Markus Tausendpfund
Simone Abendschön

# Quantitative Analyseverfahren. Eine Einführung

**Redaktion und Überarbeitung:**
**Davin Akko, Anne-Kathrin Bestgen, Felicitas Kempf und Julia Schütz**

Fakultät für
**Kultur- und Sozialwissen-schaften**

FernUniversität in Hagen

Der Inhalt dieses Studienbriefs wird gedruckt auf Recyclingpapier (80 g/m$^2$, weiß), hergestellt aus 100 % Altpapier.

# Vorwort der Modulbetreuung

Dieser Kurs bietet den Studierenden der Bildungswissenschaft eine Einführung in quantitative Analyseverfahren und einen verständlichen Einstieg, wie Daten in den Sozial- und Bildungswissenschaften quantitativ ausgewertet werden können. Die Bildungswissenschaft ist Teil der Sozialwissenschaften und beide greifen auf nahezu identische Forschungsmethoden zurück. Die empirische Bildungsforschung ist interdisziplinär zwischen Soziologie, Psychologie und auch Ökonomie aufgestellt, zentral jedoch erziehungs- und bildungswissenschaftlich verankert und untersucht mittels Methoden der empirischen Sozialforschung Bildungs-, Lern- und Reflexionsprozesse.

Mit der Bearbeitung dieses Kurses sind folgende Lernziele verbunden:[1]

- Sie können verschiedene quantitative Analyseverfahren der empirischen Bildungsforschung benennen, beschreiben und anwenden.
- Sie können Verteilungen anhand statistischer Kennwerte beschreiben sowie deskriptive Ergebnisse interpretieren.
- Sie kennen verschiedene grafische Darstellungsformen und können geeignete Formen für die Darstellung statistischer Informationen auswählen.
- Sie können Strategien Methoden und Abläufe der empirischen Datenauswertung beschreiben, erklären und anwenden.
- Sie können verschiedene quantitative Analyseverfahren der empirischen Bildungsforschung unterscheiden und für ihre Forschungsfrage geeignete Verfahren auswählen.
- Sie können die Ergebnisse uni-, bi- und multivariater Analysen auswerten und interpretieren.
- Sie sind in der Lage Analyseverfahren im Hinblick auf ihre Angemessenheit für bestimmte Fragestellungen sowie Anwendungsvoraussetzungen zu beurteilen.

Den Autor Markus Tausendpfund und die Autorin Simone Abendschön möchten wir Ihnen gerne kurz vorstellen:

**Dr. Markus Tausendpfund** studierte Sozialwissenschaften mit den Schwerpunkten Soziologie, Sozialpsychologie, Methoden der empirischen Sozialforschung, Politische Soziologie und Arbeits- und Organisationspsychologie an der Universität Mannheim. 2012 schloss er seine Promotion zum Thema „Individuelle und kontextuelle Faktoren der politischen Unterstützung der Europäischen Union“ ab und leitet seit 2014 die Arbeitsstelle Quantitative Methoden an der FernUniversität in Hagen.

**Prof.‘in Dr.‘in Simone Abendschön** studierte Geschichte, Soziologie, Politikwissenschaft und Anglistik in Freiburg und Mannheim und promovierte 2010 zur Sozialisation von politischen und

[1] Die Lernziele orientieren sich an der Lernzieltaxonomie nach Bloom, B. et al. (1956). Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain. New York, Toronto: Longmans, Green.

demokratischen Werten im jungen Kindesalter. Simone Abendschön war bereits an den Universitäten Mannheim, Bamberg und Frankfurt tätig und hat seit 2015 die Professur für Politikwissenschaft mit dem Schwerpunkt Methoden der Politikwissenschaft an der Justus-Liebig-Universität Gießen inne.

An dieser Stelle möchten wir uns insbesondere bei Markus Tausendpfund für die angenehme Kooperation und den stets interessanten Austausch bedanken.

Der Studienbrief wurde von Davin Akko, M.Sc., Dr.'in Anne-Kathrin Bestgen, Felicitas Kempf, M.A. und Prof.'in Dr.'in Julia Schütz am Lehrgebiet Empirische Bildungsforschung redaktionell überarbeitet. Dabei wurden keine inhaltlichen Änderungen vorgenommen, sondern Änderungen aufgrund eines inklusiven Sprachgebrauchs eingefügt sowie eine barrierefreie Gestaltung beachtet. Zudem werden Bildungswissenschaftler*innen explizit als Zielgruppe angesprochen. In der Moodle-Lernumgebung des Moduls werden Lehrvideos und Übungsaufgaben veröffentlicht, die die Auseinandersetzung mit den Inhalten des Kurses fördern sollen.

Wir wünschen Ihnen viel Erfolg bei der Bearbeitung und eine anregende Lektüre!

