Markus Tausendpfund
Simone Abendschön

Quantitative Analyseverfahren. Eine Einführung

Fakultät für
Kultur- und Sozialwissen-schaften

FernUniversität in Hagen

Das Werk ist urheberrechtlich geschützt. Die dadurch begründeten Rechte, insbesondere das Recht der Vervielfältigung und Verbreitung sowie der Übersetzung und des Nachdrucks, bleiben, auch bei nur auszugsweiser Verwertung, vorbehalten. Kein Teil des Werkes darf in irgendeiner Form (Druck, Fotokopie, Mikrofilm oder ein anderes Verfahren) ohne schriftliche Genehmigung der FernUniversität reproduziert oder unter Verwendung elektronischer Systeme verarbeitet, vervielfältigt oder verbreitet werden. Wir weisen darauf hin, dass die vorgenannten Verwertungsalternativen je nach Ausgestaltung der Nutzungsbedingungen bereits durch Einstellen in Cloud-Systeme verwirklicht sein können. Die FernUniversität bedient sich im Falle der Kenntnis von Urheberrechtsverletzungen sowohl zivil- als auch strafrechtlicher Instrumente, um ihre Rechte geltend zu machen.

Der Inhalt dieses Studienbriefs wird gedruckt auf Recyclingpapier (80 g/m², weiß), hergestellt aus 100 % Altpapier.

Vorwort

In der quantitativen Sozialforschung 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 Sozialwissenschaftlerinnen und Sozialwissenschaftler sind deshalb elementare Kenntnisse dieser quantitativen Analyseverfahren unverzichtbar.

Für die Sozialwissenschaften stellt die Statistik eine zentrale Hilfswissenschaft dar. Während sich Statistikerinnen – allgemeiner: Mathematikerinnen – häufig mit der Beweisführung und der Weiterentwicklung mathematischer Algorithmen beschäftigen, stehen für Studierende das Kennenlernen und die praktische Anwendung statistischer Verfahren im Vordergrund. Im Mittelpunkt der Lerneinheit steht das Verständnis quantitativer Analyseverfahren, mit denen Studierende bei der Auseinandersetzung mit quantitativen Studien konfrontiert werden.

Die vorliegende Lerneinheit behandelt vier Themenbereiche: Univariate, bivariate und multivariate Datenanalyse sowie Grundlagen der Inferenzstatistik. Das Kapitel zur univariaten Datenanalyse behandelt die Häufigkeitsverteilung einzelner Merkmale. Dabei werden unter anderem Lage- und Streuungsmaße sowie Formmaße vorgestellt. Die bivariate Datenanalyse untersucht Zusammenhänge zwischen zwei Merkmalen. 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 Sozialwissenschaften Stichproben. Deshalb behandelt der vierte Teil der Lerneinheit 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.

Ausschließlich aus Gründen der besseren Lesbarkeit wird in dieser Lerneinheit nicht durchgängig eine geschlechterneutrale Sprache verwendet. Männliche, weibliche und genderneutrale Formen wechseln sich zufallsverteilt ab. Mit den Bezeichnungen sind jeweils alle Geschlechter gemeint.

In der Moodle-Lernumgebung des Moduls M1 „Quantitative Methoden der Sozialwissenschaften“ findet sich eine Errata-Liste zur Lerneinheit. Außerdem werden dort Videos und Übungsaufgaben veröffentlicht, die die Auseinandersetzung mit den Inhalten der Lerneinheit fördern sollen. 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 Juni 2025

