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/m^2, 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 des Kurses steht das Verständnis quantitativer Analyseverfahren, mit denen Studierende 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 Sozialwissenschaften 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.

Ausschließlich aus Gründen der besseren Lesbarkeit wird in diesem Kurs nicht durchgängig eine geschlechterneutrale Sprache verwendet. Männliche, weibliche und genderneutrale Formen wechseln sich in diesem Kurs 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 zu dem Kurs. Außerdem werden dort 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 Juni 2022

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 Variablen standardisieren (z-Transformation) 36
2.6 Grafische Darstellungen 38
2.6.1 Säulen- und Balkendiagramm 38
2.6.2 Kreisdiagramm 39
2.6.3 Histogramm 40
2.6.4 Boxplot 42
3 Bivariate Datenanalyse 44
3.1 Kreuztabellen 45
3.2 Zusammenhangsmaße für nominale Merkmale 53
3.3 Zusammenhangsmaße für ordinale Merkmale 59
3.4 Zusammenhangsmaße für metrische Merkmale 63
3.5 Eta-Quadrat für metrische und nominale Merkmale 72
3.6 Zusammenfassung 77
4 Multivariate Datenanalyse 78
4.1 Einführung 78

4.2 Lineare Regression ... 80
4.2.1 Bivariate Regression ... 81
4.2.2 Multiple Regression ... 88
4.3 Logistische Regression ... 99
4.3.1 Bivariate Regression ... 100
4.3.2 Multiple Regression ... 103
5 Inferenzstatistik ... 109
5.1 Was ist das Problem? ... 109
5.2 Zentrale Konzepte der Inferenzstatistik ... 114
5.2.1 Zentraler Grenzwertsatz und Normalverteilung ... 114
5.2.2 Standardfehler ... 117
5.3 Schätzungsarten ... 123
5.3.1 Punktschätzung ... 123
5.3.2 Intervallschätzung ... 126
5.3.3 Berechnung der benötigten Fallzahl ... 134
5.3.4 Anwendungsprobleme in der Praxis ... 136
5.4 Statistisches Testen ... 138
5.4.1 Allgemeine Vorgehensweise bei einem Signifikanztest ... 140
5.4.2 Alpha- und Beta-Fehler ... 143
5.4.3 t-Test ... 144
6 Literatur ... 159

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: Säulendiagramm des Interesses an Politik (in Prozent, n = 3490) 39
Abbildung 8: Balkendiagramm des Interesses an Politik (absolute Häufigkeiten, n = 3490) 39
Abbildung 9: Zweitstimmen bei der Bundestagswahl 2017 (in Prozent)..... 40
Abbildung 10: Histogramm des Alters (absolute Häufigkeiten, n = 3486)..... 41
Abbildung 11: Elemente eines Boxplots 42
Abbildung 12: Boxplot der Interviewdauer (n = 3479) 43
Abbildung 13: IQ und Testergebnis beim räumlichen Denken – Streudiagramm 64
Abbildung 14: Weitere Arten des Zusammenhangs von zwei Merkmalen..... 65
Abbildung 15: Nettoeinkommen und Lebenszufriedenheit – Streudiagramm..... 70
Abbildung 16: Streudiagramm..... 83
Abbildung 17: Streudiagramm mit OLS-Regressionsgerade..... 85
Abbildung 18: Schematische Darstellung der vermuteten multivariaten Einflussstruktur..... 90
Abbildung 19: Streudiagramm mit Regressionskurve 102
Abbildung 20: Grundgesamtheit und Stichprobe..... 109
Abbildung 21: Rückschluss von der Stichprobe auf die Grundgesamtheit 110
Abbildung 22: Wiederholte Ziehung von Zufallsstichproben 115
Abbildung 23: Normalverteilung..... 117
Abbildung 24: Abweichungen einzelner Stichprobenmittelwerte vom wahren Mittelwert..... 118
Abbildung 25: Stichprobenverteilungen bei unterschiedlicher Fallzahl 119
Abbildung 26: Ergebnisse des Politbarometers zu zwei Zeitpunkten (in Prozent)..... 127
Abbildung 27: 95-Prozent-Konfidenzintervall..... 128
Abbildung 28: 99-Prozent-Konfidenzintervall..... 129
Abbildung 29: Fiktive Befragung zur Wahlentscheidung von 1000 Personen (in Prozent)..... 131
Abbildung 30: 95-Prozent-Konfidenzintervalle (Stichprobengröße jeweils 1000 Personen)..... 133
Abbildung 31: Schätzen und Testen im Vergleich..... 139
Abbildung 32: t-Verteilung und Normalverteilung 146
Abbildung 33: Verschiedene t-Verteilungen 147
Abbildung 34: Varianten des t-Tests 147
Abbildung 35: Einseitiger und zweiseitiger t-Test..... 149