Davin Akko, Anne-Kathrin Bestgen, Felicitas Kempf und Julia Schütz

# Vorwort der Autor*innen

In der quantitativen Sozial- und Bildungsforschung wird zur Beschreibung von Daten und zur empirischen Überprüfung von Hypothesen auf statistische Verfahren zurückgegriffen. Wer eine (quantitative) Studie verstehen und kritisch bewerten möchte, der muss die grundlegenden Prinzipien, Anwendungsvoraussetzungen und auch Probleme der verwendeten statistischen Verfahren kennen. Für Sozial- und Bildungswissenschaftler*innen sind deshalb elementare Kenntnisse dieser quantitativen Analyseverfahren unverzichtbar.

Für die Sozial- und Bildungswissenschaften stellt die Statistik eine zentrale Hilfswissenschaft dar. Während sich Statistiker*innen – allgemeiner: Mathematiker*innen – häufig mit der Beweisführung und der Weiterentwicklung mathematischer Algorithmen beschäftigen, stehen für Studierende der Bildungswissenschaft das Kennenlernen und die praktische Anwendung statistischer Verfahren im Vordergrund. Im Mittelpunkt des Kurses steht das Verständnis quantitativer Analyseverfahren, mit denen Sozial-und Bildungswissenschaftler*innen bei der Auseinandersetzung mit quantitativen Studien konfrontiert werden.

Der vorliegende Kurs behandelt vier Themenbereiche: Univariate, bivariate und multivariate Datenanalyse sowie Grundlagen der Inferenzstatistik. Das Kapitel zur univariaten Datenanalyse behandelt die Häufigkeitsverteilung einzelner Merkmale. Dabei werden Lage- und Streuungsmaße sowie Formmaße vorgestellt. Die bivariate Datenanalyse untersucht Zusammenhänge zwischen zwei Merkmalen und Unterschiede zwischen zwei Merkmalen (Mittelwertvergleiche). Dabei werden Kreuztabellen sowie wichtige Zusammenhangsmaße behandelt. Bei der multivariaten Datenanalyse werden mit der linearen und logistischen Regression zwei zentrale Analyseverfahren der Sozialwissenschaften vorgestellt, die den Einfluss mehrerer unabhängiger Variablen auf eine abhängige Variable schätzen können. Aus zeitlichen, finanziellen und forschungspraktischen Gründen dominieren in den Sozial- und Bildungswissenschaften Stichproben. Deshalb behandelt der vierte Teil des Kurses die Grundlagen der Inferenzstatistik, die Instrumente zur Verfügung stellt, um zu entscheiden, ob und wie empirische Befunde aus Zufallsstichproben auf zugehörige Grundgesamtheiten übertragen werden dürfen.

In der Moodle-Lernumgebung des Moduls werden Videos und Übungsaufgaben veröffentlicht, die die Auseinandersetzung mit den Inhalten des Kurses fördern sollen. Für die kritische Durchsicht des Kurses sind wir Christian Cleve und Daniel Saar sehr dankbar.

Über Hinweise auf Fehler, Kommentare und Verbesserungsvorschläge freuen wir uns. Senden Sie Ihre Kommentare bitte an Markus.Tausendpfund@fernuni-hagen.de. Vielen Dank.

