aus Studienteilnehmern, bei denen eine Intervention (z.B. Medikamentengabe) nicht erfolgt bzw. ein unwirksames Imitat angewendet wird (Placebo) (s. Experimentalgruppe)
Konvergente Validität:	Grad des Zusammenhangs zwischen den Ergebnissen unterschiedlicher Messverfahren für ein und dasselbe Konstrukt bei ein und derselben Stichprobe zu einem Messzeitpunkt. Ein Messinstrument ist konvergent valide, wenn ein enger Zusammenhang der Messergebnisse mit denen anderer Messinstrumente für dasselbe Konstrukt erkennbar wird (z.B. Flüssigkeits- vs. elektrisches Thermometer) (s. auch diskriminante Validität)
Korrelationskoeffizient:	standardisiertes Maß zur Abschätzung des statistischen Zusammenhangs zweier Variablen. Ein Korrelationskoeffizient kann Werte zwischen -1 und +1 annehmen, wobei Werte nahe den Extremwerten auf einen engen Zusammenhang, Werte nahe 0 auf einen schwachen bzw. keinen Zusammenhang hinweisen.
Längsschnittstudie:	Studiendesign, in dem zu mehreren Messzeitpunkten Daten von ein und derselben Stichprobe erhoben werden. Ausgehend von Längsschnittstudien können Aussagen über kausale (Ursache-Wirkungs-)Zusammenhänge getroffen werden
Ratingskala (Likertskala):	Skalenverankerung in Messinstrumenten, mit denen erhoben wird, wie sehr Studienteilnehmer einer Aussage zustimmen bzw. eine Aussage abgelehnt wird. Häufig werden sieben Bewertungsstufen vorgegeben (7-stufige Likertskala)

Literaturreview:	Prozess der Literaturrecherche und der inhaltlichen Systematisierung von Literatur mit dem Ziel, den aktuellen Forschungsstand abzubilden
Medianwert:	deskriptiv-statistische Maßzahl zur Abbildung der zentralen Tendenz ordinalskalierter Daten. Der Medianwert gibt den Wert an, den 50% der Stichprobe über- und 50% der Stichprobe unterschreiten
Mehrstufige Zufallsauswahl (cluster-sampling):	ein Verfahren zufallsgesteuerter Stichprobenauswahl bei unbekannter Grundgesamtheit, aus sinnvollen Clustern (z.B. Städten vergleichbarer Größe und sozialer Zusammensetzung) wird eines zufallsgesteuert gezogen, aus diesem Cluster wiederum zufallsgesteuert eine Stichprobe
Messinstrumente:	Anwendungen zur systematischen Sammlung von Daten
Mittelwert:	deskriptiv-statistische Maßzahl zur Abbildung der zentralen Tendenz intervallskalierter Daten. Der Mittelwert wird durch Summierung der Messwerte aller Studienteilnehmer und die anschließende Division durch die Anzahl der Studienteilnehmer gebildet.
Modalwert:	deskriptiv-statistische Maßzahl die die am häufigsten besetzte Kategorie/den am häufigsten genannten Wert abbildet.
Multiple Regression:	ein statistisches Verfahren zur Berechnung des Gewichts mehrerer unabhängiger Variablen auf eine abhängige Variablen
N:	gibt Auskunft über die Größe der Gesamtstichprobe
n:	gibt Auskunft über die Größe von Teilstichproben einer Gesamtstichprobe
Nominalskalenniveau:	Messniveau von Daten, das auf Operationalisierung von qualitativen Eigenschaften von Studienteilnehmern be-

	ruht, die keiner Rangordnung folgen (Vereinsmitgliedschaft, Geschlecht) (auch als kategoriale Variablen bezeichnet)
Nonparametrische Statistik:	statistische Verfahren zur Signifikanztestung nominal- und ordinalskalierter Daten sowie aller anderen Messniveaus mit nicht normalverteilten Daten (s. Chi-Quadrat-Test)
Nullhypothese:	Teil des statistischen Hypothesenspaars, in dem von keinem Unterschied/keinem Zusammenhang zwischen Studienvariablen ausgegangen wird (s. Alpha)
Ordinalskalenniveau:	Messniveau, das Daten in einer inhaltlich begründeten Rangfolge erhebt
Parametrische Statistik:	statistische Verfahren zur Signifikanztestung intervallskalierter, normalverteilter Daten
Pearson's r:	Korrelationskoeffizient, der die Enge des Zusammenhangs zwischen zwei intervallskalierten, normalverteilten Variablen bestimmt
Pilotstudie:	Studie mit kleiner Stichprobe zur Überprüfung des Studienplans und der verwendeten Testverfahren
Poweranalyse:	Testverfahren mit dem die benötigte Stichprobengröße ermittelt werden kann, mit der die erwarteten Unterschiede/Zusammenhänge von Studienvariablen statistisch signifikant werden
Qualitative Analyse:	Methode zur Analyse nicht-numerischer Variablen (Worte, Texte, Bilder)
Quasi-experimentelle Studie:	Studienart, in der unabhängige Variablen kontrolliert verändert werden, um beispielsweise die Wirksamkeit von Medikamenten oder therapeutischer/pflegerischer Interventionen zu untersuchen. Voraussetzung für aussagekräftige Ergebnisse ist die Bildung von mindestens zwei Gruppen, in der unterschiedliche Variationen der