Markus Tausendpfund und Simone Abendschön

Inhaltsverzeichnis

Abbildungsverzeichnis VI
Tabellenverzeichnis VII
1 Einführung 9
1.1 Einordnung im Forschungsprozess 9
1.2 Grundgesamtheit und Stichprobe 12
1.3 Klassifikationen von Variablen 14
2 Univariate Datenanalyse 17
2.1 Häufigkeitstabelle 17
2.2 Lagemaße 21
2.2.1 Modus 21
2.2.2 Median 22
2.2.3 Arithmetisches Mittel 24
2.3 Streuungsmaße 27
2.3.1 Varianz 27
2.3.2 Standardabweichung 31
2.4 Formmaße 31
2.4.1 Schiefe 32
2.4.2 Wölbung 35
2.5 Konzentrationsmaße 36
2.5.1 Lorenzkurve 36
2.5.2 Gini-Koeffizient 38
2.6 Variablen standardisieren (z-Transformation) 40
2.7 Grafische Darstellungen 41
2.7.1 Säulen- und Balkendiagramm 42
2.7.2 Kreisdiagramm 43
2.7.3 Histogramm 44
2.7.4 Boxplot 45
3 Bivariate Datenanalyse 47
3.1 Kreuztabellen 48
3.2 Zusammenhangsmaße für nominale Merkmale 56
3.3 Zusammenhangsmaße für ordinale Merkmale 62
3.4 Zusammenhangsmaße für metrische Merkmale 66
3.5 Eta-Quadrat für metrische und nominale Merkmale 75

3.6 Zusammenfassung ... 80
4 Multivariate Datenanalyse ... 81
4.1 Einführung ... 81
4.2 Lineare Regression ... 83
4.2.1 Bivariate Regression ... 84
4.2.2 Multiple Regression ... 91
4.3 Logistische Regression ... 102
4.3.1 Bivariate Regression ... 103
4.3.2 Multiple Regression ... 106
5 Inferenzstatistik ... 112
5.1 Was ist das Problem? ... 112
5.2 Zentrale Konzepte der Inferenzstatistik ... 117
5.2.1 Zentraler Grenzwertsatz und Normalverteilung ... 117
5.2.2 Standardfehler ... 120
5.3 Schätzungsarten ... 126
5.3.1 Punktschätzung ... 126
5.3.2 Intervallschätzung ... 129
5.3.3 Berechnung der benötigten Fallzahl ... 137
5.3.4 Anwendungsprobleme in der Praxis ... 139
5.4 Statistisches Testen ... 141
5.4.1 Allgemeine Vorgehensweise bei einem Signifikanztest ... 143
5.4.2 Alpha- und Beta-Fehler ... 146
5.4.3 t-Test ... 147
6 Literatur ... 162

Abbildungsverzeichnis

Abbildung 1: Idealtypischer Ablauf eines quantitativen Forschungsprojekts 10
Abbildung 2: Grundgesamtheit und Stichprobe..... 13
Abbildung 3: Normalverteilung 32
Abbildung 4: Schiefe..... 33
Abbildung 5: Empirische Verteilungen mit unterschiedlicher Schiefe..... 34
Abbildung 6: Wölbung..... 35
Abbildung 7: Lorenzkurven für drei fiktive Städte 38
Abbildung 8: Säulendiagramm des Interesses an Politik (in Prozent, n = 3490) 42
Abbildung 9: Balkendiagramm des Interesses an Politik (absolute Häufigkeiten, n = 3490) 43
Abbildung 10: Zweitstimmen bei der Bundestagswahl 2025 (in Prozent) 44
Abbildung 11: Histogramm des Alters (absolute Häufigkeiten, n = 3486) 45
Abbildung 12: Elemente eines Boxplots..... 45
Abbildung 13: Boxplot der Interviewdauer (n = 3479) 46
Abbildung 14: IQ und Testergebnis beim räumlichen Denken – Streudiagramm 67
Abbildung 15: Weitere Arten des Zusammenhangs von zwei Merkmalen 68
Abbildung 16: Nettoeinkommen und Lebenszufriedenheit – Streudiagramm..... 73
Abbildung 17: Streudiagramm 86
Abbildung 18: Streudiagramm mit OLS-Regressionsgerade..... 88
Abbildung 19: Schematische Darstellung der vermuteten multivariaten Einflussstruktur..... 93
Abbildung 20: Streudiagramm mit Regressionskurve 105
Abbildung 21: Grundgesamtheit und Stichprobe..... 112
Abbildung 22: Rückschluss von der Stichprobe auf die Grundgesamtheit 113
Abbildung 23: Wiederholte Ziehung von Zufallsstichproben 118
Abbildung 24: Normalverteilung 120
Abbildung 25: Abweichungen einzelner Stichprobenmittelwerte vom wahren Mittelwert..... 121
Abbildung 26: Stichprobenverteilungen bei unterschiedlicher Fallzahl..... 122
Abbildung 27: Ergebnisse des Politbarometers zu zwei Zeitpunkten (in Prozent)..... 130
Abbildung 28: 95-Prozent-Konfidenzintervall 131
Abbildung 29: 99-Prozent-Konfidenzintervall 132
Abbildung 30: Fiktive Befragung zur Wahlentscheidung von 1000 Personen (in Prozent)..... 134
Abbildung 31: 95-Prozent-Konfidenzintervalle (Stichprobengröße jeweils 1000 Personen)..... 136
Abbildung 32: Schätzen und Testen im Vergleich 142
Abbildung 33: t-Verteilung und Normalverteilung 149
Abbildung 34: Verschiedene t-Verteilungen 150
Abbildung 35: Varianten des t-Tests..... 150
Abbildung 36: Einseitiger und zweiseitiger t-Test..... 152