Hagen, im Januar 2020

Markus Tausendpfund und Simone Abendschön

# Inhaltsverzeichnis

# Abbildungsverzeichnis

# Tabellenverzeichnis

# 1 Einführung

Markus Tausendpfund

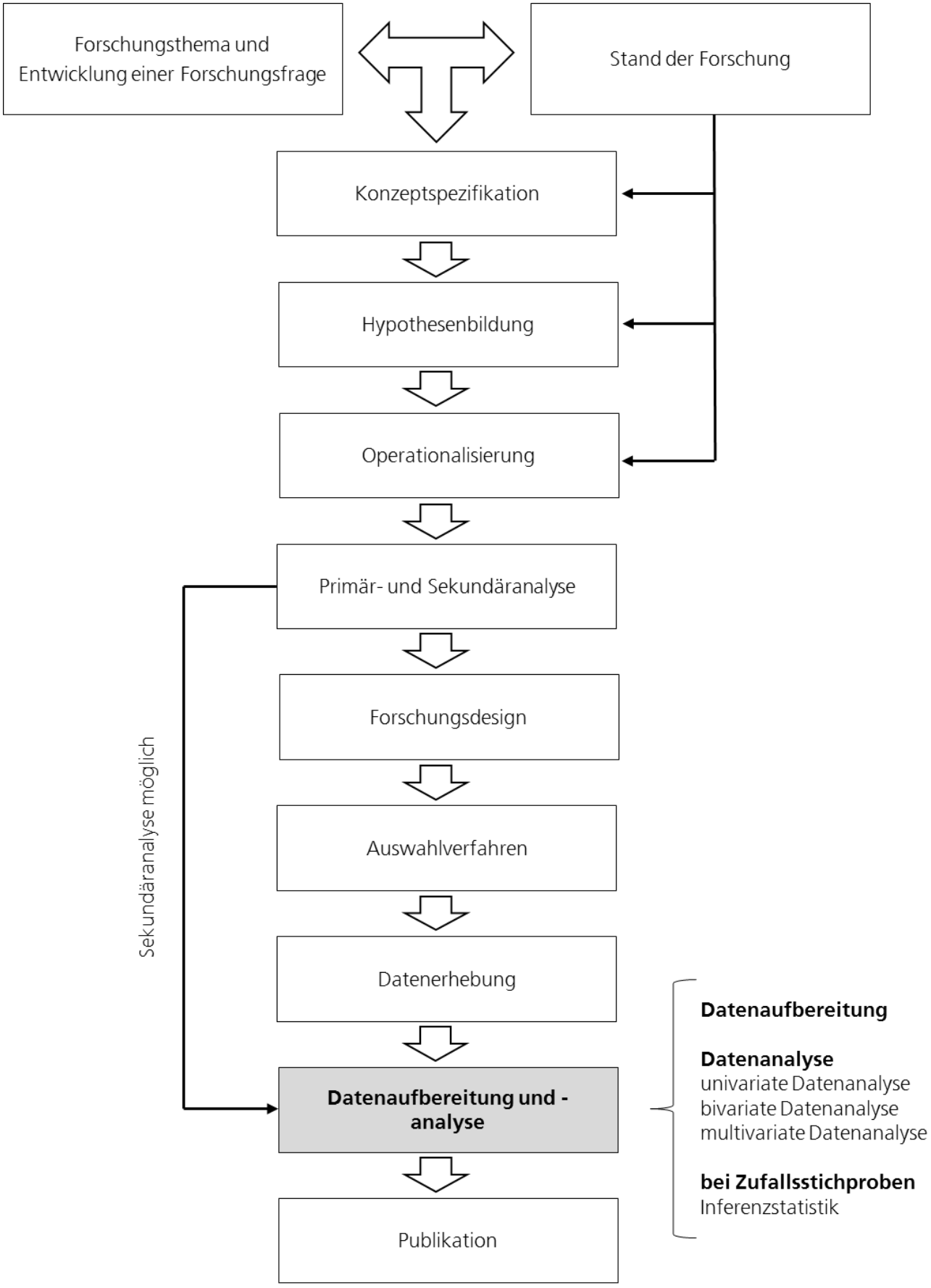

Dieses Kapitel macht Sie mit den Grundlagen der quantitativen Datenanalyse vertraut. Nach der Einordnung der Phase „Datenanalyse" innerhalb des Forschungsprozesses werden die Begriffe „Grundgesamtheit" und „Stichprobe" erläutert. Bei einer empirischen Studie werden meist Aussagen über größere Gruppen angestrebt (z.B. die wahlberechtigte Bevölkerung in Deutschland). Allerdings liegen in den meisten Studien keine Informationen über alle Elemente dieser Gruppe vor, sondern nur über eine (zufällige) Auswahl dieser Gruppe. Die Gruppe, über die eine Aussage gemacht werden soll, wird als Grundgesamtheit oder Population bezeichnet. Die Gruppe, über die empirische Informationen vorliegen, wird als Stichprobe bezeichnet. Diese Begriffe werden knapp erläutert und es werden die Voraussetzungen skizziert, unter denen Befunde einer Stichprobe auf die zugehörige Grundgesamtheit übertragen werden können. Abschließend werden typische Klassifikationen von Variablen vorgestellt. Dabei liegt der Fokus auf dem Skalenniveau von Variablen, da das Skalenniveau eine wichtige Voraussetzung für die Anwendung bestimmter Analyseverfahren ist.

## 1.1 Einordnung im Forschungsprozess

Die quantitativen Analyseverfahren werden häufig mit dem quantitativen Forschungsprozess gleichgesetzt. Quantitativ arbeitende Sozial- und Bildungswissenschaftler*innen nutzen statistische Analyseverfahren, um die theoretisch formulierten Hypothesen empirisch zu überprüfen. Sicherlich ist die Anwendung statistischer Analyseverfahren ein zentrales Merkmal des quantitativen Forschungsprozesses, aber die quantitative Datenanalyse sollte nicht isoliert betrachtet werden.

