

Markus Tausendpfund

Quantitative Datenanalyse. Eine Einführung mit R

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

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 der Modulbetreuung

Dieser Kurs bietet den Studierenden der Bildungswissenschaft eine Einführung in die quantitative Datenanalyse mit R bzw. R Studio. Das Werk dient der Vermittlung grundlegender Konzepte der Programmiersprache und eignet sich aufgrund seines hohen Praxisanteils insbesondere für den Einstieg in R. Der Studienbrief wurde von Prof. Dr. Markus Tausendpfund verfasst, der sich schwerpunktmäßig mit den Sozialwissenschaften befasst. In einigen Beispielen oder Datensätzen lässt sich deshalb kein direkter bildungswissenschaftlicher Bezug erkennen. Da die Bildungswissenschaft Teil der Sozialwissenschaften ist und beide auf nahezu identische Forschungsmethoden zurückgreifen, lassen sich die Methoden jedoch problemlos auf die bildungswissenschaftliche Forschungspraxis übertragen.

Den Autor Markus Tausendpfund möchten wir Ihnen gerne kurz vorstellen:

Prof. 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“ und im Jahre 2022 seine Habilitation ab. Er leitet seit 2014 die Arbeitsstelle Quantitative Methoden an der FernUniversität in Hagen.

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. und Prof.'in Dr.'in Julia Schütz am Lehrgebiet Empirische Bildungsforschung redaktionell überarbeitet. Dabei wurden keine inhaltlichen Änderungen vorgenommen, sondern lediglich Änderungen aufgrund eines inklusiven Sprachgebrauchs eingefügt. Zudem werden Bildungswissenschaftler*innen explizit als Zielgruppe angesprochen. In der Moodle-Lernumgebung des Moduls wurden darüber hinaus Lehrvideos zum Umgang mit R veröffentlicht. Wir empfehlen, sich zunächst mit den Lehrvideos vertraut zu machen und diesen Studienbrief als zusätzliche Quelle heranzuziehen.

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

Davin Akko und Julia Schütz

Vorwort des Autors

Die vorliegende Lerneinheit behandelt die sozial- und bildungswissenschaftliche Datenanalyse mit R. Dabei werden Kenntnisse vermittelt, um einfache Analysen selbstständig mit der Software R durchführen zu können.

Der Text entspricht dabei weniger einem klassischen Lehrbuch, sondern eher einem Begleitkurs für die Auseinandersetzung mit dem Programm R bzw. RStudio. Die Lerneinheit soll das Interesse an sozial- und bildungswissenschaftlichen Fragestellungen wecken sowie die Möglichkeiten und Grenzen der quantitativen Datenanalyse aufzeigen.

Lehrmaterial, das in erster Linie zum Selbststudium angelegt ist, profitiert insbesondere durch Rückmeldungen der Leser:innen. Deshalb möchte ich mich herzlich bei allen Personen bedanken, die mich auf Fehler und Verbesserungsmöglichkeiten hingewiesen haben. Ein besonderer Dank geht an Verena Bade, Christian Cleve, Dorothee Köstlin und Simon Stocker, die sich intensiv mit der Lerneinheit beschäftigt und mich auf Ungenauigkeiten sowie Tippfehler aufmerksam gemacht haben.

Die vorliegende Lerneinheit ist kein „Endprodukt“. Die regelmäßige Aktualisierung stellt eine Daueraufgabe dar. Deshalb freue ich mich sehr über alle Hinweise und Anregungen zur weiteren Verbesserung der Lerneinheit (E-Mail: Markus.Tausendpfund@fernuni-hagen.de).