Tabellenverzeichnis

Tabelle 1: Zulässige Rechenoperationen in Abhängigkeit vom Skalenniveau 15
Tabelle 2: Interesse an Politik 17
Tabelle 3: Subjektive Schichteinstufung 20
Tabelle 4: Lagemaße und Skalenniveau 21
Tabelle 5: Berechnung des Modus 21
Tabelle 6: Geschlecht 22
Tabelle 7: Berechnung des Medians (ungerade Fallzahl) 23
Tabelle 8: Berechnung des Medians (gerade Fallzahl) 23
Tabelle 9: Interesse an Politik 24
Tabelle 10: Berechnung des arithmetischen Mittels bei kleinen Fallzahlen 25
Tabelle 11: Berechnung des arithmetischen Mittels bei großen Fallzahlen 26
Tabelle 12: Mittelwerte und Ausreißer 26
Tabelle 13: Lebenszufriedenheit von zwei Gruppen 27
Tabelle 14: Arbeitstabelle für die Berechnung der Varianz (kleine Fallzahl) 29
Tabelle 15: Arbeitstabelle für die Berechnung der Varianz (große Fallzahl) 30
Tabelle 16: Beispieldaten für drei fiktive Städte 36
Tabelle 17: Arbeitstabelle für die Erstellung der Lorenzkurven 37
Tabelle 18: Arbeitstabelle zur Berechnung des Gini-Koeffizienten 39
Tabelle 19: Variablen standardisieren 40
Tabelle 20: Bivariate Zusammenhangsmaße in Abhängigkeit vom Skalenniveau 47
Tabelle 21: Urliste – Abendliche Bibliotheksnutzung und Studiengang (n = 9) 49
Tabelle 22: Kreuztabelle – Abendliche Bibliotheksnutzung und Studiengang (n = 9) 49
Tabelle 23: Abendliche Bibliotheksnutzung und Studiengang – Zeilenprozente (n = 100) 50
Tabelle 24: Abendliche Bibliotheksnutzung und Studiengang – Spaltenprozente (n = 100) 51
Tabelle 25: Abendliche Bibliotheksnutzung und Studiengang – Gesamtprozente (n = 100) 51
Tabelle 26: Politisches Interesse und Geschlecht (Spaltenprozente) 52
Tabelle 27: Schulabschluss und elterlicher Bildungshintergrund (Spaltenprozente) 55
Tabelle 28: Schulabschluss und elterlicher Bildungshintergrund (Zeilenprozente) 56
Tabelle 29: Politisches Interesse und Geschlecht (beobachtete Häufigkeiten) – Kontingenztafel 57
Tabelle 30: Berechnung der erwarteten Häufigkeiten 57
Tabelle 31: Politisches Interesse und Geschlecht (erwartete Häufigkeiten) – Indifferenztabelle .. 58
Tabelle 32: Arbeitstabelle zur Berechnung von Chi-Quadrat 59
Tabelle 33: Interpretation von Cramer's V 61
Tabelle 34: Interpretation von Spearman's Rho 63
Tabelle 35: Soziale Schicht und Gesundheitszustand 64
Tabelle 36: Arbeitstabelle zur Berechnung von Spearman's Rho 65
Tabelle 37: IQ und Testergebnis beim räumlichen Denken – Urliste 66
Tabelle 38: Arbeitstabelle zur Berechnung der Kovarianz 69
Tabelle 39: Interpretation von Pearson's r 70
Tabelle 40: Arbeitstabelle zur Berechnung von Pearson's r 71
Tabelle 41: Nettoeinkommen und Lebenszufriedenheit – Urliste 72
Tabelle 42: Arbeitstabelle zur Berechnung von Pearson's r 74
Tabelle 43: Zwischenergebnisse zur Berechnung von Pearson's r 74