Vor der Datenanalyse bzw. Anwendung quantitativer Analyseverfahren müssen empirische Sozial- und Bildungsforscher*innen wichtige vorgelagerte Entscheidungen treffen, die unmittelbare Auswirkungen auf die empirischen Befunde haben. Wie Abbildung 1 zeigt, stehen die Festlegung eines Forschungsthemas und die Entwicklung einer geeigneten Forschungsfrage am Beginn eines Forschungsprojekts. Auf dieser Grundlage werden die zentralen Konzepte identifiziert und theoretisch geklärt, ehe gehaltvolle Hypothesen formuliert und valide Operationalisierungen dieser Konzepte entwickelt werden. Diese Phasen in einem Forschungsprozess erfolgen in intensiver Auseinandersetzung mit dem existierenden Forschungsstand. Nur wer den Forschungsstand zu seinem*ihrem Forschungsthema kennt, kann eine gehaltvolle Forschungsfrage entwickeln. Die Auseinandersetzung mit der Fachliteratur ist aber auch für die Konzeptspezifikation und die Entwicklung von Hypothesen erforderlich. Schließlich ist auch bei der „Übersetzung" theoretischer Konzepte in empirische Indikatoren ein Überblick über existierende Operationalisierungen notwendig.

Abbildung 1: Idealtypischer Ablauf eines quantitativen Forschungsprojekts




Quelle: Eigene Darstellung

Kein Analyseverfahren kann die intensive Auseinandersetzung mit dem existierenden Forschungsstand ersetzen. Ungeeignete Konzeptspezifikationen, schwammige Hypothesen oder auch ungültige Operationalisierungen führen zwangsläufig zu schlechten Daten und kein Analyseverfahren der Welt kann aus schlechten Daten valide empirische Befunde machen. Deshalb: Die Anwendung bzw. Durchführung quantitativer Analyseverfahren kann nur dann zu belastbaren empirischen Befunden führen, wenn die vorgelagerten Phasen erfolgreich bearbeitet wurden.

Wenn für die Bearbeitung einer Forschungsfrage und die Überprüfung der Hypothesen bereits geeignetes Datenmaterial existiert (z.B. ALLBUS), dann können die Phasen „Forschungsdesign", „Auswahlverfahren" und „Datenerhebung" übersprungen werden. In einem solchen Fall führen die Sozial- und wissenschaftler*innen eine Sekundäranalyse durch. Es werden existierende Daten genutzt, um die Forschungsfrage zu bearbeiten. Falls keine geeigneten Daten zur Verfügung stehen, bietet sich eine Primäranalyse an. Bei einer Primäranalyse werden neue Daten erhoben, um die Forschungsfrage zu beantworten.

**Datenaufbereitung und -analyse**

Die Phase „Datenaufbereitung und -analyse" umfasst in der Regel mehrere Zwischenschritte (Stein 2014, S. 150; Tausendpfund 2018b, S. 50-51). Zunächst müssen die im Rahmen der Datenerhebung gesammelten empirischen Informationen systematisch in einen Datensatz aufgenommen werden (Kromrey et al. 2016, S. 217-218). Die Variablen müssen beschriftet und ein Codebuch muss angelegt werden (z.B. Lück und Baur 2011; Lück und Landrock 2014; Tausendpfund 2018b, S. 291-297). Bei der Arbeit mit qualitativ hochwertigen Sekundärdaten (z.B. ALLBUS) stehen meist „fertige" Datensätze zur Verfügung. Insbesondere bei der eigenständigen Dateneingabe, aber auch bei der Arbeit mit Sekundärdaten, sind Fehlerkontrollen (z.B. Eingabefehler) und Plausibilitätstests erforderlich.

Vor der eigentlichen Datenanalyse müssen Variablen häufig verändert oder neu erstellt werden. Dieser Prozess wird häufig als Datenmodifikation oder Datentransformation bezeichnet (Fromm 2011; Kohler und Kreuter 2017, S. 91-130; Tausendpfund 2018a, S. 62-91). Dabei wird die Kodierung von Variablen angepasst, einzelne Subgruppen werden gebildet oder es werden auf Basis der verfügbaren Informationen auch neue Variablen erstellt. Das Verändern und das Erstellen neuer Variablen dauert häufig länger als die eigentliche Datenanalyse. Eine sorgfältige Durchführung der einzelnen Schritte ist dabei eine Voraussetzung für die Gültigkeit der anschließenden Analysen.

Bei der anschließenden Datenanalyse lassen sich meist vier Schritte unterscheiden, die auch im Mittelpunkt dieses Kurses stehen: die univariate, die bivariate und die multivariate Datenanalyse sowie die Inferenzstatistik.

**Univariate Datenanalyse**

Die univariate Datenanalyse befasst sich mit einzelnen Variablen. In einem ersten Schritt werden die absoluten und relativen Häufigkeiten der einzelnen Ausprägungen einer Variable in Tabellen oder Grafiken dargestellt. In der quantitativen Sozial- und Bildungsforschung sind wir allerdings in der Regel mit vielen Untersuchungsobjekten konfrontiert. Deshalb wird in einem zweiten Schritt die Informationsmenge von mehreren tausend Beobachtungen auf wenige Kennzahlen verdichtet. Dabei lassen sich Lage-, Streuungs- und Formmaße unterscheiden. Während Lagemaße über das Zentrum einer Verteilung informieren, beschreiben

Streuungsmaße die Variation eines Merkmals in einer Verteilung. Mit Schiefe und Wölbung kann die Form einer Verteilung charakterisiert werden.