Hagen, im Dezember 2023

Markus Tausendpfund

Inhaltsverzeichnis

| | |
|---|-----|
| Vorwort der Modulbetreuung | III |
| Vorwort des Autors | IV |
| Abbildungsverzeichnis | IX |
| Tabellenverzeichnis..... | X |
| 1 Einführung | 11 |
| 1.1 Sozial- und bildungswissenschaftlicher Forschungsprozess | 12 |
| 1.2 Quantitative Datenanalyse | 15 |
| 1.3 Warum R? | 16 |
| 1.4 Struktur der Lerneinheit..... | 18 |
| 2 R und RStudio kennenlernen | 19 |
| 2.1 Installation..... | 19 |
| 2.1.1 R..... | 19 |
| 2.1.2 RStudio | 20 |
| 2.1.3 Pakete | 23 |
| 2.1.4 Aktualisierungen | 24 |
| 2.1.5 posit Cloud: eine Alternative zur lokalen R-Installation..... | 24 |
| 2.2 Ein erster Überblick..... | 25 |
| 2.2.1 Console | 25 |
| 2.2.2 Skripte | 27 |
| 2.2.3 Befehle | 28 |
| 2.2.4 Objekte..... | 30 |
| 2.2.5 Vektoren..... | 30 |
| 2.2.6 Tabellen..... | 32 |
| 2.2.7 Erste Analysen | 33 |
| 2.2.8 Fehlende Werte | 35 |
| 2.2.9 Hilfe..... | 36 |
| 3 Arbeiten mit R..... | 37 |
| 3.1 Pakete installieren und laden | 37 |
| 3.2 Daten laden..... | 39 |
| 3.3 Daten importieren | 40 |
| 3.3.1 Excel | 41 |
| 3.3.2 SPSS | 41 |
| 3.3.3 CSV | 42 |

| | | |
|-------|--|----|
| 3.4 | Objekttypen | 43 |
| 3.5 | Datenstrukturen | 47 |
| 3.6 | Saubere Skripte erstellen | 47 |
| 3.7 | Projekte in R..... | 49 |
| 4 | Beispieldatensatz | 50 |
| 4.1 | Daten und Pakete | 50 |
| 4.2 | Beispieldatensatz kennenlernen..... | 50 |
| 4.3 | Pipe-Operator | 51 |
| 4.4 | Datenmanagement mit dplyr..... | 52 |
| 4.4.1 | Variablen auswählen..... | 52 |
| 4.4.2 | Variablen umbenennen..... | 53 |
| 4.4.3 | Variablen filtern | 54 |
| 4.4.4 | Variablen verändern..... | 54 |
| 4.4.5 | Einfache Berechnungen | 55 |
| 4.4.6 | Weitere Optionen | 56 |
| 4.5 | Datenaufbereitung mit sjmisc..... | 56 |
| 4.5.1 | Variablenwerte ändern | 56 |
| 4.5.2 | Variablen zusammenfassen..... | 59 |
| 4.5.3 | Werte zählen | 61 |
| 4.5.4 | Weitere Optionen | 61 |
| 4.6 | Labels konvertieren | 62 |
| 5 | Univariate Datenanalyse | 64 |
| 5.1 | Daten und Pakete | 64 |
| 5.2 | Häufigkeitstabellen | 64 |
| 5.3 | Lagemaße | 66 |
| 5.4 | Streuungsmaße | 67 |
| 5.5 | Formmaße | 68 |
| 5.6 | Kompakte Übersichten..... | 70 |
| 6 | Bivariate Datenanalyse..... | 71 |
| 6.1 | Daten und Pakete | 71 |
| 6.2 | Univariate Statistiken nach Gruppen..... | 71 |
| 6.3 | Kreuztabellen..... | 72 |
| 6.4 | Zusammenhangsmaße | 75 |
| 6.4.1 | Nominalskalierte Merkmale..... | 76 |