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: Variablen standardisieren 37
Tabelle 17: Bivariate Zusammenhangsmaße in Abhängigkeit vom Skalenniveau 44
Tabelle 18: Urliste – Abendliche Bibliotheksnutzung und Studiengang (n = 9) 46
Tabelle 19: Kreuztabelle – Abendliche Bibliotheksnutzung und Studiengang (n = 9) 46
Tabelle 20: Abendliche Bibliotheksnutzung und Studiengang – Zeilenprozente (n = 100) 47
Tabelle 21: Abendliche Bibliotheksnutzung und Studiengang – Spaltenprozente (n = 100) 48
Tabelle 22: Abendliche Bibliotheksnutzung und Studiengang – Gesamtprozente (n = 100) 48
Tabelle 23: Politisches Interesse und Geschlecht (Spaltenprozente) 49
Tabelle 24: Schulabschluss und elterlicher Bildungshintergrund (Spaltenprozente) 52
Tabelle 25: Schulabschluss und elterlicher Bildungshintergrund (Zeilenprozente) 53
Tabelle 26: Politisches Interesse und Geschlecht (beobachtete Häufigkeiten) – Kontingenztafel 54
Tabelle 27: Berechnung der erwarteten Häufigkeiten 54
Tabelle 28: Politisches Interesse und Geschlecht (erwartete Häufigkeiten) – Indifferenztabelle .. 55
Tabelle 29: Arbeitstabelle zur Berechnung von Chi-Quadrat 56
Tabelle 30: Interpretation von Cramer's V 58
Tabelle 31: Interpretation von Spearman's Rho 60
Tabelle 32: Soziale Schicht und Gesundheitszustand 61
Tabelle 33: Arbeitstabelle zur Berechnung von Spearman's Rho 62
Tabelle 34: IQ und Testergebnis beim räumlichen Denken – Urliste 63
Tabelle 35: Arbeitstabelle zur Berechnung der Kovarianz 66
Tabelle 36: Interpretation von Pearson's r 67
Tabelle 37: Arbeitstabelle zur Berechnung von Pearson's r 68
Tabelle 38: Nettoeinkommen und Lebenszufriedenheit – Urliste 69
Tabelle 39: Arbeitstabelle zur Berechnung von Pearson's r 71
Tabelle 40: Zwischenergebnisse zur Berechnung von Pearson's r 71
Tabelle 41: Interpretation von Eta-Quadrat 74
Tabelle 42: Migrationshintergrund und politisches Wissen 74
Tabelle 43: Arbeitstabelle Migrationshintergrund und politisches Wissen 75

Tabelle 44: Arbeitstabelle Migrationshintergrund (Nein) und politisches Wissen 76
Tabelle 45: Arbeitstabelle Migrationshintergrund (Ja) und politisches Wissen 76
Tabelle 46: Unterschiedliche Bezeichnungen für Variablen der Regressionsanalyse 79
Tabelle 47: Bivariate lineare Regression mit Lebenszufriedenheit und Einkommen 82
Tabelle 48: Dummy-Kodierung für Familienstand 91
Tabelle 49: Bestimmungsfaktoren der Lebenszufriedenheit (Teil 1) 93
Tabelle 50: Bestimmungsfaktoren der Lebenszufriedenheit (Teil 2) 97
Tabelle 51: Bivariate logistische Regression mit Wahlbeteiligung und Alter 101
Tabelle 52: Bestimmungsfaktoren der Wahlbeteiligung 105
Tabelle 53: Mittelwerte in Zufallsstichproben (Stichprobengröße jeweils 1000 Personen) 111
Tabelle 54: Mittelwerte von Zufallsstichproben (Stichprobengröße jeweils 1000 Personen) 113
Tabelle 55: Vergleich zwischen Standardfehler und Standardabweichung 120
Tabelle 56: Mittelwerte von Zufallsstichproben 124
Tabelle 57: Erforderliche Stichprobengröße 135
Tabelle 58: Fehlerarten beim Hypothesentest 143
Tabelle 59: Lebenszufriedenheit von Frauen und Männern 150
Tabelle 60: Kritische Werte der t-Verteilung 152
Tabelle 61: Lebenszufriedenheit von West- und Ostdeutschen 153
Tabelle 62: Zufriedenheit mit der Demokratie 155
Tabelle 63: Beispieldaten für die Berechnung eines t-Tests bei abhängigen Stichproben 156