**Bivariate Datenanalyse**

Bei der bivariaten Datenanalyse werden immer genau zwei Variablen in Beziehung gesetzt (z.B. Bildung und Einkommen). Bivariate Analyseverfahren werden genutzt, um Zusammenhänge oder Unterschiede zwischen zwei Merkmalen zu untersuchen und Hypothesen empirisch zu überprüfen. Dafür nutzen wir Kreuztabellen und Zusammenhangsmaße. Kreuztabellen (engl. crosstabs) sind eine einfache und anschauliche Möglichkeit, um die Beziehung zwischen zwei Merkmalen in den Blick zu nehmen. Neben absoluten Häufigkeiten können auch die Anteile der einzelnen Häufigkeiten (Anteile) berechnet werden. Die Stärke einer Beziehung zwischen zwei Merkmalen (z.B. Bildung und Einkommen) kann mit Zusammenhangsmaßen – sogenannten Koeffizienten – charakterisiert werden. Die bekanntesten Zusammenhangsmaße sind sicherlich Cramér's V, Spearman's rho und Pearson's r.

**Multivariate Datenanalyse**

Mit bivariaten Analyseverfahren wird der Zusammenhang zwischen zwei Variablen untersucht. In der Realität können Merkmale wie Einkommen oder Wahlbeteiligung aber nicht durch eine Variable „erklärt" werden, sondern in der Regel muss eine Vielzahl an Variablen gleichzeitig berücksichtigt werden. Mit der Regressionsanalyse steht ein sehr mächtiges Analyseinstrument zur Verfügung, um den Einfluss einer oder mehrerer unabhängiger Variablen auf eine abhängige Variable zu schätzen. Mit der linearen und logistischen Regression werden in diesem Kurs zwei multivariate Analyseverfahren vorgestellt, die in den Sozialwissenschaften häufig verwendet werden.

**Inferenzstatistik**

Die univariate, bivariate und multivariate Datenanalyse haben das Ziel, die Verteilung von Variablen zu beschreiben und Zusammenhänge zwischen zwei oder mehr Variablen zu untersuchen. Diese Datenanalyse basiert in der Regel auf Stichproben. Das heißt, es liegen nicht von allen Untersuchungsobjekten einer Grundgesamtheit empirische Informationen vor, sondern nur von einer (zufälligen) Auswahl. Die Inferenzstatistik beschäftigt sich mit der Frage, ob und wie Befunde von Zufallsstichproben auf zugehörige Grundgesamtheiten übertragen werden können (siehe auch Abschnitt 1.2).

Abbildung 1 soll verdeutlichen, dass die Datenanalyse bzw. die Anwendung statistischer Analyseverfahren immer nur eine Phase in einem bildungswissenschaftlichen Projekt darstellt. Empirische Befunde „sprechen" niemals für sich selbst, sondern sind immer eingebunden in eine bildungswissenschaftliche Forschungsfrage. Ohne theoretische Vorarbeiten (z.B. Entwicklung von Hypothesen) kann eine quantitative Analyse nicht zielorientiert erfolgen. Deshalb muss eine quantitative Datenanalyse immer an den (theoretischen) Forschungsprozess zurückgekoppelt werden, um in Publikationen empirisch interessante und vor allem valide Schlussfolgerungen ziehen zu können.

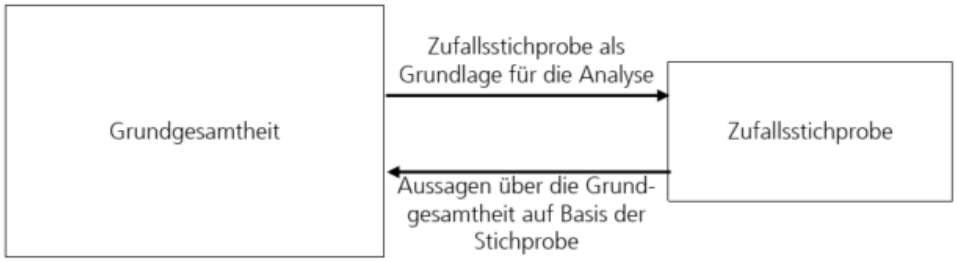
## 1.2 Grundgesamtheit und Stichprobe

Aus zeitlichen, finanziellen und forschungspraktischen Gründen dominieren in den Sozial- und Bildungswissenschaften Stichproben (Tausendpfund 2018b, S. 207-210). Bei empirischen Studien

werden in der Regel nicht alle Elemente der Grundgesamtheit untersucht, sondern nur eine zufällige Auswahl dieser Elemente. Ein Beispiel: Bei der Bundestagswahl 2017 waren nach Angaben des Bundeswahlleiters 61.688.485 Personen wahlberechtigt. Bei einer Analyse des Wahlverhaltens bei der Bundestagswahl bilden diese Personen die Grundgesamtheit. Bei einer Vollerhebung würden empirische Informationen aller Untersuchungsobjekte der Grundgesamtheit erhoben. Die Kosten der Datenerhebung und die Dauer der Erhebung sprechen allerdings gegen eine solche Vollerhebung. Für die Analyse des Wahlverhaltens (z.B. im Rahmen der German Longitudinal Election Study) wird deshalb auch keine Vollerhebung angestrebt, sondern lediglich eine Zufallsstichprobe realisiert.