Tabelle 44: Interpretation von Eta-Quadrat ... 77
Tabelle 45: Migrationshintergrund und politisches Wissen ... 77
Tabelle 46: Arbeitstabelle Migrationshintergrund und politisches Wissen ... 78
Tabelle 47: Arbeitstabelle Migrationshintergrund (Nein) und politisches Wissen ... 79
Tabelle 48: Arbeitstabelle Migrationshintergrund (Ja) und politisches Wissen ... 79
Tabelle 49: Unterschiedliche Bezeichnungen für Variablen der Regressionsanalyse ... 82
Tabelle 50: Bivariate lineare Regression mit Lebenszufriedenheit und Einkommen ... 85
Tabelle 51: Dummy-Kodierung für Familienstand ... 94
Tabelle 52: Bestimmungsfaktoren der Lebenszufriedenheit (Teil 1) ... 96
Tabelle 53: Bestimmungsfaktoren der Lebenszufriedenheit (Teil 2) ... 100
Tabelle 54: Bivariate logistische Regression mit Wahlbeteiligung und Alter ... 104
Tabelle 55: Bestimmungsfaktoren der Wahlbeteiligung ... 108
Tabelle 56: Mittelwerte in Zufallsstichproben (Stichprobengröße jeweils 1000 Personen) ... 114
Tabelle 57: Mittelwerte von Zufallsstichproben (Stichprobengröße jeweils 1000 Personen) ... 116
Tabelle 58: Vergleich zwischen Standardfehler und Standardabweichung ... 123
Tabelle 59: Mittelwerte von Zufallsstichproben ... 127
Tabelle 60: Erforderliche Stichprobengröße ... 138
Tabelle 61: Fehlerarten beim Hypothesentest ... 146
Tabelle 62: Lebenszufriedenheit von Frauen und Männern ... 153
Tabelle 63: Kritische Werte der t-Verteilung ... 155
Tabelle 64: Lebenszufriedenheit von West- und Ostdeutschen ... 156
Tabelle 65: Zufriedenheit mit der Demokratie ... 158
Tabelle 66: Beispieldaten für die Berechnung eines t-Tests bei abhängigen Stichproben ... 159