	unabhängigen Variablen erfolgen (Experimental- und Kontrollgruppe). Stichprobenziehung und Zuordnung zu Experimental- und Kontrollgruppe erfolgt hierbei willkürlich und beruht nicht auf Zufall (s. experimentelle Studie, interne Validität).
Querschnittstudie:	Studiendesign bei dem Daten einer Stichprobe zu einem Messzeitpunkt erhoben werden. Aufgrund von Querschnittstudien können keine Ursache-Wirkungsbeziehungen aufgedeckt werden, sondern lediglich das zeitgleiche Auftreten der untersuchten Zustände.
r:	Symbol des Korrelationskoeffzienten
R^2:	Symbol des quadrierten multiplen Zusammenhangs in. Gibt Auskunft über die durch alle unabhängigen Variablen im Modell erklärte Varianz der abhängigen Variablen
Range:	weist den gesamten Wertebereich einer Variablen vom kleinsten bis zum größten Wert aus
Regressionsanalyse:	Verfahren zur Vorhersage einer abhängigen Variable durch eine oder mehrere abhängige Variablen
Reliabilität (Zuverlässigkeit):	interne Konsistenz einer Skala/eines Messverfahrens, beschreibt die Zuverlässigkeit, mit der ein Messinstrument unter konstanten Bedingungen gleiche Ergebnisse liefert
Rücklaufquote (Response rate):	Anteil der an einer Studie tatsächlich Teilnehmenden an den angefragten/gezogenen Studienteilnehmern
Scatterplot Diagramm:	graphische Darstellung der Korrelation zweier Variablen (Punktwolke)
Signifikanzniveau:	Wahrscheinlichkeit, mit der ein beobachteter Zusammenhang/Unterschied auf Zufall beruht. Ein Signifikanzniveau von 5 Prozent bedeutet, dass in 5 von 100 Stu-

	dien bei denen die Nullhypothese gilt, ein Zusammenhang gefunden wird
Standardabweichung:	Maß der deskriptiven Statistik, ist der Wertebereich um einen Mittelwert, der ca. 68 Prozent der Stichprobe repräsentiert (Quadratwurzel der Varianz).
Stör- (Confounding) variable:	Einflussfaktor in Studien, der die Ausprägung der abhängigen Variable beeinflusst, jedoch vom Studienteam nicht kontrolliert wurde (s. 4.5.1 Moderator-, Kontroll- oder Mediatorvariable)
Test-Retest Reliabilität:	Verfahren zur Bestimmung der Reliabilität eines Messinstruments/einer Skala, bei dem dieselben Personen mit einem Verfahren zwei oder mehrmals befragt werden. Je größer die Übereinstimmung der Testergebnisse, desto zuverlässiger ist ein Test
t-Test:	Test zur Berechnung der Signifikanz von Mittelwertunterschieden zweier Stichproben
Unabhängige Variable:	in einem Forschungsplan die Bedingung/das Merkmal, das einen theoretisch begründeten Einfluss auf eine abhängige Variable hat
Validität (Gültigkeit):	beschreibt die Eigenschaft eines Messinstruments/einer Skala, das Konstrukt abzubilden, was es vorgibt zu messen
Variable:	Forschungsgegenstand/Merkmal, das unterschiedliche Werte annehmen kann
Varianz:	Maß der deskriptiven Statistik, quadrierte und aufsummierte Differenz zwischen Einzelmessung und Messwert in einer Stichprobe. Gibt Auskunft über die Verteilung der Messwerte in einer Stichprobe (s. Standardabweichung)
Varianzanalyse:	statistischer Test zum Vergleich arithmetischer Mittel zwischen 3 oder mehr Gruppen

Zufallsstichprobe:	ergibt sich aus einer Stichprobenauswahl, bei der alle Mitglieder der bekannten Grundgesamtheit dieselbe Chance haben, gezogen zu werden
Z-Wert:	Zahl der Standardabweichungen über bzw. unter dem Stichprobenmittelwert, den der Mittelwert einer Teilstichprobe oder eines Studienteilnehmers einnimmt.