In Abbildung 2 wird der Zusammenhang zwischen Grundgesamtheit und Stichprobe illustriert. Im Rahmen eines Forschungsprojekts sollen Aussagen über die Grundgesamtheit gemacht werden. In vielen Fällen ist allerdings eine Vollerhebung nicht möglich. Deshalb wird eine Zufallsstichprobe realisiert, die als Grundlage für empirische Analysen dient. Für die Berechnung einfacher Lage- und Streuungsmaße (univariate Datenanalyse), die Untersuchung von Zusammenhängen zwischen zwei Merkmalen (bivariate Datenanalyse) sowie die Schätzung von Regressionsmodellen (multivariate Datenanalyse) werden die Daten der Stichprobe genutzt.

Abbildung 2: Grundgesamtheit und Stichprobe




Quelle: Eigene Darstellung

Bei Zufallsstichproben sind allerdings Stichprobenfehler unvermeidlich. Der Mittel- oder Anteilswert einer Stichprobe wird vom wahren Mittel- oder Anteilswert der Grundgesamtheit abweichen. Ein Beispiel: In der Stichprobe wird ein mittleres Alter von 45,2 Jahren ermittelt. Dieses mittlere Alter wird vom mittleren Alter in der Grundgesamtheit abweichen. Das mittlere Alter in der Grundgesamtheit ist möglicherweise 45,1 Jahre oder 45,3 Jahre, aber vermutlich nicht 45,2 Jahre. Diese Abweichung wird als Stichprobenfehler bezeichnet.

Die Inferenzstatistik stellt uns Instrumente bereit, um zu entscheiden, ob und wie empirische Befunde aus Zufallsstichproben auf zugehörige Grundgesamtheiten übertragen werden dürfen. Die Grundlagen der Inferenzstatistik und ihre Instrumente werden im Kapitel „Inferenzstatistik“ behandelt. Die grundsätzliche Frage, ob Stichprobenergebnisse auf die Grundgesamtheit übertragen werden dürfen, begegnet uns allerdings bereits bei der Darstellung der univariaten, bivariaten und multivariaten Datenanalyse. Wir werden entsprechende Fragen an den erforderlichen Stellen knapp beantworten und ggf. auf das ausführliche Kapitel am Ende dieses Kurses verweisen.

An dieser Stelle möchten wir auf zwei häufige Fehler hinweisen, die wir im Zusammenhang mit Stichproben immer wieder beobachten. Erstens: Die Anwendung der Inferenzstatistik setzt eine

Zufallsstichprobe voraus. Nur bei einer Zufallsstichprobe kann innerhalb statistischer Fehlergrenzen ein Befund auf die Grundgesamtheit übertragen werden. Zweitens: Bei einem sogenannten Signifikanztest (dabei handelt es sich um ein Instrument der Inferenzstatistik) wird geprüft, ob ein in der Stichprobe gefundener Zusammenhang (sehr) wahrscheinlich auch in der Grundgesamtheit existiert. Ein Befund wird als signifikant bezeichnet, wenn er mit großer Sicherheit von der Stichprobe auf die Grundgesamtheit übertragen werden kann. Signifikant bedeutet aber nicht, dass es sich um einen wichtigen oder starken Zusammenhang zwischen zwei Merkmalen handelt.

## 1.3 Klassifikationen von Variablen

Eine Variable ist ein sozialwissenschaftliches Merkmal mit mindestens zwei Ausprägungen. Das Geschlecht, der allgemeinbildende Schulabschluss oder auch das politische Interesse einer Person sind Beispiele für sozialwissenschaftliche Variablen. Sozialwissenschaftliche Merkmale bzw. Variablen können nach verschiedenen Kriterien klassifiziert werden. Wir unterscheiden vier Kriterien: Skalenniveau, diskrete und stetige Variablen, dichotome und polytome Variablen sowie manifeste und latente Variablen.