1 Einführung

Markus Tausendpfund

Vorschau

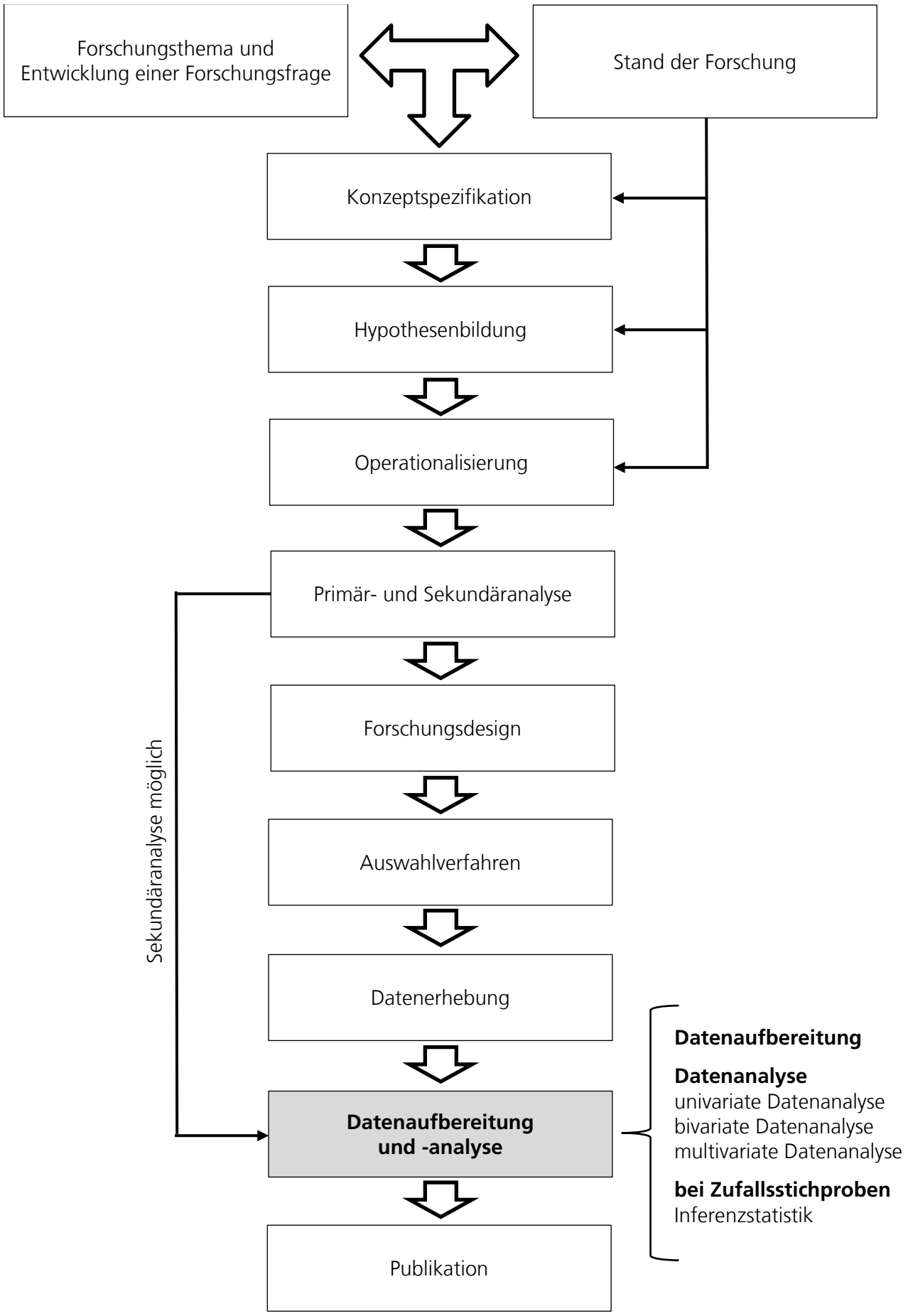
Dieses Kapitel macht mit den Grundlagen der quantitativen Datenanalyse vertraut. Nach der Einordnung der Phase „Datenanalyse" innerhalb des Forschungsprozesses werden die Begriffe „Grundgesamtheit" und „Stichprobe" erläutert. Bei einer empirischen Studie werden meist Aussagen über größere Gruppen angestrebt (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 Sozialwissenschaftlerinnen 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 Sozialforscher 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 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 führt die Sozialwissenschaftlerin 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 2022; Tausendpfund 2025, S. 48). 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 2022; Tausendpfund 2025, S. 275-279). Bei der Arbeit mit qualitativ hochwertigen Sekundärdaten (z.B. ESS) 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 2022, S. 59-64). Dabei wird die Kodierung von Variablen angepasst, einzelne Subgruppen gebildet oder es werden auf Basis der verfügbaren Informationen auch neue Variablen erstellt. Das Verändern und das Erstellen neuer Variablen dauern 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 dieser Lerneinheit 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 Sozialforschung 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.