1 Einführung

Markus Tausendpfund

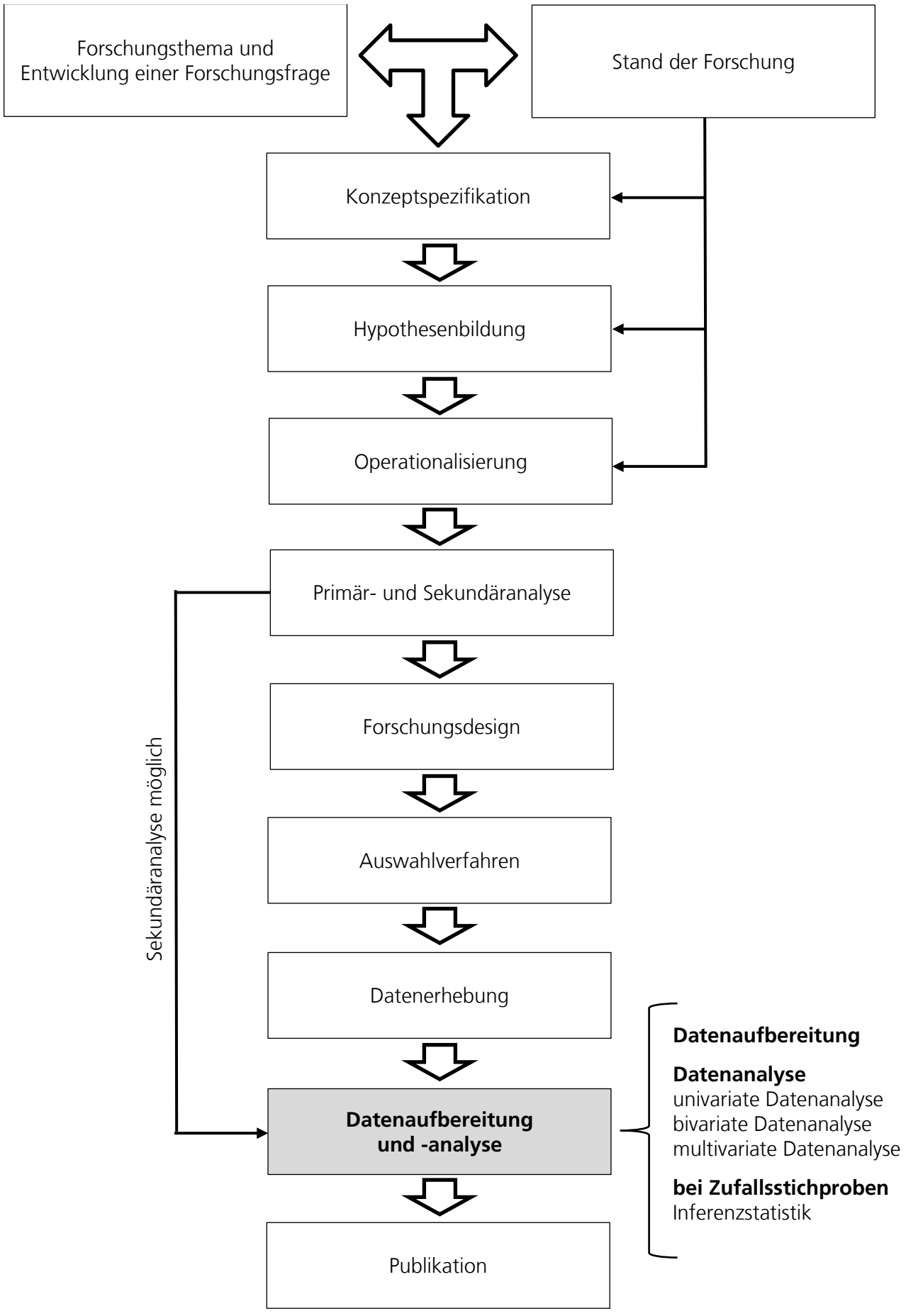
Dieses Kapitel macht Sie mit den Grundlagen der quantitativen Datenanalyse vertraut. Nach der Einordnung der Phase „Datenanalyse" innerhalb des Forschungsprozesses werden die Begriffe „Grundgesamtheit" und „Stichprobe" erläutert. Bei einer empirischen Studie werden meist Aussagen über größere Gruppen 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 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