**Verschiedene Skalenniveaus**

Eine wichtige Voraussetzung für die Anwendung bestimmter Analyseverfahren ist das Skalenniveau der Variable bzw. des Merkmals. In den Sozial- und Bildungswissenschaften werden meist die Skalenarten von Stevens (1946) verwendet, der vier Skalenniveaus unterscheiden: Nominal-, Ordinal-, Intervall- und Ratioskala. Intervall- und Ratioskalen werden auch metrische Skalen genannt (Tausendpfund 2018b, S. 119-124). Das jeweilige Skalenniveau bestimmt die zulässigen Rechenoperationen. Je höher das Skalenniveau ist, desto mehr Rechenoperationen sind möglich.

Das nominale Skalenniveau ist das niedrigste Skalenniveau. Können die Ausprägungen eines Merkmals lediglich im Hinblick auf Gleichheit oder Ungleichheit verglichen werden, liegt ein nominales Skalenniveau vor (Gehring und Weins 2009, S. 43-47). Ein Beispiel für eine nominal skalierte Variable ist das Geschlecht. In vielen sozialwissenschaftlichen Datensätzen wird der Ausprägung „weiblich" die Ziffer 1 und der Ausprägung „männlich" die Ziffer 2 zugeordnet. Aber diese Zuordnung ist eine Konvention. Man könnte auch 1 für „männlich" und 2 für „weiblich" verwenden.[2] Bei einer nominalskalierten Variable stellen die Ziffern lediglich eine Kennzeichnung dar, die nicht richtig oder falsch, sondern allenfalls mehr oder weniger sinnhaft ist. Die Möglichkeiten der quantitativen Datenanalyse bei nominalskalierten Variablen sind daher begrenzt.

Das ordinale Skalenniveau ist das nächsthöhere Skalenniveau. Bei einer ordinalskalierten Variable können die verschiedenen Ausprägungen einer Variable in eine Rangfolge gebracht werden. Beispiele für ordinalskalierte Variablen sind der Schulabschluss oder auch das politische Interesse. Die allgemeine Hochschulreife ist ein höherer Schulabschluss als die Mittlere Reife und die Mittlere Reife ist ein höherer Abschluss als ein Hauptschulabschluss. Ein „sehr starkes" Interesse für Politik ist ein größeres Interesse als ein „mittleres" Interesse für Politik. Bei einer ordinalskalierten Variable können zwar die einzelnen Ausprägungen in eine Rangfolge gebracht werden, aber die Abstände

[2] Eine kritische Auseinandersetzung mit der dichotomen Operationalisierung des Geschlechts bieten Berner, N., Rosenkranz, L., & Schütz, J. (2019).

zwischen den Ausprägungen (z.B. Abstand zwischen „Hauptschulabschluss“ und „Mittlere Reife“ sowie zwischen „Mittlere Reife“ und „Allgemeine Hochschulreife“) sind nicht gleich. Über die Abstände zwischen den Ausprägungen von ordinalskalierten Variablen sind daher keine Aussagen möglich.

**Pseudometrische Variablen**

In der Praxis werden ordinale Variablen ab etwa fünf Ausprägungen häufig als pseudometrische Variable behandelt. Neben der Mindestanzahl von fünf geordneten Ausprägungen ist allerdings entscheidend, dass angenommen wird, dass die Abstände zwischen den Ausprägungen gleich sind (Baur 2011; Faulbaum et al. 2009, S. 26; Urban und Mayerl 2018, S. 14).

Variablen sind intervallskaliert, wenn deren Ausprägungen nicht nur in eine Rangfolge gebracht werden können, sondern auch die Abstände zwischen den Ausprägungen sinnvoll interpretiert werden können. Ein Beispiel ist die Temperaturmessung in Celsius. Der Abstand zwischen 15 und 20 Grad Celsius ist genau so groß wie der Abstand zwischen 20 und 25 Grad Celsius (jeweils fünf Grad Celsius). Intervallskalen besitzen allerdings keinen natürlichen Nullpunkt. Der Nullpunkt bei der Celsius-Skala wurde lediglich unter pragmatischen Gesichtspunkten gewählt; auch Temperaturen im negativen Bereich der Celsius-Skala sind immer noch eine „Temperatur“. Bei einer Intervallskala sind die Abstände zwischen den Merkmalsausprägungen interpretierbar, aber es können keine Verhältnisse berechnet werden.

Bei einer Ratioskala (auch Verhältnisskala genannt) existiert ein natürlicher (echter) Nullpunkt. Die Temperaturmessung in Kelvin erfolgt auf einer Ratioskala, da bei 0 Kelvin keine Temperatur (keine Bewegungsenergie) mehr feststellbar ist. Auch das Einkommen und das Alter sind Beispiele für ratioskalierte bzw. verhältnisskalierte Variablen. Dabei können nicht nur die Abstände zwischen zwei Ausprägungen, sondern auch die Verhältnisse von zwei Ausprägungen interpretiert werden. Ein Einkommen von 5000 Euro ist doppelt so hoch wie ein Einkommen von 2500 Euro. Eine 60-jährige Person ist doppelt so alt wie eine 30-jährige Person.