| | | |
|-------|-------------------------------------|-----|
| 6.4.2 | Ordinalskalierte Merkmale | 79 |
| 6.4.3 | Metrische Merkmale | 81 |
| 7 | Multivariate Datenanalyse..... | 84 |
| 7.1 | Einführung | 84 |
| 7.2 | Lineare Regression..... | 86 |
| 7.2.1 | Das Grundmodell..... | 86 |
| 7.2.2 | Daten und Pakete | 94 |
| 7.2.3 | Lineare Regression mit R | 94 |
| 7.2.4 | Interpretation der Ergebnisse | 97 |
| 7.2.5 | Weitere Möglichkeiten | 100 |
| 7.2.6 | Anwendungsvoraussetzungen | 106 |
| 7.2.7 | Praktische Hinweise | 108 |
| 7.3 | Logistische Regression | 109 |
| 7.3.1 | Das Grundmodell..... | 109 |
| 7.3.2 | Daten und Pakete | 115 |
| 7.3.3 | Logistische Regression mit R..... | 116 |
| 7.3.4 | Interpretation der Ergebnisse | 118 |
| 7.3.5 | Weitere Möglichkeiten | 120 |
| 7.3.6 | Praktische Hinweise | 126 |
| 8 | Inferenzstatistik | 128 |
| 8.1 | Daten und Pakete..... | 128 |
| 8.2 | Konfidenzintervalle..... | 129 |
| 8.3 | Mittelwertvergleiche (t-Test) | 130 |
| 9 | Grafiken | 136 |
| 9.1 | Einführung | 136 |
| 9.2 | Ausgewählte Diagramme | 138 |
| 9.2.1 | Säulen- und Balkendiagramm..... | 138 |
| 9.2.2 | Kreisdiagramm..... | 140 |
| 9.2.3 | Histogramm | 141 |
| 9.2.4 | Boxplot | 142 |
| 9.2.5 | Streudiagramm | 144 |
| 9.2.6 | Liniendiagramm | 146 |
| 9.3 | Weitere Pakete | 147 |
| | Pakete im Überblick..... | 148 |

| | |
|--------------------------------------|-----|
| Beispieldatensatz im Überblick | 150 |
| Literaturverzeichnis | 154 |

Abbildungsverzeichnis

| | |
|--|-----|
| Abbildung 1: Phasen eines quantitativen Forschungsprojekts | 14 |
| Abbildung 2: The Comprehensive R Archive Network (CRAN) | 19 |
| Abbildung 3: Startbildschirm von R | 20 |
| Abbildung 4: RStudio mit drei Fenstern | 21 |
| Abbildung 5: RStudio mit vier Fenstern | 22 |
| Abbildung 6: Global Options bei RStudio | 23 |
| Abbildung 7: posit Cloud | 25 |
| Abbildung 8: Streudiagramm des Einkommens in Abhängigkeit des Alters | 35 |
| Abbildung 9: Installierte Pakete | 38 |
| Abbildung 10: „Import Dataset“-Funktion in RStudio | 40 |
| Abbildung 11: Excel-Datensatz importieren | 41 |
| Abbildung 12: SPSS-Datensatz importieren | 42 |
| Abbildung 13: Normalverteilung | 69 |
| Abbildung 14: Auswahl von regressionsanalytischen Verfahren | 84 |
| Abbildung 15: Streudiagramm von Alter und Einkommen | 87 |
| Abbildung 16: Streudiagramm von Einkommen und Alter mit Regressionsgerade | 88 |
| Abbildung 17: Empirischer und geschätzter Wert einer linearen Regression | 89 |
| Abbildung 18: Grafische Darstellung einer multiplen Regression | 93 |
| Abbildung 19: Lineare Regressionsfunktion | 111 |
| Abbildung 20: Logistische Regressionsfunktion | 113 |
| Abbildung 21: Logistische Regression (Beispieldaten) | 115 |
| Abbildung 22: Grafische Darstellung der logistischen Regression I | 125 |
| Abbildung 23: Grafische Darstellung der logistischen Regression II | 126 |
| Abbildung 24: Aussage über die Grundgesamtheit auf Basis einer Zufallsstichprobe | 132 |
| Abbildung 25: Säulen- und Balkendiagramm | 139 |
| Abbildung 26: Kreisdiagramm | 140 |
| Abbildung 27: Histogramm | 141 |
| Abbildung 28: Elemente eines Boxplots | 142 |
| Abbildung 29: Boxplots | 143 |
| Abbildung 30: Plotsymbole bei R | 145 |
| Abbildung 31: Streudiagramm | 145 |
| Abbildung 32: Liniendiagramm | 146 |