In Tabelle 1 sind die zulässigen Rechenoperationen in Abhängigkeit vom Skalenniveau dokumentiert. Wie Tabelle 1 zeigt, steigt mit dem Skalenniveau auch die Anzahl der möglichen Rechenoperationen. Bei einem nominalskalierten Merkmal können die Ausprägungen nur ausgezählt werden, bei einem ordinalskalierten Merkmal können die Ausprägungen in eine Reihenfolge gebracht werden. Bei intervallskalierten Variablen können Differenzen, bei ratioskalierten Variablen auch Verhältnisse gebildet werden.

Tabelle 1: Zulässige Rechenoperationen in Abhängigkeit vom Skalenniveau

| | Auszählen | ordnen | Differenz bilden | Quotienten bilden |
|---|---|---|---|---|
| Nominalskala | Ja | Nein | Nein | Nein |
| Ordinalskala | Ja | Ja | Nein | Nein |
| Intervallskala | Ja | Ja | Ja | Nein |
| Ratioskala | Ja | Ja | Ja | Ja |

Quelle: Mittag (2017, S. 20)

Die Kenntnis des Skalenniveaus einer Variable ist eine wichtige Voraussetzung für die Wahl eines geeigneten Analyseverfahrens. Je höher das Skalenniveau einer Variable ist, desto mehr (und leistungsfähigere) Analyseverfahren stehen den Sozial- und Bildungswissenschaftler*innen zur Verfügung. Die Kenntnis des Skalenniveaus einer Variable ist wichtig, um bei der Datenanalyse nur die zulässigen Analyseverfahren auszuwählen. Viele statistische Verfahren sind nur zulässig, wenn die Variable mindestens intervallskaliert ist bzw. als pseudometrisch behandelt werden kann.

**Diskrete und stetige Variablen**

Die Einteilung als diskrete oder stetige Variable basiert auf der Anzahl der möglichen Ausprägungen. Eine diskrete Variable ist eine Variable, die nur endlich viele Ausprägungen oder höchstens „abzählbar" unendlich viele verschiedene Ausprägungen besitzt (Diaz-Bone 2018, S. 22; Mittag 2017, S. 18). Bei einer diskreten Variable sind keine Zwischenwerte zwischen zwei aufeinander folgenden Ausprägungen möglich. Beispiele für diskrete Variablen sind der Familienstand einer Person, die Anzahl der Fachsemester oder auch die Kinderzahl einer Familie. Bei diesen Variablen sind Zwischenwerte wie 5,6 Fachsemester oder 2,3 Kinder keine möglichen Ausprägungen. Eine stetige Variable ist dadurch gekennzeichnet, dass auch Zwischenwerte möglich sind. Typische Beispiele für stetige Variablen sind Zeit- und Größenangaben, aber auch monetäre Größen wie Einkommen oder Mietpreise. In der Praxis wird bei solchen Merkmalen aber nur eine begrenzte Anzahl an Nachkommastellen erfasst, beispielsweise werden bei Größenangaben meist nur zwei Nachkommastellen angegeben. Grundsätzlich sind allerdings auch mehr Nachkommastellen möglich.

**Dichotome und polytome Variablen**

Eine diskrete Variable mit nur zwei Ausprägungen, wird als „dichotom" bezeichnet und ist dem kategorialen Merkmalsspektrum zuzuordnen. Typische Beispiele für dichotome Variablen sind der Tabakkonsum oder auch die Wahlbeteiligung, bei denen nur die Ausprägungen „Ja" und „Nein" möglich sind. Eine diskrete Variable mit mehreren Ausprägungen wird als polytome Variable bezeichnet. Ein Beispiel für eine polytome Variable ist die Zugehörigkeit bzw. Nicht-Zugehörigkeit zu einer Religionsgemeinschaft mit den Ausprägungen „römisch-katholische Kirche", „evangelische Kirche (ohne Freikirchen)", „evangelische Freikirche", „eine andere christliche Religionsgemeinschaft", „eine andere, nicht-christliche Religionsgemeinschaft" und „keine Religionsgemeinschaft".

**Manifeste und latente Variablen**

Schließlich lassen sich auch manifeste und latente Variablen unterscheiden. Bei manifesten Variablen handelt es sich um Merkmale, die direkt beobachtbar sind. Eine manifeste Variable ist beispielsweise das Geschlecht oder die Haarfarbe einer Person. Dagegen handelt es sich bei latenten Variablen um Merkmale, die sich der direkten Beobachtung entziehen. Latente Variablen sind beispielsweise Intelligenz, Einstellungen wie die Zufriedenheit mit der Demokratie oder auch das soziale Vertrauen. Für eine empirische Untersuchung müssen latente Variablen erst „beobachtbar" gemacht werden. Dieser Vorgang wird als Operationalisierung bezeichnet (Tausendpfund 2018b, S. 107-137).