Tabellenverzeichnis

| | |
|--|-----|
| Tabelle 1: SPSS, Stata und R im Überblick..... | 17 |
| Tabelle 2: Mathematische Funktionen in R | 26 |
| Tabelle 3: Logische Abfragen in R..... | 27 |
| Tabelle 4: Installieren und Laden von Paketen..... | 38 |
| Tabelle 5: Beispieldatensatz (peanuts_r)..... | 39 |
| Tabelle 6: Ausgewählte Funktionen zum Testen und Konvertieren von Objekten | 45 |
| Tabelle 7: Objekttypen in den Datensätzen..... | 46 |
| Tabelle 8: Ausgewählte Funktionen von dplyr..... | 52 |
| Tabelle 9: Auswahl von Filter-Möglichkeiten | 54 |
| Tabelle 10: Ausgewählte Möglichkeiten des rec-Arguments | 58 |
| Tabelle 11: Interpretation von Schiefe und Wölbung | 70 |
| Tabelle 12: Kreuztabelle zwischen Wahlbeteiligung und Bildung | 74 |
| Tabelle 13: Zusammenhangsmaße bei der bivariaten Datenanalyse | 76 |
| Tabelle 14: Arbeitstabelle für die Berechnung von Chi-Quadrat | 77 |
| Tabelle 15: Interpretation von Cramer's V | 79 |
| Tabelle 16: Interpretation von Spearman's Rho..... | 81 |
| Tabelle 17: Interpretation des Korrelationskoeffizienten nach Pearson | 83 |
| Tabelle 18: Beispieldaten für Alter und Einkommen..... | 86 |
| Tabelle 19: Bivariate Regression (Beispieldaten) | 91 |
| Tabelle 20: Beispieldaten für Alter, Einkommen und Berufserfahrung | 92 |
| Tabelle 21: Multiple Regression (Beispieldaten)..... | 93 |
| Tabelle 22: Bestimmungsfaktoren der Lebenszufriedenheit I..... | 95 |
| Tabelle 23: Bestimmungsfaktoren der Lebenszufriedenheit II..... | 96 |
| Tabelle 24: Bestimmungsfaktoren der Lebenszufriedenheit III | 97 |
| Tabelle 25: Bestimmungsfaktoren der Lebenszufriedenheit IV..... | 101 |
| Tabelle 26: Bestimmungsfaktoren der Lebenszufriedenheit V (standardisierte Koeffizienten).. | 102 |
| Tabelle 27: Bestimmungsfaktoren der Lebenszufriedenheit V (mit Faktoren) | 104 |
| Tabelle 28: Bestimmungsfaktoren der Lebenszufriedenheit VI (Regressionstabelle mit sjPlot).. | 105 |
| Tabelle 29: Informationen zu einer linearen Regression | 106 |
| Tabelle 30: Beispieldaten für Rauchen und Alter in Jahren | 110 |
| Tabelle 31: Bestimmungsfaktoren des Tabakkonsums | 114 |
| Tabelle 32: Bestimmungsfaktoren der Wahlbeteiligung | 117 |
| Tabelle 33: Bedeutung der Asteriske | 119 |
| Tabelle 34: Logistische Regressionskoeffizienten und Odds Ratio im Vergleich..... | 122 |
| Tabelle 35: t-Test der Lebenszufriedenheit in Abhängigkeit des Geschlechts..... | 133 |
| Tabelle 36: t-Test der Demokratiezufriedenheit in Abhängigkeit des politischen Interesses..... | 134 |
| Tabelle 37: Argumente beim t-Test | 135 |
| Tabelle 38: Argumente bei Grafikbefehlen | 136 |
| Tabelle 39: Pakete zur Erstellung von Grafiken | 147 |
| Tabelle 40: Pakete im Überblick..... | 148 |
| Tabelle 41: Beispieldatensatz im Überblick..... | 150 |

1 Einführung

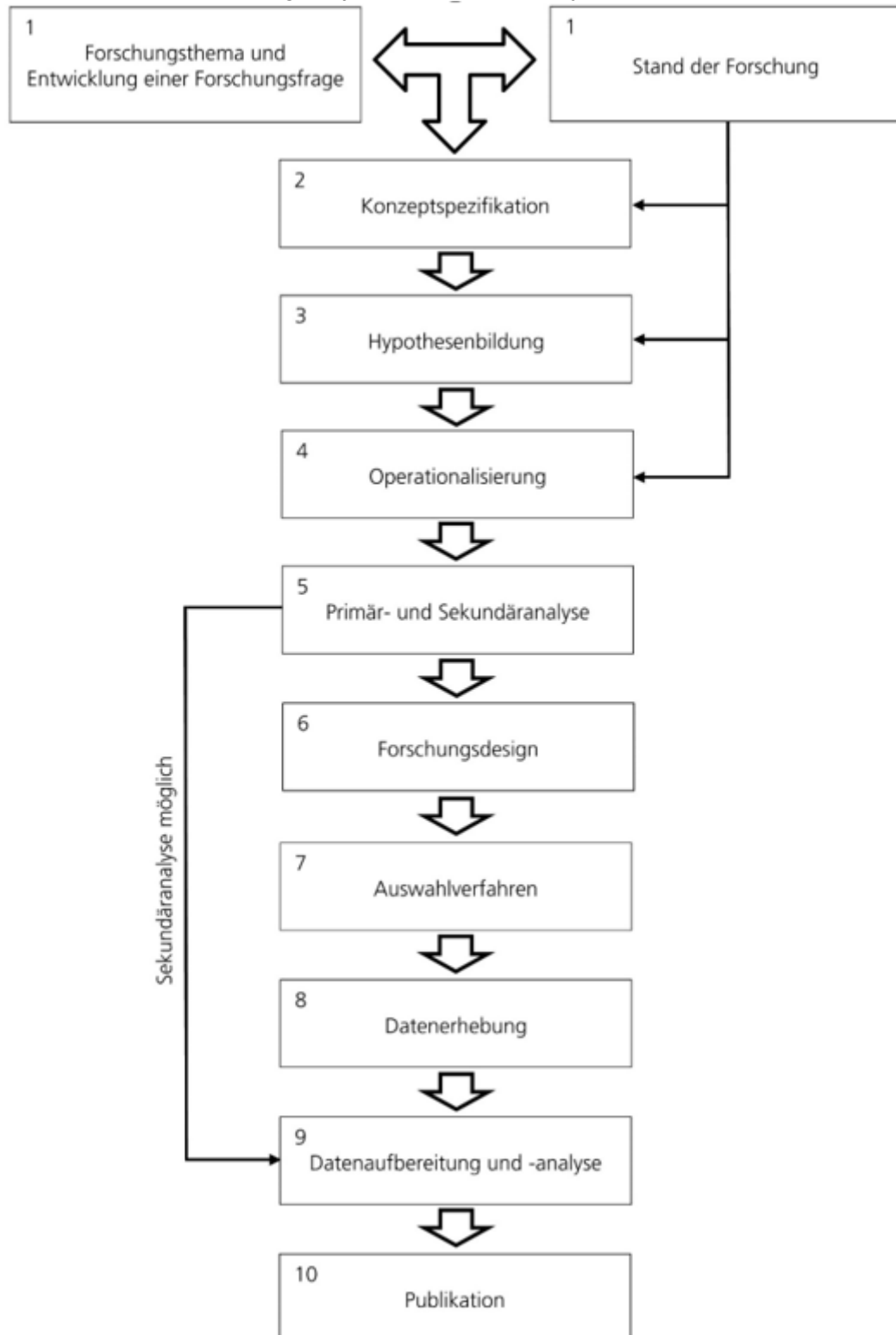
Die quantitative Datenanalyse ist die Phase im sozial- und bildungswissenschaftlichen Forschungsprozess, in der die theoretisch entwickelten Hypothesen empirisch geprüft werden. In diesem Kapitel werden die einzelnen Phasen des Forschungsprozesses knapp skizziert, die Bedeutung der Methodenkompetenz für die Auseinandersetzung mit empirischen Studien dargelegt und Statistikprogramme vorgestellt. Die Einführung schließt mit einem Ausblick auf die weiteren Kapitel dieser Lerneinheit und verweist auf ergänzende Materialien in der Moodle-Lernumgebung.

Vorschau



1.1 Sozial- und bildungswissenschaftlicher Forschungsprozess

In einem quantitativen Forschungsprojekt lassen sich idealtypisch mehrere Phasen unterscheiden



(siehe

Abbildung 1). Nach der Entscheidung für ein Forschungsthema und der Entwicklung einer Forschungsfrage (1) müssen zunächst die zentralen Konzepte der Forschungsfrage identifiziert und theoretisch geklärt werden (2). Auf dieser Grundlage können Hypothesen formuliert (3) und Operationalisierungen der Konzepte (4) entwickelt werden (ausführlicher Tausendpfund 2018).